HUAWEI

Version 2024

Intelligent World 2030

Building a Fully Connected, Intelligent World

Contents

Executive Summary

We are making strides towards an intelligent world. When looking ahead to 2030, we hope that the future will bring improved quality of life, sustainable and green diets, and more comfortable living spaces. We also look forward to the end of traffic congestion and pollution in cities, fully green energy, and a wide range of new digital services. We dream of robots that can do repetitive and dangerous work for us so that we can devote more time and energy to more valuable, creative work, and to our personal interests. These are the goals that drive exploration in every industry.

Huawei has held in-depth discussions with well-known scholars, customers, and partners in the industry to explore the intelligent world. We have found that the rapid advancement of the intelligent world has given rise to an increasing number of new technologies and scenarios and the exponential growth of related industry parameters. Therefore, Huawei has systematically updated the Intelligent World 2030 that was released in 2021 to show our latest vision for the scenarios and trends in 2030 and adjust relevant forecast data accordingly.

Huawei is committed to bringing digital to every person, home and organization for a fully connected, intelligent world. In this report, we examine the prospects for the intelligent world over the next decade by analyzing macro trends in healthcare, food, living spaces, transportation, cities, enterprises, energy, and digital trust.

We believe in the infinite possibilities of the intelligent world, but constant collaboration and exploration among many different industries will be required to build a better future.

Outlook for Healthcare: Smart Health Services Enhance Quality of Life

By 2030, ICT technology will further shift the healthcare system from traditional diagnosis and treatment to full-lifecycle health management. With ongoing AI innovations, accessible healthcare services will significantly enhance quality of life.

Huawei predicts that by 2030

- Global general computing power (FP32) will reach 3.3 ZFLOPS, a 10-fold increase over 2020.
- AI computing power (FP16) will reach 864 ZFLOPS, a 4,000-fold increase over 2020.

Outlook for Food: Data-driven Food Production for More Bountiful, Inclusive, and "Green" Diets

By 2030, visualized data graphs for agricultural production will make precision farming possible. Based on collected data, people will be able to control factors that affect crop growth, such as temperature and humidity, and create vertical farms unaffected by the uncertainties of climate and weather.

Huawei predicts that by 2030

- There will be 200 billion connections worldwide.
- 1YB of data will be generated annually worldwide, a 23-fold increase over 2020.

Outlook for Living Spaces: Personalized Spaces with Novel Interactive Experiences

By 2030, we will no longer have to live with clutter. We will manage our possessions with a digital catalog powered by a 10-gigabit network, holograms, and other technologies. Automatic delivery systems will bring household items from shared warehouses to our doors whenever we need them. Intelligent management systems that control our physical surroundings for automatic interactions will mean that the buildings where we live and work may produce net zero carbon. Next-generation IoT operating systems will enable people to live and work in adaptive environments that understand their needs.

Huawei predicts that by 2030

There will be 1.6 billion fiber broadband subscribers.

25% of homes will have access to 10 gigabit fiber broadband.

Outlook for Transportation: Smart, Low-carbon Transportation Opens up the Mobile Third Space

In 2030, new energy, autonomous driving, and vehicle-road-cloud synergy will enter the fast lane, making vehicles into a mobile third space outside our homes and workplaces.

Huawei predicts that by 2030

82% of new vehicles sold will be electric vehicles.

In addition, by 2030, the whole-vehicle computing power will exceed 5,000 TOPS.

Outlook for Cities:
New Digital Infrastructure Makes Cities More Human and Livable

Digital technologies ranging from digital infrastructure, cloud computing, and trustworthy data spaces will make cities more livable and city governance more efficient.

Huawei predicts that by 2030

84% of companies will have access to 10 gigabit Wi-Fi networks.

Outlook for Enterprises:
New Productive Forces, New Production Models, New Resilience

In 2030, new productive forces like collaborative robots, autonomous mobile robots (AMRs), and digital employees will enter numerous industries, and the wide adoption of new productive forces like industrial humanoid robots will significantly improve quality and boost efficiency.

Huawei predicts that by 2030

Every 10,000 workers will work with 390 robots.

One million companies are expected to build their own 5G private networks (including virtual private networks).

Cloud services are forecast to account for 87% of enterprises' application expenditures.

AI computing will account for 7% of a company's total IT investment.

Outlook for Energy: Intelligent, Green Energy for a Better Planet

The energy world will be centered on electricity, with green hydrogen becoming a major player by 2030. The solar PV and energy storage industries will develop rapidly, expanding from a few countries to the entire world. Power plants will generate electricity from renewable sources in lakes and near-shore marine areas. An "energy Internet" will emerge, utilizing digital technologies to connect generation, grid, load, and storage, including virtual power plants and an energy cloud. Network-wide intelligence will be a reality.

Huawei predicts that by 2030

Renewables will account for 65% of all electricity generation globally.

Outlook for Digital Trust: Technologies and Rules Creating a Trusted Digital Future

By 2030, technologies such as digital identities, post-quantum cryptography (PQC), digital watermarking, privacy-enhancing computation (PEC), and AI provenance and verification will lay a solid foundation for the sustainable development of digital civilization.

Huawei predicts that by 2030

Privacy-enhanced computing technologies will be used in more than 50% of computing scenarios.

100% of ICT systems will have a quantum-safe capacity or the capacity to migrate to quantum-safe solutions

Healthcare

Smart Health Services
Enhance Quality of Life

Over the past decade, there have been significant improvements in people's health and well-being. According to a United Nations report, global life expectancy at birth reached 73.3 years in 2024, an increase of 8.4 years since 1995. [1] As people live longer, finding ways to enhance the quality of life is becoming a top priority.

As the global population ages, the demand for healthcare is expected to surge. By 2030, there will be about 994 million elderly individuals, and this figure is expected to rise to 1.6 billion by 2050. [2] According to the World Health Organization (WHO), global spending on healthcare is already outpacing global economic growth, and by 2030, there will be a shortage of 10 million healthcare workers worldwide, [3] 5.7 million of which will be nurses. [4]

Additionally, the negative impact of chronic diseases and subhealth in general is increasing. World Health Statistics 2024 reports that seven of the top 10 causes of death are non-communicable diseases (NCDs), and the premature mortality rate is rising. The probability of dying between the ages of 30 and 70 from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases is now 22.7% [5].

Population growth is outpacing the production of healthcare resources. According to the World Population Prospects 2024, the global population is expected to reach 10.3 billion by the mid-2080s, up from 8.2 billion in 2024. This growing population will place additional strain on healthcare resources, making it harder to achieve the Sustainable Development Goals (SDGs). For example, in 2023, Africa's population was twice that of Europe, and by 2050, it is expected to be 3.5 times larger. [6] However, healthcare resources remain unevenly distributed—Germany currently has 16 times more hospital beds per 10,000 people than Nigeria. [7] The aging population will put even more pressure on healthcare resources. By 2080, 2.2 billion people will be at or above the age of 65, and they will outnumber the children under the age of 18. [8] To address these challenges, countries and regions must invest in human capital to ensure that healthcare and quality education are accessible to all. [9]

In the future, healthcare must shift from being treatment-focused to comprehensive, life-cycle health management. By prioritizing universal access and improving the quality of care, the future healthcare systems will be able to offer continuous and efficient services. This will involve building an integrated and seamless health management network and establishing a city-centric smart health system to improve public health.

Direction for exploration: Unlocking the value of health data, and shifting the focus from reactive treatment to proactive prevention

According to the WHO, 60% of related factors to illnesses are correlated to lifestyle, [10] making healthy habits essential for well-being. With user consent, wearable devices can collect and analyze real-time health data, and offer predictive insights and medical guidance with the assistance of a unified AI architecture. This shift toward proactive prevention integrates health management into daily life, connecting disease control, hospitals, health centers, and families to reduce the risk of a more serious health condition arising.

Snapshot from the future: Building a knowledge graph to achieve real-time and efficient health management

Thanks to the advancements in the Internet, IoT, and AI, as well as the widespread adoption of wearable devices and home monitoring equipment, by 2040, at least a quarter of outpatient care, preventive care, long-term care, and health services will move online. [11]

Specifically, these technologies will be used to analyze real-time health data, medical responses, and clinical outcomes to identify potential health risks. For example, AI can detect the early signs of heart disease or pre-diabetic conditions by analyzing and warning users about anomalous heart rate and blood pressure measurements, allowing for early intervention. Additionally, through interventions in nutrition, exercise, and sleep, users can be guided to gradually change their unhealthy habits and develop a healthier lifestyle, reducing the likelihood of illness. For instance, a study by Stanford University showed that continuous monitoring of heart rate and skin temperature through smartwatches and other wearables could help AI detect early signs of infection, as these powerful monitoring devices can take and analyze up to 250,000 measurements per day. [12]

Moreover, a comprehensive health management platform could allow hospitals, doctors, users, and families to collectively access and view health data. This data-sharing mechanism ensures that doctors can stay updated on their patients' health in real time, both inside and outside the hospital, leading to more accurate diagnoses and better treatment decisions.

Snapshot from the future: Using intelligent disease prediction and prevention to enhance public health response capabilities

By integrating electronic health records (EHRs), wearable device data, lab results, and public health data from a variety of sources, AI systems can build comprehensive health databases. Using machine learning and deep learning, AI will analyze this data to identify patterns related to the incidence of disease and develop predictive models. These models will provide early warnings about the potential risk of being affected by certain diseases at both the individual and population levels, allowing for timely preventive action.

For example, researchers are using AI and big data to analyze global public health data and epidemiological information and to develop models capable of predicting outbreaks of infectious diseases such as influenza. These models can identify potential hotspots and transmission pathways, allowing health organizations to implement preventative measures before outbreaks spread. [13]

The intelligent disease prediction and prevention system enables public health organizations to swiftly respond to epidemic threats and adopt effective control measures, significantly reducing the societal impact of outbreaks. It not only enhances public health response capabilities but also improves individual health management, and this contributes to the improvement of overall public health.

Direction for exploration: Enabling healthcare with digital intelligence to improve quality and accessibility

The integration of AI into healthcare not only safeguards public health but also drives improvements in economic and social development. [14] By using advanced technologies, digital healthcare significantly enhances the quality and accessibility of medical services. AI and machine learning algorithms can analyze vast amounts of medical data and assist doctors in making precise diagnoses and developing personalized treatment plans. This improves treatment outcomes, enhances patient satisfaction, and reduces the likelihood of a misdiagnosis.

Digital healthcare systems also optimize resource allocation by ensuring balanced coverage across regions, including in remote and underserved areas. Through intelligent scheduling and resource management, these systems ensure that healthcare services are available to more patients, and this improves access to timely and high-quality medical care.

Snapshot from the future: Smart healthcare innovations for enhanced diagnostic efficiency and precision

In the future, digital and AI-powered medical care will lead to significant improvements in efficiency and precision. AI and big data will play a crucial role in medical imaging in particular. Using deep learning algorithms, AI can analyze large datasets of medical images, such as X-rays, CT scans, and magnetic resonance imaging (MRI) scans, and detect lesions and provide accurate diagnostic recommendations. This will not only enhance screening efficiency but also reduce the risk of misdiagnosis. For example, research shows that AI outperforms traditional methods in early breast cancer detection. [15]

Furthermore, AI will be integrated with electronic medical records (EMRs) to continuously update patient health data in real-time. This real-time data will help AI develop personalized treatment plans, predict disease risks, and assist doctors in making more effective treatment decisions, ultimately improving patient outcomes. [16]

AI will also play a significant role in pathology screening, where it can be used to analyze pathology slides and detect abnormal cell and tissue changes. This assists pathologists in making faster and more accurate diagnoses. AI's ability to detect cancer cells in pathology images has already been widely recognized. [17]

Big data and AI will also play an important role in helping healthcare organizations and insurance companies manage expenses more intelligently and thus help to keep healthcare costs reasonable for patients. By analyzing vast amounts of medical data, AI can predict healthcare cost trends, identify unnecessary expenses, and detect potential fraud. This smart cost-control method not only protects patients' interests but also optimizes the use of healthcare resources.

Snapshot from the future: Enabling AI-driven, all-domain, collaborative healthcare to optimize resource allocation

The future of healthcare will be revolutionized by modern communications and information technologies. They will extend medical services to remote monitoring, consultation, and treatment. AI and foundation models will be central to this revolution. Clinical decision support systems (CDSs) are particularly important. By leveraging deep learning and machine learning, CDSs analyze medical images and EHRs to provide accurate and efficient diagnoses and personalized treatment solutions. These systems can quickly supply essential diagnostic information to doctors in remote areas, leading to more accurate decisions and treatments. For example, in Türkiye there were plans to develop a web- and mobile-based application which would allow doctors to remotely monitor patient data in real time and which would unify and enable the screening, diagnosis, treatment, and monitoring of diabetes diseases. [18]

Personalized health management is another key area. By analyzing both historical and real-time health data, AI can predict health risks and offer tailored recommendations to help individuals adjust their lifestyles and prevent diseases. These technologies show a lot of promise in chronic disease management; for example, remote monitoring of hypertension allows for timely interventions and significantly improves health outcomes. In all-domain healthcare collaboration, continuous data monitoring and analysis enable remote doctors to provide more precise health guidance.

Large language models (LLMs) help patients by answering questions in real time, thereby easing the burden on remote doctors. These models provide medical advice, psychological support, and can also assist in making appointments with doctors, making healthcare more efficient. For instance, virtual assistants can analyze conversations with patients to offer personalized suggestions which can help patients improve their mental health. [19]

Video communication technologies also play a vital role, as they allow doctors to conduct remote consultations and diagnostics, increasing the convenience and accessibility of medical services. For example, the Remote Hypertension Improvement Program uses video calls and remote monitoring to efficiently lower patients' blood pressure. [20]

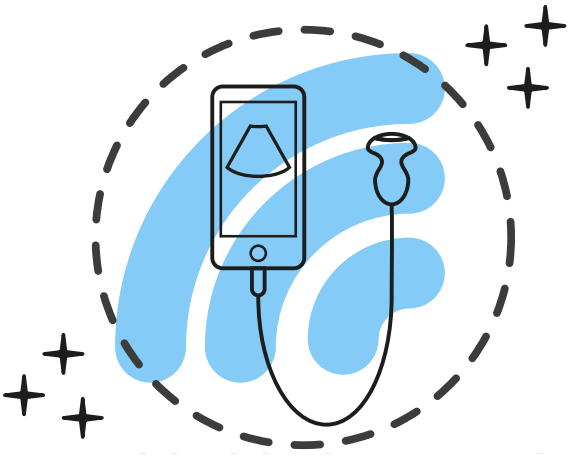
Overall, AI and foundation models elevate the quality and efficiency of healthcare services, enabling personalized health management and optimized resource allocation, and driving the development of more inclusive healthcare services.

Direction for exploration: Multidisciplinary integration and its role in driving innovation in medical research

The integration of multiple disciplines is a key frontier to drive medical research innovation today. Advancements in biomedical engineering, information technology, AI, and big data, are driving a significant transformation in medical research. For instance, nanotechnology in drug delivery systems enhances treatment precision and reduces side effects. [21] Big data analysis allows researchers to uncover new disease patterns and develop novel treatment methods which advance personalized medicine. Meanwhile, AI can be used in combination with medical images and EHRs to accelerate and improve the accuracy of diagnostics. [22] This multidisciplinary approach accelerates the research-to-clinical-practice process, drives rapid advancements in medical technology, and establishes a solid foundation for the future of healthcare.

Snapshot from the future: Multimodal data integration for more advanced precision medicine

Multimodal data integration is set to revolutionize precision medicine. By combining genomics, proteomics, metabolomics, imaging data, and clinical information, researchers can gain a more comprehensive understanding of disease mechanisms. For example, AI can predict disease risks more accurately and develop personalized treatment plans with multimodal data. A study in Nature Medicine has shown how the convergence of generative AI and LLMs in medical imaging has opened up new ways to harness the power of both visual and textual information. Integrating these advanced technologies enables multimodal data integration, representation learning, and improved clinical decision support systems. Multimodal models built upon generative AI and LLMs can integrate the visual features of medical images with contextual information from radiology reports or EHRs to facilitate various medical-image processing tasks. LLMs can process radiology reports to extract pertinent information, match them with the corresponding images, and generate natural-language summaries that can enhance communication between healthcare professionals and facilitate better decision-making when it comes to patient care. By leveraging patient-specific information, such as genetic data, medical history, and lifestyle factors, and evaluating it in conjunction with medical images, these AI models can facilitate more efficient treatment and diagnoses for patients. [23]



Snapshot from the future: The rise of intelligent medicine and its role in driving industry transformation

The future of medicine is poised to move from a one-size-fits-all approach to a more bespoke and patient-centric approach. Key factors such as the physical condition of the patient, appropriate drug types, timing, dosage, and the treatment duration must be taken into consideration when designing drug treatment plans. These plans, once formulated, also need to be regularly updated based on the treatment effect and progress. This puts significant pressure on doctors, and they are often forced to rely on general expertise and experience rather than the patient's specific symptoms and indicators to quickly formulate a general treatment plan. However, with AI and foundation model technologies, vast amounts of pathological data can be analyzed in real time, enabling doctors to offer more personalized treatment recommendations. For example, a research institute in Singapore has developed an AI-powered platform that evaluates medication effectiveness. The platform can quickly analyze a patient's clinical data, provide the patient with a personalized prescription, and modulate tumor sizes or biomarker levels in the patient's profile based on available data. [24]

In drug development, AI has moved beyond the proof-of-concept stage. Advanced machine learning technologies are accelerating the pace of innovation, reducing evaluation times, and enabling the exploration of new areas of medicine. In practice, AI R&D tools improve the speed of ingesting, structuring, and extracting inferences from scientific literature by a factor of 1,000. AI-driven simulations run 2 to 40 times faster, and AI models can propose new hypotheses 10 times faster than before. Autonomous AI-powered laboratories can conduct experiments 100 times faster than before. This reduction in manual data processing and information handling has increased the overall speed of drug discovery tenfold. [25]

For example, Insilico Medicine, an AI pharmaceutical company, launched the world's first automated AI-assisted decision-making laboratory. The lab integrates AI with automation, robotics, and biological capabilities, and can complete the entire cycle of target discovery and validation within 14 days. [26]

Looking ahead, AI will drive cost reduction and efficiency gains in the pharmaceutical industry, accelerate drug development, and bring unprecedented changes to this industry.

Conclusion:

Smart healthcare services are the key to a better quality of life

By 2030, advanced technologies like AI, foundation models, cloud computing, and big data will form the backbone of efforts to enhance global health. These innovations will permeate every aspect of healthcare, from early disease diagnosis to precise treatment. They will drive full-scale intelligence and higher efficiency in medical services.

In health management, AI-powered unified architectures will enable personalized, end-to-end health management in more cities and remote areas and will enable many more people to quickly and easily access medical services.

In diagnostics and treatment, integrated data platforms that handle vast amounts of structured and unstructured medical data will improve both the accuracy and speed of disease diagnosis. These platforms will also consolidate global medical research and clinical trial data, aiding in new medical breakthroughs and discoveries. Cloud computing will provide robust support for data storage and processing, ensuring that these intelligent technologies can be seamlessly integrated and run efficiently.

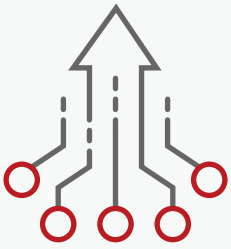
In medical research, multimodal technologies featuring big data, AI, genomics, advanced imaging technologies, and clinical decision support systems will significantly enhance the precision and efficiency of clinical diagnostics and drug development.

The application and integration of these technologies will dramatically improve the quality and accessibility of healthcare services, optimize the allocation of medical resources, and foster innovative applications. This will lead to better health outcomes and longer life expectancies, and take the global healthcare industry to new heights.

By 2030, advanced ICT technologies will play a vital role in driving global health development. Advancements in large-scale computing power will enable a variety of intelligent health applications, and will also make it easier for more people to access efficient medical services. Huawei predicts that by 2030, there will be 3.3 ZFLOPS of general-purpose computing power (FP32) available, a 10-fold increase over 2020; and 864 ZFLOPS of AI computing power (FP16), a 4,000-fold increase over 2020.

global industry vision

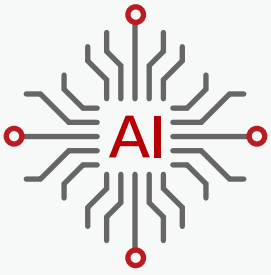
Huawei predicts that by 2030,



Global general computing power (FP32)

will reach **3.3** ZFLOPS,

a **10**-fold increase over 2020.



AI computing power (FP16)

will reach **864** ZFLOPS,

a **4,000**-fold increase over 2020.

References

1 UN, World Population Prospects 2024, https://population.un.org/wpp/Publications/.

2 UN, World Population Ageing 2022, https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

3 WHO, Health workforce, https://www.who.int/health-topics/health-workforce#tab=tab_1.

4 WHO, State of The World's nursing 2020, https://www.who.int/publications/i/item/9789240003279.

5 WHO, World Health Statistics 2024, https://www.who.int/publications/i/item/9789240094703.

6 UN, World Population Prospects 2024, https://population.un.org/wpp/Download/Standard/MostUsed/.

7 China's Health Statistics Yearbook 2022, http://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202305/6ef68aac6bd14c1eb9375e01a0faa1fb.shtml.

8 UN, World Population Prospects 2024, https://population.un.org/wpp/Publications/.

9 UN, World Population Prospects 2022, https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/wpp2022_summary_of_results.pdf.

10 Dariush D. FARHUD, "Impact of Lifestyle on Health," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4703222/#B1.

11 Deloitte, "The Future of Virtual Health,"
https://www2.deloitte.com/cn/en/pages/life-sciences-and-healthcare/articles/the-future-of-virtual-health.html.

12 "Stanford Medicine scientists hope to use data from wearable devices to predict,"
https://med.stanford.edu/news/all-news/2020/04/wearable-devices-for-predicting-illness-.html.

13 "Could an algorithm predict the next pandemic?,"
https://www.nature.com/articles/d41586-022-03358-4.

14 Tsinghua University Press, Intelligent Industry Architecture and Its Implementation in Cities and for Public Utilities.

15 "AI shows promise for breast cancer screening,"
https://www.nature.com/articles/d41586-019-03822-8.

16 "Can AI Fix Medical Records?,"
https://www.nature.com/articles/d41586-019-03848-y.

17 "How AI is improving cancer diagnostics,"
https://www.nature.com/articles/d41586-020-00847-2.

18 Intechopen, "Clinical Decision Support Systems for Diabetes Care,"
https://www.intechopen.com/chapters/84540.

19 "Is the world ready for ChatGPT therapists?," https://www.nature.com/articles/d41586-023-01473-4.

20 "Telemedicine: Past, present, and future," https://www.ccjm.org/content/85/12/938.

21 "Nanotechnology in leukemia: diagnosis, efficient-targeted drug delivery, and clinical trials," https://eurjmedres.biomedcentral.com/articles/10.1186/s40001-023-01539-z.

22 "Revolutionising Medical Imaging with AI and Big Data Analytics," https://openmedscience.com/revolutionising-medical-imaging-with-ai-and-big-data-analytics/.

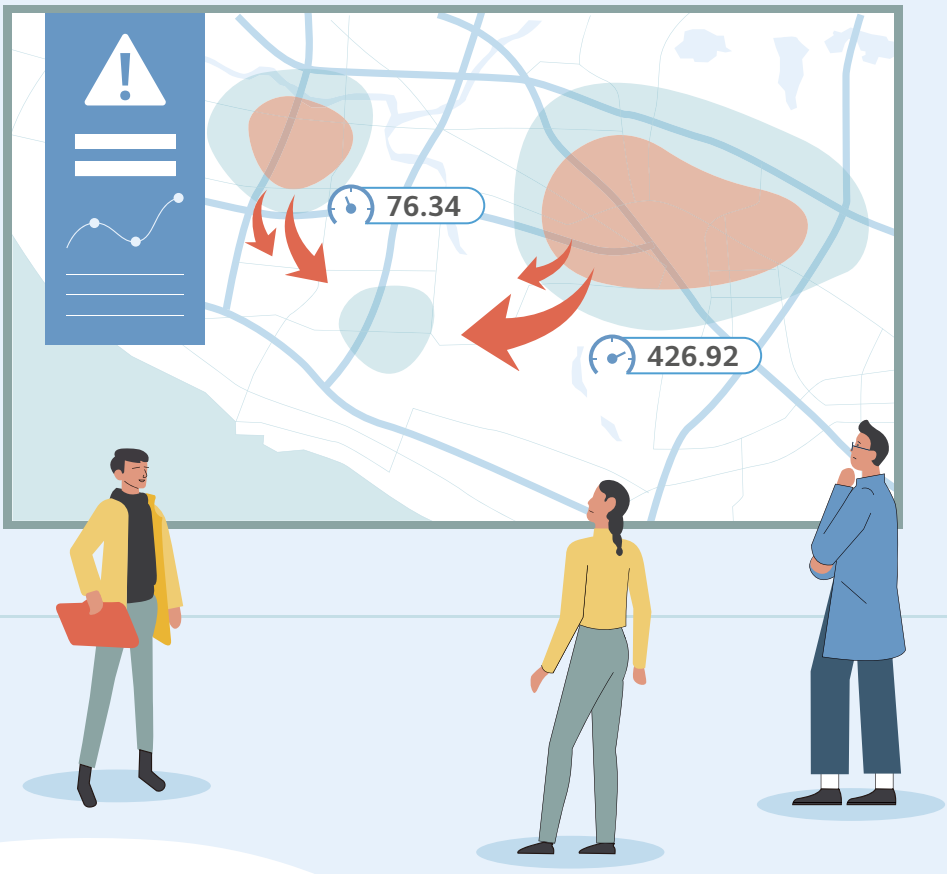
23 "Generative AI in Medical Imaging," https://link.springer.com/article/10.1007/s10916-023-01987-4.

24 "Harnessing AI to see a patient's unique patterns," https://www.nature.com/articles/d42473-023-00384-2.

25 "How AI is accelerating and transforming drug discovery," https://www.nature.com/articles/d43747-023-00029-9.

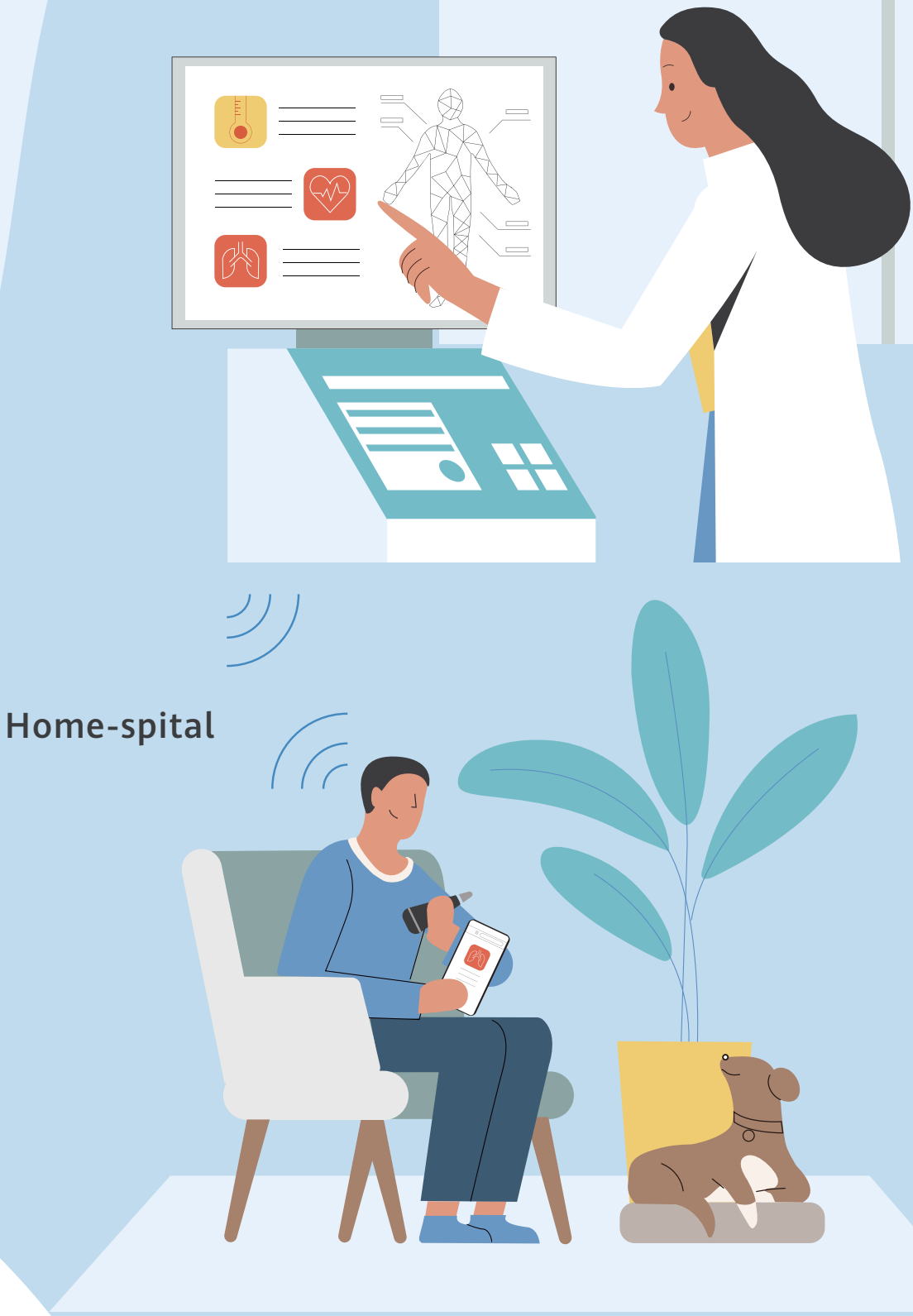
26 "Insilico Medicine launches an AI-assisted decision-making laboratory," https://www.thepaper.cn/newsDetail_forward_25222299.

Proactive prevention



Precision medicine

Home-spital



Food

Data-driven Food Production for More Bountiful, Inclusive, and "Green" Diets

Food is a necessity for everyone, so the UN has made "Zero Hunger" one of its Sustainable Development Goals (SDGs) for 2030. Current estimates show that nearly 690 million people are hungry, and if recent trends continue, the number of people affected by hunger would surpass 840 million by 2030. [1]

The agriculture workforce is shrinking: According to the International Labor Organization, the proportion of the world population working in agriculture dropped from 43.699% in 1991 to 26.757% in 2019. [2]

Arable land per capita is decreasing: According to World Bank data, arable land per capita fell from 0.32 hectares in 1968 to 0.18 hectares in 2021 – a drop of 44%. [3]

Overuse of pesticides is causing severe soil pollution: According to statistics, 64% of global agricultural land (approximately 24.5 million square kilometers) is at risk of pesticide pollution, with 31% at high risk. [4]

Simultaneously, the focus of people's diets worldwide is shifting from "Does this taste good?" to "Is this good for me?" This has resulted in more nutrition and food safety standards. For example, 13,316 food products in China received some kind of green certification in 2018. This number increased to 14,699 in 2019, up 10.4% YoY. [5] This higher demand for green-certified products results in higher requirements on agricultural conditions and technologies.

As we move towards 2030, global food supply faces new challenges and demands. It is clear that technology is key to empowering agriculture, and will help overcome traditional growth constraints, increase food production across the board, and bring "green" food to every table around the world.

Direction for exploration: Guiding cultivation with data, not experience

As the saying goes, there is “a time to plant and a time to pluck up that which is planted ”. Farmers typically rely on calendars to determine the best times to sow. They also rely heavily on personal experience to decide when to sow, fertilize, and use pesticides. However, this leaves a great deal of uncertainty, and whether or not any given year will yield a good harvest is still ultimately up to fate.

Snapshot from the future: Precision farming based on visualized data graphs

Within any single crop field, the moisture content, available nutrients, and crop conditions can vary. But with modern tools like sensors and mobile devices, farmers can remotely and accurately monitor soil moisture, ambient temperatures, and crop conditions in real-time. This makes it possible to flexibly adjust agronomic measures, like sowing, irrigation, fertilizing, and seed adjustment, based on diverse data sets, to better align crops with the available soil. If we look at maize, for example, data-powered adaptive sowing can increase crop yield by 300 to 600 kilograms per hectare of land. [6]

Precision agriculture also relies on in-depth analysis of collected data and cloud-based visualized data graphs. Data from these graphs help farmers make informed decisions regarding soil fertility, water, and nutrient delivery throughout the key stages of crop growth. These graphs can also help farmers better understand information such as local topographical characteristics, climate conditions, and crop diseases or pests so that they can better estimate crop yields, implement agricultural measures, and adjust budgets accordingly.

Furthermore, visualized data graphs can be used to monitor and manage agricultural production in real time, helping farmers respond proactively, quickly, and precisely to changes in their environment. For example, in the event of extreme weather, farmers can use this data to rapidly locate affected areas, develop solutions, and mitigate negative impacts on their yields.

Direction for exploration: Adopting a "factory-like" approach to protect agricultural production from environmental conditions

While precision agriculture is an effective tool for increasing agricultural yields, it is not sufficient to meet the surging food demands caused by population booms, shrinking arable land per capita, pesticide pollution, and worsening climate change.

Under precision agriculture, data is used for analysis and calculation so that the best cultivation solution can be found. However, the environment is constantly changing, meaning data can only be used at the moment it is collected. Therefore, the results of agricultural data analysis cannot be used iteratively.

In addition to precision agriculture, we can adopt a more "factory-like" approach to agriculture, creating "vertical farms" in enclosed environments. Not only can vertical farms be used to collect more data, they can also allow farmers to directly adjust parameters to allow crops to grow in an optimal environment. Both countries with little arable land, like Japan, South Korea, and Singapore, and countries with abundant land resources, like the US, are proactively developing vertical farm technologies.

Snapshot from the future: A new form of agriculture in intelligent vertical farms

Within any single crop field, the moisture content, available nutrients, and crop conditions can vary. But with modern tools like sensors and mobile devices, farmers can remotely and accurately monitor soil moisture, ambient temperatures, and crop conditions in real-time. This makes it possible to flexibly adjust agronomic measures, like sowing, irrigation, fertilizing, and seed adjustment, based on diverse data sets, to better align crops with the available soil. If we look at maize, for example, data-powered adaptive sowing can increase crop yield by 300 to 600 kilograms per hectare of land.

Precision agriculture also relies on in-depth analysis of collected data and cloud-based visualized data graphs. Data from these graphs help farmers make informed decisions regarding

Vertical farms have three main advantages:

- They don't need pesticides or soil, and reduce agricultural water waste: Hydroponics and aeroponics, which are common in vertical farms, utilize solutions that are more efficient at delivering nutrients to plants, with any remaining nutrients being recaptured together with water. These methods use less than one tenth of the water used in traditional agriculture, making the entire process more eco-friendly and reducing pollution.

- They are not affected by climate, providing consistent and ideal conditions for fresh produce: As vertical farms create closed environments, automatic control systems can be used to ensure reliable, large-scale cultivation. This makes it possible to grow vegetables in a wider variety of locations and climates. Vertical farms can be built on rooftops, inside office buildings, abandoned warehouses, or basements, and even in deserts, rivers, or seas.

- They provide smart agricultural models that are globally replicable: The ICT control system and data model used in one vertical farm can be used anywhere in the world to achieve almost the same results. The vertical farm model allows anyone to emulate the environment in which the finest wine-making grapes grow, and even regions that see little daylight can be used to grow sun-loving cherries.

Recent pilot programs for vertical farms have found that, if harvested every 16 days, an area of 7,000 square meters can yield a staggering 900,000 kilograms of vegetables every year. [7]

Conclusion:

Transforming data into food to solve global hunger

In the future, we can use the Internet of Things (IoT) to monitor and analyze soil conditions and crop growth, and to increase yields based on collected data. We can also use historical data to predict changes in the natural environment so as to take proactive intervention measures that reduce the risk of yield reduction. Science-based precision agriculture systems powered by big data, AI, and agricultural know-how will make it possible for farmers to precisely water and fertilize crops and also use drones to more accurately apply pesticides.

Intelligent farming models, such as data-based vertical farms will free agricultural production from the constraints of climate dynamics. These models can be replicated worldwide for more inclusive, green diets.

By 2030, ICT technology will enable us to connect key agricultural production factors, such as farmland, farm tools, and crops, and collect and utilize data on climate, soil, crops, etc., to increase yield. Huawei predicts that by 2030, the data generated worldwide will reach 1 YB each year, a 23-fold increase over 2020. There will be 200 billion connections worldwide, and IPv6 adoption will reach 90%. With the wider application of data in agriculture, we will build a more resilient and greener food system.

global industry vision

Huawei predicts that by 2030,

There will be **200 billion** connections worldwide.

IPv6 adoption will reach **90%**.

1YB of data will be generated annually worldwide, a **23**-fold increase over 2020.

References

1 UN, https://www.un.org/sustainabledevelopment/hunger/

2 International Labour Organization
https://data.worldbank.org/indicator/SL.AGR.EMPL.ZS

3 World Bank,
https://data.worldbank.org.cn/indicator/AG.LND.ARBL.HA.PC

4 Risk of pesticide pollution at the global scale, Fiona H. M. Tang et al.,
https://www.nature.com/articles/s41561-021-00712-5

5 China Green Food Market Research Report 2021 – Industry Trend and Market Prospects, Insight & Info Consulting

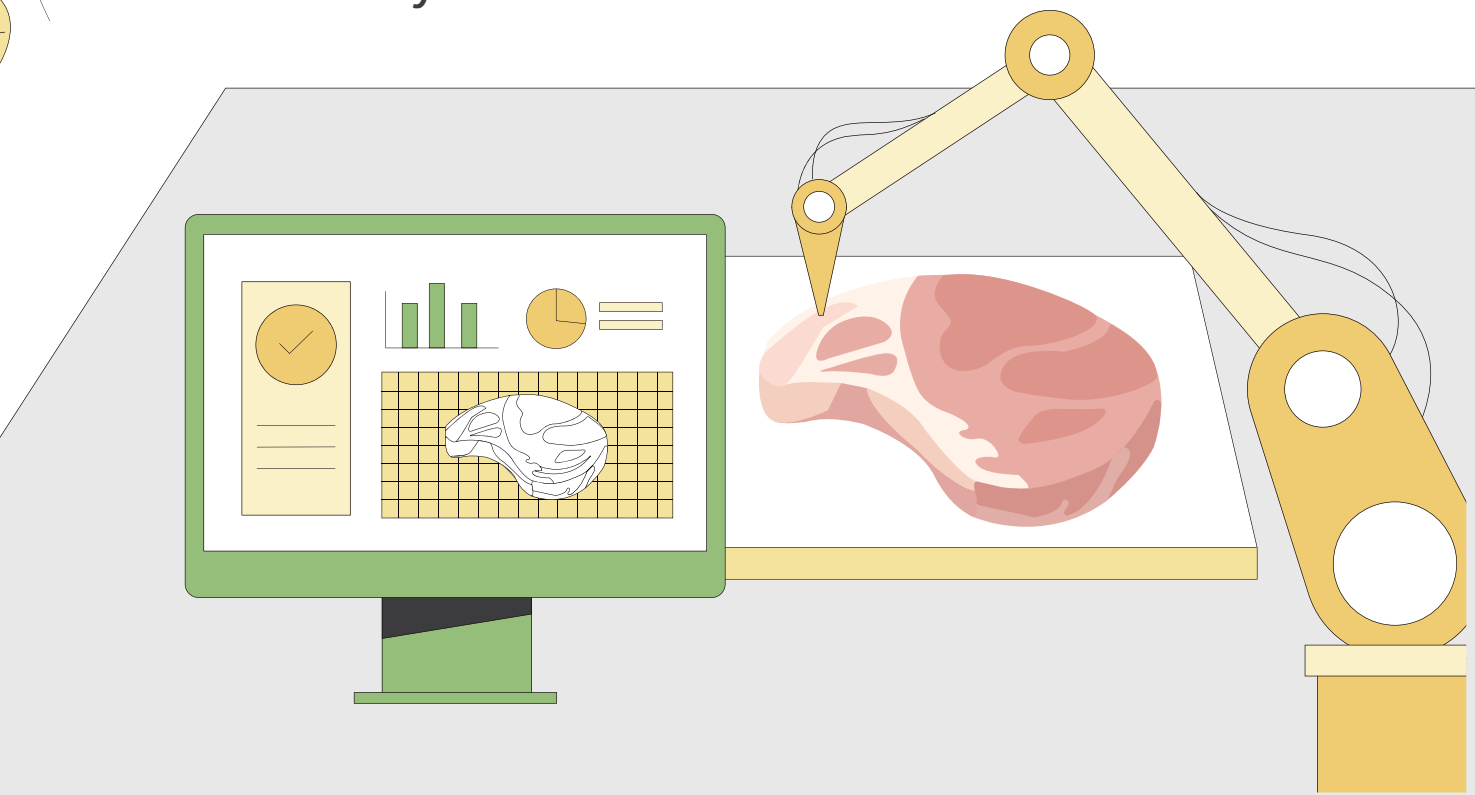
6 Big Data Enables Agricultural Development — Farmers' Harvest Festival & Trade Fair: 10 Popular Case Studies Worldwide, CHINA NEWS PRODUCT NETWORK, https://m.caijing.com.cn/api/show?contentid=3750868

7 AeroFarms, https://www.aerofarms.com/

Precision farming

"Factory-like" agricultural production

New healthy meat



Living Spaces

Personalized Spaces with
Novel Interactive Experiences

The spaces people live in have dramatically changed throughout history. Over time, ancient caves have been replaced by modern buildings, and people are continuing to move from rural communities to dense urban districts. The role of living spaces has also transformed. What used to simply be a shelter from the elements has become the vault for our most precious possessions.

Industrial advancements have led to an abundance of material wealth, and while material possessions bring us joy, they have also filled our living spaces to the brim. The average American home has more than 300,000 items [1], and 1 out of every 10 Americans rents offsite storage [2]. Similarly, a British study found that the average 10-year-old child owns 238 toys but plays with just 12 on a daily basis [3]. While many find buying new objects satisfying, the downsides of hyper-materialism are becoming increasingly apparent to society as a whole. This dilemma opens up new possibilities for future home design.

It has been reported that in 2019, CO_2 emissions from the operation of buildings worldwide reached 10 $GtCO_2$, or 28% of total global energy-related CO_2 emissions [4]. Moreover, the number of new buildings all over the world is set to explode. The International Energy Agency estimates that global building stock will rise by 5.5 billion square meters per year on average until 2050 [5]. The UN Environment Programme's Global Status Report 2017 predicts that the world will add 230 billion square meters (2.5 trillion square feet) in new construction by 2060, which is almost equal to current global building stock. This means an amount of construction equal the size of New York City will be added to the planet every 34 days over the next 40 years [6]. In response, the World Green Building Council has stated that, if we want to meet the goals of the Paris Agreement, all new buildings should operate at net zero carbon by 2030 and all buildings should achieve this by 2050 [7] .

As demand for personalized home experiences continues to rise, ICT-enabled smart home technology is gaining popularity. A survey found that about 80% of millennials and 69.2% of baby boomers are interested in smart home technologies [8]. In the UK, 80% of consumers are now aware of smart home technology and it is second only to mobile payments in consumer awareness of a basket of tech trends. Interoperability has risen as one of the most important buying considerations [9]. Interest in smart living spaces that offer enhanced convenience and safety is also on the rise.

Direction for exploration: New infrastructure provides comprehensive services for communities

Smart doors, smart smoke detectors, falling object alerts, delivery notifications, and many other smart services are becoming increasingly widespread. This means that residents are much more closely connected with their communities and local authorities. In the future, new communities will deliver comprehensive services to residents, powered by the Internet of Things (IoT), 10-gigabit fiber networks, and other new advanced infrastructure. Services such as virtual community events and smart pet management will bring residents and their communities more closely together. Groundbreaking new design concepts will also start changing the way our homes look at the household level.

Snapshot from the future: Digital cataloguing and automated delivery for offsite storage

One potential solution to the overwhelming amount of possessions that now fill households is offsite storage. Some proposed solutions include digitalization and cataloguing of all household items, with technologies like 3D scanning, and then storage in local shared warehouses. This would mean when you decide to go to a party, you can flick through a 3D hologram menu to pick out the dress and accessories that you want. At the touch of a button, those items can be delivered to your door, either by robot [10] or through the building's internal delivery system [11].

A similar system could power a shared "library" of useful household items. For example, in a typical household, electric drills are not needed often, so instead of buying your own, you could search for one in the shared library's online resource catalogue and borrow it for a few days. Automated delivery systems will bring the drill to you and take it back when you are finished.

Direction for exploration: Net-zero-carbon buildings with IoT and intelligent management systems

According to the World Green Building Council, a net zero carbon building is "a highly energy efficient building that is fully powered from on-site and/or off-site renewable energy sources and offsets." Net zero carbon is achieved when the amount of carbon dioxide emissions released on an annual basis is zero or negative. Minimizing building energy consumption through new designs and eco-friendly materials is the first step to achieving net zero carbon [12]. The next step requires not only clean energy sources, but also information and communications technologies (ICT).

One day, net-zero-carbon buildings will be able to automatically interact with their environment through sensors.

- Sensors monitor and generate data about the building in real time, including its environment and condition.
- The Internet of Things connects sensors, cloud-based control systems, and core systems such as lighting, electricity meters, water meters/pumps, heaters, fire alarm systems, and water chillers.
- Intelligent, cloud-based systems utilize sophisticated algorithms and real-time data to automatically decide how the building can minimize energy use. For example, a complete automated system could use IoT devices to check the number of people in a building in real time, and then decide when to switch air conditioners and lights on or off in different parts of the building. Such a system would also be able to manage elevators, hallways, and shutters, depending on actual human activity.

In addition to the environmental benefits, net zero-carbon buildings will also make people's lives more comfortable. Automated systems can keep indoor temperatures at agreeable levels, while soundproofing materials can keep outside noise down to a minimum. There will also be health benefits: Automated systems can decide how much sunlight should pass through a window, to help limit UV exposure, encourage natural sources of vitamin D, support regular sleeping patterns, and combat seasonal affective disorder.

Snapshot from the future: Automated building management systems for museums

Some key museums are already upgrading their energy systems with automated controls. For example, one museum in Australia installed a building management system that constantly monitors 3,000 different indoor environment data points, and automatically adjusts utilities to provide the right conditions for visitors and the objects on display. Heating, ventilation, air conditioning, lighting, and water efficiency have all been upgraded, reducing the museum's GHG emissions by 35%, and electricity costs by 32% [13].

Direction for exploration: Adaptive home environments that understand your needs

Today, we expect more from our homes than ever before. Homes should be more than just a place to live: They should also offer superb experiences. The homes of the future will intuitively understand all of our needs. The moment we arrive home after a long exhausting day, the lights, sound systems, air filters, and television will switch on automatically. When we walk into the kitchen, the refrigerator will push healthy meal suggestions adapted to our personalized dietary needs. In the bedroom, air conditioners will check air quality and automatically adjust temperature and humidity based on what we are doing. From the comfort of the sofa or bed, we will be able to share incredible photos and videos with our loved ones, or finish some mundane paperwork anytime we want. In the event of an emergency, like a fall, where someone is unable to call for help, the home system will come to help by notifying family members, doctors, or security guards.

There are perceptible and imperceptible factors that determine how comfortable our home is. Perceptible factors are those we can instinctively feel, such as temperature, humidity, lighting, ease of access to household items, and ease of information sharing. Imperceptible factors usually include indoor air quality and safety. Intelligent automated systems can enable real-time control of both types of factors.

Snapshot from the future: Whole-house intelligence that understands usage and creates intuitive experiences

Smart home systems collect data from a wide range of smart appliances and sensors, over highly-reliable, high-speed networks that reach every corner of your home. These systems synchronize data on cloud and local storage to make smooth data flow possible. They also use AI engines to determine what is happening in your home and run appropriate applications. The AI engines, in turn, need distributed processing and computing to understand your behavior, indoor environment, and hardware systems, and then make smart decisions to configure your home appliances. These steps could be taken independently or in collaboration with other systems, to meet your needs. When implemented properly, smart home systems deliver immersive, personalized, and intelligent experiences that evolve as your usage needs change.

The variety of smart home appliances we will see in the coming years is expected to explode. They will work together to intelligently anticipate and meet your needs in different situations. Everything, from smart beds and pillows to lights and audio devices, will be able to collaborate. A sleep support solution could easily be created for the bedroom by designing a system that automatically adjusts the softness of your mattress and pillow to suit your body and sleeping habits, and changes your bedroom lighting to stimulate the production of melatonin – the hormone that helps you fall asleep. Bedroom speakers could play music to relax you, and air conditioners could keep track of temperature, humidity, and oxygen levels. Such a system could even identify snoring and curb it by rapidly adjusting the softness of your mattress and pillow. Temperature and humidity regulation could also be achieved to stop you from tossing and turning in bed [14]. In the future, the way we interact with home appliances will also change through touch panels, apps, voice commands, and gestures. Sometimes interactions will be so subtle that we won't even be aware of them.

All members of the family will be able to store videos and photos together in a cloud-connected storage system, and access the system on any device from anywhere. Huawei predicts that by 2030, 35% of homes will use cloud storage. Traditional computers at some homes will become cloud computers and work seamlessly with gadgets and home appliances with screens, delivering a consistent experience across all these devices. Huawei predicts that by 2030, cloud computers will be in use in 17% of homes.

AI-powered home cameras and optical sensing devices will be able to recognize people's movements, and identify if someone takes a fall or is in danger so that help can be quickly notified. Huawei predicts that by 2030, cameras with 3D radar optical sensors will be adopted in 8% of homes to support home nursing while ensuring privacy. These systems could also identify intruders and send alerts to police or security guards. Huawei predicts that by 2030, in China, 24% of homes will be equipped with surveillance cameras; and globally, the percentage is 15%.

Conclusion:

Personalized spaces with novel interactive experiences

In the future, new intelligent infrastructure will play an integral role in smart communities. With the help of ICT technologies, community management systems will aggregate massive amounts of data generated by smart equipment, and use that data to holistically manage how communities operate in real time, delivering superior services to residents.

Net-zero-carbon buildings will be made possible by eco-friendly designs and clean energy sources. Passive design for energy conservation cannot achieve zero carbon goals alone, and energy management systems can contribute to these efforts by effectively managing energy sources and accurately controlling indoor environments to minimize energy consumption.

5G and Artificial Intelligence of Things (AIoT) will help smart home systems autonomously adapt to user needs. These systems will rely on superfast network connections and sophisticated algorithms that enable them to promptly sense user needs and provide intuitive services.

In the future, ICT technologies, especially sensors, IoT, and AI, will reshape our living spaces, from communities to buildings to homes. The result will be safer, personalized living spaces that offer all kinds of new interactive experiences: a space that knows you as well as you know it.

At that time, your home may be full of smart appliances that bring a new level of interactivity to your lifestyle and entertainment. The building you live in may be supported by a great variety of smart control systems, and smart functions may be more widely available in your local community. However, none of this will be possible without connections that deliver high bandwidth and extremely low latency. Huawei predicts that by 2030, there will be 1.6 billion fiber broadband subscribers and 25% of homes will have access to 10-gigabit fiber broadband.

global industry vision

Huawei predicts that by 2030,

1.6 billion

fiber broadband subscribers.

25% of homes will have access to 10 gigabit fiber broadband.

References

1 For many people, gathering possessions is just the stuff of life, Los Angeles Times, https://www.latimes.com.

2 The Self-Storage Self, New York Times, http://nytimes.stats.com/nba/scoreboard.asp.

3 Ten-year-olds have £7,000 worth of toys but play with just £330, The Telegraph, https://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/8074156/Ten-year-olds-have-7000-worth-of-toys-but-play-with-just-330.html.

4 2020 Global Status Report for Buildings and Construction, UN Environment Programme and Global Alliance for Buildings and Construction.

5 Net Zero by 2050: A Roadmap for the Global Energy Sector, IEA, https://iea.blob.core.windows.net/assets/beceb956-0dcf-4d73-89fe-1310e3046d68/NetZeroby2050-ARoadmapfortheGlobalEnergySector_CORR.pdf.

6 Global Status Report 2017, UN Environment, https://www.worldgbc.org/sites/default/files/UNEP%20188_GABC_en%20%28web%29.pdf.

7 WGBC, https://www.worldgbc.org/news-media/every-building-planet-must-be-%E2%80%98net-zerocarbon%E2%80%99-2050-keep-global-warming-below-2%C2%B0c-new.

8 Smart Home Technologies Reshape Real Estate Preferences in 2020, realtor.com, https://www.realtor.com/research/smart-home-tech-2020/.

9 The state of the Connected Home 2021: a year like no other, GfK, https://www.gfk.com/home.

10 2021 Research Report on Commercial Applications of the End-point Delivery Sector Powered by Autonomous Driving in China, iyiou, https://www.iyiou.com/t/yiouzhiku/.

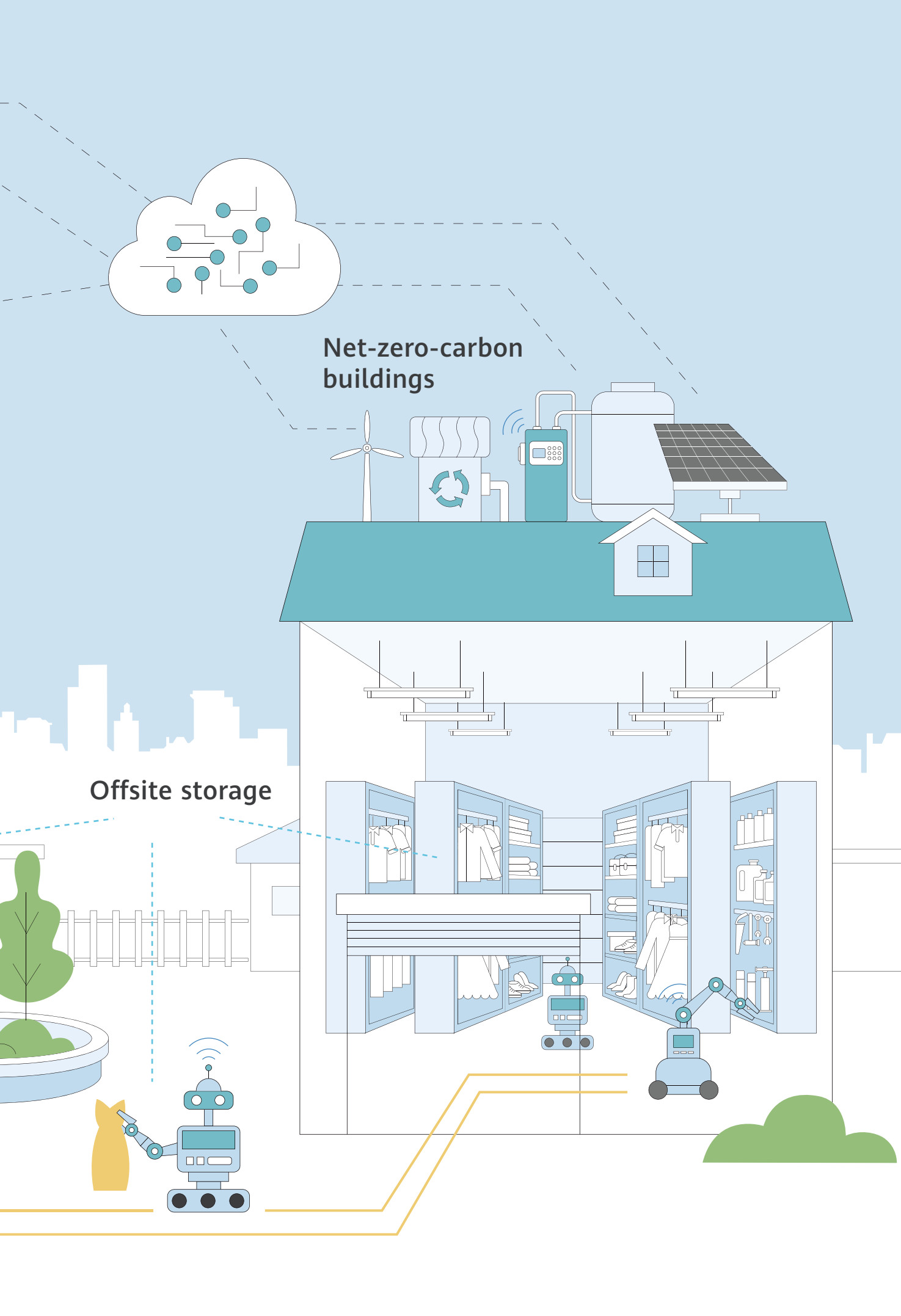
11 Toronto Tomorrow: A new approach for inclusive growth, Sidewalk Labs, https://www.sidewalklabs.com.

12 WGBC, https://worldgbc.org/thecommitment.

13 Siemens, https://new.siemens.com/cn/zh/products/buildings/references/victoria-museum-melbourne.html.

14 Huawei, https://consumer.huawei.com/cn/.


Adaptive home environments
The room temperature is controlled at 25 degrees for you.


Net-zero-carbon buildings
Offsite storage


71
24 °C
Autonomous Mode
21 °C

Transportation

Smart, Low-carbon Transportation Opens up the Mobile Third Space

Travel is an important part of the modern world, with different modes of transport, particularly private cars, becoming increasingly common. For example, the US has long been known as a "nation on wheels", and in 2020, the vehicle-miles traveled within the US totaled 2.83 trillion [1], which is more than 30,000 times the distance between the Earth and the sun. In Europe, vehicles travel more than 12,000 kilometers a year on average.[2] As urban areas continue to expand, more and more commuters face the challenge of a long daily commute. Car-based mobility is set to grow globally by around 70% by 2030 (passenger kilometers, as well as predicted vehicle stock in a business-as-usual use scenario). [3] Existing transportation systems will continue to face many challenges.

More congestion, lower efficiency: Traffic congestion is an increasingly serious and frequent problem worldwide. The average person wastes at least 15 minutes every day in traffic jams. In Colombia's capital Bogota, the world's most congested city, drivers wasted an average of 272 hours, the equivalent of more than 11 days, in traffic jams in 2018. [4] Congestion also causes significant economic losses. For example, congestion caused losses of US$70.4 billion in the US in 2023. [5]

The huge amount of travel also has a significant impact on the environment: According to the International Energy Agency (IEA), the transportation industry accounted for 26% of global carbon emissions in 2020, and this figure far exceeded the emissions from the

manufacturing and construction industries. The adoption of circular economy practices combined with accelerated electrification in the automotive industry has the potential to reduce carbon emissions by up to 75% and resource consumption by up to 80% per passenger kilometer by 2030. [6] This will accelerate low-carbon development.

The transportation systems of the future will not only be congestion-free and low-carbon, but also hassle-free. Travelers will no longer need to think about their routes, but only about what to do during their journeys. They will enjoy a quiet and personal space where they can watch TV, concentrate on work, or even relax with an in-seat massage. Commuters will no longer feel exhausted at the end of their journeys, and will instead feel a sense of satisfaction for having made the most of their time.

In order to realize these changes, transportation systems will need to be further upgraded. All of the key elements (vehicles, traffic lights, pedestrians, etc.) need to be connected through ICT so that each phase of a journey can be automated.

▶ Direction for exploration: How digital and intelligent comprehensive transportation systems improve travel experiences and reduce logistics costs

In the future, comprehensive transportation systems that integrate and collaboratively schedule different modes of transport will significantly improve passenger experiences and greatly reduce the overall cost of cargo transportation.

Technologies like big data, AI, and IoT will be more widely used in these transportation systems. Intelligent logistics distribution systems, self-driving transportation vehicles, and traffic management platforms that provide real-time updates will emerge to further improve transportation and logistics efficiency and quality.

Favorable policies will also be released to support the development of these transportation systems, such as those that promote cross-department and cross-region cooperation; improve the planning of infrastructure construction; improve relevant laws, regulations, and standards; and attract more investment from the private sector.

To summarize, the comprehensive transportation systems of the future will be integrated, intelligent, and efficient, and will enable unprecedented convenience and opportunities which will improve people's lives, drive economic growth, and create a more prosperous and sustainable future for everyone.

Snapshot from the future: Centralized and collaborative scheduling for transportation hubs

In the future, we will see increased sharing and integration of data between transportation hubs like ports, airports, and railway stations. By accurately analyzing different types of information, such as passenger traffic, cargo volumes, and weather changes, these hubs will be able to plan resource allocation in advance and schedule resources in a centralized and collaborative manner.

For example, ports will be able to dispatch ships and cargo more efficiently and align these dispatches with nearby railways and highways, thereby improving the efficiency of logistics and operations. Airports will be able to optimize processes for flight take-off and landing, transit, and passenger services, and connect to nearby highways and railways. This will reduce delays and allow passengers to depart and arrive more easily. Railway stations will be able to improve train timetables and enable quick transit to nearby transportation facilities, providing passengers with smoother travel experiences.

To summarize, transportation nodes will no longer be isolated, but closely connected through multimodal transportation and centralized, intelligent scheduling. Ports will no longer be just places for cargo loading and unloading, but will be deeply integrated with other industries, realizing collaborative development between ports, industries, and cities. Airports will not just provide air transportation, but will be further integrated into urban infrastructure and provide complementary functions to nearby urban facilities. Railway stations will be able to intelligently schedule resources and enable seamless transfer to other modes of transport, making transit more convenient and efficient for passengers. These integrated and collaborative systems will make more efficient use of resources, reduce redundancy and waste, and lower overall logistics costs.

Snapshot from the future: Intelligent connected vehicles and vehicle-road-cloud integration

Currently, many countries around the world are striving to develop intelligent connected vehicles in order to capitalize on opportunities in this relatively new industry. China was the first to publish a strategy for vehicle-road-cloud integration, and has been leveraging its advantages in cross-industry collaboration, infrastructure construction, and ICT to facilitate the high-quality development of intelligent connected vehicles. Integrated vehicle-road-cloud systems use next-generation ICT to merge the physical and virtual worlds, including people, vehicles, roads, and cloud. These systems provide functions like systematic and collaborative sensing, decision-making, and control to enable the secure, efficient, eco-friendly, and stable operations of intelligent connected vehicles and transportation systems. Integrated vehicle-road-cloud systems include all road users, vehicles, roadside infrastructure, cloud-controlled platforms, supporting platforms, and communications networks.

These systems will completely change our perceptions of transportation and travel. In these systems, vehicles will no longer run independently, but act as intelligent units closely connected to road infrastructure and cloud platforms. Vehicles will be able to obtain information about road conditions and traffic signals in advance through real-time communications with other vehicles and with roads, so that they can optimize their routes and speeds to improve travel efficiency and reduce congestion. Road infrastructure will be equipped with lots of sensors and intelligent devices to monitor road conditions and traffic in real time, and these devices will quickly transfer this data to the cloud. Cloud platforms will then use their powerful computing capabilities and data analytics technologies to process and analyze the data, inform the decision-making of traffic management agencies, and enable intelligent traffic control.

As a key part of the vehicle-road-cloud systems, autonomous driving technologies will be more widely used. Guided by intelligent systems, self-driving vehicles will be able to run more safely and efficiently, reducing the pressure on drivers and the risk of traffic accidents. Vehicle-road-cloud integration will also improve the collaboration between transportation and other domains, like energy and urban planning. For example, the real-time adjustment of street lamp brightness based on traffic flows can save energy and reduce carbon emissions. The optimization of road layouts based on urban development plans can drive the sustainable development of cities.

In conclusion, vehicle-road-cloud integration will bring unprecedented changes to highways, make transportation more intelligent, efficient, safe, and eco-friendly, and deliver better travel experiences for everyone.

Direction for exploration: Electric vehicles for green transportation

As transportation consumes increasing amounts of energy, many countries and regions around the world are making efforts to promote low-carbon travel. Energy conservation and emissions reduction in transportation are the key to carbon neutrality. In July 2021, the European Commission officially launched the European Green Deal, specifying the goal of reducing greenhouse gas emissions to 55% below 1990 levels by 2030, and achieving carbon neutrality by 2050. The Deal also set a target to reduce carbon emissions from transportation to 90% below 2021 levels by 2050. Land transportation is the main focus of this policy, as it makes up 20.4% of the EU's greenhouse gas emissions. The EU and the UK have announced plans to accelerate the deployment of renewable energy (Climate Action Tracker, 2022 [54]). The European Commission's RepowerEU plan proposed an increase of the target proportion of renewable energy in the overall energy mix by 2030 from 40% to 45%. (EC, 2022 [30]).

To save energy and cut emissions from land transportation, countries are working tirelessly to develop vehicles that use new energy, including pure electric vehicles (EVs), plug-in hybrid EVs, and fuel-cell vehicles. Many countries and regions have set schedules for reducing the number of internal-combustion-engine vehicles (ICEV) on their roads. For instance, the EU has set the goals of reducing emissions from passenger cars and vans by 55% and 50%, respectively, by 2030 (the previous goals were 37.5% and 31%, respectively). It has also set a new goal of ensuring that by 2035, all new vehicles sold will be zero-emission vehicles, which is equivalent to banning the sales of ICEVs from 2035 onwards. [7] Japan set the goal that 50% to 70% of the cars on Japanese roads should be next-generation vehicles by 2030. [8] China has proposed to start phasing out ICEVs in 2030. [9]

Snapshot from the future: New energy for green mobility

Some countries have made significant progress in adding EVs, such as buses and taxis, to their urban public transportation systems. For example, by 2017, all of Shenzhen's 16,000 buses were electric. This made Shenzhen the world's first city with an entirely electric bus fleet. [10] In Europe, EVs make up over 78% of Denmark's new buses, and about two-thirds of new buses in Luxembourg and the Netherlands are zero-emission. [11]

There are two reasons behind these rapid advancements in public transportation. First, public transportation vehicles are replaced relatively frequently, which provides the opportunity to plan and implement the deployment and adoption of new energy vehicles. Government subsidies and efficient O&M solutions can also reduce the operating costs of electric fleets to levels close to or even lower than those of conventional vehicles, and this reduces the barriers to introducing new energy vehicles.

Second, these publicly-owned vehicles are centrally stored and maintained in specialized facilities that can easily be upgraded into multi-functional spaces with charging piles for EVs. Therefore, lack of charging facilities is not a major obstacle for the electrification of public transportation.

Public transportation vehicles also travel longer distances every day and generate more carbon emissions than private vehicles. Therefore, the wide adoption of electric public buses and taxis is an effective and efficient way to reduce vehicle emissions. In Beijing, for example, 71,000 private pure EVs saved 89 million liters of gasoline and 199,000 tons of CO_2 emissions in 2018. By contrast, just 9,400 electric taxis in Beijing made a similar contribution, saving 65 million liters of gasoline and 145,000 tons of CO_2 emissions. [12]

According to the IEA, although the global automotive market shrank by 16% in 2020, following the COVID-19 pandemic, the number of newly registered EVs hit a new high of 3 million, up 41% from the previous year. There were more than 10 million EVs in use worldwide in 2020, and that strong momentum has continued. The sales of EVs in the first quarter of 2021 were almost 1.5 times higher than in Q1 2020. Moreover, consumer spending on EVs grew by 50% in 2020, reaching US$120 billion, while government subsidies were just US$14 billion. Government subsidies as a percentage of the total spending on EVs have declined for five years in a row. This indicates that although government subsidies played an important part in stimulating demand, the sales of EVs are now increasingly driven by consumer choice.

By 2030, it is predicted that there will be 145 million EVs in use around the world, including cars, vans, heavy trucks, and buses. If governments accelerate efforts to achieve the global climate and energy goals, the global EV fleet could even reach 230 million vehicles by 2030.[13] The IEA also predicts that more than 300 million new energy vehicles will be in use by 2040, and this will reduce oil consumption by 3 million barrels per day. [14]

Snapshot from the future: New energy aircraft trials

In an effort to reduce pollution, protect the environment, and cut O&M costs, the aviation industry is actively developing new energy aircraft, both small planes and larger passenger aircraft. The continual development of urban air mobility (UAM) is also driving the aviation industry to use more electric power, setting it on a greener path.

The aviation industry accounted for about 2% of global anthropogenic CO_2 emissions in 2022. If such emissions are not effectively curtailed, this percentage is expected to increase to 25% by the middle of this century. [15]

The industry's costs of global maintenance, repair, and operations (MRO) were US$69 billion in 2018, accounting for 9% of airlines' total operating costs. Maintenance costs for engines accounted for 42% (US$29 billion) of the total MRO costs. The global aviation industry spent US$215 billion on fuel in 2023, and this made up 30% of their total operating costs.

At present, three main types of new energy aircraft are being developed: hybrid-electric, pure electric, and hydrogen-powered. In addition to increasing energy efficiency and reducing pollution and noise, new energy aircraft also present an opportunity to trial new designs, such as the blended wing body design. This design can significantly reduce aircraft's drag and energy consumption and improve their flight performance. In addition, this design can increase the amount of space in the cabin, which is a very valuable upgrade, as it increases aircraft's carrying capacity.

In June 2020, France announced that it would invest EUR1.5 billion in the development of large new energy passenger aircraft. France plans to complete the maiden flight of their new energy aircraft before 2035, and expects more than 1,300 companies working in the aviation industry to participate in this project. For this, France has drawn up a clear roadmap which starts with a revamp of Airbus's A320 product line to develop a hybrid electric "successor" model to the A320. The prototype of this new model will be unveiled between 2026 and 2028 and make its maiden flight by 2035.

Direction for exploration: Autonomy opening up the mobile third space

From horse-drawn carts to modern cars, people have developed many generations of vehicles to help us move faster and farther than we could ever run. However, the upcoming autonomous driving era means that vehicles will soon have their own "brains". Autonomous driving technologies will reshape our travel experiences, and transform business models in the transportation industry.

As AI will be making decisions for vehicles, drivers will no longer have to keep their hands on the wheel, their feet on the pedals, or their eyes on the road. This will give rise to a whole new sector of mobile entertainment, social interaction, shopping, and remote work on the go, turning vehicles into our mobile third spaces.

The US Society of Automotive Engineers and the US National Highway Traffic Safety Administration classify autonomous driving technologies into six levels – L0 to L5. Specifically, L0 refers to traditional human driving with no automation; L1 to L3 include AI-assisted driving with low or moderate automation; and L4 and L5 represent vehicles that can be entirely controlled by AI systems, with no human operation needed.

The development of autonomous driving technologies involves cross-industry collaboration between many industries such as ICT, manufacturing, and transportation. This can in turn stimulate economic growth. When deployed at scale, autonomous driving will significantly improve road safety and transportation efficiency, and create positive socioeconomic benefits through energy conservation and emissions reduction.

Autonomous driving is expected to create about US$3.2–6.3 trillion in economic benefits for the US by 2050, leading to nearly US$800 billion in annual social and consumer benefits. [16]

In the report On the road to automated mobility: An EU strategy for mobility of the future, the European Commission set out the goal of making self-driving vehicles commonplace across the EU by 2030. The report stated that driverless vehicles, when fully integrated into the wider transportation system, would play a significant role in achieving Vision Zero, i.e. no road fatalities on European roads by 2050. [17] In China, 11 ministries, including the National Development and Reform Commission, jointly issued the Smart Car Innovation and Development Strategy in February 2020. The strategy set the goal of achieving large-scale production of vehicles ready for autonomous driving and commercialization of high-level autonomous vehicles for specific scenarios by 2025, as well as formulating a mature and standardized intelligent vehicle system by 2050. [18] In December 2023, China's Ministry of Transport issued a document titled Guidance for Road Engineering Infrastructure to Support Autonomous Driving Technologies, which stated that road infrastructure must support the development of autonomous driving. In January 2024, China's Ministry of Industry and Information Technology, Ministry of Transport, and three other ministries jointly issued the Notice on Piloting Vehicle-Road-Cloud Integration for Intelligent Connected Vehicles, which further strengthened policy support for autonomous driving and set guiding principles for vehicle-road-cloud integration.

Snapshot from the future: Autonomous driving and vehicle-road-cloud synergy in the fast lane

As we approach 2030, we are seeing self-driving vehicles go from the lower L2 and L3 levels of automation to the higher L4 and L5 levels of automation. We also expect to see more breakthroughs in autonomous driving technologies. As technologies such as AI, big data, cloud computing, and sensing continue to evolve, self-driving systems will be able to sense and make decisions faster and more accurately. This means that self-driving vehicles will be able to operate more safely and efficiently in more complex traffic environments. Buses, taxis, low-speed logistics services, and industrial transportation (logistics and mining) are likely to be the first commercial applications of autonomous driving.

The development of autonomous driving technologies relies on intelligent individual vehicles and vehicle-road-cloud synergy, which poses higher requirements on computing chips, 24/7 all-weather sensing, cloud computing, big data analytics, low-latency and high-reliability wired and wireless networks, and AI algorithms. An intelligent vehicle needs to be able to make automotive-grade driving decisions, such as emergency collision avoidance and unmanned driving. Vehicle-road-cloud synergy needs to provide functions that cannot be provided by the intelligent vehicle, such as safety beyond line-of-sight and early access to road network control information. Together, they are driving the universal adoption of autonomous driving.

Low- and medium-speed public roads: Self-driving vehicles have delivered positive results in fields such as logistics and distribution, cleaning and disinfection, and patrolling. Unmanned vehicles for logistics and distribution can successfully drive at low speeds on roads with less complicated conditions. This means they can provide safe unmanned delivery services on public roads. Low-speed unmanned vehicles have provided valuable support during the fight against COVID-19, especially in the transportation and distribution of medical supplies, cleaning and disinfection, patrolling, and temperature checks. These vehicles have proved their practical value, laying a foundation for their adoption in other markets.

High-speed semi-closed roads: Heavy trucks are expensive, so the price of sensors is not a limiting factor. Sensors such as lidar can be installed in these trucks for better sensing of their environment. Heavy trucks are mainly used in high-speed cargo transportation, ports, and logistics parks, which means the driving environment is less complex and routes are generally fixed. Heavy trucks are rarely seen on complex urban roads. This means that the driving environment that autonomous driving systems have to handle is not particularly complex. Truck drivers are expensive, and they frequently breach rules by overloading their vehicles and working overtime. Autonomous driving of heavy trucks would quickly help industries cut costs and work more efficiently, making this a compelling business case. According to a Deloitte report on smart logistics in China, technologies like unmanned trucks and artificial intelligence will mature in a decade or so, and will be widely used in warehousing, transportation, distribution, and last mile delivery. [20]

Special non-public roads: Autonomous driving is playing an increasingly important role in environments like mines and ports. Some companies are working with ports to test self-driving container trucks. We have already seen unmanned trucks working in multiple fleets and

even during night shifts at mines. Since 2023, 92 intelligent guided vehicles (IGVs) have been operating autonomously at the Second Container Terminal of Tianjin Port in China. This was made possible thanks to 5G, the Beidou navigation system, and automatic driving technologies. We are expected to see autonomous horizontal transportation at 30% of terminals by 2030.

With its high level of safety and efficiency, autonomous driving will first demonstrate its commercial value in mining. While working autonomously, many mechanical vehicles, such as mining trucks, excavators, and bulldozers can work together. In the past, one driver took care of each mining truck, but now a commander will take care of an entire group of trucks. In the event of a fault or danger, the commander can remotely pilot the vehicle to a safe area from the control center, and send warnings to nearby vehicles.

Public roads: Autonomous driving technologies can make driving safer for the general public and help local authorities manage roads more efficiently. For example, they can quickly detect traffic incidents and access the relevant information, issue warnings about secondary accidents, select better routes to avoid traffic jams, send traffic alerts to vulnerable road users, and provide information about construction sites and other areas with temporary traffic controls. Autonomous driving can significantly reduce the number of traffic accidents.

Autonomous driving technologies will lead to more innovative changes in the designs of car bodies. All possible configurations will be explored and exploited. Cars can become the mobile third space, catering to many different scenarios. This will give rise to new business models in industries like catering. Self-driving food trucks may become the standard of the future, and dinner with friends and family may take on a whole new form: After you book a lunch, a self-driving food truck will pick you up and carry you along whatever scenic route you choose. You can enjoy the views while dining and chatting, all within a private space. This model would eliminate the need to visit restaurants and ensure privacy during the meal. For a restaurant, the size of the premises would no longer be a limiting factor for the size of its business, and location would no longer restrict the clientele it could attract. Business results could be disconnected from footfall.

Snapshot from the future: Urban air mobility

In the future, airspace will become an important resource for urban transportation. An efficient air-based urban transportation network will significantly reduce congestion and travel times, and improve the efficiency of logistics and emergency services. Urban air mobility (UAM) and electric vertical takeoff and landing (eVTOL) aircraft have emerged as promising fields in the global aviation industry.

As a new productive force, the low-altitude economy presents huge potential and could grow to be worth trillions of yuan [21]. As an important enabler of the low-altitude economy, eVTOL is expected to be commercially deployed at a faster rate.

The research and development of eVTOL aircraft has attracted investment from innovative companies around the world, and their performance has seen solid improvements. Currently, the five-seat aircraft which are being manufactured by several companies have a cruising range of about 250 kilometers. Some companies are working on eVTOL aircraft with seven seats or more. [22] Some are exploring hydrogen-fueled aerial vehicles [23] for longer ranges (more than 600 kilometers). These new aircraft may be used in various scenarios, including emergency medical services, urban air mobility, regional air mobility (RAM), air freight transportation, and personal aircraft.

Air emergency rescue systems: Over the past decade, skyscrapers have sprung up in major cities around the world. The number of skyscrapers will continue to grow over the next decade as global urbanization continues. The rapid rise of skyscrapers may make for impressive skylines, but they also create safety risks. Providing firefighting and emergency medical services to people in skyscrapers will be a new challenge for cities. Air emergency rescue offers a new solution to these challenges. It allows firefighters and medical personnel to get to the higher floors faster so that they can better protect people and property within the critical timeframe.

Air metro/air taxis: Convenient and efficient transportation is one of the core needs of urban residents. Batteries with higher energy density are enabling electric aircraft to work for longer periods of time and have a larger capacity. eVTOL will prove to be an effective tool to improve the urban transportation experience. Pilot projects have begun for air passenger transportation services and significant progress has been made so far.

Air express: UAV express delivery is fast, efficient, and flexible, and it can reduce delivery times and improve service experience, especially when it comes to deliveries during emergencies, in remote areas, or in other special circumstances. Technologies like big data, AI, and 5G-A powered harmonized communication and sensing can enable multi-source data fitting, airspace situational awareness, black fly identification and warning, and low-altitude route planning. High-precision and grid-based low-altitude airspace computing can help achieve fine-grained low-altitude airspace management, properly plan and dynamically manage the flight paths of UAVs, and support large-scale, high-density, and complex low-altitude operations. At the same time, technologies like 5G-A, radar, and radio detection can enable real-time and continuous

situational awareness during low-altitude flights and promptly identify and combat black flies to ensure low-altitude security in cities.

Advances in battery technologies have dramatically increased the range and capacity of eVTOLs. The latest lithium-ion and solid-state battery technologies, for example, allow eVTOLs to fly farther and carry more passengers and cargo. Several cities around the world have already run pilot projects for urban air mobility services.

In 2019, a Chinese tech company launched the world's first urban air mobility service in Zhejiang, cutting road trips that normally took 40 minutes to a five-minute air hop. [24] In 2023, a Chinese tech company based in Shanghai announced that its eVTOL aircraft had set a global record for a two-ton eVTOL by flying 250.3 kilometers on a single charge. During the 2024 Olympic Games in Paris, a German tech company provided trial urban air mobility services. On February 27, 2024, a Chinese tech company conducted China's first ever cross-city and cross-sea eVTOL demonstration flight from Shenzhen to Zhuhai. After flying about 55 kilometers over 20 minutes, the aircraft successfully landed at Jiuzhou Port in Zhuhai. According to NASA's projections, air metro will support 740 million passenger trips by 2030.

Of course, such travel requires fast and stable space-air-ground integrated networks and positioning systems, cost-effective and reliable visual sensors and lidars, secure and stable automatic flight algorithms, and efficient, real-time command and dispatch platforms.

By 2030, airworthiness regulations, infrastructure, and air traffic control systems will be established for low-altitude transport, and technological advances will make aircraft safer and more reliable and increase their range. Therefore, low-altitude transport will emerge as one of the main modes of transport.

Direction for exploration: Connected vehicles for safer, faster, and larger-scale autonomous driving

Safety, operation design domains (ODDs), and affordability are three key but conflicting factors affecting the scaling of autonomous driving. ODD refers to the operating conditions under which a given driving automation system is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics. To improve the safety of autonomous driving, it is necessary to set limitations on ODDs and optimize the system so that it strives to approach its upper limits. This is the only way to commercially deploy autonomous driving on a small scale. Another option is to use more expensive equipment to improve the safety of each intelligent, self-driving vehicle, but this will make self-driving less affordable. To commercialize autonomous driving on a large scale and increase the level of automation, we need to strike a balance between safety, ODD restrictions, and affordability.

Vehicle-infrastructure cooperative autonomous driving (VICAD) is a promising solution. It equips each vehicle with a perfect perspective to ensure safety and also helps efficiently allocate road resources, so that every part of the transportation system runs efficiently and collaboratively. To make VICAD a reality, we need to integrate advanced technologies such as sensing, computing, communications, and decision-making controls and build a closed-loop enablement system that can connect the digital space and physical space and offer situational awareness, real-time interactions, well-informed decision making, and precise execution based on data flows [27].

Connecting vehicles requires continuous network coverage. Currently, global mobile communications services cover only about 20% of the total land area (land only covers 29% of the earth's surface), and less than 6% of the earth's surface. For example, more than 95% of the sea area of China is not covered by terrestrial mobile communications networks. [28] So a space-air-ground integrated network is needed to provide continuous coverage. As in-vehicle and in-flight entertainment on large screens and holographic conferences are becoming more popular, terrestrial networks alone will not give users the consistent experience they demand for entertainment and work. A space-air-ground integrated network will be needed to provide large bandwidth and high availability.

Snapshot from the future: Safer, more efficient dispatch services

Over the past decade, pioneers have begun exploring the use of elevated rails to transport containers in busy ports. Containers are sent to rails similar to cable railways. The railway system dispatches the containers based on their destination and sends them to railway stations, truck warehouses, or even waterless ports in inland cities. This makes container transportation much faster at a very low cost. In the future, we will see a comprehensive transportation system that supports the coordinated scheduling of different modes of transport. This system will ensure smooth traffic, speed up the distribution of goods, and drive the development of port-related industries. When this system is up and running, transport facilities will be fully connected and

different modes of transport will work together seamlessly, which will help boost logistics efficiency, form industry clusters, and drive urban development. In other words, ports, industries, and cities will work more closely than ever for shared success.

Snapshot from the future: Broadband in the air, just as at home

Moving forward, broadband coverage will extend beyond the ground into the air and beyond. Broadband connections will be available to devices at various heights, such as drones less than 1 kilometer above the ground, aerial vehicles 10 kilometers above the ground, and low-orbit spacecraft hundreds of kilometers above the ground. The integrated network will consist of small cells covering hotspots within a radius of 100 meters, macro cells with a radius of 1 to 10 kilometers, and low-orbit satellites with coverage over a radius of 300 to 400 kilometers, providing users with unbroken access to broadband of up to 10 Gbit/s, 1 Gbit/s, and 100 Mbit/s, respectively. [29]

Conclusion:

Smart, low-carbon transport opens up the mobile third space

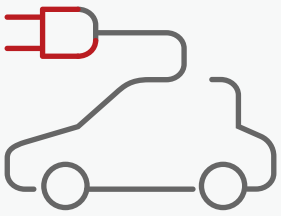
In the future, transport will be a multi-dimensional and innovative system. The shift to electric, autonomous, shared, and connected vehicles will create an intelligent, convenient, low-carbon transport experience. To make this happen, we need innovative applications of new energy technologies; secure and stable autonomous driving algorithms; cost-effective, reliable sensors; a high-speed, stable space-air-ground integrated network; and a traffic management brain based on great computing power.

The mobile third space will reshape the transport experience, incubate innovative mobility services, and drive the emergence of new business models. An intelligent urban transport management system can optimize resource allocation, enable more efficient sharing of transport resources, alleviate traffic congestion, and reduce the environmental pollution caused by traffic. This is how we will resolve the conflict between the surging demand for transport and the urgent need to decarbonize.

Huawei predicts that by 2030, 82% of new vehicles sold in China will be powered by new energy, and 30% of these new vehicles will come with L3 or higher levels of autonomous driving systems. In addition, by 2030, the whole-vehicle computing power will exceed 5,000 TOPS, and 60% of new vehicles sold will support C-V2X.

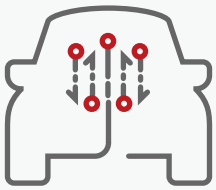
global industry vision

Huawei predicts that by 2030,



82% of new vehicles sold will be electric vehicles.

30% of new vehicles sold in China will be L3 and above autonomous vehicles.



Whole-vehicle computing power will exceed **5,000** TOPS.

60% of new vehicles sold will support C-V2X.

References

1 U.S. Department of Transportation Federal Highway Administration, https://highways.dot.gov/.

2 ACEA, https://www.acea.auto/fact/passenger-cars-what-they-are-and-why-they-are-so-important/.

3 Commuting Monitoring Report of Major Cities in China 2021, China Academy of Urban Planning and Design, http://www.chinautc.com/upload/fckeditor/2021tongqinjiancebaogao.pdf.

4 Raising Ambitions: A new roadmap for the automotive circular economy, WEF, http://www3.weforum.org/docs/WEF_Raising_Ambitions_2020.pdf.

5 The INRIX Global Traffic Scorecard, INRIX, https://static.poder360.com.br/2019/02/INRIX_2018_Global_Traffic_Scorecard_Report__final_.pdf.

6 INRIX, https://inrix.com/press-releases/2019-traffic-scorecard-us/.

7 Raising Ambitions: A new roadmap for the automotive circular economy, WEF, http://www3.weforum.org/docs/WEF_Raising_Ambitions_2020.pdf.

8 A European Green Deal, EU, https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en.

9 Strategies of the Next Generation Vehicles (NGV) in Japan, Chowdhury Mahbubul Alam, https://www.kitakyu-u.ac.jp/law/kenkyu/pdf/46-3_4choedhry2.pdf.

10 Study on the Chinese Traditional Fuel Vehicle Exit Schedule, Innovation Center for Energy and Transportation, http://www.icet.org.cn/admin/upload/2019052339423961.pdf.

11 Shenzhen achieved 100% pure electric bus fleet in September, Shenzhen Transport Bureau, http://jtys.sz.gov.cn/zwgk/jtzx/gzdt/content/post_4317723.html.

12 Denmark, Luxembourg, Netherlands lead the way on emissions-free buses, Pierre Dornier, https://www.transportenvironment.org/press/denmark-luxembourg-netherlands-lead-way-emissions-free-buses.

13 Annual Report on the Big Data of New Energy Vehicles in China, https://www.ssap.com.cn/c/2018-11-23/1073981.shtml.

14 Global EV Outlook 2021, IEA, https://www.iea.org/reports/global-ev-outlook-2021.

15 World Energy Outlook, IEA, https://www.iea.org/topics/world-energy-outlook.

16 Exploration on how to achieve carbon neutrality in the air transport and aviation manufacturing industry, Yu Zhanfu, http://att.caacnews.com.cn/zsfw/ysfw/202108/t20210811_58483.html.

17 America's Workforce and the Self-Driving Future Realizing Productivity Gains and Spurring Economic Grow, SAFE, https://avworkforce.secureenergy.org/wp-content/uploads/2018/06/Americas-Workforce-and-the-Self-Driving-Future_Realizing-Productivity-Gains-and-Spurring-Economic-Growth.pdf.

18 On the road to automated mobility: An EU strategy for mobility of the future, EU, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2018:0283:FIN:EN:PDF.

19 Strategy for Innovation and Development of Intelligent Vehicles, National Development and Reform Commission of China, http://www.gov.cn/zhengce/zhengceku/2020-02/24/content_5482655.htm.

20 China's Logistics Industry, Deloitte, https://www2.deloitte.com/global/en.html.

21 Implementation Plan for Innovative Application of General Aviation Equipment (2024-2030), www.gov.cn.

22 Lilium, https://lilium.com/.

23 Skai, https://www.skai.co/.

24 EHang, https://www.ehang.com/cn/.

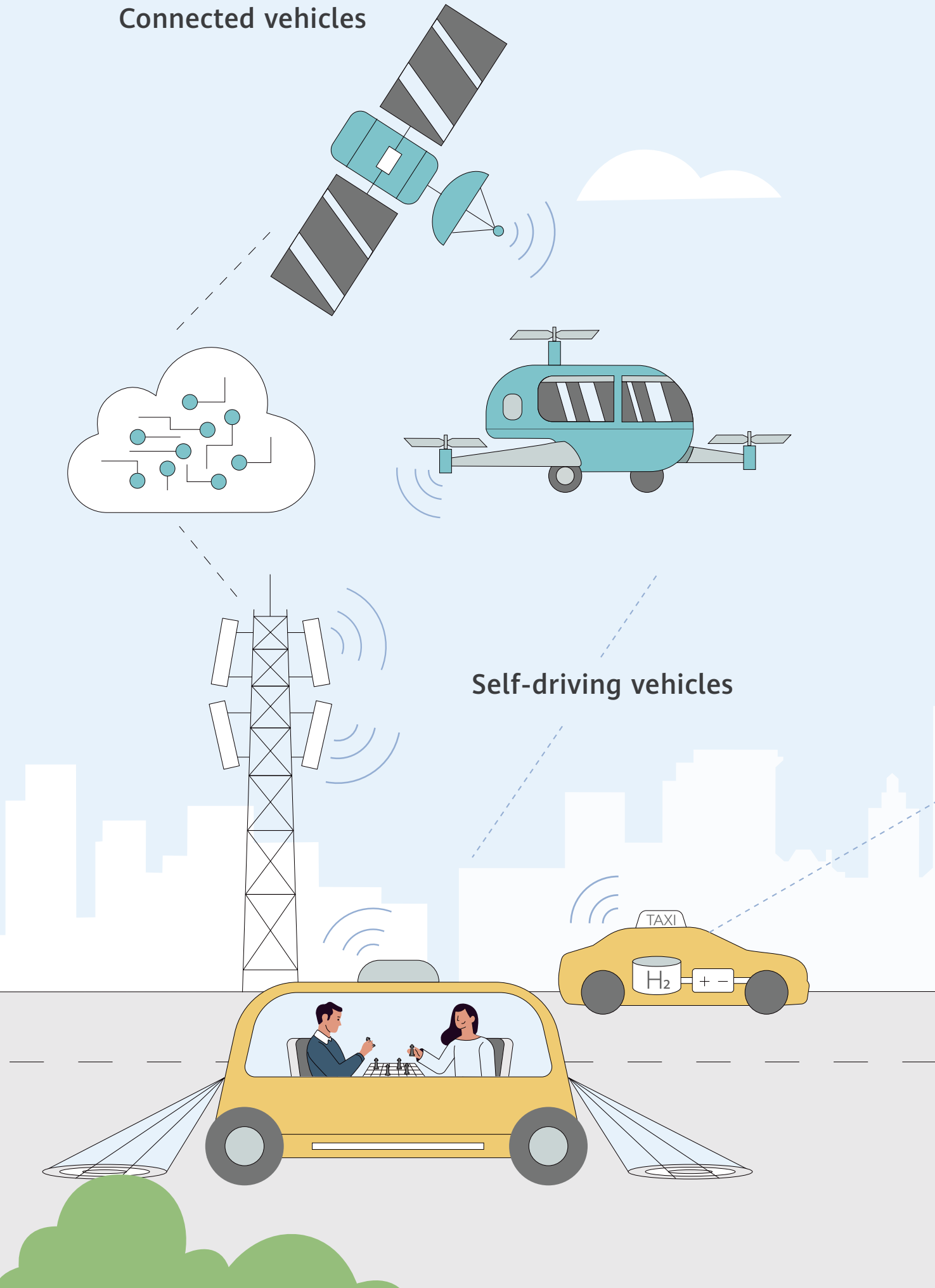
25 IRU, https://www.iru.org/who-we-are/where-we-work/europe/maas-mobility-service.

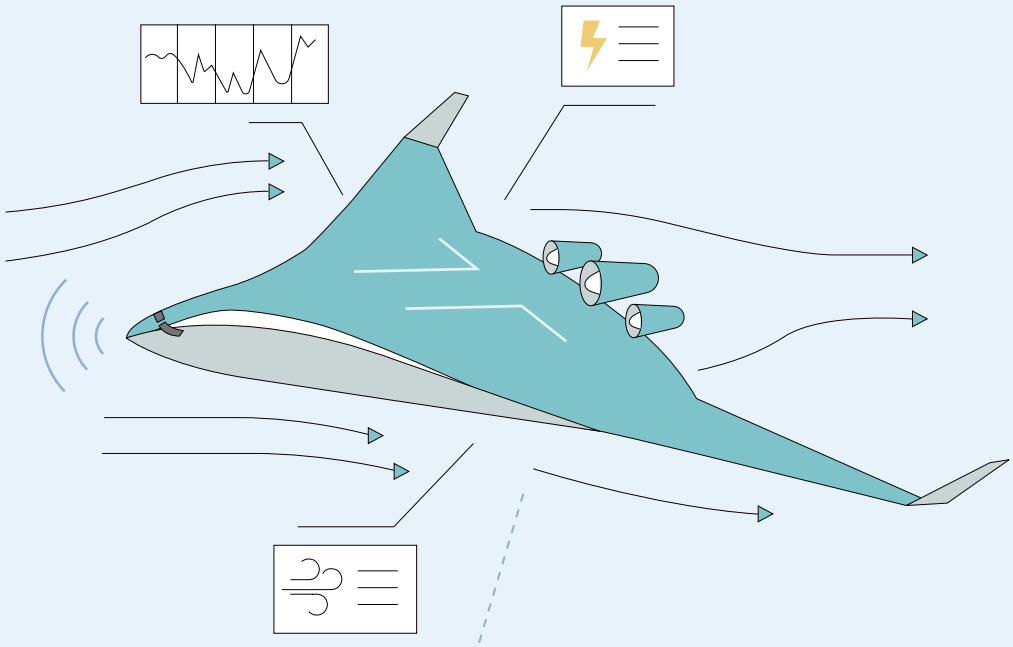
26 ITS Deployment Evaluation, https://www.itskrs.its.dot.gov/its/benecost.nsf/ID/d1c266bdeefee830852581bb005d27be.

27 Key Technologies and Prospects of Vehicle-Infrastructure Collaboration for Autonomous Driving 2.0, Institute for AI Industry Research at Tsinghua University and Baidu Apollo, https://jiaotong.bj.bcebos.com/cms/whitepaper/《面向自动驾驶的车路协同关键技术与展望 2.0》.pdf.

28 Vision, Requirements, and Technology Trend of 6G: How to Tackle the Challenges of System Coverage, Capacity, User Data-Rate and Movement Speed, Shanzhi Chen et al., https://ieeexplore.ieee.org/document/9003618.

29 Network 2030: A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond, FG-NET-2030, https://www.itu.int/en/ITU-T/focusgroups/net2030/Documents/White_Paper.pdf.


Connected vehicles
Self-driving vehicles
TAXI
H2



Electric vehicles

Shared vehicles

Cities

New Digital Infrastructure
Makes Cities
More Human and Livable

Urbanization is one of the global mega trends of this century. More than half of the world's population now lives in urban environments, and this figure is expected to rise to 60% by 2030 [1] . Take China as an example: 53.1% of China's population permanently lived in urban areas in 2012, but that percentage increased to 66.2% by 2023. This has driven a significant expansion in the coverage of basic public services in cities and towns [2]. Huawei predicts that China's urbanization rate will rise to 70% by 2030. China has planned the development of five super-city clusters, such as the Guangdong–Hong Kong–Macau Greater Bay Area and the Yangtze River Delta [3]. By 2030, the world is predicted to have 43 megacities with populations of 10 million or more [4].

As urbanization accelerates, cities worldwide will be forced to strike a balance between increasing size and limited resources. Typical urban problems, such as high energy consumption, pollution, traffic jams, and unequal access to digital infrastructure, will become even more pressing.

According to UN-Habitat, cities consume about 75% of the world's primary energy and emit between 50% and 60% of total greenhouse gases [5]. By 2030, the world is expected to generate 2.59 billion tons of waste annually [6], and up to 53 million tons of plastic could end up in rivers, lakes, and oceans every year [7]. Another worrying statistic is that air pollution kills an estimated seven million people worldwide every year [8].

The conflict between growth and limited resources will be the biggest headache for cities. They will need to use available resources as efficiently as possible. Rapid advances in new technologies, such as 5G, 5G-A, cloud computing, AI, data spaces, intelligent sensing, augmented reality (AR), and virtual reality (VR), are opening up more possibilities for cities, which represent the best places to create and incubate new applications for these technologies.

Over the past decade or so, countries across the globe have been making their cities more digital and exploring ways to leverage technologies to support sustainable development. In 2020, nearly 1,000 pilot smart city projects were underway worldwide. China was home to about 500 of these, with another 90 in Europe and 40 in the US [9]. Global spending on smart city initiatives is also increasing every year. In 2020, this spending totaled nearly US$124 billion, an increase of 18.9% over 2019 [10]. According to IDC, China's total ICT spending on smart cities will grow at a compound annual growth rate of 8% and exceed CNY1.1 trillion by 2027 [11]. Huawei predicts that China's spending on smart cities will surpass CNY1.5 trillion by 2030. Advancing digital and intelligent transformation has become one of the key pathways to sustainable development for the world's leading cities.

Direction for exploration: New digital infrastructure as the new engine of digital cities

The continuous expansion of cities is exerting pressure on the environment and our limited resources. The biggest challenge for the future of urban development will be how to use technology to significantly improve urban governance. Cities need advanced, refined systems that deliver sustainability while minimizing resource use.

In the past, cities provided physical public infrastructure – including water, electricity, gas, and road networks – to support rapid industrialization. Moving forward, a major direction for exploration will be new digital infrastructure that supports the development of digital and intelligent urban systems.

We believe that the digital infrastructure of new digital cities will consist of four layers. The bottom layer is an intelligent sensing system that can accurately sense dynamic situations and track the heartbeat of the city in real time. The second layer is intelligent connectivity. High-speed wired and wireless connections will connect the city, creating an organic whole. The third layer is an intelligent hub, which will serve as the city's "brain" and decision-making system. This layer aggregates massive amounts of data, supports key services that underpin the city's sustainable development, and enables city-wide data sharing, so that AI systems can maximize the value they provide, making refined, data-driven, highly-automated city development a reality. The fourth and top layer is intelligent application. A comprehensive ecosystem of intelligent city applications will be built on top of the underlying digital infrastructure, covering the "last mile" of service delivery and creating infinite possibilities for a smart city. These four layers will interconnect and support each other, ultimately creating city intelligent twins that embody the new goal of ubiquitous intelligence citywide. This will help create an intelligent era.

Snapshot from the future: Nanosensors tracking the pulse of the city

Digital cities depend on data, which comes from a wide array of sensors scattered throughout the city. Just as people perceive their surroundings using their senses of sight, hearing, smell, taste, and touch, a city needs its own sense organs – deployed around the city – to sense changes. These sensors provide the data that underpins the growth of a digital city.

The MIT Technology Review listed "Sensing City" as one of the 10 Breakthrough Technologies 2018. We believe that in the cities of the future, isolated sensing systems will be merged into a comprehensive sensing network, with sensors connected using various transmission systems. Analytics using these massive flows of comprehensive data will generate a more accurate picture of the latest developments within a city. At the same time, breakthroughs and advances in sensing technology will also drive leapfrog advances in sensing cities.

One particularly cost-effective and revolutionary technology – nanosensors – is expected to drive the next sensor revolution. Nanosensors have huge potential and can be deployed in massive numbers to form a wireless nanosensing network.

This will greatly enhance a city's ability to sense, leading to major advances in climate monitoring, health monitoring, environmental protection, and other domains.

Nanosensors are very small and precise, and will vastly improve sensing performance. Working at the atomic scale, they are expanding our understanding of what sensors can be, driving new advances in sensor manufacturing, and opening up new fields of application. Early applications of nanosensors cover many different fields, including biology, chemistry, mechanics, and aerospace.

Graphene gas nanosensors, for example, are ultra-sensitive to odors. A nano-coating on the sensor's surface where it makes contact with the gas improves sensitivity and performance. The sensor collects odor molecules with a metal-organic film, and then amplifies the chemical signals using plasma nanocrystals. The most common application of these sensors is in carbon dioxide detection, but they can also be used to quickly detect hazardous and toxic gases.

One American university has created a novel type of nano-coating using graphene, and when the coating is applied as a nanofilm on gas sensors, it delivers a 100-fold increase in molecular response compared to the best available sensors that use carbon-based materials [12].

In the near future, these sensors will be able to accurately identify hazardous, toxic, or explosive gases in the air, greatly improving safety in scenarios like manufacturing and customs inspections.

Nanocrack-based acoustic sensors are able to recognize specific frequencies of sound. They are more sensitive than other acoustic sensors, as the spacing between cracks in nanocrack-based

sensors can be as little as just a few nanometers. Researchers build the frames of these sensors by adding a 20-nanometer-thick platinum layer to the surface of a viscoelastic polymer. When the platinum layer deforms and stretches, it pulls away from the underlying polymer. Researchers have then been able to measure the conductivity of the sensor surface. In an environment with 92 decibels of noise, the nanocrack sensor performed far better than conventional microphones at separating out sounds in a given frequency range. For example, when these nanosensors are placed on the surface of a violin, they can accurately record every note of a tune and "translate" it so that a connected device can accurately recreate an electronic version of the tune. When someone puts this kind of sensor on their wrist, it can even accurately monitor their heartbeat. Naturally, breakthroughs in this technology will greatly enhance the acoustic monitoring of a city [13].

Snapshot from the future: All-optical, 10-gigabit cities

The digital transformation of cities requires massive flows of information. The newest generation of connectivity technologies, such as 5G, F5G, and 10-gigabit Wi-Fi, is enabling high-speed networks with universal coverage in urban spaces. All these technologies enable high-speed information flows, which require all-optical networks. A foundation of all-optical networks will allow cities to merge their operational infrastructure into their communications infrastructure. As a result, new types of people-oriented government services will be more accessible and affordable to individuals, homes, and organizations.

Several major cities have already conducted preliminary research into this area and found tentative signs of all-optical cities' tremendous value and growth potential.

In April 2021, Shanghai became the world's first All-optical Smart City. With its F5G optical network, the city is able to deliver stable connections across the city that have a latency below 1 millisecond anywhere. The deployment of this high-speed optical network has laid a solid foundation for Shanghai's future digital transformation [14]. Major cities around the world are now racing to release their own 10-gigabit action plans, with over 20 provinces and cities having already released one. Beijing has announced its Optical Network Capital, 10-Gigabit City Action Plan (2023–2025) while Shanghai has unveiled its Action Plan for Further Promoting New Infrastructure Construction (2023–2026) to connect the whole city with 10-gigabit 5G-A networks and 10-gigabit optical networks. Qinghai has also released Guidelines on Dual 10 Gigabit (5G-A & F5G) Industry Development and Application Innovation. Saudi Arabia has also unveiled The 10 Gbps Society white paper which echoes the Middle East's vision for 10-gigabit experience.

In Adelaide, Australia, more than 1,000 buildings are now connected to 10-gigabit networks. Companies in these buildings can access cloud services at a speed of 10 Gbit/s, creating huge opportunities for industries such as education, video, IT, and software engineering [15].

We believe that all-optical infrastructure will drive leapfrog improvement in terms of capacity, bandwidth, and user experience for the communications networks of future cities: Uplink and downlink rates will reach 10 Gbit/s; latency will be reduced to microseconds; and the number of connections will see a 100-fold increase.

The future architecture of an all-optical city will consist of four parts:

All-optical access: All network connections will be optical, including home, building, enterprise, and 5G base station connections. All-optical transmission networks will be extended into edge environments like large enterprises, buildings, and 5G base stations. 10-gigabit access networks will enable digital transformation across industries, support access to cloud and computing networks within one millisecond, and drive F5G application in diverse scenarios as well as 5G adoption in business settings. By 2030, optical transport networks will cover all government agencies, financial institutions, key universities and scientific research institutions, large hospitals, large industrial enterprises, as well as development zones and industrial parks above the county level.

All-optical anchors: Connections originating in home broadband, enterprise broadband, 5G networks, and data centers will be routed and transmitted through all-optical networks. All-optical anchors will support multiple technologies, enable service-based traffic steering, and provide one-hop connections to cloud and computing networks. By 2030, every 10,000 people will have four all-optical OTN anchors, and 25% of these anchors will deliver 100G services.

All-optical bearing: Urban optical networks support one-hop access to services. All-optical cross-connect, optical-electrical hybrid automatically switched optical network (ASON), and other technologies will be used to build multi-layer optical networks that support one-hop access to services, highly reliable networking, high-speed inter-cloud transmission, and high synergy between optical and computing networks. More than 50% of data centers will be connected with each other through optical networks, with single-wavelength bandwidth higher than 400 Gbit/s. Mesh fiber networks can deliver 99.9999% reliability thanks to their ability to withstand two or more cut fibers, without services being interrupted. All data center networks will adopt a hybrid optical-electrical design that uses all-optical switching technologies to connect switches and routers within individual data centers.

Intelligent and automated O&M: Real-time sensing of network status with proactive, preventive O&M will support elastic network resources, and automated service provisioning, resource allocation, and O&M.

By 2030, many cities will have 10-gigabit connectivity, with 10-gigabit services available to organizations, homes, and individuals.

Snapshot from the future: Intelligent and digital city upgrade underpinned by intelligent computing centers

As a key pillar of urban information infrastructure, intelligent computing centers are essential to satisfy the fast-growing demand for AI computing power in cities. They play an important role in promoting AI industrialization, industry intelligence, intelligent governance, and industry clusters. They are also key to driving comprehensive digital city transformation, optimizing industry structures, and improving the competitiveness of individual cities in the digital economy era. According to the Action Plan for High-quality Development of Computing Power Infrastructure jointly released by China's Ministry of Industry and Information Technology (MIIT) and other five ministries, by 2025, China's computing power is expected to exceed 300 EFLOPS, with intelligent computing accounting for over 35% of that power. According to Huawei, by 2030, global intelligent computing will exceed 864 ZFLOPS.

Intelligent computing centers will support cross-industry collaboration by providing diversified computing power and integrated services. They will also consume less energy and make computing resources more affordable and accessible. Intelligent computing centers will also be connected to AI computing networks. Computing infrastructure, especially intelligent computing centers and computing networks, can effectively promote AI industrialization and industry intelligence, and serve as the foundation for intelligent cities that underpin the digital economy. Intelligent computing centers provide cost-effective, affordable, and secure computing resources that will make AI computing power as easily available to governments, enterprises, and the public as water, electricity, gas, and other urban public utilities.

Wuhan AI Computing Center, which officially entered operation on May 31, 2021, currently has a capacity 1,500 PFLOPS and plans to expand its capacity to 6,800 PFLOPS. The center is China's first AI computing infrastructure intended for public services, and, as its name indicates, it is based in Wuhan from where it will serve the neighboring Hunan, Hubei, and Jiangxi provinces. It has become the largest intelligent computing center in central China and delivers immense computing power to enterprises in Wuhan and the surrounding areas. By making computing power more affordable and accessible, the center supports the R&D of innovative applications in many fields, including smart manufacturing, autonomous driving, smart cities, smart agriculture, weather forecasting, and pharmaceutical development.

With the help of the intelligent computing center, Wuhan Bosheng Technology's SiFold protein structure prediction platform can narrow the scope of its searches for target drugs and shorten its prediction process to one day. In addition, the protein structure predictions made by this platform are 10-times more efficient than previous predictions and lower the cost to just a few hundred Chinese yuan.

Intelligent computing power is being used in many different scenarios, from intelligent manufacturing plants to cutting-edge labs, addressing real-world issues and creating new solutions that will benefit people for years to come. As a core driver of productivity in the digital economy, computing power is becoming increasingly essential in Wuhan's ambitions to test cutting-edge technologies, incubate new applications, and nurture up-and-coming enterprises with explosive potential to advance the digital economy.

Snapshot from the future: Intelligent hubs cutting out the human factor from urban management

During the digital transformation of cities, data barriers will be broken down, and all data will ultimately form a single data lake. Therefore, AI will play an increasingly important role, and many public service and regulatory issues will be decided by expert algorithms instead of being left up to human judgment. Narrow AI applications will expand until they can deliver data-driven intelligence for all city governance scenarios.

Technological advances are driving important changes in the approaches we take to city governance. For example, there has been a shift from reactive services to proactive services, from broad-brush to clearly defined regulations, and from post-hoc response to real-time response and even incident prediction and prevention.

These changes will pose new challenges: For example, new public governance agencies will have to be created to keep AI in check. Algorithmic authority, which means the authority to wield algorithms and massive data volumes, will be woven into the city governance system – a move that will reshape city governance itself. With AI, the city governance system can extend its reach, with AI-enabled sensors across the city providing citizens the urban resources they need, right when they need them. AI ethics, including human-centered AI, equality, and fairness, must be constantly observed to ensure technology evolves in a way that truly benefits all of humankind.

Huawei believes that as the goals of city governance evolve, cities will need powerful intelligent hubs to aggregate data, support applications, and help us overcome the challenges created by technology. They will also need the capacity to iterate and improve themselves. These hubs will aggregate massive amounts of data from every corner of a city, and mine that data for insights to support better city governance. This will benefit every industry and greatly improve the efficiency of city governance and the experience of users of government services.

Early-stage explorations by Toyota: In Toyota's plan for the city of the future, each house, building, and vehicle will be equipped with sensors. Data from these sensors will then be aggregated into a city's data operation system, through which people, buildings, and vehicles will be connected. When the data is aggregated to the intelligent hub, AI will be used to analyze people's surroundings and then guarantee the safety of pedestrians and drivers by keeping them separated. In addition to new technologies like indoor robots, residents will be able to use AI to check their health at home. Wearable medical sensors for home use will transfer data to the data operating system, which will provide instructions for healthcare and healthy living[16] .

Snapshot from the future: Smart ecosystems spreading intelligent services across all city use cases

When cities have ubiquitous, high-speed connectivity, intelligent hubs, and massive, real-time data from city sensors, smart applications will emerge in every area of urban life.

Starting with government services, AI will expand to enable industry development and smart lifestyles. The key to this process will be an ecosystem for innovation in smart applications that bridges the last mile in services for residents. These services will be vital to unleashing the value of new digital infrastructure.

Some major cities are already early adopters of this idea. In China, Huawei has partnered with Shenzhen's Guangming Science City to build a showcase for green, all-optical, smart districts. Once it kicks into full gear, this project will accelerate innovation in services for key urban industries, such as smart manufacturing, life sciences, and optical networks.

Huawei and Guangming will also build China's first innovation center for life sciences and intelligent manufacturing. This center will include two service platforms – EI Health and Fusion Plant. By bringing together both upstream and downstream players to serve their respective industries, Huawei and Guangming will accelerate their modernization and use of AI.

These industry platforms will provide two key services to support business innovation: a powerful public cloud computing platform and massive storage. In the field of life sciences, these services will support trained algorithms for image analytics, gene analysis, and drug R&D data analytics. For manufacturers, they will offer industrial Internet services. This will accelerate digital transformation for biomedical and industrial enterprises, and support the emergence of smart applications and application providers.

Thanks to this new digital infrastructure, cities will also be able to create their own smart innovation ecosystems. These ecosystems, in turn, will take advantage of infrastructure to bridge gaps in digital innovation. We expect to see the positive cycle between ecosystems and industry development. This will unleash tremendous value of digital infrastructure across cities, benefiting countless industries. Smart ecosystems are key to spreading intelligent services across all use cases.

Direction for exploration: Smart government services making cities more human

In China, accessing government services often involves many visits to different agencies and administrative departments. Today however, most Chinese cities have government service halls and most government services can be delivered within these halls. Since the outbreak of COVID-19, more government services have also been delivered through smartphones without the need to visit service halls.

Major cities in China and other parts of the world are bringing the idea of people-centric government service to life. This trend, combined with the evolution of cloud, AI, intelligent sensing, blockchain, and other advanced technologies, will unleash tremendous amounts of power that we can only begin to imagine. Moving forward, cities need to figure out how they can get this combination right, before they can truly deliver swift and easy access to government services and become friendlier places to live.

Snapshot from the future: Proactive, precise, data-based government services

Machine recognition technology enables contactless services. Today, in most of China's developed provinces, residents no longer need to go to government service halls to access government services. Instead, they can now simply use their smartphones. Over the next decade, more and more intelligent digital government services will surface.

- Digital identity authentication is expected to be widely adopted. The ID cards, drivers' licenses, social security cards, and bank cards that people carry at all times will be digitalized, creating a total addressable market for global electronic identity authentication services worth US$18 billion by 2027 [17].

- Digital credit will underpin and restructure many public service processes and the customer experience. It will be one of the founding technologies for digital government. Most residents are already familiar with some services like electronic library cards, social security cards, and car rental services that require a credit rating. As these services continue to improve, they will deliver a better experience to millions of residents.

- Universal access to one-stop, e-government services will soon be realized. In the future, all government services will be remotely accessible and government service halls may cease to exist.

Technological advances will give rise to new digital government practices and government services. Today's centralized digital government model, or "One-net Management", is a great example. In many Chinese cities, the local government has established this government structure using big data and IoT. This model enables complex operations involving multiple different departments, regions, and levels of government. It offers an architecture that addresses the needs of the municipal government and of local residents. In the future, as governments aggregate more massive data and their AI technologies mature, they will be able to deliver government services in

a more precise and proactive way; manage their municipalities more efficiently; and improve their service experience.

Let's look at smart care for the elderly as an example. Some communities in Shanghai have installed smart water meters at the homes of elderly people who live alone, with their consent. If the total water used within a 12-hour period falls below 0.01 cubic meters, the meter will send an alarm to the One-net Management Platform of the local community workers. These workers will then arrange a welfare check for the elderly resident. Such initiatives add a human touch and improve support for the elderly [18].

Snapshot from the future: Urban data spaces that promote the circulation and unlock the value of data elements

Data has become the fifth key factor of production in addition to land, labor, capital, and technology. It has also become a new driver of the digital economy. Efforts should be made to accelerate data circulation and transactions, as this is crucial for unlocking the value and potential of data. By 2030, the global data transaction market is expected to reach US$301.1 billion, with China's market reaching CNY515.59 billion at a fast compound annual growth rate of about 20.3% [19].

Digital societies produce massive amounts of data and open up new spaces for urban data. The amount of data generated by cities is growing from terabytes to petabytes, and data is becoming the new oil for cities. Through urban data spaces and technologies, especially big data, blockchain, AI, privacy computing, and security and trustworthiness technologies, cities will be able to effectively process and analyze massive amounts of data, maximize the value of data elements, and use data to inform urban management and decision making. This means data has the potential to create greater economic and social value, and become a new driver of urban development. However, different players in society have different requirements for data aggregation, circulation, and security in different scenarios. To meet these diverse requirements, urban data spaces will need to support one-stop data collection, processing, development, and application, and provide a complete portfolio of capabilities, from data aggregation, storage, governance and processing, to development, utilization, and intelligent application. Urban data spaces will need to meet the development requirements for socialized and comprehensive governance in complex scenarios, and provide a unified platform for all participants to interact with each other and match supply and demand. This can promote data openness, sharing, and authorization, and create new value with data in multiple service scenarios.

As the entity authorized to manage public data in Shanghai, Shanghai Data Group has been tasked with creating a data element market, unleashing the potential of data elements, and ensuring data security. By integrating urban data spaces, industry data spaces, enterprise data spaces, and personal data spaces, Shanghai Data Group is leveraging innovative technologies to identify high-value scenarios for data elements, unleash the productivity of data elements, and help government agencies, local enterprises, and residents in Shanghai unlock the value of different data elements. These efforts aim to create new urban data spaces and facilitate smooth circulation of data across the city.

Direction for exploration: Intelligent environmental protection for livable cities

Urban development always puts pressure on the environment. Cities are major sources of air pollution, carbon dioxide emissions, solid waste, and water pollution. Infrastructure for protecting the environment usually lags far behind the economic development and population growth of a city. Therefore, a key aspect of constructing future cities relates to addressing the conflict between urban growth and the environment. This leads us to examine how digital and intelligent technologies can be used to protect the environment more efficiently and make cities more livable for every resident.

Snapshot from the future: Automatic waste disposal for zero waste cities

Cities produce huge amounts of solid waste every day, and effective waste disposal has always been a daunting challenge for city administrations. The zero waste city is a concept that emerged from building innovative, green cities, and sharing the benefits of city development. Under a zero waste city initiative, cities grow and operate in a greener way, reduce solid waste, use resources more efficiently, and minimize the impact of solid waste on the environment.

"Zero waste city" initiatives have already been launched around the world. For example, the European Union (EU) has kicked off the European Green Deal. Key elements of the waste proposal in this deal include:

- A common EU target for recycling 65% of municipal waste by 2030;
- A common EU target for recycling 75% of packaging waste by 2030;
- A binding landfill target to reduce landfill to a maximum of 10% of all waste by 2030;
- A ban on landfilling of separately collected waste;
- Concrete measures to promote re-use and stimulate industrial symbiosis – turning one industry's by-product into another industry's raw material [20].

C40 Cities also issued an Advancing Towards Zero Waste Declaration, pledging to achieve zero waste cities by:

- Reducing municipal solid waste per capita by at least 15% by 2030 compared to 2015;
- Reducing the amount of municipal solid waste disposed through landfills and incineration by at least 50% by 2030 compared to 2015; and
- Increasing diversion away from landfill and incineration to at least 70% by 2030 [21].

In 2019, China started piloting its own zero waste city initiative in "11 + 5" cities [22] to explore how to construct a solid waste classification and recycling system.

With so many countries setting goals to construct zero waste cities, solid waste processing technologies and innovations are set to develop rapidly, generating many new methods and best practices in this sector.

The Songdo Smart City Hub in South Korea has introduced an automated waste disposal system which uses negative pressure to suck domestic waste to a waste processing center through underground pipes. In Malaysia, one company has developed a waste disposal system [23] that transports municipal or domestic solid waste through underground pipes at high speeds, taking the waste from waste chutes and outdoor load stations into sealed containers located up to 2.5 km away. Once the containers are full, they are collected at specific times by an arm-roll truck. This system dramatically speeds up the entire waste collection process whilst reducing manpower.

One company in Europe has also developed an automated waste sorting robot. Powered by AI, this robot can automatically identify different types of waste on a conveyor belt, classify the waste as required by customers, and then process and recycle the waste. The robot can sort waste several times faster than a normal worker and can run 24/7, significantly speeding up the entire waste sorting process. In the future, with the help of such intelligent waste sorting robots, waste sorting in

cities is set to be much faster, as it becomes fully automated and unmanned.

With the help of AI, the entire waste management process in a future city, from collection and transportation to sorting and processing, will be automated and intelligent. Intelligent waste recycling bins, driverless garbage trucks, automated waste sorting robots, automated garbage recycling devices, and other innovative applications will emerge one after another. Hopefully, this will help to make more and more zero waste cities possible.

Snapshot from the future: Optical detection making water sources safer

The uneven distribution of water resources and water pollution have long been problems for cities. Many face water shortages. At present, slightly less than one half of the global population – about 3.6 billion people – live in areas that suffer from water scarcity [24]. Industrial wastewater and agrochemical water pollution are far more serious today than in the past.

The management and use of water resources in most cities worldwide is still siloed, split across several different industrial sectors, and lacking any kind of centralized overview. In the future, cities will need to better manage their full water cycle, from intake and use to discharge. A holistic, AI-managed system will need to be introduced to manage water resources, and cities will need to rebuild their water facilities. An AI system will be able to maximize the use of water resources within a city by refining every stage of water intake, use, and discharge through forecasts for weather and water consumption. This process will also involve precise, scheduled use of water resources and a reduction in the total energy consumed.

Water quality monitoring, especially during the treatment of industrial wastewater, is another key concern for the conservation of city water resources. Many emerging technologies can be applied in this sector. Wastewater often goes through a chemical treatment process. However, this approach usually takes a long time and has many restrictions. In contrast, new optical detection technologies do not suffer from any of the drawbacks of chemical treatment. As different substances produce unique spectral patterns, new technologies could be used to monitor water quality in real time throughout the entire process, detecting wastewater whenever it is present.

One research team in the US has developed optical sensors to detect sewage contamination. [25]

Statistical relations between optical properties and genetic bacteria markers are being used to calibrate field sensors for the detection of sewage contamination, including the sources and timing of contamination .

Optical technologies can also be further integrated with analytics from the IoT, AI, and cloud computing. Sensors for water quality and deep data analytics will move us closer to 24/7, efficient, real-time, automated, intelligent water quality monitoring and enable faster warnings in cases of water contamination.

Optical technology can also be used alongside AI to explore the hidden relationships between water quality parameters and treatment processes, so as to upgrade and transform urban sewage treatment processes more methodically.

All of this can create a closed loop of prevention, control, monitoring, treatment.

Snapshot from the future: Real-time AI air quality monitoring

In recent years, air pollution has posed an increasing threat to people's health, and urban air quality is attracting more attention than ever before. According to the World Health Organization, nearly 90% of cities worldwide fail to meet its air quality standards. This problem is only getting worse. Industrial exhaust, coal burning, automobile exhaust, and other types of air pollution have all become major challenges to public health.

Most cities will likely soon opt to deploy cost-effective and reliable air quality sensors and build new monitoring networks. This will allow them to monitor air quality and weather across the entire city and take targeted measures to improve air quality and the urban environment. One company has developed a highly-integrated, real-time air monitoring system [26] that uses integrated sensors and software to monitor the concentrations of environmental pollutants in urban environments,

such as PM2.5, PM10, CO, NOx, SOx, and O3, as well as other environmental parameters like noise, temperature, humidity, air pressure, rainfall, and flooding. The system's data is wirelessly transmitted in real time to a cloud platform that in turn provides a real-time visual dashboard for effective management of the overall environment in key areas of the city.

In the future, we can consider integrating these sensors with AI. With the support of machine learning, sensors will be able to better detect their surroundings and make preliminary judgments regarding potential changes. This intelligent upgrade of sensors could substantially improve the ability of cities to automatically sense environmental changes in real time.

The use of AI-powered sensing during the COVID-19 pandemic was a prime example of how such technologies can be applied. Every time a person exhales, small droplets are released into the air that can transmit viruses to others. The lower the humidity or temperature around a person, the longer an infectious aerosol can stay in the air and the higher the probability of infecting another person. An AI sensing system can monitor volatile organic compounds, humidity, and air temperature, and determine whether an environment is conducive to virus transmission. In addition, such systems can automatically and centrally control ventilation and air conditioning systems to reduce the risk of infection.

Novel applications in this sector will boost the ability of cities to monitor and improve air quality and simplify the process of managing our atmospheric environment.

Conclusion:

New digital infrastructure making cities more human and livable

ICT technologies like 5G, optical networks, AI, cloud, blockchain, and intelligent sensing will be rolled out rapidly over the next decade. Cities will soon enter an era of 10-gigabit connectivity, with 10-gigabit services becoming available to organizations, homes, and individuals. Huawei predicts that by 2030 worldwide, 84% of companies will have access to 10-gigabit Wi-Fi services and 25% of homes will have access to 10-gigabit broadband services.

The application of these ICT technologies in cities will result in fundamental changes, allowing cities to go digital and intelligent on all fronts. Urban infrastructure will become increasingly intelligent, data will become increasingly valuable, business will increasingly apply AI models, and applications will become increasingly smart. These four trends will enable new intelligent architectures for cities and new models for urban infrastructure construction.

There will be structural changes in how cities pursue sustainable development and economic growth. City intelligent twins and smart resource planning and utilization will help significantly improve governance efficiency, city resource utilization, and user experiences. The integration and innovation in new forms of productivity will drive the development of urban economies. This can ultimately help cities achieve sustainable development goals and make them more human and livable.

global industry vision

Huawei predicts that by 2030,

84% of companies will have access to 10 gigabit Wi-Fi networks.

References

1 Sustainable Development Goals, UN, https://www.un.org/sustainabledevelopment/cities/.

2 Chinese governments website: https://www.gov.cn/zhengce/202408/content_6967695.htm.

3 China's Supercities Boom and the Next Era of Urban Growth, Morgan Stanley.

4 The World's Cities in 2018, UN, https://www.un.org/en/events/citiesday/assets/pdf/the_worlds_cities_in_2018_data_booklet.pdf.

5 UN-Habitat, https://unhabitat.org/topic/energy.

6 "Li Jinhui: The Concept of Waste-free City Promoting Sustainable Development," China's Ministry of Ecology and Environment, https://www.mee.gov.cn/home/ztbd/2020/wfcsjssdgz/wfcsxwbd/ylgd/201912/t20191203_745330.shtml.

7 Amy Arthur, "Annual plastic water pollution could reach 53 million tonnes by 2030," PA Science, https://www.sciencefocus.com/news/annual-plastic-water-pollution-couldreach-53-million-tonnes-by-2030/.

8 UN, https://news.un.org/zh/story/2019/03/1029531.

9 Super Smart City: Happier Society with Higher Quality, Deloitte, https://www.henandaily.cn/content/2019/0504/163011.html.

10 Worldwide Smart Cities Spending Guide, IDC, https://www.idc.com/tracker/showproductinfo.jsp?containerId=IDC_P37477.

11 Predictions for China's Smart City Market (2023–2027), IDC, https://www.idc.com/getdoc.jsp?containerId=prCHC51776624

12 Mohammad Mehdi Pour et al., "Laterally extended atomically precise graphene nanoribbons with improved electrical conductivity for efficient gas sensing," https://www.nature.com/articles/s41467-017-00692-4.

13 Daeshik Kang et al., "Ultrasensitive mechanical crack-based sensor inspired by the spider sensory system," https://www.nature.com/articles/nature14002.

14 "What implications does the world's first full-optical, intelligent city bring?," Xinhuanet, http://www.xinhuanet.com/info/2021-04/30/c_139916374.htm.

15 Smartcitiesworld.net, https://www.smartcitiesworld.net/news/news/adelaides-ten-gigabitnetwork-connects-its-1000th-building-5781.

16 Woven Planet Holdings, Inc., https://www.woven-city.global/.

17 Shanhong Liu, "Identity verification market value worldwide 2017-2027," https://www.statista.com/statistics/1036470/worldwide-identity-verification-market-revenue/.

18 "Tech-powered elderly care: Low water usage sends a welfare alert," Xinhuanet, http://www.xinhuanet.com/local/2020-12/10/c_1126846671.htm.

19 Research Report on China's Data Trading Market in 2023, Shanghai Stock Exchange.

20 Towards a circular economy – Waste management in the EU, ERPS, https://www.europarl.europa.eu/RegData/etudes/STUD/2017/581913/EPRS_STU%282017%29581913_EN.pdf.

21 Jessica L. Leath, "Is Bitcoin Reminiscent of Past Bubbles?," https://www.c40knowledgehub.org/s/article/Advancing-to wards-zero-wastedeclaration-Planned-actions-to-deliver-commitments?language=en_US.

22 China's Ministry of Ecology and Environment, https://www.mee.gov.cn/home/ztbd/2020/wfcsjssdgz/wfcsxwbd/wfcsmtbd/202003/t20200324_770316.shtml.

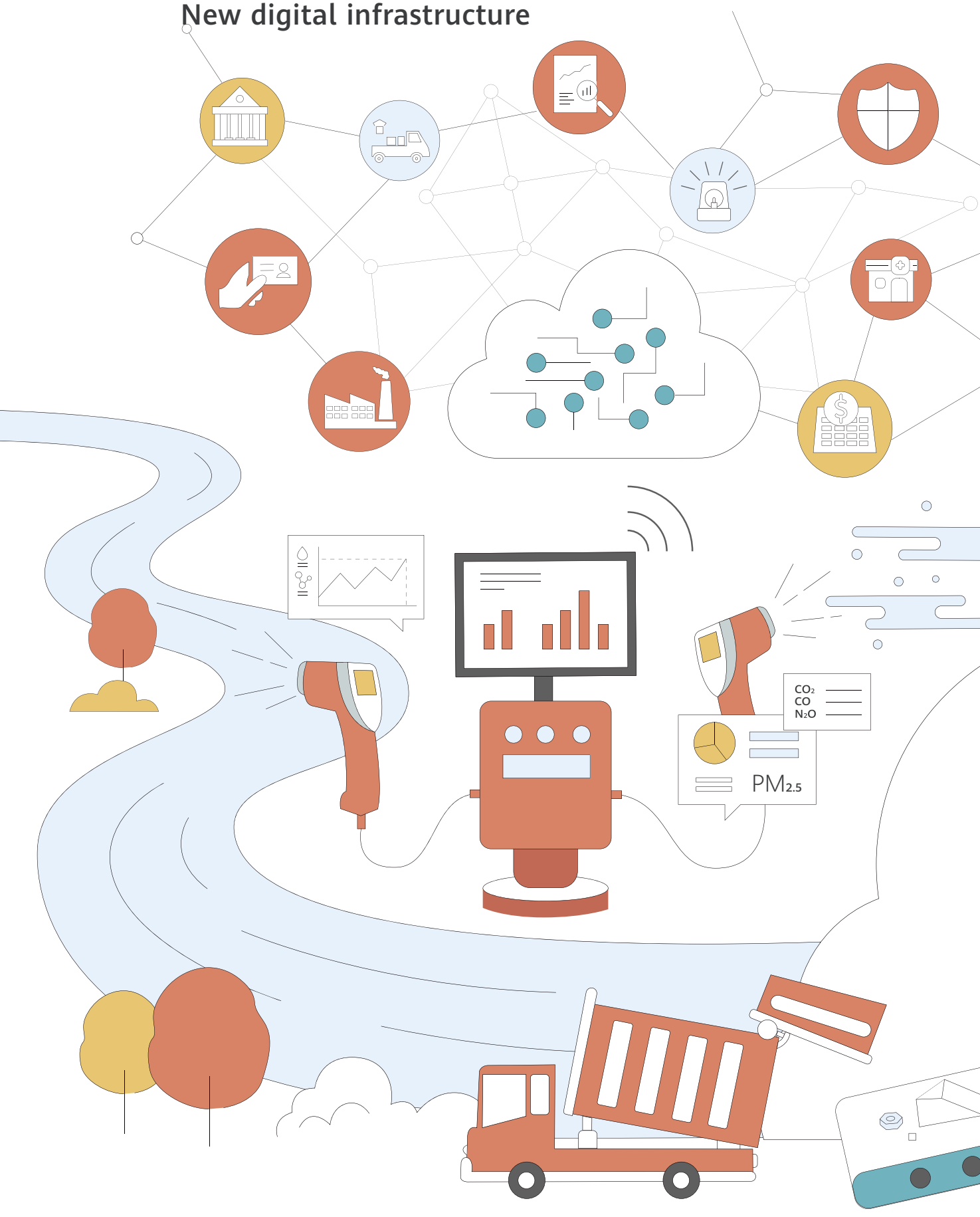
23 Stream, https://stream-environment.com/what-we-do.

24 UN World Water Development Report, UN-Water.

25 "Using optical sensors to detect sewage contamination in the Great Lakes," USGS, https://www.usgs.gov/centers/umid-water/science/using-opt cal-sensors-detect-sewage-contamination-great-lakes?qt-science_center_objects=0#qt-science_center_objects.

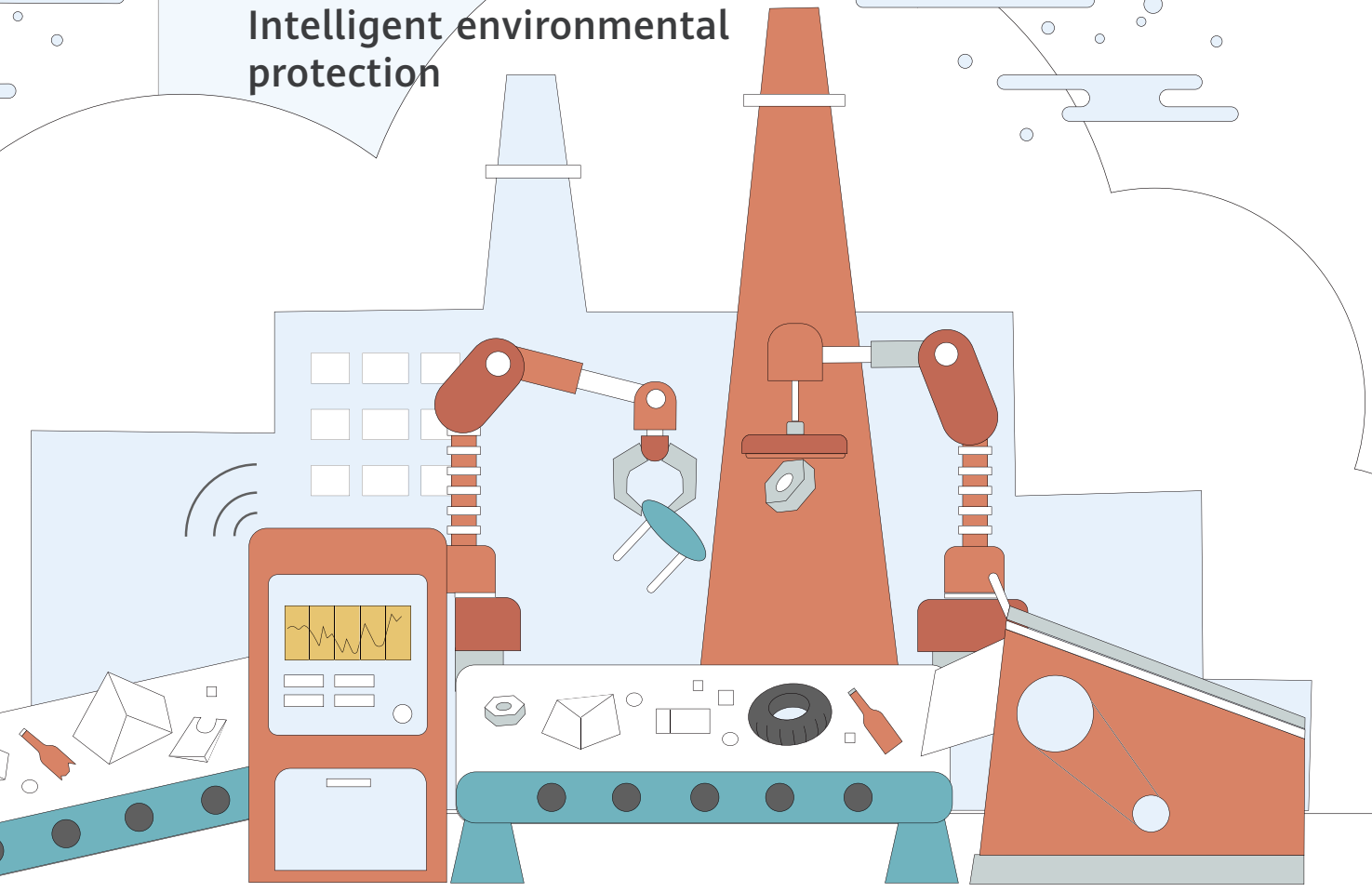
26 Oizom, https://oizom.com/.

New digital infrastructure



Smart government services

Intelligent environmental protection



Enterprises

New Productive Forces,
New Production Models,
New Resilience

Over the next decade, the world's population will age significantly and irreversibly. According to a report published by the UN [1], the global population aged 65 and over is projected to exceed 12% of the total population by 2030, while the global population aged under 25 will decrease from 41% in 2020 to 39% in 2030.

Population ageing will lead to a huge worldwide labor shortage. By 2030, we can expect a deficit of 85.2 million workers around the world – more than the current population of Germany [2]. Labor shortages are soon expected in 12 of the world's 15 biggest economies – which collectively account for 70% of global GDP. The size of the workforce is an important factor in economic growth for every country. Take manufacturing for example. By 2030, this sector is estimated to face a global labor shortage of 7.9 million workers, leading to unrealized output of US$607.14 billion [3].

Co nsumer demand is set to become much more diverse, which will profoundly change production models and force businesses to innovate. Companies that aim to grow and expand their business will need to capture, stimulate, and nurture increasingly diverse consumer demand. Companies of the future must rapidly respond to new consumer demands by launching products with innovative functions. For example, as the "singles economy" gains traction, companies can rapidly adjust their products by targeting areas like solo dining, small home appliances, and mini karaoke booths. In addition, companies need to take the initiative and stimulate demand through emotional appeal, and rapidly produce combinatorial designs for product appearance, images, and implications. For example, they can customize limited-edition products or launch co-branded products within the shortest time possible.

Black swan events pose new challenges to business continuity. For example, the global spread of COVID-19 negatively impacted the economy, resulting in factory shutdowns, material shortages, and disruptions to global logistics and supply chains. It is estimated that the pandemic cut global trade by between US$1.7 trillion and 2.6 trillion [4], highlighting the importance of enhancing value chain resilience.

Direction for exploration: Bringing unmanned operations to manufacturing and services to make up for labor shortages

In order to expand, companies need to promptly seize business opportunities. When they receive a large order with a tight deadline, they must quickly expand their production capacity. However, more and more companies are missing out on opportunities due to labor shortages. This is where new productive forces come in.

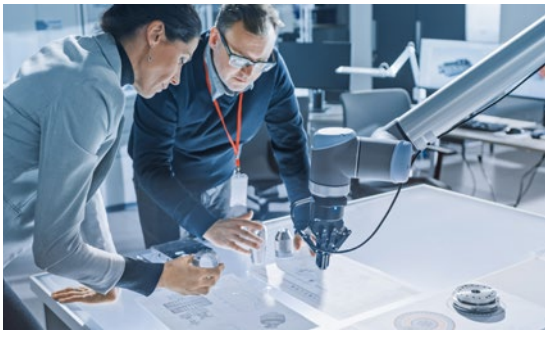
People are also trying to introduce new productive forces to help solve chronic issues in education, healthcare, and many other industries, such as unequal distribution of resources and shortages of professional talent.

Snapshot from the future: Collaborative robots

Collaborative robots are a type of industrial robot. They were initially designed to meet the customized and flexible manufacturing requirements of small- and medium-sized enterprises, and perfectly align with the development trends of the manufacturing industry. Collaborative robots are suitable for jobs that people don't want to do, such as highly repetitive work like sorting and packaging. Collaborative robots have several unique advantages:

Safer: Collaborative robots are compact and intelligent, and their sophisticated sensors enable them to stop in an instant. Unlike traditional industrial robots, collaborative robots do not need physical fences, as the scope of their movement is restricted by virtual digital fences. This means they can be placed at any location along a production line on demand, and work closely together with human workers on the production line to get work done.

Faster and more flexible deployment: Traditional industrial robots require professionals to plan and program their movement paths and actions, so they take a long time to deploy and are very costly. In contrast, collaborative robots feature user-friendly programming, such as programming by demonstration, natural language processing, and visual guidance. They can be placed in new positions at any time, and programing and commissioning can be completed rapidly, so they can start working very quickly.



Lower total cost of ownership (TCO) and shorter payback period: The price and annual maintenance cost of collaborative robots are significantly lower than those of traditional industrial robots. According to China's Forward Industry Research Institute, the average selling price of collaborative robots has halved over the past several years [5]. As collaborative robots are adopted more widely, their price will fall even further, meaning investment in these robots will quickly result in positive cash flow.

Collaborative robots are currently most widely used in the manufacturing of computers, communications equipment, consumer electronics products, and automobiles. They are also starting to be used in the medical industry for analysis and testing, liberating medical professionals from repetitive and time-consuming procedures (e.g., urinalysis) and reducing the risk of infection among medical workers by taking care of tasks like throat swabs.

Snapshot from the future: Autonomous mobile robots

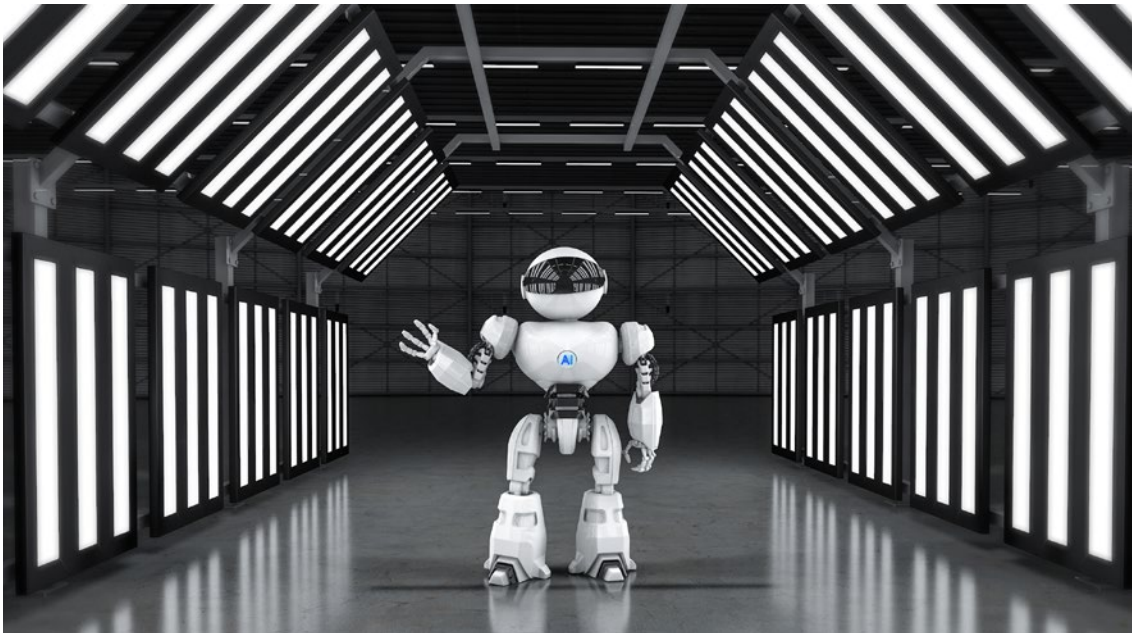
Autonomous mobile robots (AMRs) are a key enabler to help the manufacturing industry become flexible and intelligent. They will reshape production, warehousing, and logistics processes.

AMRs generally need rich environmental awareness. They feature dynamic route planning, flexible obstacle avoidance, and global positioning. The AMRs used in industrial manufacturing and logistics are mainly powered by simultaneous localization and mapping (SLAM) technology to enable autonomous navigation. Their environment does not need to be tagged to enable them to navigate [6].

On production lines, AMRs make automated and unmanned logistics possible. This includes unmanned execution; unmanned interaction between AMRs and other equipment for material collection, feeding, and unloading; and unmanned material handling.

In warehouses, AMRs implement goods-to-person picking and execute intelligent picking, movement, and stock-in and stock-out procedures. In this model, the control system receives an order and assigns an AMR, which then lifts the shelf containing the required goods, moves it to the operator console, and unloads the goods to complete the order. After picking is completed, the robot moves the shelf back to its original position.

The distribution and picking of materials are not confined to factory buildings; AMR systems can be expanded to an entire campus. For example, when goods are unloaded, robots can automatically move them into their designated warehouses. Goods will be automatically logged into and out of warehouses, and the movement of goods between factories or warehouses will be automatically registered. These functions require the robots to support outdoor autonomous navigation, using features such as laser navigation, visual navigation, and satellite positioning.



Snapshot from the future: Industrial humanoid robots

Humanoid robots are designed to have human-like forms and functions, including anthropomorphic limbs, mobility skills, sensory perception, and learning and cognition capabilities. They will likely become the most valuable carriers of “embodied AI”. Combined with rapidly developing general artificial intelligence and AI foundation models, humanoid robots will enable machines to interoperate and interact with their environments in a more intuitive and intelligent manner, and perform a wide variety of complex tasks just like humans.

Industrial humanoid robots can flexibly carry out different operations, move agilely, and independently learn and make decisions. Unlike traditional industrial robots, humanoid robots can complete specific tasks without requiring advanced planning. They can autonomously perceive, understand, learn, and make decisions when completing production line tasks and are capable of powerful autonomous decision making, operations, and interactions. Humanoid robots can currently work in a number of positions in factories, including material handling, quality inspection, labeling, assembly, and high-risk operations.

Humanoid robots can also work around the clock without rest, meaning they can significantly improve both production output and product quality, solve the long-standing challenges caused by labor shortages, and usher in a new era of intelligent manufacturing.

Snapshot from the future: AI-powered adaptive teaching

Conventional education uses the same model to deliver the same course content to different students. AI can transform this industry by analyzing learning models and individual differences between students. This improves the quality of education and makes it possible to teach students in accordance with their aptitudes.

For example, as technologies such as big data, cloud computing, Internet of Things (IoT), virtual reality (VR), and augmented reality (AR) evolve, AI-assisted education will break down learning and teaching behavior in a more granular way and build more robust and precise education models. VR and AR technologies can be used to present materials in a more engaging manner and deliver interactions that suit students' personal preferences, helping students better master their course content.

AI liberates teachers from the repetitive and tedious grading of exam papers and administration, allowing them to focus on the creative work of educational research and one-on-one communication with students. Supported by huge amounts of data generated through educational activities, AI will help teachers better understand the effectiveness of their teaching, and provide key recommendations on the most effective teaching methods and the best way to organize course content.

In schools, AI can be deployed anywhere, and can simulate the best teachers of any subject, bringing the highest-quality education and content to the most remote schools. AI-based education offers multi-channel engagement with students, including video and audio, which can help make up for the scarcity of teaching resources in some areas (for example, in understaffed schools, a teacher may have to teach four or even five different subjects). In this way, AI promotes educational equity [7].

Direction for exploration: New production models geared towards personalized needs

The role of consumers in the production process is changing remarkably. Consumers are being given more say in upstream activities, and can engage in more and more steps during the production process. Under the old model, large-scale production is the norm, where companies design and manufacture products, and consumers simply choose what they want from a range of finished products. With companies now better understanding what consumers really want, they offer more and more product categories to provide consumers with more choices, but this often results in overstocking.

New models such as e-commerce and livestreaming enable companies to more directly and accurately assess customer demand and promptly adjust the size of their production runs to avoid overstocking. Companies can even plan how many resources they will need in advance to avoid overcapacity.

In the future, consumers will become directly involved in design processes. They will be able to express their opinions and even make design decisions. For example, modular designs in the flexible manufacturing process can allow consumers to mix and match and decide on the form factors or style of the products they want. Companies then only start production once the customer has made their choice. This will make the whole production model truly personalized. As modular manufacturing offers more options, consumers will be given more freedom to choose exactly what they want, ultimately leading to a fully personalized production model.

Snapshot from the future: ICT-powered flexible manufacturing

To respond to changing market conditions and set themselves apart in the face of fierce competition, companies must take the initiative and embrace new production models. That's why an increasing number of companies are looking to concepts like flexible manufacturing.

Flexible manufacturing is an advanced production model characterized by on-demand production. It helps companies become more flexible and enables them to rapidly respond to ever-changing market demand. In addition, flexible manufacturing shortens the R&D cycle, cuts R&D costs, and ensures equipment is not left idle, all while reducing inventory risks and speeding up capital turnover. Therefore, it allows companies to seize market opportunities and grow sustainably. Flexible manufacturing involves the following areas:

Flexibility of product design and production line planning: After receiving an order for a new category of product, companies need to quickly

conduct R&D and design, and rapidly adjust factors such as production line equipment, working procedures, processes, and batch size. This is where ICT comes in, as simulation, modeling, VR, and other ICT technologies can be used to simulate the entire new manufacturing process. This will reduce the cost of new product development and design, and support more accurate adjustment cost projections and capacity projections.

Flexibility of process: In flexible manufacturing, companies can design products based on the personalized needs of customers, or invite customers to directly participate in product design (e.g., using modular systems to enable customers to define what a product will ultimately look like). Both models require an intelligent scheduling system. Such a system will make automatic adjustments and provide an optimal production plan based on known features such as the factory's production capacity, order complexity, and delivery deadlines.

After a company receives an order, the scheduling system will automatically identify all universal components, custom components, and procedures and materials required to manufacture these components. By coordinating production tasks and the provisioning of materials and tools, the scheduling system maximizes the productivity of all equipment and workers in the factory so that no component will create a bottleneck in order delivery.

Flexibility of equipment: As the number of customizations and small-batch orders increases, factories must be able to switch between production processes in real time. Conventional manufacturing equipment can generally only be reconfigured by trained engineers using specific programming devices and languages. This makes switchover processes time-consuming, and does not support the kind of rapid responsiveness that companies need. In the future, ICT technologies such as visual programming, natural language interaction, and action capture will help factories reprogram equipment quickly and easily. This will help promptly meet companies' demand for flexible manufacturing.

Flexibility of logistics: One of the keys to flexible manufacturing is modularization, through which a large number of finished components are manufactured. This requires automated ICT technology to effectively manage warehousing and logistics, which prevents omissions and other errors in the shipment process. Take furniture producers as an example. With large-scale customization, every board, decorative strip, and handle may need its own identification code or radio frequency identification (RFID) tag to facilitate automated packing and loading, and to support traceability throughout the whole transportation and distribution process.

Traditional manufacturing follows a "product > place > people" model that forces sales to start with production. As manufacturing becomes flexible, we can reverse that model to "people > place > product" so that production is based on demand, or even reduce it to a "people > product" approach. This will create a new, truly "people-centric" production model.

Direction for exploration: Resilient and intelligent supply chains that help enterprises respond to crises

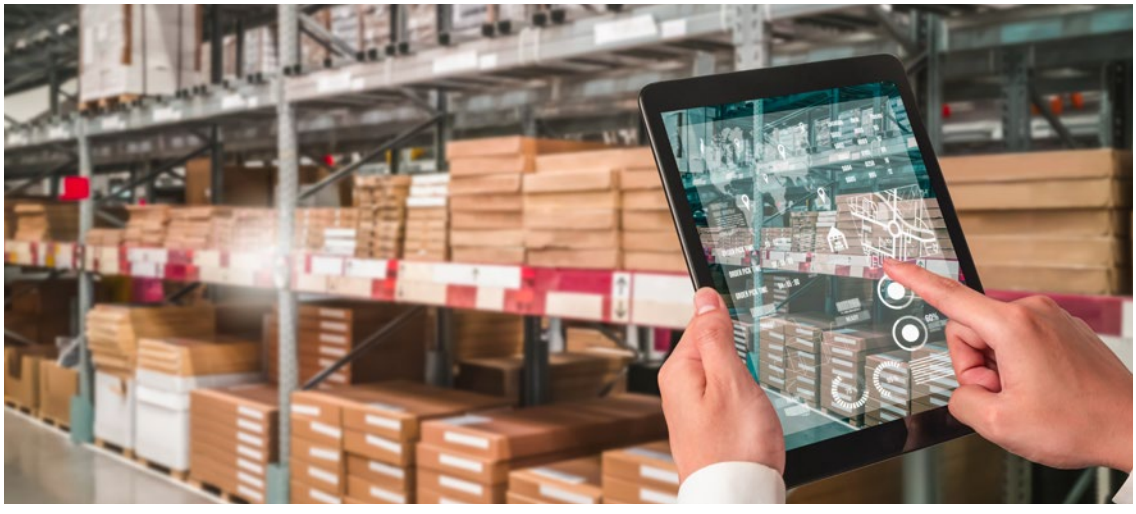
Recent years have seen frequent black swan events, which pose new challenges to traditional supply chains. Facing constant uncertainty, companies want to consolidate their supply systems to enhance their resilience and ensure business continuity. A global supply chain survey conducted by Allianz Research found that 94% of companies reported disruptions to their supply chains because of COVID-19, and 62% of companies said they were considering looking for new suppliers in the long term [8]. More and more companies regard building a resilient and intelligent supply chain as one of their most important strategic priorities.

Snapshot from the future: Supply chain visualization powered by digital technologies

Supply chain visualization is about using ICT technology to collect, transmit, store, and analyze upstream and downstream orders, logistics, inventories, and other relevant supply-chain information, and graphically display such information. Such visualization can effectively improve the transparency and controllability of the whole supply chain and thus greatly reduce supply chain risks.

Supply chain visualization supports the tracking of materials and equipment in upstream activities. Logistics information is displayed in real time, including information on packing, goods logged in, goods logged out, and quality inspections; goods can even be traced throughout the production process.

With supply chain visualization, the operation data of various transportation vehicles in the logistics system is also available, and the status of these vehicles can be displayed in real time. Global Positioning System (GPS), AI, 5G-A, IoT, and other technologies are used to monitor the transportation process and the status of goods while in transit. There is also a visualized



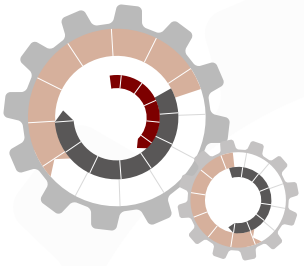
scheduling center that enables the consolidation or splitting of orders at any time, and the optimization of transportation resources and routes. This enables companies to detect and rapidly respond to any logistics emergency by promptly adjusting logistics routes to ensure the timely and safe delivery of goods.

A remote monitoring system monitors the environment in warehouses in real time. This system uses various sensors to graphically display operations and maintenance (O&M) information such as temperature, humidity, dust, and smoke. This allows the timely detection of any signs of fire or water leakage, enabling prompt intervention and preventing material losses. Goods can also be tracked in real time as they are logged into and out of warehouses. As goods are moved, IoT, RFID, and QR code technologies are used to automatically identify and register goods, and the warehousing status data of goods can be accessed remotely in real time.

Snapshot from the future: From supply chain to supply network

In the traditional supply chain model, each link in the chain depends on the previous link delivering as expected. Each link could be a bottleneck that prevents the normal flow of goods down the chain. For example, if the supply of an upstream raw material provider is disrupted, downstream manufacturers will definitely be affected, resulting in inefficient operations or even a standstill for the entire supply chain [9].

With the adoption of ICT technologies such as cloud computing, IoT, big data, and AI, the supply chain will transform into a supply network. In this network, the upstream materials required by every link have multiple alternative sources, and they can be sourced through multiple routes. A multi-contact collaborative supply ecosystem will be created by enhancing the internal and external interconnectivity of enterprises. The failure of any single link will not paralyze the entire supply network.



Conclusion:

New productive forces will reshape production models and enhance resilience

By 2030, digital technologies will be transforming companies. Technologies such as AI, sensors, IoT, cloud computing, 5G-A, and AR/VR are poised to create new productive forces. They will help make up for labor shortages, so that companies can seize new business opportunities and expand their possibilities.

Huawei predicts that by 2030, every 10,000 workers in manufacturing companies will work with 1,000 robots, and the number of VR and AR users will reach 1 billion. Furthermore, one million companies are expected to build their own 5G-A private networks (including virtual private networks). In addition, cloud services are forecast to account for 87% of enterprises' application expenditures, while AI computing will account for 7% of a company's total IT investment.

In the future, product design, process design, equipment functions, logistics, and distribution will all be reshaped to become more flexible and serve new people-centric production models. As 3D printing advances and becomes more widely adopted in commercial settings, mold manufacturing, production line adjustment, and many other activities can be eliminated. This will give consumers much more say in the design and production process, and personalized production models will be created. Powered by digitalization, supply chains will be visualized and expand into supply networks. Production sites will be upgraded to serve as the foundation of new industrialization and come fully equipped with new industrial production networks and other advanced technologies. The manufacturing industry will enter the new production era, enabling companies to develop the resilience they need in the face of volatile markets.

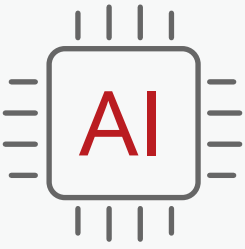
global industry vision

Huawei predicts that by 2030,

Every **10,000** workers will work with **390** robots.

1 million companies are expected to build their own 5G private networks (including virtual private networks).

Cloud services are forecast to account for **87%** of enterprises' application expenditures.



AI computing will account for **7%** of a company's total IT investment.

References

1 World Population Prospects 2019, UN.

2 Future of Work – The Global Talent Crunch, Korn Ferry, https://www.kornferry.com/content/dam/kornferry/docs/pdfs/KF-Future-of-Work-Talent-Crunch-Report.pdf.

3 Future of Work – The Global Talent Crunch, Korn Ferry, https://www.kornferry.com/content/dam/kornferry/docs/pdfs/KF-Future-of-Work-Talent-Crunch-Report.pdf.

4 An Updated Assessment of the Economic Impact of COVID-19, Cyn-Young Park et al., https://www.adb.org/publications/updated-assessment-economic-impact-covid-19.

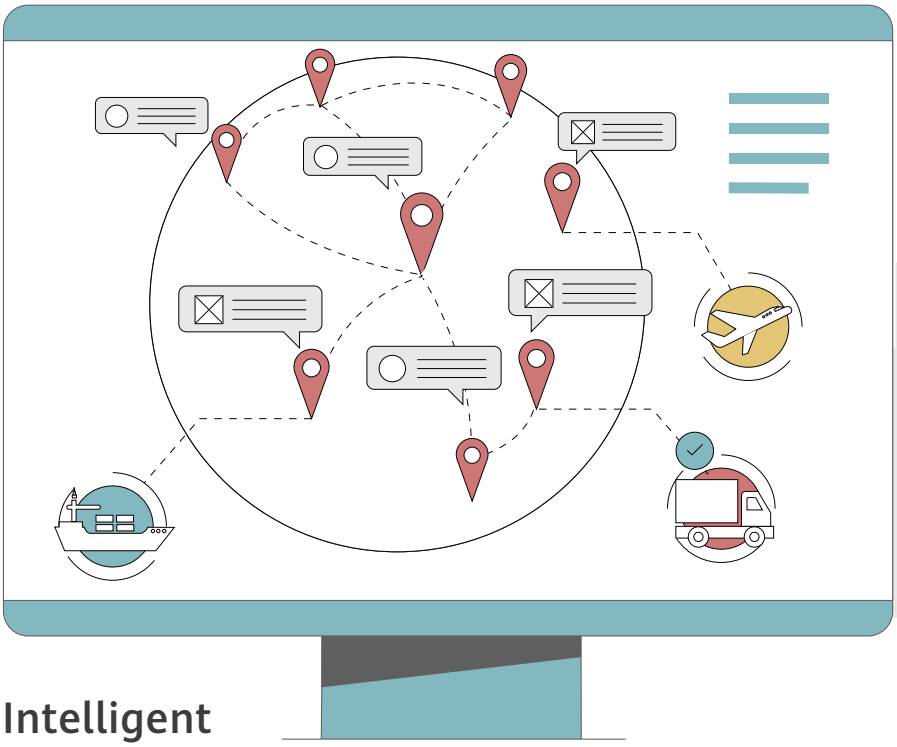
5 Analysis of the Current Market Conditions and Competitive Landscape of the Chinese Collaborative Robot Industry in 2020, Forward Industry Research Institute.

6 Research Report on the 2020–2021 Development of the AGV/AMR Industry in the Industrial Manufacturing Domain, China Mobile Robot (AGV/AMR) Industry Alliance.

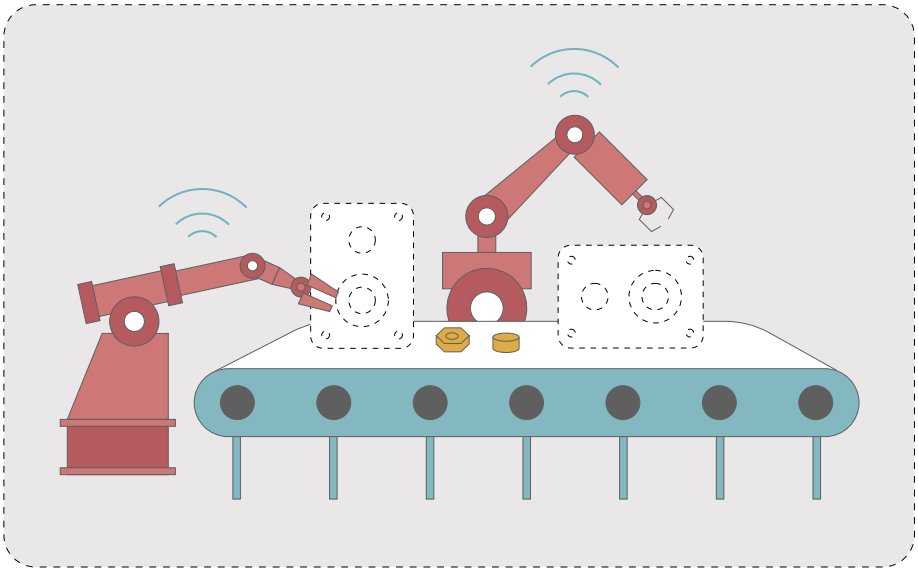
7 Research Report on the 2020–2021 Development of the AGV/AMR Industry in the Industrial Manufacturing Domain, China Mobile Robot (AGV/AMR) Industry Alliance.

8 Global Supply Chain Survey in Search of Post-COVID-19 Resilience, Allianz Research, https://www.eulerhermes.com/content/dam/onemarketing/ehndbx/eulerhermes_com/en_gl/erd/publications/pdf/2020_10_12_SupplyChainSurvey.pdf.

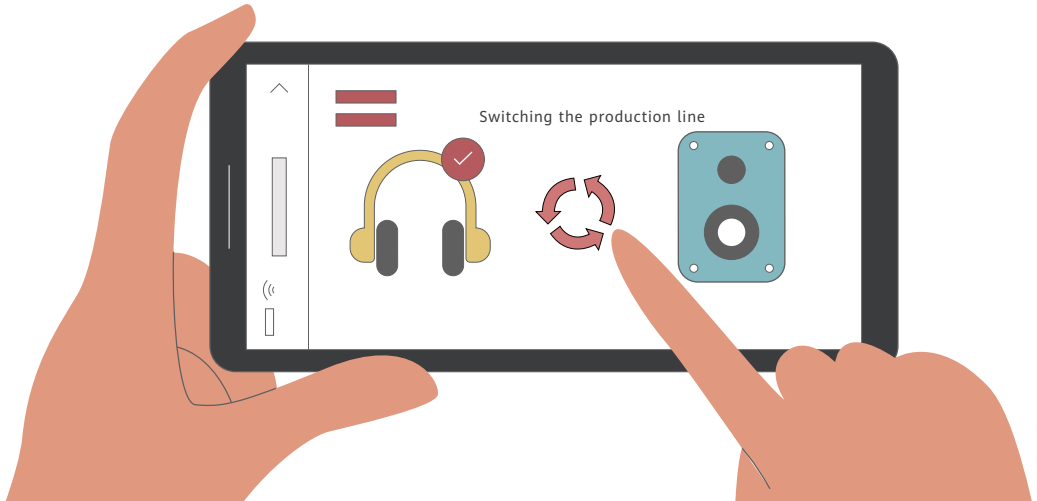
9 Six Secrets to Supply Chain Digitalization, Accenture, https://www.accenture.com/_acnmedia/pdf-106/accenture-supply-chain-services-pdf.pdf.



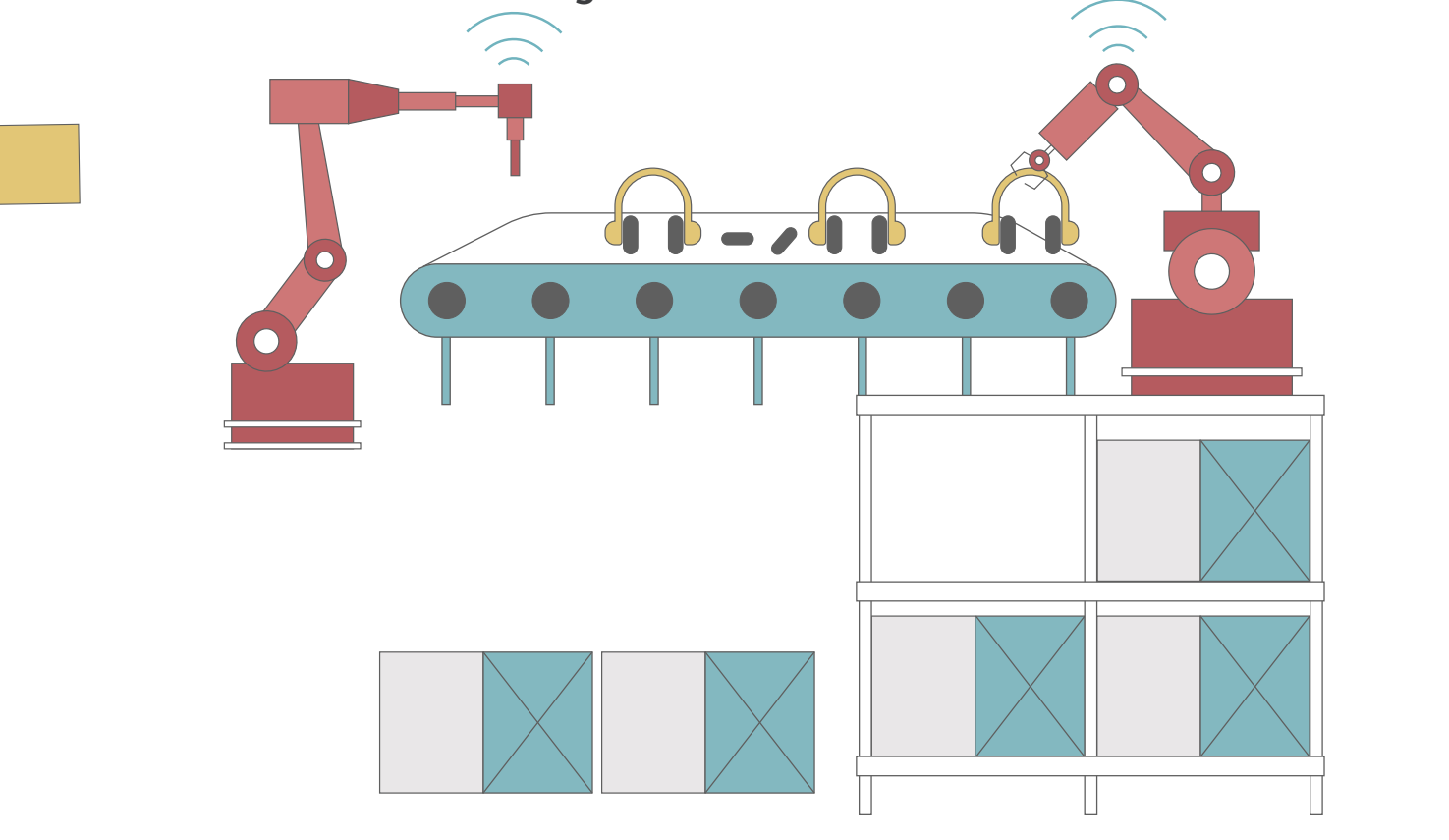
Intelligent supply chains

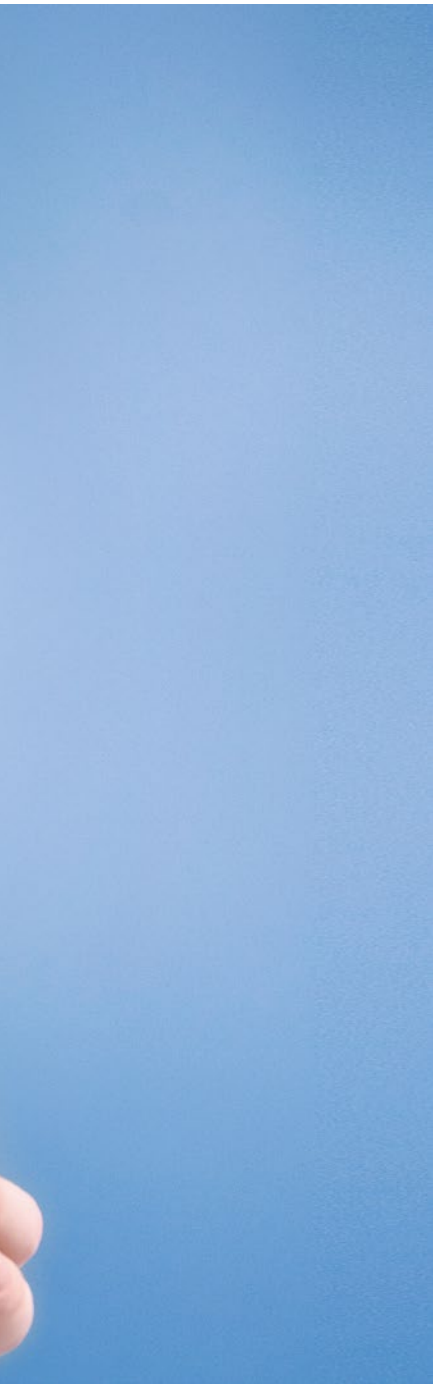


New production models



Unmanned operations in manufacturing and services





Energy

Intelligent, Green Energy for a Better Planet

Climate change is becoming an increasingly pressing concern each and every day. According to the State of the Global Climate 2023, released by the World Meteorological Organization in March 2024, records were once again broken regarding several climate change indicators, such as global greenhouse gas concentrations, surface temperatures, ocean heat and acidification, sea level increases, and the retreat of Antarctic sea ice coverage and glaciers. The global mean near-surface temperature in 2023 was 1.45°C (±0.12°C) above pre-industrial levels, and the past 10 years have been the warmest decade on record. Furthermore, the global mean sea surface temperature has reached a record high since April 2023, and Antarctic sea ice coverage fell to a record low. At the end of winter in 2023, the maximum sea ice coverage was 1 million square kilometers less than the previous lowest coverage recorded. Concentrations of three major greenhouse gases, namely, carbon dioxide (CO_2), methane, and nitrous oxide, continued to rise at record levels in 2022, with CO_2 concentrations 50% higher than the pre-industrial levels [1]. Climate change is also closely linked to economic development. The International Monetary Fund found that for a medium- or low-income developing country with an annual average temperature of 25°C, each 1°C increase in temperature leads to a decrease in economic growth of 1.2% [2].

Climate change is a global challenge that many countries have come together to tackle. At the COP 21 UN Climate Conference in 2015, parties to the Paris Agreement agreed to intensify efforts to limit global warming to below 2°C, preferably 1.5°C, compared to pre-industrial levels, and set the goal of reaching net zero CO_2 emissions globally by around 2050. In other words, by the middle of this century, the CO2 emitted by human activities needs to be matched by the amount of CO_2 deliberately taken out of the atmosphere. At the UN General Assembly in September 2020, China pledged to peak its carbon emissions by 2030 and achieve carbon neutrality by 2060. According to the UN Environment Programme (UNEP), annual emissions need to be 15 gigatons of CO_2-equivalent (GtCO2e) lower than current unconditional Nationally Determined

Contributions (NDCs) by 2030 in order to even hit the 2°C goal. Limiting the rise in global temperatures to below 1.5°C would thus require an even greater decrease in global emissions by 2030 [3]. According to the Global Stocktake Report released at the COP 28 UN Climate Change Conference, global greenhouse gas emissions must be cut by 43% by 2030, compared to 2019 levels, to limit global warming to below 1.5°C. The stocktake calls on all parties to take actions toward achieving, on a global scale, a tripling of renewable energy capacity and doubling energy efficiency improvements by 2030 [4].

Concerted efforts are needed to combat climate change and drive the transformation of the global energy mix in three key areas: energy supply, consumption, and carbon fixation. On the supply side, renewable energy should be used wherever possible as a cleaner alternative to fossil fuels, such as in electricity generation and hydrogen production. This means a shift in the energy production model. The International Energy Agency (IEA) predicts that the renewable energy share in electricity generation, which currently stands at 30%, will increase to 68% by 2030 [5].

On the consumption side, fossil fuels will need to give way to electricity in the transport, industrial, agricultural, and construction sectors, changing the way energy is used. The IEA predicts that the share of electricity in final energy consumption will increase from the current 22% to 29% by 2030 [6]. Carbon fixation is another option if some carbon emissions prove unavoidable. In such cases, technologies like soil carbon sequestration and carbon capture and storage can be utilized, alongside ecological improvement efforts, to remove carbon from the atmosphere.

As the share of renewables in energy networks continues to increase, challenging the conventional architecture of the energy industry and energy supply chains, a paradigm shift is occurring. With the increasing complexity of energy networks and increasing digitalization of the energy sector, digital and power electronics technologies have become an important part of decarbonization solutions. Today, the key questions for global warming are: How can we further increase the share of renewables in the energy mix? How can we adapt to the new energy mix? How can we fully harness the power of technologies?

▶ Direction of exploration: Solar PV and energy storage industries are developing at an accelerated pace, expanding from a few countries to the whole world

The International Renewable Energy Agency (IRENA) announced that the globally installed capacity of renewables increased by 473 gigawatts (GW) or 14% in 2023, compared to 2022. Of this, 347 GW came from solar power, representing an increase of 33%, and 114 GW from wind, an increase of 13% [7]. By 2050, renewables will account for over 90% of the world's total electricity generation [8]. In 2020, the Chinese government declared its goal to raise the total installed capacity of solar and wind power to above 1.2 billion kW by 2030 [9], and this was achieved in 2024, ahead of schedule. According to the goals of the German federal government, the share of renewables in the country's electricity mix will exceed 80% by 2030 [10].

Among renewables, the overall electricity generation cost of solar PV and energy storage is decreasing. Distributed solar PV and energy storage have achieved grid parity in 50% of countries around the world, and the internal rate of return (IRR) in 30% of countries has exceeded 10%. In 2015, only seven countries or regions achieved an annual installed PV capacity of over 1 GW. This number rose to 11 by 2018, and is expected to reach 38 by 2024 [11]. In the future, solar PV and energy storage will see rapid deployment across the globe.

Snapshot from the future: Safety, stability, and reliability are critical to utility-scale renewable power plants that feature a synergy of wind, solar, hydro, thermal power, storage, and hydrogen

Utility-scale power plants achieve economies of scale, reduce unit energy costs, and improve energy utilization through centralized management and optimized energy configuration. Such power plants are typically constructed in areas that are rich in water, sunlight, and land resources. These plants are equipped with efficient energy transmission networks to transmit clean energy to other regions over long distances, meeting extensive energy demands. Advanced technologies are usually applied and promoted in utility-scale power plants, driving the technological progress and upgrade of energy-related industries.

China is currently following its plan to build nine utility-scale renewable power plants in a number of regions, including Xinjiang, the upper reaches of the Yellow River, and the integrated renewable power plants of hydro, wind, and solar in Sichuan, Yunnan, Guizhou, and Guangxi, from 2021 to 2025. The total capacity of renewables in Southwest and West China will ultimately exceed 600 GW [12]. One renewable power plant located in Morocco, North Africa, consists of a 10.5 GW wind and PV power system and a 5 GW/20 GWh energy storage system (ESS). This plant transmits electricity to the UK through submarine cables over a distance 3,800 km, meeting 7.5% of the UK's electricity demands [13].

Utility-scale renewable power plants that feature a synergy of wind, solar, hydro, thermal power, storage, and hydrogen are attracting increasing attention. The plants preferably harness clean energy sources like wind and solar, leverage the adjustment performance of hydro and coal-fired power, and collaborate with energy storage facilities and hydrogen energy systems. This allows multiple types of energy resources to be developed in a coordinated manner and scientifically configured. In March 2022, China released the 14th Five-Year Plan: Modern Energy System Planning (2021–2025). The document proposed to promote the construction of renewable power plants, scientifically optimize the energy mix, and preferentially use existing conventional power supplies to implement mutual complementation of energy sources including wind, solar, hydro, thermal power, and storage. The Hydro-Solar Hybrid Power Plant at Lianghekou Dam on the Yalong River in Sichuan Province, China has the world's largest installed capacity and is located at the highest altitude (4,000–4,600 m) among all projects of its kind anywhere in the world. The plant has increased its development scale to GW-level for the first time, generating 2 billion kWh of electricity and reducing carbon emissions by 1.6 million tons each year. [14]

Utility-scale renewable power plants face numerous safety challenges. High proportions of renewables and power electronic equipment make it challenging to transmit, integrate, and consume renewables that are variable, intermittent, and fluctuating. Large land footprints and remote locations further increase the challenges related to plant operations & maintenance (O&M), equipment reliability, and PV+ESS safety. Effective solutions should be developed to address these challenges and ensure that renewables can be produced and utilized in a safe, stable, and reliable manner.

Snapshot from the future: Comprehensive development of solar PV and energy storage in utility-scale, commercial & industrial (C&I), and residential scenarios

At the early stages of PV development, subsidies were provided to promote PV construction around the world due to the high levelized cost of electricity (LCOE) of PV. For example, the National Development and Reform Commission of China released the Notice on Using Price Leverage to Promote the Healthy Development of the PV Industry in 2013. According to the document, feed-in tariff or subsidy standards shall be implemented once PV power generation projects are initiated. In policy-driven market development, most PV projects were utility-scale plants, accounting for more than 60% of total installed PV capacity.

The LCOE of PV power has been reduced by more than 90% thanks to technological advances, and PV power has achieved grid parity in most regions. The return on investment (ROI) in C&I and residential PV scenarios has also been rapidly increasing. PV business is booming in China, Europe, and Japan, but the utility-scale PV plant market remains the most important, with the related installed capacity accounting for about 50% of the total. The industry has been jointly driven by both policies and markets.

As PV and energy storage costs continue to decrease, all-scenario commercialization and co-development are becoming the mainstream business model. The installed capacity of utility-scale plants now accounts for about 40% of the total, and new scenarios like clean energy bases and city-level microgrids are emerging. The growth of C&I businesses is accelerating in all walks of life, and C&I plants are now being applied to new scenarios like mines and islands. Furthermore, residential businesses are expanding from the US and Europe to emerging markets such as Asia Pacific, Latin America, and Africa. According to BloombergNEF, the compound annual growth rates (CAGRs) of the installed capacity in utility-scale, C&I, and residential scenarios reached 6.2%, 18.8%, and 12% respectively within five years from 2015 [15].

Direction for exploration: Renewable electricity generation: Floating power plants

The rapid development of onshore wind and solar projects is forcing us to confront problems such as land shortages, distances from electrical load centers, reduced efficiency of solar photovoltaic (PV) systems under high temperatures, and biodiversity loss. A new trend for the future, particularly apparent in island nations, is building wind and solar power installations offshore to take advantage of the excellent geographical features and abundant space of near-shore locations.

Snapshot from the future: Offshore wind, a promising energy source for the future

Some countries are actively using offshore power generation. In 2023, the global installed offshore wind capacity increased by 10.8 GW, an increase of 24% over the previous year. China took the lead with a growth of 6.3 GW, followed by Europe with a growth of 3.8 GW. In spite of the large number, offshore wind energy feeds just 0.3% of electricity consumption globally, meaning great potential to be unleashed for development [16]. Thanks to a large number of technological innovations that have reduced the installation and operating costs of offshore wind farms, offshore wind is expected to see rapid growth.

Offshore locations offer higher wind intensity and offshore wind turbines are productive for a greater proportion of time. Furthermore, thanks to new technologies, offshore wind turbines can be larger than their onshore counterparts, and consequently have a higher capacity factor.

$$P=1/2\ \rho A V^3\ Cp$$

The equation above is used to calculate the power output of a wind turbine. The generated power, P, is proportional to both the cube of the wind speed, V, and to the swept area of the turbine, A.

Offshore wind is better than onshore wind, because when wind flows over rough ground surfaces or obstacles, it changes speed and direction. Sea surfaces are less rough and there are fewer obstacles. On average, the wind 10 km offshore is 25% faster than wind at the shoreline [17]. In addition, offshore wind is less turbulent and wind direction is more consistent. As a result, turbines suffer less fatigue, and the service life of offshore wind power equipment is longer. The swept area of a wind turbine depends on the diameter of the rotor. In 2021, offshore wind turbines with a rotor diameter of 164 meters and a capacity of 10 megawatts (MW) became available. By 2030, an offshore wind turbine is expected to have an average rotor diameter of 230–250 meters and a capacity of 15–20 MW, which is three to four times the capacity of an onshore wind turbine. There are fewer periods of calm sea, so offshore wind turbines can generate power for 3,000 hours a year, compared to 2,000 hours a year for onshore counterparts [18]. This makes for more efficient use of generator capacity. Furthermore, with advances in technology, the capacity factor of offshore wind power can be 40–50%, higher than that of onshore wind and twice that of PV systems. In many areas, the capacity factor of offshore wind is close to that of natural gas and coal [19], meaning

offshore wind energy generation has the potential to become a baseload technology.

Currently, offshore wind turbines are mainly deployed in shallow water areas less than 40 meters deep, within 80 kilometers of coastlines, and are fixed by single-standing piles. However, new floating turbine technologies offer an alternative with simpler installation and lower costs. Floating turbines can be installed in water up to 60 meters deep and are supported by the new high-voltage direct current (HVDC) technology that offers a more cost-effective solution for transmission at a distance of 80–150 kilometers from coasts. These innovations have greatly expanded the potential of offshore wind .

Diverse innovations have led to significant reductions in the costs of offshore wind installations. This means we are about to experience a boom in offshore wind power. The IEA forecasts that the cost of offshore power generation in 2040 will be 60% lower than that in 2019 [21]. The Global Wind Energy Council (GWEC) forecasts that the global offshore wind capacity will increase from 75 GW in 2023 to 275 GW by 2030. The offshore wind capacity is also expected to grow by 25% per year over the next five years [22]. The IEA forecasts that offshore wind will become Europe's largest source of electricity by 2040 [23].

Snapshot from the future: Floating power plants, a new deployment model

According to the International Renewable Energy Agency (IRNEA), the total globally installed capacity of solar PV at the end of 2023 was 1,411 GW [24]. Onshore PV power plants are the most common form of PV installation, but there are numerous problems associated with onshore solar plants: land acquisition, high costs, and low efficiency under high temperatures. As a result, floating PV (FPV) is a new direction for solar PV.

FPV plants can be installed in near-shore marine areas, ponds, small- and medium-sized lakes, reservoirs, river basins, or flooded mining pits. There are three types of FPV installations: thin film, submerged, and floating arrays. Thin-film modules are lightweight and do not require the support of a rigid pontoon structure. Submerged PV installations can be supported with or without a pontoon structure, while floating arrays must be supported by rigid pontoons.

Compared with land-based PV (LBPV) systems, installation of FPV systems on water saves land for agricultural use. The lack of obstacles on the surface of the water means less shading loss and less dust. In addition, the natural cooling potential of the body of water may enhance PV performance, thanks to higher wind speeds offshore, along with the presence of water. In 2020, a research team from Utrecht University in the Netherlands simulated an FPV system on the North Sea. They found that the apparent temperature of PV modules at sea was much lower than that on land, due to higher relative humidity and higher wind speeds. The average ambient temperature difference between the two locations was 5.05°C, but the apparent temperature difference of PV modules was nearly doubled—9.36°C. This study also found that an FPV system could output about 12.96% more energy on average, on an annual basis, than an LBPV system [25].

As technologies mature, a rapid growth is anticipated in FPV. On July 14, 2021, Singapore's Sembcorp Industries unveiled a floating solar plant deployed on the Tengeh Reservoir. With 122,000 solar panels spanning 45 hectares (equivalent to about 45 football fields), the 60 MW solar farm is one of the world's largest inland floating PV systems [26]. According to Rethink Energy, the global FPV market capacity will exceed 60 GW by 2030 [27] and the estimated potential global capacity is 400 GW [28] . The floating solar market is set to accelerate as technologies mature, presenting new opportunities to scale up global renewables.

Direction for exploration: The future energy world will be centered on electricity, and green hydrogen will become a big player

Replacing fossil fuels with electric energy is crucial in reducing carbon emissions. As the costs of renewables such as wind and solar power decrease, the global electrification process continues to accelerate. Technological progress has also been made in terms of hydrogen production, storage, transport, and use. With emerging applications in sectors like the chemical industry, transportation, and power generation, green hydrogen will become an important supplement to electric energy. According to IRENA, the proportion of green hydrogen for direct consumption will reach 3% by 2030, and that of traditional fossil fuels will decrease from 63% in 2020 to 47% [29].

Snapshot from the future: Accelerated adoption of electricity in sectors such as industry and transportation

Sectors like industry and transportation are the main sources of carbon emissions through energy consumption. According to IEA, the electric power sector accounts for about 40%, the industrial sector accounts for about 24%, and the transportation industry accounts for about 21% [30]. In the industrial sector, CO_2 emissions mainly come from traditional manufacturing industries such as ferrous metals, non-metallic minerals, petrochemicals, and non-ferrous metals. To reduce emissions, priority should be given to green transformation in traditional sectors by promoting green electricity and electric manufacturing. For example, the iron and steel industry can adopt electric arc furnaces to reduce the use of fossil fuels, while the non-ferrous metal sector can apply green electricity to reduce emissions. The building material and cement sector can also increase electrification through raw material substitution and waste heat recovery for electricity generation. Furthermore, the heating sector can use heat pumps instead of natural gas or fuel-fired boilers to heat buildings.

Vehicle-based road traffic accounts for nearly 74.5% of the total carbon emissions in the transportation sector [31]. The number of new energy vehicles (NEVs) on roads has grown faster than expected in recent years. In China alone, this number exceeded 18 million by the end of 2023, and is expected to reach 180 million by 2034, representing a 10-fold increase over the next 10 years. By the end of 2023, the number of commercial NEVs on the road in China had reached 2.44 million and is expected to surpass 22 million, a 9-fold increase, over the next 10 years. In a word, mobility electrification has become an irreversible trend [32]. Railway transportation will also be further electrified to optimize overland transport. Currently, maritime transport is partially electrified, and the electrification level of ports and terminals is constantly being improved. Thanks to short transportation distances, inland waterway transport is set to adopt more electrification and fuel cell technologies. We should therefore focus on optimizing the transportation structure, developing green mobility, and constructing more renewable energy infrastructure to build an intelligent and electric transportation system while applying technologies such as smart grids, 5G, and AI. This will reduce carbon emissions and contribute to green and low-carbon cities.

Snapshot from the future: Promoting an energy balance across time and distance through diverse energy storage technologies

With more electricity being generated from renewables, power systems are experiencing ever-increasing challenges in terms of safety and stability because solar and wind energy sources fluctuate and are intermittent, especially in extreme weather. By combining energy storage technologies with grid forming and grid following technologies, ESSs can be integrated with power systems based on renewables like solar and wind to effectively improve the quality and controllability of renewable energy output. This improves renewable integration and reduces carbon emissions. Moreover, ESSs can offer diverse ancillary services, such as frequency and phase regulation, reserve, and black start, improving the stability and flexibility of power grids.

Global electricity demand continues to grow, with peak load becoming increasingly prominent and extensive. Peak load is characterized by a short duration, low frequency, and small amount of electricity. This makes it costly and inefficient to balance the electricity demand by simply increasing power grid investment. However, ESSs can store electric energy during off-peak hours and discharge that energy during peak hours for peak shaving and load balancing, thus improving the operating efficiency and reliability of power grids while cutting power system investment. In scenarios with various energy demands, such as industrial parks and urban public buildings, ESSs can work alongside distributed renewable power systems through flexible deployment and efficient management, thus achieving friendly interactions and efficient operations regarding generation-grid-load-storage.

As a key measure for improving the regulation capacity of power systems, pumped storage is expected to remain the main force of adjustable resources in power systems until 2030. By the end of 2023, the globally installed capacity of pumped storage had reached 194 million kW, and that of new-type energy storage reached 91.3 million kW [33]. As planned by China's National Energy Administration, the installed capacity of pumped storage in China will exceed 120 million kW by 2030 [34].

Multiple new energy storage technologies will coexist, such as compressed-air energy storage, electrochemical energy storage, and thermal (cold) energy storage. Based on business models such as system-friendly "renewables + energy storage" power plants, energy storage for utility-scale renewable energy plants, standalone energy storage for power grids, shared energy storage, and user-side energy storage, diverse installations will be deployed on the generation, grid, and load sides to meet system regulation requirements.

Electrochemical energy storage products, such as lithium-ion and sodium-ion batteries, are widely used across energy storage scenarios on the generation, grid, and user sides thanks to advantages such as high energy density, long cycle life, high efficiency, and quick response. Among them, sodium-ion batteries are applauded for high safety and large material reserves. Flow batteries apply to scenarios that require long-duration discharge, including large-scale energy storage and long-duration energy storage. Compressed-air energy storage applies to large-scale energy storage scenarios, such as energy transfer and peak shaving. Flywheel energy storage can respond rapidly and thus is mainly used in scenarios like power grid frequency regulation, short-duration power support, and power quality improvement.

Supercapacitor energy storage applies to scenarios involving short-duration and high-power load smoothing and peak power management, such as startup support and dynamic voltage recovery of high-power DC motors. Thermal energy storage technologies, which include sensible heat storage, phase-change heat storage, and thermochemical heat storage, can be widely used in fields such as photothermal power generation, clean energy-based heating, and thermal power flexibility enhancement.

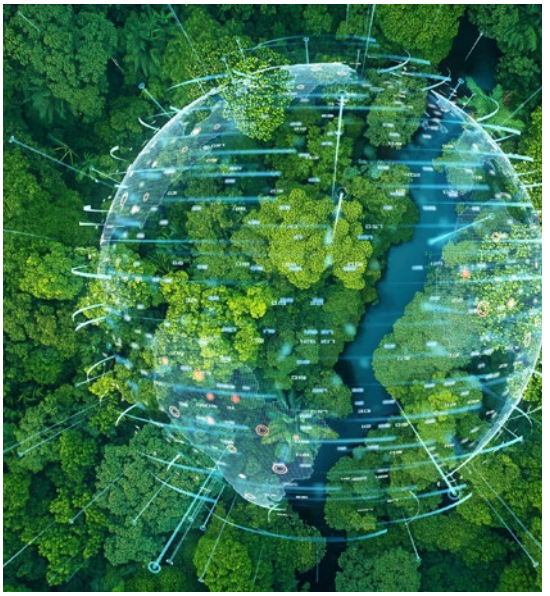
Snapshot from the future: Green hydrogen will see more extensive application

New technologies and business models, such as green transportation, hydrogen metallurgy, hydrogen production from renewables, ammonia/methanol synthesis by green hydrogen, and hydrogen-based power generation, will all be widely promoted. Electricity plays a pivotal role in energy systems. It interacts with secondary energy sources like hydrogen through electricity-to-hydrogen conversion and electric fuel production, helping build a multi-energy complementary system that interconnects multiple energy sources with electric energy. In fields such as metallurgy, chemical industry, transportation, and power generation, hydrogen, as a reacting substance or raw material, has become essential to clean electricity. Together with electric energy, hydrogen is used to build an energy consumption system that focuses on electricity-hydrogen collaboration, helping the global community move closer to net zero.

Countries worldwide consider hydrogen to be an integral part of the green energy transition for sustainable development. They have intensified efforts to deploy hydrogen production equipment, hydrogen fueling facilities, hydrogen storage and

transportation systems, and hydrogen fueling networks, develop key technologies, and cultivate hydrogen professionals. China plans to build a comprehensive hydrogen technology innovation system and renewables-based hydrogen production and supply system by 2030. These systems will help establish a well-managed industry landscape and expand the application of renewables-based hydrogen production to hit carbon emission peak earlier than initially planned. By 2035, a hydrogen industry system will be established to shape a diversified hydrogen application ecosystem covering fields like transportation, energy storage, and industry. The share of renewable hydrogen in final energy consumption will increase significantly, boosting the green energy transition. It is estimated that by 2030, the installed capacity of electrolyzers in China will exceed 100 GW, and the annual production of green hydrogen will exceed 7.7 million tons [35]. In September 2022, the U.S. Department of Energy released the U.S. National Clean Hydrogen Strategy and Roadmap, proposing that clean hydrogen will contribute to about 10% of carbon emissions reduction by 2050, and that U.S. demand for clean hydrogen will reach 10, 20, and 50 million tons per year by 2030, 2040, and 2050 respectively [36]. In 2022, Europe released the REPowerEU Plan, which proposed multiple strategies for promoting hydrogen energy development. By 2030, Europe will both produce and import 10 million tons of green hydrogen to ensure hydrogen energy supply. The European Hydrogen Energy Bank has also been established, which will invest 3 billion euros to develop the hydrogen energy industry [37]. According to the Global Hydrogen Review 2023 by the International Energy Agency (IEA), the global installed capacity of green hydrogen will reach 51 million tons by 2030 and that of electrolyzers will reach 500–600 GW [38].

Direction for exploration: Digital and intelligent generation-grid-load-storage-consumption through the Energy Internet



As generation-grid-load-storage synergy accelerates and deepens, the boundaries of the traditional value chain will be broken and power systems will no longer adjust electricity generation simply based on plans and loads. In addition, electricity supply and demand will become more flexible and random. Therefore, to drive the digital and intelligent transformation of electricity, next-generation technologies for enabling digital transformation will be vigorously developed and extensively applied in numerous areas including digital edge (edge-device collection and control), ubiquitous communication networks (terrestrial communication and satellite communication), compute and storage (cloud platform, cloud-edge-device collaboration, spatial computing, and blockchain), and algorithms and applications (such as artificial intelligence, graph computing, and advanced analytics). The physical world and digital space will be fully connected, device information and production processes in power systems will be converted into digital expressions, and digital mirrors of power systems will be built within the virtual space. In addition, alongside the advancement of digital capabilities such as digital surveillance, intelligent analytics, and digital and intelligent autonomy, the physical world and digital space will evolve from virtual-physical mapping to in-depth synergy, creating digital twins of entire power systems.

The digital twins of power systems can take three forms:

1) Digital surveillance: The purpose of this is to comprehensively and accurately reflect the running process and status of power equipment assets in the digital space in real time through ubiquitous sensing, high-speed communication, and platform storage. Then, equipment assets can be dynamically monitored and diagnosed throughout their lifecycle based on multi-dimensional data. This allows "bits to be used to sense watts" in various electricity scenarios. Sensing networks and mechanism models can also be built to facilitate the efficient digital monitoring of power systems. In addition, data interworking and ubiquitous IoT require data encryption technologies to ensure information security.

2) Intelligent analytics: The purpose is to analyze, predict, and simulate future operation changes regarding generator sets, power transmission and distribution networks, and power loads based on the determined operation modes, mechanisms, and rules, thus providing decision-making support for operation optimization and system control based on the existing system. This allows "bits to be used to manage watts" in various electricity scenarios. Compute and algorithms are core

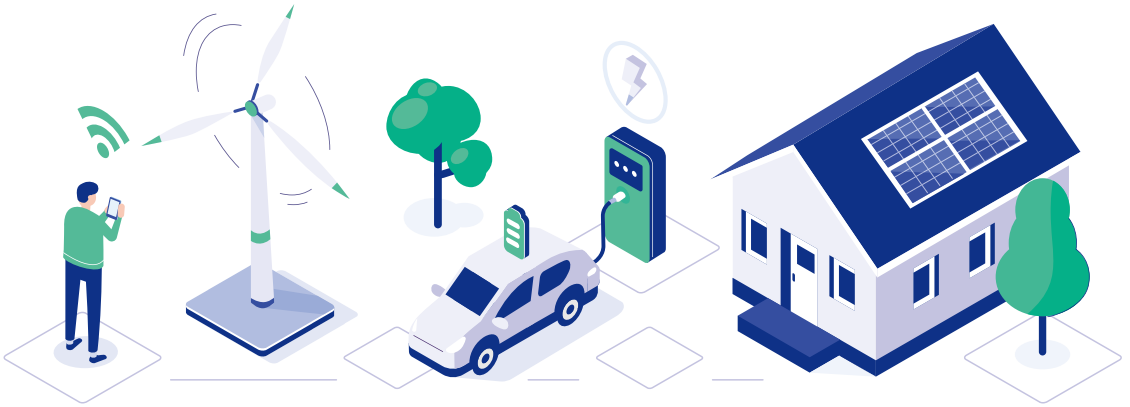
technologies for improving the accuracy of intelligent analytics in power systems. By building complex data models that cover multiple domains and disciplines and simulating the digital space, optimization and decision-making can be enabled for physical entities through an effective closed-loop process.

3) Digital and intelligent autonomy: Massive data across systems and modules, adaptive and self-evolving complex algorithm models, and intelligent achievements shared by the digital space can be leveraged to proactively identify bottlenecks in the current operation mode of the physical world. Then decision-making instructions or predictive reconstruction solutions can be proposed to promote in-depth interactions between the physical world and the digital space through decision-making autonomy and achievement feedback through the digital space. This allows "bits to add more value for watts." As a large amount of cross-system data needs to be exchanged and shared, advanced technologies like blockchain and privacy computing, in addition to AI technologies such as advanced analytics, are key.

Snapshot from the future: Virtual power plants, a paradigm shift for the power value chain

The emergence of virtual power plants (VPPs) is redrawing the boundaries between power producers and power consumers, and they are poised to reshape the power generation value chain. IRENA defines a VPP as "a system that relies on software and a smart grid to remotely and automatically dispatch and optimize distributed energy resources. By orchestrating distributed generation, solar PV, storage systems, controllable and flexible loads, and other distributed energy resources, VPPs can provide fast-ramping ancillary services, replacing fossil fuel-based reserves.

VPPs aggregate distributed heterogeneous energy sources. Distributed energy sources include innovative renewable energy generation systems, such as rooftop PV plants and small-scale wind power plants. They also include industrial and household energy systems, such as heating, ventilation and air conditioning (HVAC) systems, electric heating pumps, and battery-based hydrogen production systems. To offset the variability of renewable energy generation, VPPs may also be connected to conventional energy sources like small gas-fired power plants, small



hydropower plants, and diesel generators. As electric vehicles and household energy storage continue to develop, they will be incorporated into the heterogeneous energy equipment connected to VPPs.

VPPs do not change physical network structures. Instead, they aggregate scattered and independent power supply devices, energy storage devices, and controllable loads through the use of a software platform. In addition, through flexible dispatching management and efficient interactions with the power grid, VPPs supply power to the power system and receive surplus power from the system to keep power balanced over time and across regions, thus improving the security and renewable integration of the power grid. In addition, AI and big data technologies contribute to optimal dispatching. Intelligent dispatching decision-making is one of the key capabilities of VPPs. Reasonable and effective dispatching plans are made for aggregated power supplies and energy storage devices to ensure the balanced operation of the power grid, improve energy utilization, and boost the economic benefits and participation enthusiasm of each aggregator.

VPPs make dispatching decisions based on two aspects: (1) Supply: VPPs can participate in power trading in two ways. First, VPPs determine the adjustable amount of power that directly participates in spot trading based on the power generation capability of adjustable power supplies. Second, VPPs participate in electricity ancillary services such as peak shaving and frequency regulation based on the demand-side response requirements of the power grid to suppress power grid fluctuations and ensure power grid balance. With this method, accurately judging changes in the output of adjustable power supplies or the demand of controllable loads is essential to dispatching decision-making. AI technologies can also significantly improve the accuracy of prediction models. By establishing the impact relationship between climate, power generation performance, and energy consumption demands through sparse modeling, ensemble learning, or other machine learning methods, VPP operators can accurately predict supply and demand curves based on future weather changes.

In addition, as multi-source distribution networks are complex, the operating status of distributed power supplies or energy storage devices may be changed and adjusted at any time. Therefore, VPP operators must also monitor the status of the aggregated adjustable power supplies in real time. When the device status or output changes, the prediction model parameters can be updated and adjusted in a timely manner, and the output policy can be dynamically optimized based on the real-time prediction result in order to realize flexible dispatching. (2) Pricing: The output curves of various types of power supplies aggregated in VPPs differ due to their varying characteristics. Based on the output prediction of the adjustable power supplies and the price prediction of the power market, VPP operators can implement differentiated dispatching for different generators through data modeling to improve the economic benefits for both aggregators and operators.

Commercially, VPPs will leverage economies of scale to realize the commercial model that distributed energy producers cannot achieve alone. In order to participate in the future energy market and generate profits, distributed energy producers should be capable of sensing market prices in real time. Furthermore, distributed renewable energy devices will need to respond to market changes and power grid fluctuations in real time. This will require ICT infrastructure such as interconnected networks and edge gateways or edge computing. Producers will incur the kinds of transaction costs that come with being part of the market, such as insurance and compliance costs. These additional costs represent a barrier to market entry for distributed energy

producers, but by aggregating the large number of distributed energy sources, VPPs reduce costs and generate profits through economies of scale. VPPs may operate in either grid-oriented or user-oriented models. In the grid-oriented business scenario, VPPs provide services for power grid operators, aggregating power from distributed heterogeneous resources. Typical services include providing frequency responses for power grids through aggregated power generation systems, energy storage equipment, and thermal storage facilities. In this case, VPPs treat aggregated distributed resources as a whole and are rewarded for enabling demand-side flexibility in the power grid. In the user-oriented business scenario, VPPs track energy market prices and provide users with paid services such as peak shaving, reducing their users' power bills. In short, VPPs provide flexibility to power grids by aggregating distributed energy assets, automatically and remotely scheduling and managing distributed energy resources, and tracking energy markets in real time. VPPs enable small producers of distributed energy to save costs through less power consumption and profit by delivering generated electricity to the grid. At the same time, they provide greater flexibility to innovative power systems based on renewables.

The VPP model requires the collaboration of different players who bring different skills to the table. These include software, renewable energy, fossil fuel, and electric power companies. In a VPP project in Huangpu District, Shanghai, adjustable loads in commercial buildings were aggregated. The project has developed 60,000 kW of adjustable resources, involving numerous pieces of controllable electrical equipment, such as chillers, air-cooled heat pumps, electric boilers, mechanical equipment, lighting facilities, and

chargers, in a total of 130 buildings. Each building provides an average response capacity of 100-500 kW, and 20% of the commercial buildings can provide automatic demand response in minutes. In addition, diversified response resources in two residential areas and three electric vehicle charging platforms have been connected to the VPP. Applications have also been developed and launched for VPP production, operations, and dispatching of the commercial buildings. The project functions as a routine resource dispatching center for responding to electricity demands in Shanghai. To date, it has participated in dispatching more than 2,000 times, shaving peak load up to 50,000 kW each time, which exceeds 10% of the regional capacity of peak flexible load control [39]. Although there are currently a number of technical and commercial problems standing in the way of the ultimate success of the VPP model, VPPs are expected to have a place in the power systems of the future.

Snapshot from the future: Energy cloud as the operating system for the energy Internet

Conventional energy networks are typically built with centralized architecture. The operator builds up equipment capacity, operates higher voltages, and expands the network to profit from economies of scale. Energy production, transmission, and consumption are separated, and there is no way to implement the end-to-end management and scheduling of electricity production, transmission, and consumption. Different energy networks, such as electricity, gas, heat, and cooling supplies, are separated from each other, which hinders overall energy efficiency. With distributed energy being deployed on a growing scale, energy consumers that also have production capacity will simultaneously act as producers and consumers, or prosumers, blurring the once clear boundary between energy production and consumption. Demand-side responsiveness is another area that is becoming increasingly important. The interconnection of multiple types of energy can improve comprehensive energy efficiency and contribute to the consumption of renewable energy. There is an urgent need for integrated platforms that address these issues, and energy cloud may be a solution.

Energy cloud is a new multidisciplinary concept which is very much still evolving and crystallizing. It can be understood as the operating system of the energy Internet, and is typically characterized by convergence, openness, and intelligence. Generation, grids, storage, and consumption of power all need to be converged in an end-to-end manner. Generators now include a large number of distributed new energy sources, such as solar energy, wind energy, and biomass, as well as fossil fuel sources like gas. The most important entity in the grids is the energy router that can direct energy flows free from constraints. Consumers include various industrial, commercial, and household facilities, such as HVAC systems and electric heating pumps. Meanwhile, storage entities include various fixed energy storage devices at the generation, grid, and consumption sides, as well as mobile energy storage devices like electric vehicles. The energy cloud will also break down the boundaries between electricity, gas, heat, and cooling. By connecting multiple systems of energy sources, such as heat, gas, and cooling, the energy cloud offers a comprehensive converged system that can optimize total energy use through the synergy of multiple types of energy.

Supported by the energy cloud, the energy Internet of the future will be an open system. Energy cloud users will include individuals (e.g. electric vehicle owners and residential power equipment), businesses (e.g. zero-carbon campuses and VPPs), and governments (e.g. zero-carbon cities). The number of users will far exceed that of traditional energy users. In addition, the energy cloud needs to interconnect with third-party systems, such as carbon trading systems. Therefore, the energy cloud should be an open ecosystem. It will offer developers a variety of open energy data and programing interfaces, enabling them to implement apps for different scenarios. In addition, an energy app store will be built on the cloud to distribute apps to users so that developers can commercially benefit from their work. Open decoupled programing will enable interconnection with third-party ecosystems, such as energy and carbon trading ecosystems, creating the potential for the emergence of new business models for the energy industry.

To enable convergence and openness, the energy cloud must be an intelligent platform. AI algorithms will make energy assets smarter. For example, AI technologies can be used to control the angle of solar panels to increase energy yield. But intelligence will also be built into the fabric of the energy cloud itself. The energy cloud will build data assets based on massive data of distributed energy sources and end-to-end information on energy generation, grids, load, storage, and consumption. It will have a user- and developer-oriented data platform based on data assets and big data modeling capabilities. Algorithms can be used to forecast distributed generation of energy and energy demand based on historical data, to dynamically respond to demand, and to analyze energy market prices in real time. With the support of efficient, intelligent technologies, such as AI and big data, the energy cloud aims to enable a frictionless flow of energy from producers to consumers on demand. Ultimately, this will create a green, low-carbon, safe, stable, and diverse energy system.

Green and digital technologies both drive economic transformation. Energy is the foundation of the digital world, and digital technologies will help create a smarter energy industry. Building an energy Internet operating system and promoting the modernization of the energy industry through digital technologies will help reduce emissions across the industry.

▶ Direction for exploration: Efficient power use for ICT: Saving energy and cutting more emissions

According to the EU's digital strategy, Shaping Europe's Digital Future, digital solutions can increase energy efficiency and cut the use of fossil fuels by tracking when and where electric power is needed most. However, the ICT industry itself also needs to undergo a green transition. It is estimated that the ICT industry accounts for 5–9% of the world's total electricity consumption and more than 2% of total emissions [40]. Data centers and telecom networks must improve their energy efficiency, reuse waste energy, and increase the share of renewables. The EU requires that all data centers be climate neutral, energy efficient, and sustainable by 2030, and that telecom operators adopt more transparent measures to track their environmental footprint.

Snapshot from the future: Low-carbon data centers and communication networks

According to IEA, with the rapid growth of global Internet users and traffic, the electricity consumption of data centers and transmission networks has increased significantly. In 2022, the global electricity demand from data centers was about 460 terawatt-hours (TWh), accounting for 2% of global electricity demand. By 2026, the electricity demand of data centers is expected to exceed 1,000 TWh, accounting for one third of the electricity demand increment [41]. Furthermore, data networks consumed about 250 TWh in 2019, accounting for about 1% of global electricity consumption, with mobile networks making up two-thirds of this figure [42]. Data center electricity consumption in China alone is expected to exceed 400 billion kWh in 2030, accounting for 3.7% of the country's total electricity consumption [43]. If data center power usage effectiveness (PUE) improves by just 0.1, the result will be 25 billion kWh of power saved and about 10 million fewer tons of carbon emissions. If all data centers use green power, carbon emissions will be reduced by 320 million tons each year [44]. Green power and PUE optimization are essential for low-carbon data centers.

Large ICT companies have been the biggest purchasers of green power, as they strive to reduce carbon emissions in data centers and telecom networks. Google, Facebook, Amazon, and Microsoft were the world's top four buyers of green power in 2019. According to BloombergNEF [45] , global green power buying reached a record high in 2023. In the first two months of 2024, Google and Microsoft announced the signing of 609 MW and 295 MW renewable power purchase agreements (PPAs), respectively. In May 2024, Microsoft and Brookfield Corporation announced the signing of the world's largest ever corporate clean energy procurement agreement and that they would invest more than US$10 billion into developing renewable power generation capabilities, looking to meet the growing requirements of AI and data centers. Google has announced plans for "24/7 zero-carbon" global operations by 2030. If successful, this means that the company will achieve zero carbon on an hourly statistics basis instead of a yearly basis [46]. Facebook plans to achieve net zero emissions across the supply chain by 2030 [47]. Microsoft declared that it will achieve negative carbon emissions by 2030 and offset all historical carbon emissions by 2050 [48].

According to Uptime, the average PUE in data centers around the world in 2022 was 1.55,

meaning that about 36% of the power drawn by data centers was used for cooling and other auxiliary functions [49]. As an increasing number of high-temperature-proof servers are put into use, cooling using natural air instead of traditional chillers and air conditioners will become possible. This will reduce the energy consumption of cooling systems, thereby decreasing PUE. There have been a number of practices in the industry. One type of cooling system for a data center is powered entirely by seawater. Another data center uses cold outdoor air to ensure its equipment stays at optimal temperatures. Meanwhile, a submarine data center keeps its PUE as low as 1.07 [50].

In addition to applying renewable energy and free cooling, AI is another effective way to make data centers more efficient and save energy. Sensors in data centers collect data such as temperature, power levels, pump speed, power consumption rate, and settings, all of which are analyzed by AI. Then, the data center operations and control thresholds are adjusted accordingly, reducing costs and increasing efficiency. AI is used in data center cooling to reduce the energy used for cooling by 40% [51]. China Unicom (Henan) employs Huawei's iCooling@ AI solution, which leverages big data and AI technologies to automatically optimize energy efficiency in data centers. This can improve data center PUE by about 8–15% [52]. According to Datacenter Dynamics, the Boden Type Data Center (BTDC), an experimental data center built in Sweden with funding from the EU's Horizon 2020 programme, has achieved a PUE level of 1.01 [53] by using AI algorithms to achieve synergy between the cooling system, computing loads, server fans, and temperatures, in addition to free cooling. As AI becomes increasingly widespread, diversified computing is on the rise and data center power density is growing. Large data centers require holistic systems. Balancing power supply, servers, and workloads based on AI algorithms may be the next technological step required to reduce data center PUE while continuing to support density increase.

In terms of communications networks, the ITU, GeSI, GSMA, and SBTi published a target-based roadmap, consistent with the Paris Agreement, to reduce the greenhouse gas emissions of the ICT industry by 45% before 2030 [54]. In addition to using green power like we do in data centers, we can use solar-grid hybrid solutions and a simplified network architecture to build greener communications networks with lower carbon emissions.

Communications and computing equipment share the same fundamental technologies. Today, Moore's Law is breaking down, and optoelectronic integration is the next step for the industry's structural reform towards higher energy efficiency. The use of optoelectronic hybrid technologies in networks, devices, and chips can constantly improve the energy efficiency of communications devices. The green communications networks of the future will be built to support more than 100 times today's capacity, but their total energy consumption will be no higher than that of today's networks. Conventional communications networks are defined by their specialist functions, which makes for fragmented operation and maintenance (O&M) and means they cannot keep pace with the latest network automation and intelligence. Networks need to be reconstructed to deliver essential services through a simplified architecture that consists of three layers: basic telecom network, cloud network, and algorithms. This simplified network architecture will greatly reduce the complexity of the algorithms in autonomous driving networks, lower the demand for computing, and cut O&M costs, contributing to greener, low-carbon networks.

Using green power, innovative architecture, AI algorithms, and other effective approaches, data centers and communications networks will be more energy-efficient, and will allow us to finally achieve net zero emissions.

Conclusion:

Technologies drive the energy revolution for a better, greener future

The world needs to halve its emissions by 2030. New technologies, such as digital and power electronics technologies, will accelerate the production of renewables such as wind and solar to increase the share of electricity in the energy mix. The technologies will also improve the quality of various energy equipment, ensure the safety and security of power systems and other energy systems, and help various industries reduce carbon emissions.

Huawei predicts that by 2030, renewables will account for 65% of all electricity generation globally, about 6,000 GW of PV plants will be in place, the LCOE of PV power will decrease to US$0.01 per kWh of electricity, and renewables will power 80% of all digital infrastructure.

In 2030, technologies will make green energy smarter, enable a wide variety of industries to further reduce emissions, and promote an energy revolution. Technologies will support green, low-carbon transformation and sustainable development for the global economy.

global industry vision

Huawei predicts that by 2030,

Renewables will

account for **65%**

of all electricity generation globally.

References

1 WMO, "State of the Global Climate 2023," https://wmo.int/zh-hans/news/media-centre/2023nianqihoubianhuazhibiaodachuangjilushuipingwmo

2 WMO, "The Global Climate in 2015-2019," https://library.wmo.int/doc_num.php?explnum_id=9936

3 UNEP, "Emissions Gap Report 2020," https://www.unep.org/zh-hans/emissions-gap-report-2020

4 UN, https://unfccc.int/zh/cop28/5-key-takeaways#end-of-fossil-fuels

5 IRENA, "World Energy Transformation Outlook 2023"

6 IRENA, "World Energy Transformation Outlook 2023"

7 IRENA, "Renewable_Energy_Statistics_2024"

8 IRENA, "World Energy Transformation Outlook 2023"

9 http://www.gov.cn/gongbao/content/2020/content_5570055.htm

10 IRENA, "Remap 2030,Renewable Energy Prospects:Germany"

11 IRENA, "Renewable_Energy_Statistics_2024"

12 China's National Development and Reform Commission, "14th Five-Year Plan: Modern Energy System Planning"

13 https://www.afrik21.africa/en/morocco-xlinks-to-bring-10-5-gw-of-solar-and-wind-power-to-the-uk/

14 Xinhua News Agency, http://www.news.cn/politics/2023-06/25/c_1129715025.htm

15 BloombergNEF, New Energy Outlook Series

16 GWEC, GLOBAL OFFSHORE WIND REPORT 2024

17 Zheng Chongwei et al. " Progress on Global Offshore Wind PowerResearch," http://www.haiyangkaifayuguanli.com/ch/reader/create_pdf.aspx?file_no=20140607&year_id=2014&quarte r_id=6&falg=1

18 Zhao Zhenzhou, Wang Tongguang, and Zheng Yuan, Principles of Wind Turbines

19 IEA, "Offshore Wind Outlook 2019"

20 IRENA, "Future of wind 2019"

11 IEA, "Offshore Wind Outlook 2019"

22 GWEC, GLOBAL OFFSHORE WIND REPORT 2024

23 IEA, "Offshore Wind Outlook 2019"

24 IRENA, "Renewable_Energy_Statistics_2024"

25 S. Zahra Golroodbari Wilfried van Sark, "Simulation of performance differences between offshore and land-basedphotovoltaic systems,"
https://onlinelibrary.wiley.com/doi/full/10.1002/pip.3276

26 Chen Lin, "Singapore unveils one of the world's biggest floating solar panel farms," https://www.reuters.com/business/energy/singapore-unveils-one-worlds-biggest-floating-solar-panelfarms-2021-07-14/

27 Harry Morgan, "Floating Solar set for 60-Fold boom by 2030,"
https://rethinkresearch.biz/articles/floating-solar-set-for-60-fold-boom-by-2030/

28 WORLD BANK GROUP, "Where Sun Meets Water Floating Solar Market Report,"
https://openknowledge.worldbank.org/handle/10986/31880

29 IRENA, "World Energy Transformation Outlook 2023"

30 IEA, Energy Statistics Data Browser, CO2 emissions by sector, World, 1990-2021

31 http://www.sutpc.com/news/jishufenxiang/814.html

32 "Top 10 Trends of Charging Network,"
https://digitalpower.huawei.com/attachments/index/81b0f79b9641433a9bcedf16c6f75001.pdf

33 CNESA, "Energy Storage Industry Research White Paper 2024"

34 China National Energy Administration, "Blue Book on the Development of New Power Systems"

35 RMI and China Hydrogen Alliance Research Institute, China 2030 "Renewable Hydrogen 100" Development Roadmap, https://rmi.org.cn/wp-content/uploads/2022/07/Chinas-Green-Hydrogen-New-Era-2030-Chinas-Renewable-Hydrogen-100GW-Roadmap.pdf

36 U.S. National Clean Hydrogen Strategy and Roadmap, https://www.hydrogen.energy.gov/library/roadmaps-vision/clean-hydrogen-strategy-roadmap

37 EU Hydrogen strategy, https://energy.ec.europa.eu/topics/energy-systems-integration/hydrogen_en

38 IEA, "Global Hydrogen Review 2023," https://www.iea.org/reports/global-hydrogen-review-2023

39 ZHAO Jianli, XIANG Jiani, TANG Zhuofan, et al., "Practice exploration and prospect analysis of virtual power plant in shanghai [J]," Electric Power, 2023

10 European Commission, "Shapping Europe's Digital Future," https://www.consilium.europa.eu/en/press/press-releases/2020/06/09/shaping-europe-sdigital-future-council-adopts-conclusions/

41 IEA, "Electricity 2024"

42 IEA, "Data Centres and Data Transmission Network"

43 China Communications Standards Association, http://www.ccsa.org.cn/detail/4319?title=“碳达峰”“碳中和”与数据中心的关系

44 Jiang Junmu, "Reducing PUE with Innovative Solutions: Huawei Helps Win the Battle of Data Center Green Development," https://m.c114.com.cn/w3542-1166194.html

45 Bloomberg NEF, "Corporate Clean Energy Buying Grew 18% in 2020, Despite Mountain of Adversity"

46 Google, https://sustainability.google/commitments/#

47 Facebook, https://sustainability.fb.com/

48 Microsoft, https://blogs.microsoft.com/blog/2020/01/16/microsoft-will-be-carbon-negative-by-2030/

49 HARRY MENEAR, "The Uptime Institute releases annual data centre findings," https://datacentremagazine.com/data-centres/uptime-institute-releases-annual-data-centrefinding

50 Peter Judge, "Project Natick: Microsoft's underwater voyage of discovery," https://www.datacenterdynamics.com/en/analysis/project-natick-microsofts-underwatervoyage-discovery/

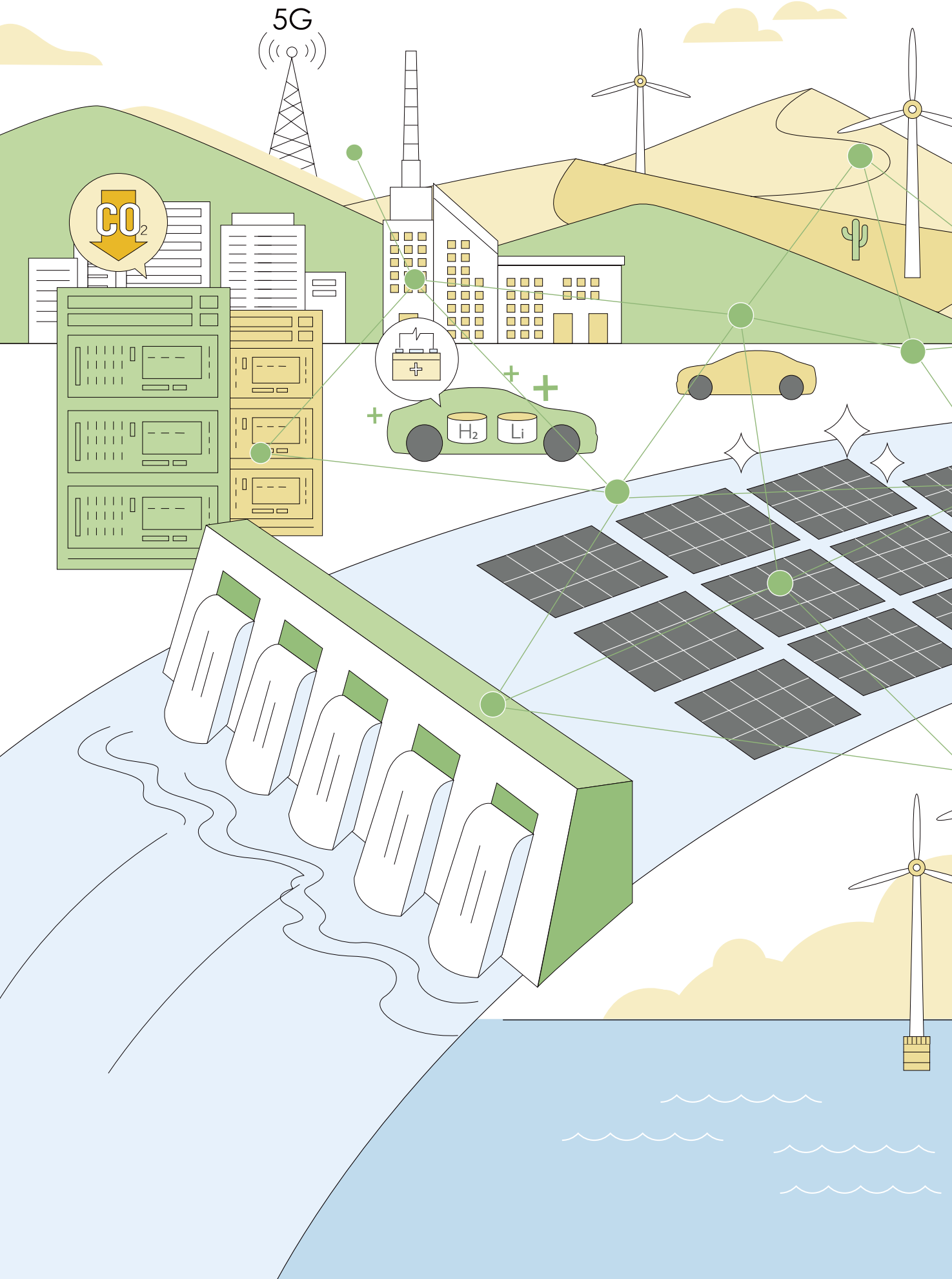
51 Joe Devanesan, "Has Google cracked the data center cooling problem with AI?" https://techwireasia.com/2020/05/has-google-cracked-the-data-centre-cooling-problem-with-AI?"

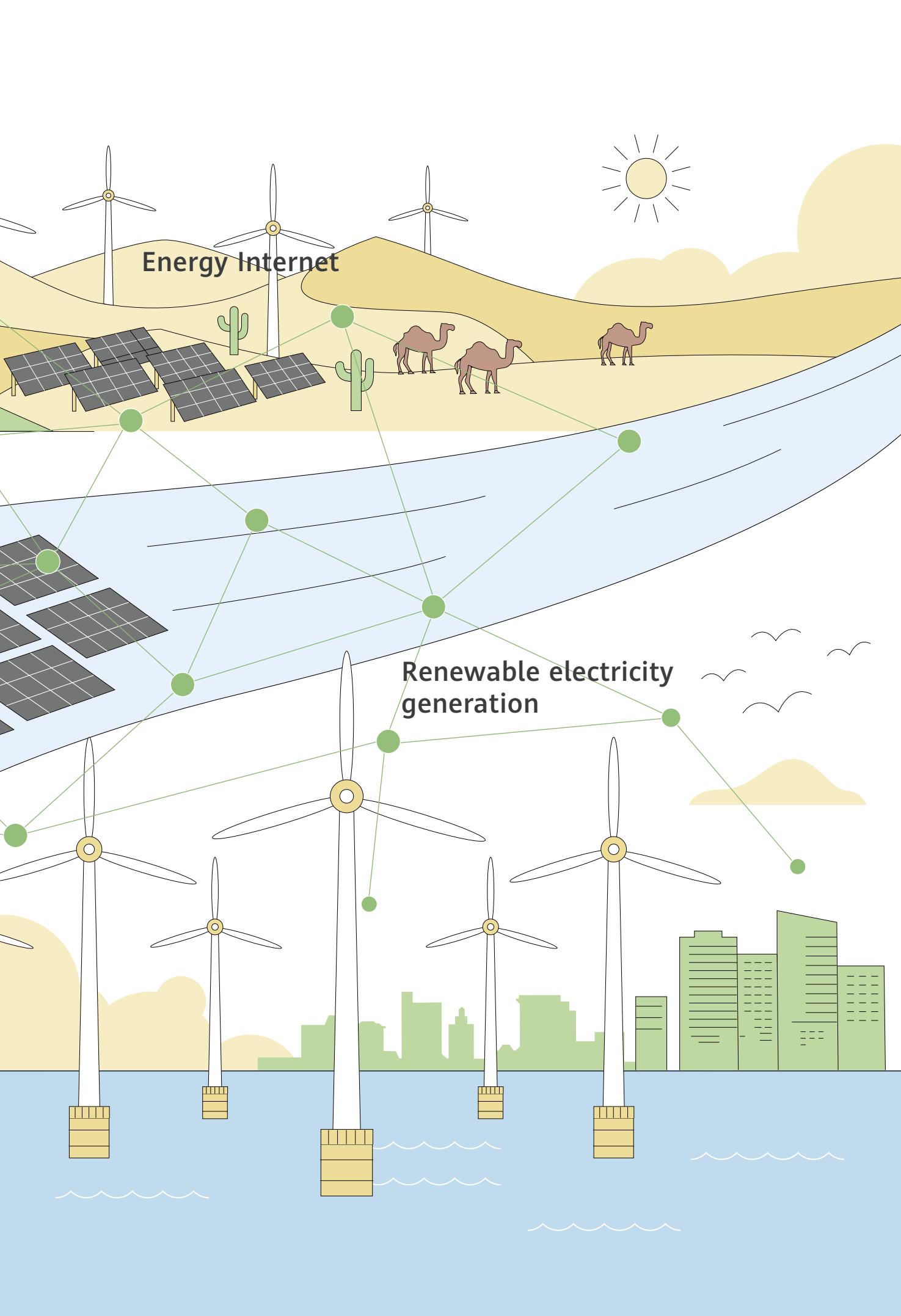
52 HUAWEI, https://www.huawei.com/en/technology-insights/cases/henan-unicom-icooling

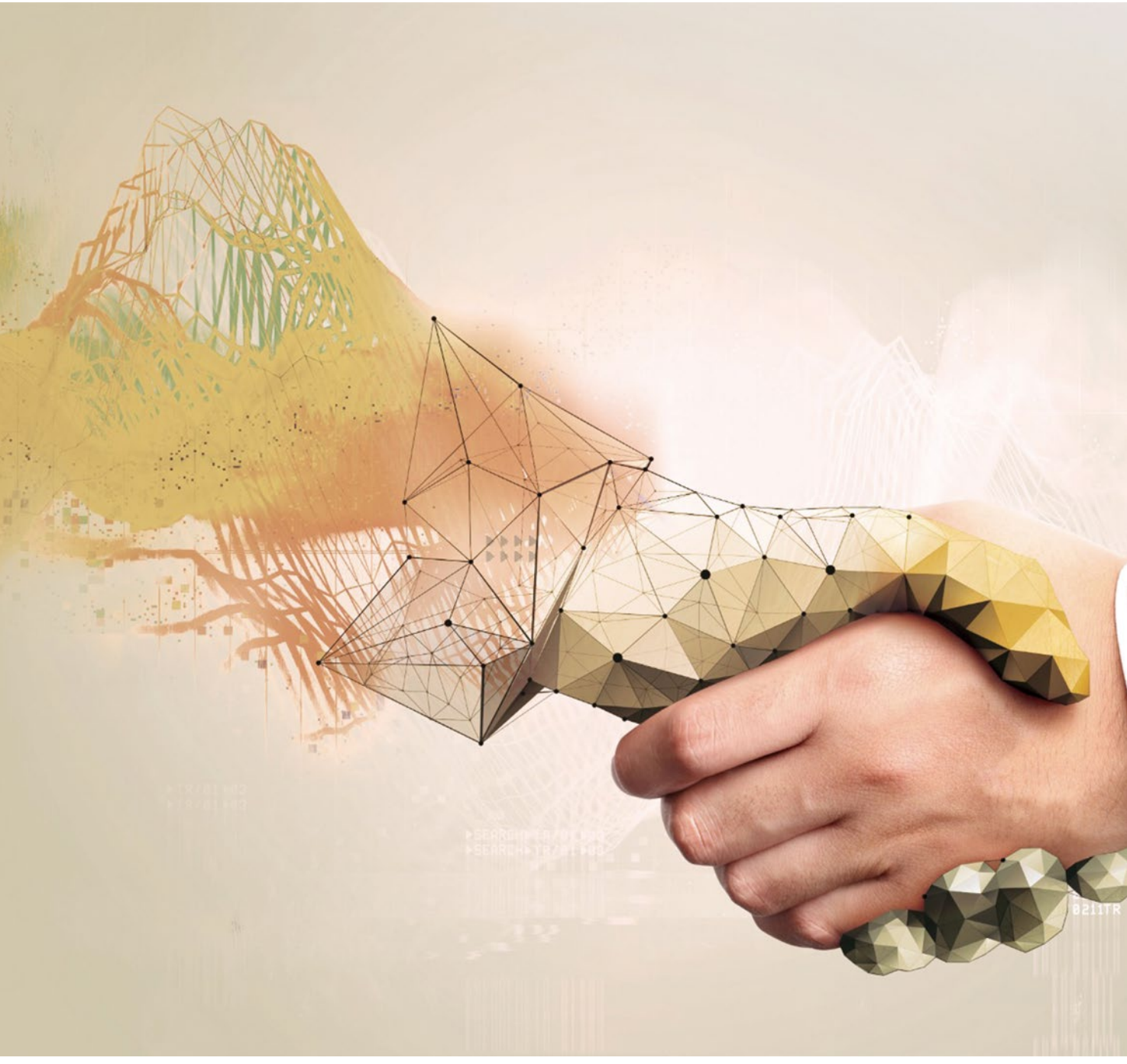
54 HARRY MENEAR, "Hive partners with world's most efficient data centre, BTDC" https://datacentremagazine.com/it/hive-partners-worlds-most-efficient-data-centre-btdc

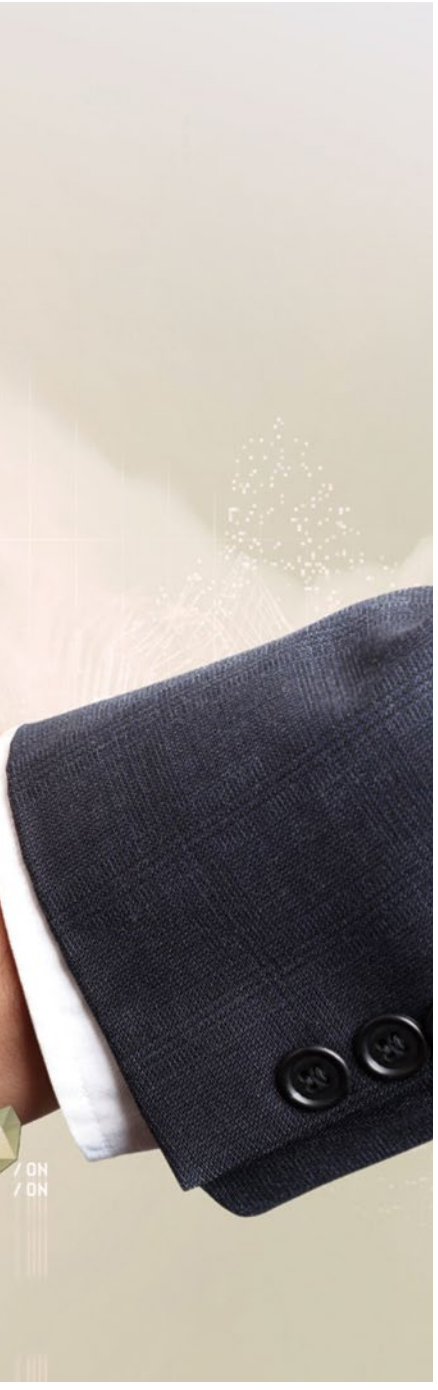
54 ITU, "ICT industry to reduce greenhouse gas emissions by 45 per cent by 2030"

Efficient power use for ICT




Energy Internet
Renewable electricity generation





Digital Trust

Technologies and Rules Creating a Trusted Digital Future

Collaboration drives progress, and the foundation of collaboration is trust. All interactions in the business world, including cross-enterprise cooperation, enterprise operations and management, and supply chain operations, are ultimately based on trust. Digital technology is reshaping these interactions. The exponential growth of AI and digital twins, as well as the emergence of digital humans and metaverse, is increasing interactions between humans, between humans and machines, and between machines themselves. Digital trust is therefore becoming a prerequisite for the connectivity of everything, which makes it a paramount strategic objective for organizations. As digital transformation continues to gather momentum, more and more products will go digital. Privacy breaches and information corruption erode trust, which will then jeopardize business operations, business value (such as brand and market value), reputation, and credibility.

AI is helping more and more people efficiently accomplish tasks, and seeing rapid uptake in sectors like healthcare and autonomous driving. As we get closer to artificial general intelligence (AGI), the role of AI will only grow. The rapid iteration of generative AI and AI-generated content (AIGC) has created a vast range of new models and applications which are generating even more vast amounts of text, audio, and video data. It is also increasingly difficult to verify the authenticity of such content. A responsible AI system needs to not only guarantee system and model security, but also ensure the data and content it produces is transparent and traceable. Such a system must provide mechanisms to verify the authenticity and origin of AIGC. Only by building an accountable and evolvable digital system can we foster enduring digital trust.

Digital trust is a complex field that covers a range of areas, including privacy, security, identity, transparency, traceability, data ownership, data integrity, governance, and compliance. [1] The realization of digital trust must address all of these different dimensions and use a variety of tools, such as digital identities, digital watermarking, privacy-enhancing computation (PEC), and AI. Thankfully, the new technologies and new rules constantly emerging are helping create this trusted digital future.

Direction for exploration 1: ICT enabling digital trust

For both organizations and individuals, data assets are being used to unlock unprecedented efficiency and convenience. However, this increases the risks associated with information theft. Data security and integrity relies on a variety of technologies, including access controls, risk prediction, digital ledgers, encryption, and digital authentication.

Research into ICT technologies for digital trust will help define data ownership paradigms which will make data sharing and data transactions more traceable and verifiable. This will help organizations and individuals fully realize the value of their data, effectively manage digital assets, and better protect core data.

Snapshot from the future: Intelligent agents with digital identities and trusted identifiers

The explosive development of large language models (LLMs) has increased AI penetration in all sectors. Consequentially, various AI-based intelligent agents are emerging in both the digital and physical worlds, increasing productivity, facilitating creativity, and turbocharging economic development. Digital identities build trust between people and intelligent agents, allowing them to engage and collaborate in new ways.

Digital identifiers are essential for interactions between intelligent agents and between people and intelligent agents. In the future, intelligent agents will be popular in all sectors and may even exceed the number of human beings, without being subject to the limitations of geographical location. A decentralized digital identity system that delivers a wide range of authentication methods will be the way forward.

The identifier of an intelligent agent works in the same way as a person's name or ID card. People can only be assured that they are interacting with the right intelligent agents if they have access to accurate identifiers. Blockchain or other distributed storage mechanisms are ideal here, as they can facilitate public query and registration of identifiers.

Snapshot from the future: AI security to ensure trustworthiness in the digital life

Advancements in machine learning and deep learning have catalyzed the development and refinement of foundation models, such as ChatGPT and Sora. The growing volume and quality of AIGC is also blurring the boundaries between AIGC and real-world content. AI will only acquire even more human-like cognitive and learning capabilities as we approach AGI.

This rapid AI development has been a double-edged sword. On the positive side, AI is facilitating better lives and work for many. On the other hand, AI makes deepfakes much easier to generate and much more difficult to identify, with more and more people mistaking fake audio and video as real. For instance, the Wall Street Journal has reported on at least one instance of criminals using AI to mimic the voice of a company's CEO over the phone. They used this fake recording to order an executive of the company's UK-based subsidiary to transfer US$243,000 to a Hungarian supplier. The money was eventually moved from the supplier's Hungarian bank account to one in Mexico. All losses were eventually borne by insurance companies. [2]

It is difficult for people to identify deepfake audio and video, even with the help of traditional technological protection tools. Guaranteeing AIGC provenance and verification is an effective way to address this problem. Such technologies could add metadata, such as identity and digital watermarks, to AIGC, to assure recipients that content comes from a trustworthy source.

Some governments have already put forth standards, laws, and executive orders on AIGC labeling and provenance. In October 2023, Joe Biden issued the US's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. In Article 4.5 "Reducing the Risks Posed by Synthetic Content", the order calls on stakeholders to, "foster capabilities for identifying and labeling synthetic

content produced by AI systems." China initiated a mandatory standards project named Labeling Method for Content Generated by Artificial Intelligence In June 2024. And the EU's Artificial Intelligence Act, which officially took effect on August 1, 2024, also highlights that certain AI systems that generate data must provide the ability to identify the data.

The industry has also already established some technical standards for AI identification and tracking. For instance, in June 2022, ISO/IEC set up a Joint Photographic Experts Group for media trustworthiness (JPEG Trust). JPEG Trust then released the draft ISO/IEC 21617 standard for AIGC provenance and authenticity verification. The draft standard was passed by national representatives in July 2024 and will be released as an official international standard shortly. In the future, the trust model and core ideas of JPEG Trust will serve as a reference for international AIGC governance to hopefully guide the formulation and application of standards and the establishment of a mechanism for worldwide interconnection. The end goal of such efforts is to eventually form a comprehensive content tracing and authenticity verification system and redefine trust for digital content.

AI systems are playing an increasingly important role in various sectors, exponentially magnifying the impact of attacks on AI systems. In the future, AI systems will increasingly require the ability to support provenance and authenticity verification. Concurrently, stringent technical requirements must be put in place to ensure the security of AI systems and models, as well as the trustworthiness of AI input and output data, thereby mitigating the risk of AI system misuse and safeguarding the value generated by AI.

Snapshot from the future: Digital watermarking to support information provenance

In the digital world, data will be the most important asset, meaning it will face heightened risk of theft and misappropriation. The unprecedented capabilities provided by AIGC and similar technologies, like natural language processing and image generation, are revolutionizing the industry. The theft or abuse of data assets or generated data can have far-reaching consequences for both individuals and enterprises. Consequently, a robust mechanism for tracing and safeguarding data assets and generated data must be established.

Digital watermarking–based provenance technology is an effective method for addressing this problem. By processing the dataset with digital watermarking, it ensures traceability and ownership, and prevents unauthorized use. Embedding watermarks during model training also allows owners to identify and prove ownership in the event of theft or abuse.

AI has also revolutionized traditional digital watermarking, enabling the development of excellent semantic watermarks that deliver a level of security similar to cryptography. Digital watermarking–based provenance will play an important role in building trust in the future. It can even be applied in real-time audio and video communications to locate the source of information disclosers. In the future, it is probable that only image and audio datasets that have undergone digital watermarking processing will be eligible for publication. Creatives can also adopt digital watermarking to protect their data from unauthorized collection and use.

Snapshot from the future: PEC technologies that improve computing security

In the era of big data, data will be the new oil. But unlike oil, data will never run out. The value of data will be realized again and again in different scenarios and regions by all kinds of enterprises and organizations. However, data sharing presents new challenges to security and privacy. Data mining and analytics driven by machine learning is becoming increasingly prevalent. Many sectors such as finance, healthcare, and retail in particular need to guarantee data privacy as they seek to mine data, obtain its value, and share it for collaboration. As data analytics and data warehouse environments become increasingly complex, traditional data desensitization technologies will no longer be sufficient. Therefore, PEC technologies are being explored as an alternative.

PEC technologies are data security technologies used to protect and enhance privacy and security during the collection, storage, search, and analysis of private information. PEC supports efficient, high-quality services by protecting personal data from abuse, while allowing effective use of the data. This allows us to realize the data's full business, scientific, and social value. PEC technologies are being explored in the following areas:

Differential privacy: Random noise is injected into databases to be mixed with personal data, while statistical estimation can still be performed using the data. This method guarantees personal privacy even when data is shared, because the original data has been scrambled.

Homomorphic encryption: This technique allows users to perform computations on encrypted data without decrypting it. When data is homomorphically encrypted, the computations on the data are also executed in an encrypted form. When the output is then decrypted, it is identical to the answers that would have been obtained if the computations had been performed on the unencrypted data.

Federated learning: This method allows data to stay in a company's local servers for machine learning. Separate learning models are built there after encrypted samples have been aligned, and a virtual joint model is then developed based on these models. The performance of this joint model is almost identical to a model trained on data directly gathered in the conventional way.

In addition to the above-mentioned technologies, PEC technologies include a trusted execution environment (TEE), zero-knowledge proofs, k-anonymity, and l-diversity. In the future, PEC will be supported by more algorithms and widely used in more applications, helping us to find the right balance between privacy and data value.

Snapshot from the future: Quantum security that ensures trustworthiness in the digital life

Quantum computing is maturing rapidly and is expected to be capable of breaking traditional security algorithms by 2030. Therefore, it is imperative we transition to post-quantum cryptography (PQC) and quantum key distribution (QKD) which are resistant to quantum computing attacks.

The first batch of standardized quantum-resistant algorithms selected through the NIST PQC Standardization competition has already been officially released. The first batch of ISO/IEC PQC standards has also already been released. In light of the attack on Supersingular Elliptic Curve Isogeny (SIKE) – a NIST candidate algorithm – a preliminary consensus has been reached on the evolution towards diverse PQC algorithms. For example, a current draft ISO/IEC standard in this field includes an algorithm submitted by the US NIST and the German FrodoKEM algorithm. Meanwhile, multiple countries and industry organizations have released strategies and planning guidelines to guide the industry through PQC migration.

QKD is a secure key distribution method implemented by using quantum mechanics and transmission measurement of quantum superposition states. QKD has provable security and is recognized in both China and many European countries as an effective method to ensure quantum security in security-critical scenarios.

Direction for exploration 2: Rules that redefine digital trust

The rapid growth of digital technologies has been accompanied by numerous new security challenges. Emerging technologies like cloud, IoT, and AI are advancing at an unprecedented pace as more sectors go digital, however this transformation is also creating new cyber security risks. Ensuring the security of digital environments is vital to ensuring the impact ICT has on the economy remains positive. As a global technology provider, Huawei is acutely aware of how important cyber security is for ensuring trust in the digital world we all share.

In the international community, cyber security is increasingly intertwined with political concerns and trade barriers. However, political concerns have done nothing to help address core cyber security issues. Cyber security has been used as an excuse for new trade barriers, but the true nature of this security conundrum remains obscure.

To resolve trust issues in the future digital world, more governments and organizations will need to establish clear fair and just rules. Establishing unified neutral technical standards, implementing security assurance measures, and creating a digital trust system will be essential for future progress.

Snapshot from the future: Unified rules that enhance data protection and mitigate data monopoly

Digital trust involves many organizations and stakeholders. Notably, in the realm of personal information protection, the EU has enacted the General Data Protection Regulation (GDPR), while other countries and regions have subsequently adopted similar laws, such as China's Personal Information Protection Law. GDPR is the world's most stringent privacy and security law for personal data. It came into force on May 25, 2018 and has since played an active role in protecting personal data and enforcing compliance on data giants that misuse data. As of July 31, 2024, a total of 1701 GDPR fines totaling EUR4 billion had been imposed, with the single largest fine amounting to EUR1.2 billion.

A global trend in governments standing up against corporate data monopolies has also unfolded. In 2019, the US launched an anti-monopoly investigation into giant companies suspected of monopolizing the market, suppressing competition, and infringing user privacy. On May 27, 2020, the Japanese Senate officially adopted the Act on Improving Transparency and Fairness of Specified Digital Platforms (TFDPA), which seeks to regulate specific digital platforms and increase their disclosure obligations. On January 19, 2021, Germany's Tenth Act Amending the Act Against Restraints of Competition came into effect, expanding the application scope of their existing competition rules to prevent and curb enterprise abuse of dominant market positions. In 2021, the State Council of China issued the Anti-monopoly Guidelines of the Anti-monopoly Commission of the State Council in the Field of Platform Economy.

In the future, global data protection and anti-monopoly legislation will continue to improve and be implemented to prevent large platforms from illegally obtaining, abusing, and trading personal privacy data. This will bolster digital security, ensure fair competition, and foster a robust digital credit ecosystem.

Snapshot from the future: Impartial standards that drive healthy ICT industry development

Governments and industry organizations must establish unified cyber security standards that are technology-neutral and equally applicable to all enterprises and all ICT products. The telecom industry has a long history of developing shared standards to promote the continuity, reliability, and interoperability of telecom networks. Nicholas Negroponte, the founder of the MIT Media Lab, once even wrote in a Fast Company article, "Telecommunications policy should be based on objective standards, not geopolitical issues."

Huawei has been an active contributor to the standards organizations involved in the development of ICT security standards. In the connectivity field, we have contributed more than 300 proposals to 3GPP and GSMA, maintaining our longstanding position as a leader in the industry. In the computing field, we have contributed security proposals, including an AI computing platform security framework, general confidential computing framework, and remote attestation procedures (RATS) architecture to ISO/IEC, ETSI, and the National Information Security Standardization Technical Committee of China (TC260), and worked with industry partners to drive the development and application of computing security technologies.

Clear, fair, and unified cyber security technologies and verification standards will ensure ICT products can be independently and comprehensively verified and evaluated. This will enable organizations to select products that meet their specific security requirements, based on the verification and evaluation results, and foster healthy development within the ICT industry.

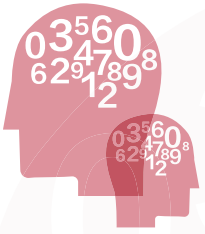
Snapshot from the future: A cyber security and privacy assurance system that boosts digital trust

Building and implementing an end-to-end global cyber security and privacy protection assurance system is one of Huawei's key strategies. In compliance with applicable laws and regulations in countries where we operate and international standards, we are continuing to invest in an effective, sustainable, and reliable cyber security and privacy protection assurance system that addresses the requirements of both regulators and customers, as well as industry best practices. Additionally, we actively work with governments, customers, and industry partners to address cyber security and privacy challenges.

Huawei's own cyber security values include integrity, trustworthiness, accountability, capability, openness, and transparency. We do not and will never implant backdoors into our equipment or allow others to do so, and will never illegitimately collect intelligence for any individual or organization, including government organizations, agents, and entities. We are fully aware of the importance of privacy protection and we are committed to protecting personal data of our consumers, customers, suppliers, partners, employees, and other relevant entities, in compliance with all privacy and personal data protection laws and regulations everywhere we operate.

Huawei has published a Statement on Establishing a Global Cyber Security Assurance System and Huawei General Privacy Protection Policy to explicitly state our position, general principles, and requirements concerning cyber security and privacy protection.

- Our business departments are required to identify cyber security and privacy risks based on business scenarios and high-risk groups, develop management requirements and incorporate these requirements into related business processes, and IT systems and tools.
- We have established an end-to-end cyber security and privacy protection verification system, regularly carry out measurement, inspections, and internal audits, set up organizations independent of business systems to verify Huawei products and services, and make constant management improvements to address identified issues. We also work with third parties on tests, certifications, and external audits to continuously improve Huawei's cyber security and privacy protection management.
- We organize regular cyber security and privacy protection awareness training, education, and examination for all staff, and provide targeted training for managers and high-risk groups. We have also established an accountability mechanism to hold violators accountable in accordance with the Accountability Rating Criteria for Cyber Security and Privacy Protection Violations.



Conclusion:

Building an intelligent world with digital trust together

Looking forward towards 2030, it is clear to us that the world will be using a number of technologies such as digital identities, AI, and quantum-safe algorithms to better protect user privacy and data assets; to fight against digital fakes and fake news more effectively; and to reduce the risk of fraud and data theft. Digital watermarking, PEC, and other technologies will enable data sharing with secure encryption, maintaining the value generated by data flows while protecting privacy. Huawei predicts that by 2030, PEC technologies will be used in more than 50% of computing scenarios, and 100% of ICT systems will have a quantum-safe capacity or the capacity to migrate to quantum-safe solutions.

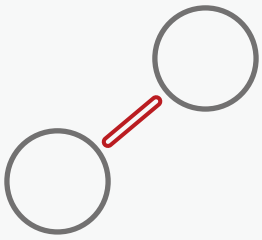
At the same time, digital security laws and regulations like GDPR and anti-monopoly action in the data domain will be seen in more countries. This will help build trust between individuals and organizations, and help organizations maintain legal compliance in the area of digital trust.

Robust digital trust ecosystems need the cooperation of multiple parties. Enterprises will need to maintain their own legal compliance and manage the compliance of their partners. They will also need to work with regulators to combat information security violations and data monopolies, as well as to protect user data security. This means enterprises will need to set up their own cyber security and privacy assurance systems, and proactively support public education and training to help build digital literacy and awareness of data and privacy issues within their communities. Working together, we can build an intelligent world with digital trust.

global industry vision

Huawei predicts that by 2030,

Privacy-enhanced computing technologies will be used in more than **50%** of computing scenarios.

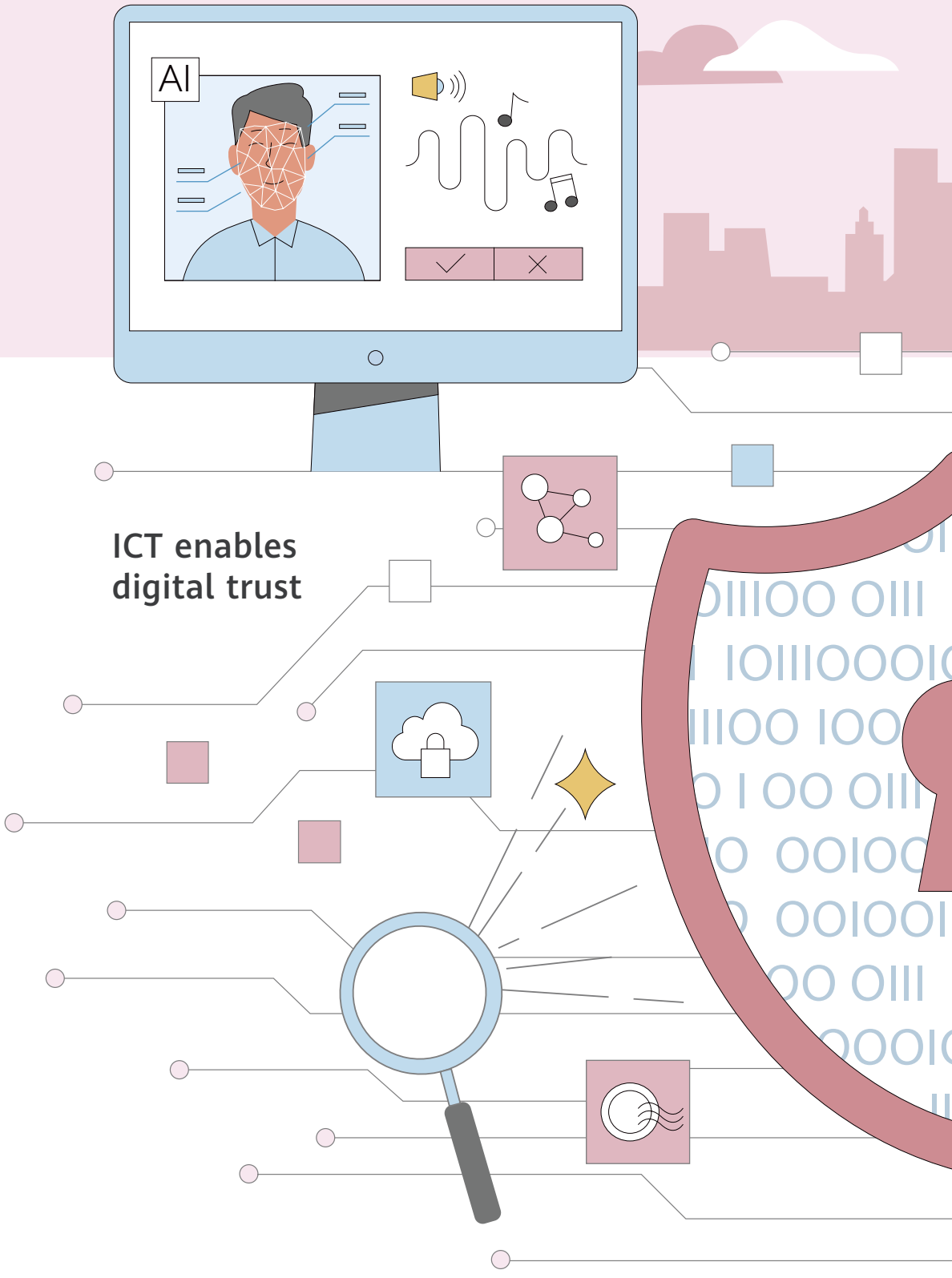


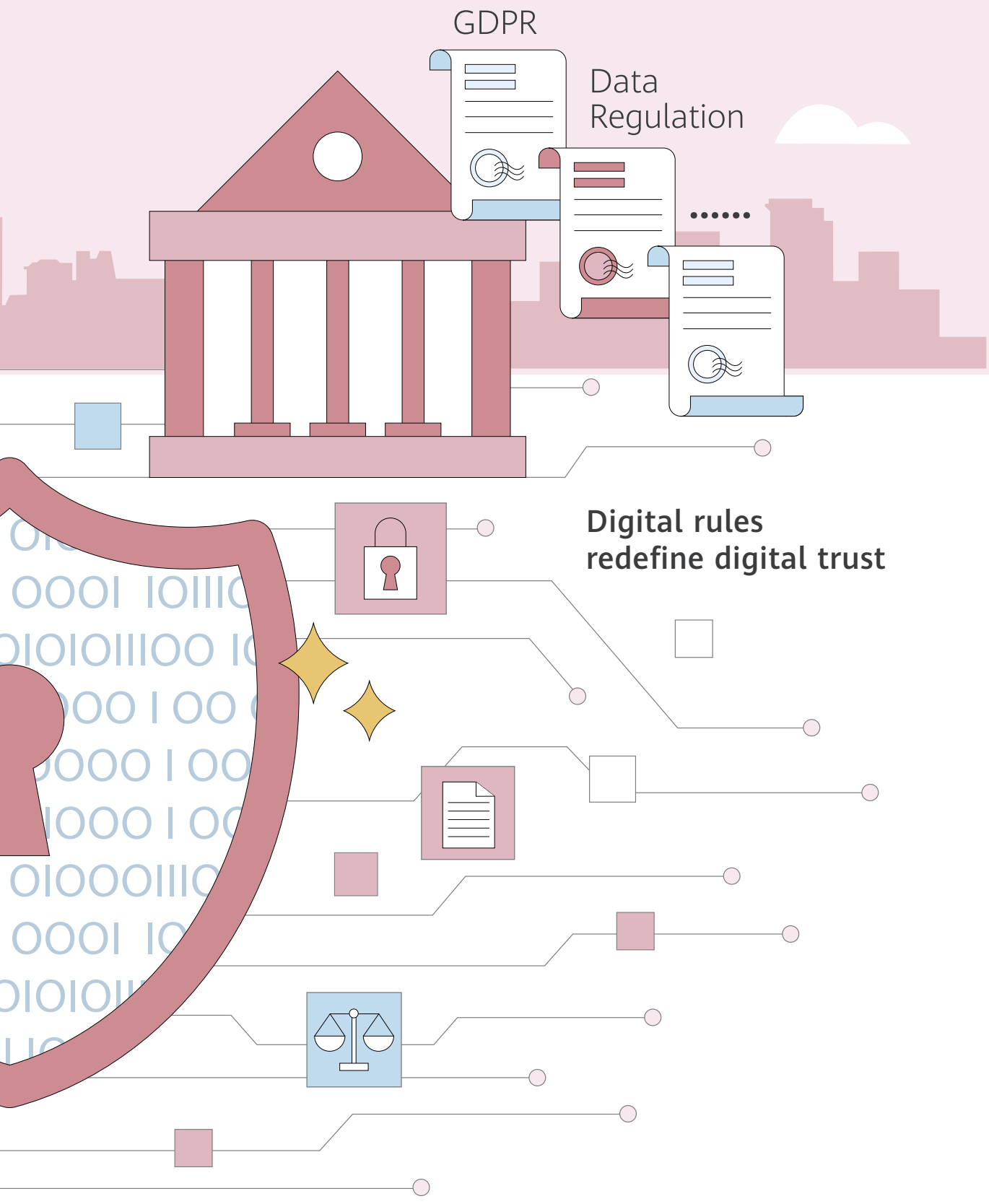
100% of ICT systems will have a quantum-safe capacity or the capacity to migrate to quantum-safe solutions

References

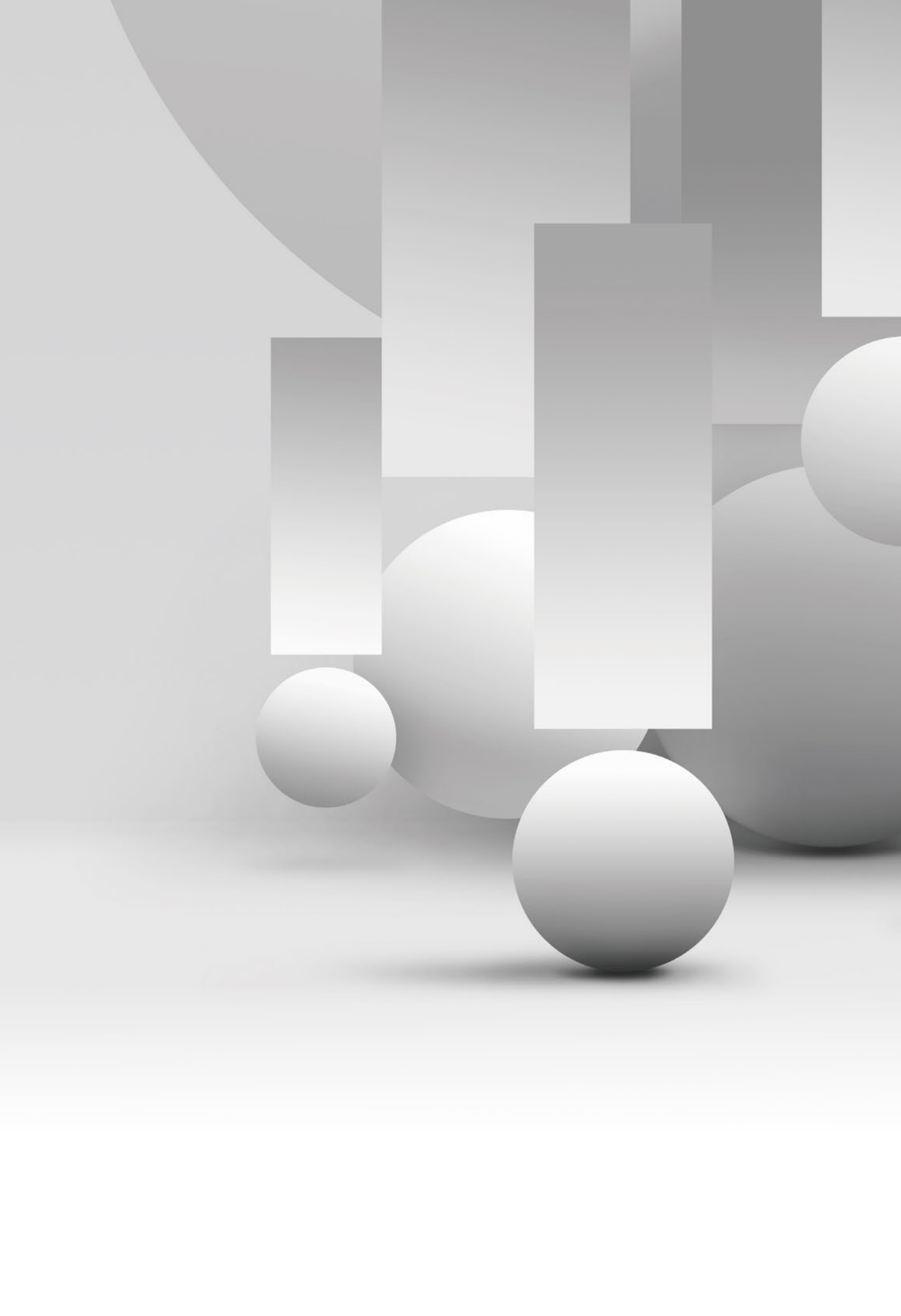
1 Omar Abbosh and Kelly Bissell, "Securing the Digital Economy: Reinventing the Internet for Trust," Accenture, https://www.accenture.com/content/dam/accenture/final/a-com-migration/pdf/pdf-94/accenture-securing-the-digital-economy-reinventing-the-internet-for-trust.pdf.

2 Catherine Stupp, "Fraudsters Used AI to Mimic CEO's in Unusual Cybercrime Case," Wall Street Journal, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.


AI
ICT enables
digital trust



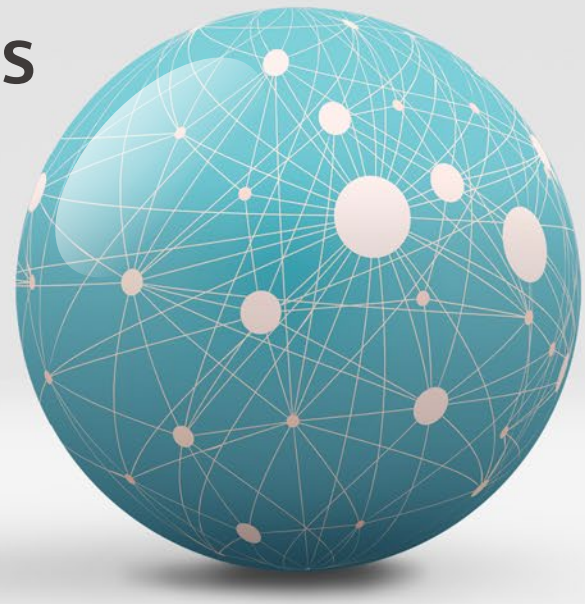
Digital rules redefine digital trust



HUAWEI

Version 2024

Communications Network 2030



Building a Fully Connected, Intelligent World

01

Industry Trends

Going intelligent has become the general direction that the world is heading in over the coming decade. China, the EU, and the US have all published their new visions for this area.

In its Outline of the 14th Five-Year Plan (2021–2025) for National Economic and Social Development and the Long-Range Objectives Through the Year 2035, China prioritizes industry intelligence as an important area of development, and sets clear development goals for industries including manufacturing, energy, agriculture, healthcare, and education, as well as for government management.

In its 2030 Digital Compass plan, the EU articulates the following targets: By 2030, 75% of European enterprises will have taken up cloud computing, big data, and Artificial Intelligence (AI) services, and more than 90% of European small- and medium-sized enterprises (SMEs) will reach at least a basic level of "digital intensity". To achieve these targets, the EU announced an increase in investment into energy and digital infrastructure.

In its Vision 2030 report, the US National Science Board (NSB) recommends increasing investment in AI while continuing investment in corresponding digital infrastructures such as data, software, computing, and networks over the next decade. These recommendations aim to help maintain the US's competitiveness in the intelligent era.

The intelligent development of industries requires enterprises to upgrade their networks. In its Industrial Internet Innovation and Development Action Plan (2021–2023), China's Ministry of Industry and Information Technology (MIIT) puts forward the following measures: (1) Accelerate the network-based development of industrial equipment, drive the upgrade of enterprise Intranet, and promote the integration of information technology (IT) networks and operational technology (OT) networks to build

industrial Internet campus networks. (2) Explore the deployment of new technologies such as cloud-network synergy, deterministic networking, and Segment Routing over IPv6 (SRv6). In its Digitizing European Industry platform plan, the EU considers nanophotonics, AI, 5G, and Internet of Things (IoT) to be key enablers of future industrial networks, and plans to increase investment in these technologies in order to stay ahead in the future.

In recent years, generative AI (GenAI) has garnered much attention around the world, and today is regarded as one of the key enablers of industry intelligence. Serving as "pipes" that connect the massive computing infrastructure on which GenAI relies, networks play a pivotal role in the efficient utilization of computing power. To drive the high-quality development of computing infrastructure, the MIIT and other government departments propose the following targets in Action Plan for High-Quality Development of Computing Infrastructure: (1) Improve the coverage of optical transport networks (OTNs) in key application places. (2) Increase the adoption of innovative technologies such as SRv6. (3) Implement low-latency connections between data centers at hub nodes.

As industries increasingly adopt intelligent technologies, leading telecom carriers around the world are taking action and beginning to explore how they can fully unleash the potential of connectivity in this process. For example:

- China Mobile has unveiled a "5G + AICDE" development strategy, where AICDE stands for AI, IoT, cloud computing, big data, and edge computing.

- China Telecom has set out the goal of building an integrated cloud-network architecture by 2030.

- China Unicom published its CUBE-Net 3.0 strategy, which articulates a new development direction that combines connectivity, computing, and intelligence.

- In its outlook for 2030, Deutsche Telekom aims to become the leading digital enabler in the business to business (B2B) market, providing comprehensive network, IoT, cloud, and digital services.

A survey conducted by GSM Association (GSMA) shows that carriers worldwide can fully unleash the potential of their connectivity portfolios by focusing on B2B, cloud, and IoT services that target the industry, finance, health, energy, and agriculture sectors. Carrier networks are also undergoing intelligent transformation, with all leading carriers proposing their "AI+" strategies. For example, China Mobile comprehensively promotes the "AI+" action and aims to reach L4 autonomous networks (ANs) by 2025.

By 2030, many amazing things that we can only dream of today will be a reality. For example, highly sensitive biosensors and intelligent hardware connected through broadband networks will enable us to monitor the indicators of our physical health in real time. And we will be able to analyze massive amounts of historical health data securely stored on terminals and clouds through AI. This will allow us to proactively manage our own health and reduce our dependence on doctors, thereby improving our health and quality of life.

New technologies, such as home broadband that supports speeds of over 10 Gbit/s and holographic communications, will enable more intuitive human-machine interactions. An air-ground cubic network will connect all means of transportation, facilitating easy, smart, and low-carbon travel. Sensing technology, 10-gigabit wired and wireless broadband, inclusive AI, and applications that target numerous industries will be available everywhere, allowing us to build urban digital infrastructure that improves the quality of city life.

With Harmonized Communication and Sensing (HCS), automation, and intelligence technologies, we will be able to efficiently protect our environment. New types of labor, such as collaborative robots, automated mobile robots (AMRs), and digital labor, can be adopted in tandem with the industrial Internet to increase accuracy and decrease costs throughout the whole process from demand to production and delivery, while also improving the resilience of the manufacturing industry.

Energy IoT can be integrated into smart grids to form a green energy Internet and fully digitalize all activities, including generation, grid, load, and storage. Zero-carbon data centers and zero-carbon communications sites may soon become a reality. We can also guarantee digital security and trustworthiness by combining blockchain, digital watermarking, AI-driven anti-counterfeiting, privacy-enhancing computing, and endogenous network security.

In 2030, communications networks will evolve from connecting billions of people to connecting hundreds of billions of things, and face many challenges along the way.

First, the scale of communications networks will continue expanding. This means network management will become even more complex. Over the next decade, how can we innovate in software technology to enable self-configuring, self-healing, and self-optimizing networks and prevent operation & maintenance (O&M) costs from rising in step with the continuous expansion of network scale? This poses a daunting challenge.

Second, IoT scenarios such as unattended operations in industrial and agricultural settings, end-to-end (E2E) self-driving vehicles, low-altitude manned flight, and low-altitude drone take-outs and freight movement will require carriers to further improve the coverage, quality assurance, security, and trustworthiness of their networks. Over the next decade, how can we innovate in

protocols and algorithms to enable networks to carry multiple types of services while meeting the requirements for high quality and flexibility? This will be a very challenging task.

Third, although Moore's law has held true for decades, the semiconductor industry is now struggling to maintain that pace of improvement, and new technologies like quantum computing are not yet mature. Meanwhile, demand for computing power, storage capacity, and network energy efficiency continues to grow, and these factors are increasingly becoming bottlenecks. Over the next decade, how can we innovate in fundamental technologies to build a green, low-carbon network and increase network capacity by dozens of times without increasing energy consumption? This is another extremely challenging task that lies ahead of us.

Communications networks are one of the major forces driving the world forward. The development of communications networks kicked off during the first Industrial Revolution and, unlike traditional industries, it still shows no signs of slowing down after nearly two centuries. In fact, the pace of development of communications technologies has been particularly rapid in recent decades. Both the evolution from 2G to 5G and the shift from the asymmetric digital subscriber lines (ADSL) to gigabit optical home broadband took just 30 years. Over the next decade, we will witness the emergence of new use cases and scenarios for communications technologies and fully embrace an intelligent world.

Communications networks are one of the major forces driving the world forward. The development of communications networks kicked off during the first Industrial Revolution and, unlike traditional industries, it still shows no signs of slowing down after nearly two centuries. In fact, the pace of development of communications technologies has been particularly rapid in recent decades. Both the evolution from 2G to 5G and the shift from the asymmetric digital subscriber lines (ADSL) to gigabit optical home broadband took just 30 years. Over the next decade, we will witness the emergence of new use cases and scenarios for communications technologies and fully embrace an intelligent world.

02

Future Network Scenarios

Communications networks have come a long way since Samuel Morse invented the electric telegraph in 1837. They have evolved from connecting individuals and homes to connecting organizations and from wired to wireless. In today's environment of diverse and rapidly changing services, it takes continuous innovation for communications networks to keep up with the needs of customers. To meet the rich and diverse business needs that will arise in the intelligent world of the next 10 years, communications networks will need to go beyond connecting individuals. They will also need to connect multiple perception, display, computing resources, and AI agents related to each individual. In the near future, networks will have to connect home users as well as home appliances, vehicles, and content resources, while organizations will expect networks to do more than just create connections between employees — they must also connect an organization's machines, edge computing nodes, and cloud resources.

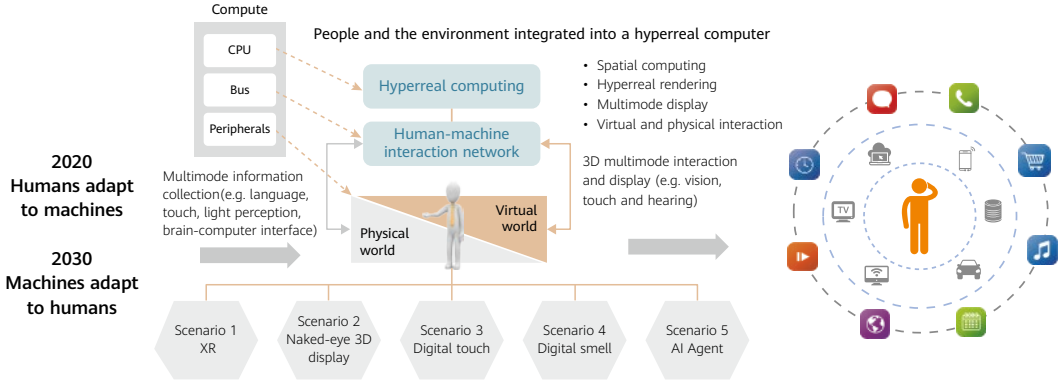
The scope of network connections is expanding, business needs are changing, and the industry has reached a consensus that, over the next 10 years, networks will evolve from 5G to 5G-A/6G, from 5th Generation Fixed Network (F5G) to F5G-A/F6G, and from Net5G to Net5.5G/Net6G, and Autonomous Networks (AN) will evolve from L2 to L4+. In addition, new use cases will continue to emerge.

2.1 Next-Generation Human-Machine Interaction Network: A Human-centric Hyperreal Experience

In a world of cold machines, it is up to human beings to adapt to the machines. With the wide use of the automobile, we learned to work with pedals and a gearstick. In the PC era, we learned to use the mouse and keyboard. In the smartphone era, we learned to use touchscreens.

However, with sufficiently advanced levels of intelligence, it is possible to turn this paradigm on its head and have machines adapt to the needs of their human users. Intelligent machines (e.g., smart screens, smart home appliances, intelligent vehicles, and smart exoskeletons) will be able to understand natural language, gestures, and eye movement, and even read human brain waves, enabling more intuitive integration between the virtual and physical worlds and bringing a hyperreal sensory experience to human-machine interaction. (Figure 1 Hyperreal human-machine interaction experience)

Figure 1 Hyperreal human-machine interaction experience



Over the course of the coming decade, communications networks must evolve to support brand-new human-machine interaction experiences such as XR, naked-eye 3D display, digital touch, digital smell, and AI agent.

2.1.1 XR: An Intuitive Interaction Experience Through a Perfect Synthesis of the Virtual and Physical Worlds

Virtual Reality (VR) is about rendering packaged digital visual and audio content. Augmented Reality (AR) refers to the overlaying of information or artificially generated content onto the existing environment. Mixed Reality (MR) is an advanced form of AR that integrates virtual elements into physical scenarios. eXtended Reality (XR), which covers VR, AR, and MR, is a catchall term that refers to all real and virtual combined environments and human-machine interactions generated by computer technology and wearables. Characterized by three-dimensional environments, intuitive interactions, spatial computing, and other features that set it apart from existing Internet devices, XR is considered the next major platform for personal interactions.

In 2020, due to the impact of social distancing caused by COVID-19, demand for VR games, virtual meetings, and AR-assisted temperature taking increased exponentially. The number of active VR users on the US video game digital distribution service Steam doubled. Some manufacturers have unveiled new AR headsets that are more portable. With the wide adoption of 5G, wireless fidelity 6 (Wi-Fi 6), and fiber broadband, all of which can deliver gigabit speeds, XR services are set to boom over the next decade. Huawei predicts that the number of VR/AR users is expected to reach 1 billion by 2030.

In its Virtual Reality/Augmented Reality White Paper, the China Academy of Information and Communications Technology (CAICT) divides the technical architecture of XR into five parts: near-eye display, perception and interaction, network transmission, rendering processing, and content creation. The white paper also predicts the development stages of XR. The CAICT's conclusions have, to some extent, been endorsed by the ICT industry. (Table 1 Network requirements of XR services)

Table 1 Network requirements of XR services

Technical System	Technical Index	Partial Immersion 2021	Deep Immersion 2022–2025	Full Immersion (XR) 2026–2030
Near-eye display	Monocular resolution	2K	4K	8K
	Field of view (FOV)	120°	140°	200°
	Pixel per degree (PPD)	20	30	60
	Varifocal display	No	Yes	Yes
Content creation	360° panoramic resolution: Weak interaction	8K	12K	24K
	Gaming: Strong interaction	4K	8K	16K
Network transmission (Average value)	Weak interaction (Mbit/s)	90	290	1,090
	Round-trip latency: Weak interaction	20	20	20
	Round-trip latency: Strong interaction	5	5	5
	Transmission medium	Wired/Wireless	Wireless	
Rendering processing	Rendering computing	4K/90 FPS	8K/120 FPS	16K/240 FPS
		/	Fixation point rendering	
Perception and interaction	Eye interaction	/	Eye tracking	
	Voice interaction	Immersive sound	Personalized immersive sound	
	Tactile interaction	Tactile feedback		Refined tactile feedback
	Mobile interaction	Virtual mobility (Movement redirection)		High-performance virtual mobility

Currently, XR is entering the deep immersive experience phase. At the beginning of 2024, Apple Vision Pro was officially launched. Its glasses' viewfinders have 23 megapixels between them, providing greater than 4K resolution in each eye and achieving excellent definition, color accuracy, and visual experiences. We predict that XR will reach the stage of full immersion by 2030, by which time it will be supported by 8K monocular resolution, 200° FOV, and a gigabit-level bitrate.

The improvement of the XR display poses higher requirements on content. If content rendering is implemented on the cloud, the device-cloud motion-to-photon (MTP) interaction requires a network transmission round-trip time (RTT) of 20 ms for carrying XR services. For weak-interaction streaming services with few motions, the requirement for lower than 20 ms RTT latency can be met. For strong-interaction gaming services with frequent motions, the RTT latency must be controlled within 5 ms.

Therefore, to support the development of XR services over the next 10 years, networks must have bandwidth of higher than 1 Gbit/s and latency of lower than either 5 ms or 20 ms, depending on the scenario.

2.1.2 Naked-Eye 3D Display: A Brand-new Visual Experience Through Lifelike Image Reproduction

The implementation of naked-eye 3D display involves three major phases: the digitalization of 3D objects, network transmission, and optical or computational reconstruction and display.

There are two types of naked-eye 3D display technology: light field display (through lenslets) and the use of spatial light modulators (SLMs).

Light field display leverages the binocular parallax to create 3D visual effects. It uses parallax barriers, lenticular lenses, and directional backlight, all of which impose fairly inflexible requirements in terms of viewing angles. Their adoption would require real-time capturing of user location and dynamic adjustment.

An alternative approach would be to use SLMs. An interferometric method is used to store all amplitude and phase information of light waves scattered on the surface of a 3D object in a recording medium. When the hologram is irradiated with the same visible light, the original object light wave can be reproduced thanks to diffraction, providing users with a lifelike visual experience. (Table 2 Network requirements of naked-eye 3D display)

Table 2 Network requirements of naked-eye 3D display

Technical System	Technical Index	Lenslet (2021-2025)	SLM (2025-2030)
Maturity prediction		Large-scale deployment and high maturity	Sporadic application
Display	Size	70-inch screen	10-inch to 70-inch screens
	Resolution	16K	16K
Network transmission	Bandwidth	Around 1 Gbit/s	10 Gbit/s - 1 Tbit/s (4K,60 frames, and 10 Gbit/s are required forobjects with a size of 10 x10 cm.)
	Round-trip network latency	Weak interaction: 20 ms Strong interaction: 5 ms	Weak interaction: 5 ms Strong interaction: 1 ms
	Transmission medium	Wired/Wireless	
Interaction design	Voice interaction	Location tracking and spatial sound	
	Gesture interaction	Gesture recognition	
	Mobile interaction	Location tracking and spatial computing	
Availability		Audio: 99.9% Video: 99.999%	

References: IEEE 1981.1 Tactile Internet and Digital Holography and 3D Display

In recent years, naked-eye 3D display featuring light field display has developed rapidly, in step with the development of user location awareness and computing technologies. Some manufacturers have commercialized their innovative products. We predict that a large number of use cases will emerge in the entertainment and commercial sectors by 2025. This type of 3D display requires higher than 1 Gbit/s bandwidth and real-time interaction. In strong interaction scenarios, the network latency must be less than 5 ms, and commercial applications will require network availability of 99.999% (this means annual downtime must be less than 5 minutes and 15 seconds).

Over the past several years, breakthroughs have also been made in holographic technology, which is based on optical reconstruction. Product prototypes have been developed with a thickness of 10 cm and a projection size of around 100 cm^2. We predict that these small-scale holographic products will become commercially available at exhibitions, for teaching purposes, and as personal portable devices over the next 10 years. They will require bandwidth of around 10 Gbit/s, latency of no more than 5 ms or as low as 1 ms, and network availability of more than 99.999%, the same as that required in commercial settings. True-to-life holographic products will require higher bandwidth (over 1 Tbit/s), but we do not expect them to be ready for large-scale commercial deployment by 2030.

Therefore, the naked-eye 3D display products coming to market over the next decade will need to be supported by networks capable of delivering 1–10 Gbit/s bandwidth per user, latency of 1–5 ms, and 99.999% availability.

2.1.3 Digital Touch: Tactile Internet Made Possible Through Multi-dimensional Sensory Interaction

In IEEE's tactile Internet architecture, digital tactile technology is divided into three layers: user layer, network layer, and avatar layer. The user layer enters information such as location, speed, force, and impedance. After being digitalized over the network, the information is converted into instruction data and provided to the avatar layer. The avatar layer then collects tactile, auditory, and proprioception data and provides the data to the user layer through the Internet to inform users' real-time decision making.

Digital tactile technology has two interaction modes. The first is machine control. Use cases include remote driving and remote control. The second is hyperfine interaction, and use cases include electronic skin and remote surgery. (Table 3 Network requirements of digital touch)

Table 3 Network requirements of digital touch

Interaction Mode	Direction of Traffic	Traffic Type	Reliability	Latency (ms)	Bandwidth
Machine control	User-Avatar	Touch	99.999%	1-10	2Mbps
	Avatar-User	Video	99.999%	10-20	1-100Mbps
		Audio	99.9%	10-20	512Kbps
		Tactile feedback	99.999%	1-10	20Mbps (100 DOFs)
Hyerfine interaction	Avatar-User	Tactile feedback	99.999%	1-10	1~10Gbps (Electronic skin)

Active coqnitive capability: The network layer also needs to support services such as dynamic performancemonitoring,task awareness, and 3D mapping.

Reference: IEEE 1981.1 Tactile Internet

Machine control has numerous use cases in industrial settings, and has high requirements for network availability (above 99.999%). Some industries even require availability to reach 99.99999%. The required bandwidth is generally less than 100 Mbit/s, and the maximum permissible latency varies from 1 to 10 ms, depending on the specific circumstances.

Electronic skin powered by flexible electronics in hyperfine interactions has the most development potential. Electronic skin integrates a large number of high-precision sensors such as pressure and temperature sensors. According to a study by the University of Surrey in the UK, each square inch of electronic skin will require bandwidth of 20 to 50 Mbit/s, meaning that an average hand would require bandwidth of 1 Gbit/s. The wearers of electronic skin won't all be humans; intelligent machines present another class of potential users. The user layer may perform analysis, computing, and decision making based on the massive amounts of data collected by the electronic skin on the avatar layer to control the avatar layer. The user layer can also be directly connected to humans through brain-computer interfaces or myoelectric neural interfaces to deliver an immersive remote interaction experience. We predict that network bandwidth of 1 to 10 Gbit/s will be required in hyperfine interaction scenarios.

Therefore, to support digital touch, networks will need to deliver 1–10 Gbit/s bandwidth per user, availability greater than 99.999%, and latency below 10 ms, or as low as 1 ms in certain use cases.

2.1.4 Digital Smell: Internet That Enables Us to Smell Through Deep Sensory Interaction

Among our five senses, two of them – touch and taste – require direct contact, while three – sight, hearing, and smell – do not. Of the latter three, smell involves the deepest interaction.

Digital smell includes three technical phases: odor perception, network transmission, and smell reproduction.

There have been some use cases for odor perception, such as using composite materials to form a barcode, which can generate chemical reactions according to the odor and create color changes. The relationship between the barcode and odor can then be identified through Deep Convolutional Neural Network (DCNN) algorithms. Use cases can be found in specific scenarios like detection of dangerous goods and detection of food freshness.
There are already some commercial odor reproduction products available in the industry, such as smelling generators for VR games, which use five odor cartridges and selectively release odors from the cartridges. They emit scents such as the ocean, gunpowder, wood, and soil, deepening the immersion of the gaming experience. However, some research reports suggest that the future of smell in VR won't rely on these odor cartridges, but will instead work through brain-computer interfaces to enable people to sense odors more directly and accurately.

The combination of odor perception (using electronic noses) and odor reproduction can help create an Internet that enables us to not only hear and see, but also smell. It is not yet clear what kind of network bandwidth and latency this function will require, but the computing requirements are already relatively well understood.

In a nutshell, the next-generation human-machine interaction network will support brand-new experiences including XR, naked-eye 3D display, digital touch, and digital smell. Making these technologies work will require networks capable of delivering bandwidth of 10 Gbit/s and 99.999% availability, with latency as low as 1 ms for some use cases.

2.1.5 AI Agent: Independent Personal Assistant for Near-Human Interaction

Advancements in AI are driving AI application development toward agents. Once given a task, an AI agent will break it down into sub-tasks and create a prompt for each sub-task based on external feedback and autonomous thinking to complete the sub-tasks, and ultimately, fulfil the task it was assigned and the user intent.

The introduction of AI agents will directly result in four changes in the physical world:

- Change in objects: AI agents constitute a new type of connected object on our networks — independent silicon-based entities. The range of interactions occurring in the physical world is broadening from human-to-human alone to include interactions such as human-to-digital human, human-to-robot, human-to-household robot, and robot-to-robot interactions.
- Change in experience: Conventional network design prioritizes coverage and capacity for downlink services. However, AI agents are more sensitive to network latency and uplink speed, and in the future, network design will have to take this into account.
- Change in content: Interaction modalities are expanding from 2D audio and video to more sophisticated modalities such as environment information and 3D calling. For example, AI can generate a virtual 3D calling setup in real time in which two people in different locations meet each other as if in person, in an immersive shared environment in which participants even experience the same temperature.

Change in scope: Today's network services are predominantly human-centered, with service capabilities determined based on the scope of human activities. Future network service design will

factor in the activities and activity scope of human-like AI agents to provide 24/7 services covering every conceivable domain.

During interactions between AI agents and people, user experience will hinge on RTT. RTT is the sum of the time the AI spends processing and network transmission latency. Huawei estimates that a typical AI agent will require that RTT be no more than 400 ms to deliver a human-level face-to-face communication experience. GPT-4o, despite being an LLM and not an agent, is still an instructive example. Launched by OpenAI in 2024, GPT-4o requires that RTT be kept below 700 ms in order to deliver near-human interactions.

Huawei predicts that by 2030, human-to-AI agent interactions will entail transmission of three images together with voice streams per second on average. This requires a guaranteed network speed of at least 10 Mbit/s to 20 Mbit/s for excellent experience and at least 32 Mbit/s to 64 Mbit/s for superior experience. (Table 4 Network latency and bandwidth requirements of AI agents)

Table 4 Network latency and bandwidth requirements of AI agents

	Image Size	Network Latency	Guaranteed Bandwidth
Excellent Experience	200 KB (small)	200 ms	10 Mbit/s
	400 KB (large)		20 Mbit/s
Superior Experience	200 KB (small)	70 ms	32 Mbit/s
	400 KB (large)		64 Mbit/s

In the future, the average person may own several AI agents, just like most of us own several computers today. Huawei predicts that globally by 2030, there will be 6 billion active wireless AI agent users, including those using digital twins in the virtual world and embodied AI in the physical world, such as industrial robots, service robots, companion robots, autonomous drones, and autonomous vehicles. These AI agents will run as independent entities and become independent participants in society.

2.2 Networks That Deliver a Consistent Experience for Homes, Offices, and Vehicles: The Third Space with the Same Broadband Experience

With the large-scale commercial use of Huawei's Advanced Driving System (ADS) and Tesla's Full Self-Driving (FSD) system and the widespread use of Baidu's Apollo Go robotaxi in Wuhan, it is foreseeable that end-to-end autonomous driving will become a new norm by 2030. Vehicles will automatically pick up passengers from parking lots, drive along the road, and park at the destination, and the brain, eyes, hands, and feet of drivers will be freed. When we envision the future of self-driving cars, the most appealing feature for many is that we will be able to enjoy the immersive entertainment, social, and work experience we get at home while on the go. Multi-screen collaboration has been used both at home and in cars, and 3D display and holograms will be used in the future. 8K and 16K smart screens will be gradually adopted at home and MR will be widely used in cars.

With 5G-A, F5G-A, and Net5.5G, mobile and fixed broadband basically enters the ultra-gigabit era at the same time, making it possible to deliver the same level of experience to users regardless of whether they are at home, in the office, or on the go. In the future, self-driving cars will become the "third space" beyond homes and offices, and users will enjoy the same broadband service experience in all three scenarios. (Table 5 Network requirements for delivering a consistent experience at home, in the office, and on the go)

Table 5 Network requirements for delivering a consistent experience at home, in the office, and on the go

Scenario Type	Commercial Deployment	Home			Vehicle
		Service	Peak Bandwidth	Round-Trip Latency	Service
Cinema	Within 10 years	16K video (180-inch screen)	1.6 Gbit/s	50 ms	1.6 Gbit/s, 20 ms (16K XR)
Gaming	Within 10 years	360° 24K 3D VR/AR	4.4 Gbit/s	5 ms	4.4 Gbit/s, 5 ms (24K XR)
Holographic teaching	Within 10 years	10-inch hologram	12.6 Gbit/s	20 ms	12.6 Gbit/s, 20 ms
Holographic meeting	Within 10 to 20 years	True-to-life hologram (70-inch)	1.9 Tbit/s	1–5 ms	12.6 Gbit/s, 1–5 ms (Miniature hologram, 10-inch)
Autonomous driving	Within 10 years	Home robots	10 cm positioning	99.999% availability	5–20 cm positioning Availability: 99.999% to 99.9999%
Cloud PC	Within 2 to 3 years	Ultra-fast GPU cloud PC (shallow compression encoding, 4K 60 FPS)	≥ 500 Mbit/s	≤ 15 ms	—
Storage	Within 1 to 2 years	Ultra-fast converged storage (localization-like experience)	≥ 5 Gbit/s	≤ 5 ms (edge cloud deployment)	—
Home security	Within 2 to 3 years	3D optical sensing (1024*768 30 FPS)	About 1 Gbit/s	≤ 20 ms	—

Over the next decade, common home and office services will include smart screens, multi-screen collaboration, 3D, holographic teaching, and XR. With the continuous development of embodied intelligence and humanoid robot technologies, robots will be smarter and more human-like, as well as perform more physical tasks. As the "eyes" of robots, visual sensing is an important part of multimodal interaction and environment sensing. To implement full environment sensing, machine vision requires high spatial resolution, a high frame rate, and a wide light sensing range, which sharply increase the amount of image information. Additionally, to meet the real-time requirements of human-like interaction, network bandwidth needs to be 100 times higher, and the network needs to meet the ultra-low latency requirement. Considering that the penetration rate of true-to-life holographic conferencing will be low in 2030, the mainstream broadband requirements of home and office services will still be 1–10 Gbit/s bandwidth and lower than 5 ms latency. In the future, home and office networks will not only provide seamless broadband coverage, but also support brand-new scenarios such as working from home, premise security, and robotics. Based on HCS capabilities, home networks will be able to sense user locations, indoor space, and environment security, and create a more user-friendly living and work environment for people. By 2030, the average monthly home network traffic will reach 1.3TB.

Services like 3D, holographic teaching, and XR will also be available in our self-driving cars. Over the next decade, their key requirements for network bandwidth will be 1 to 10 Gbit/s, and latency requirements will be less than 5 ms. As autonomous driving will require vehicle-road collaboration, it will require network availability greater than 99.999% and positioning precision of 10 cm. Moreover, with the continuous improvement of scenarios such as automatic parking and passenger pick-up, clear requirements are imposed on the network coverage and rate at vehicles' start and stop locations such as parking lots.

In addition to immersive entertainment, future homes will also have a wide range of services, including cloud PC, home security, cloud storage, and NAS. Cloud PC is an important cloud service under the cloud-network synergy trend. It uses cloud rendering technology to transfer computing and rendering from terminals to the cloud. In this way, users can use lightweight terminals to enjoy computer services. In addition to existing smart cameras for home security, other sensing technologies are upgrading and converging. New security solutions, such as 3D optical sensing for healthcare, are gradually emerging. Home storage is developing towards high speed, convergence, and application integration. Ultra-fast cloud storage provides localized experiences of basic services such as data storage and backup, and supports ultra-fast speed for operations such as online file editing and video-on-demand playback. In addition, it can integrate various online applications such as document collaboration and smart album.

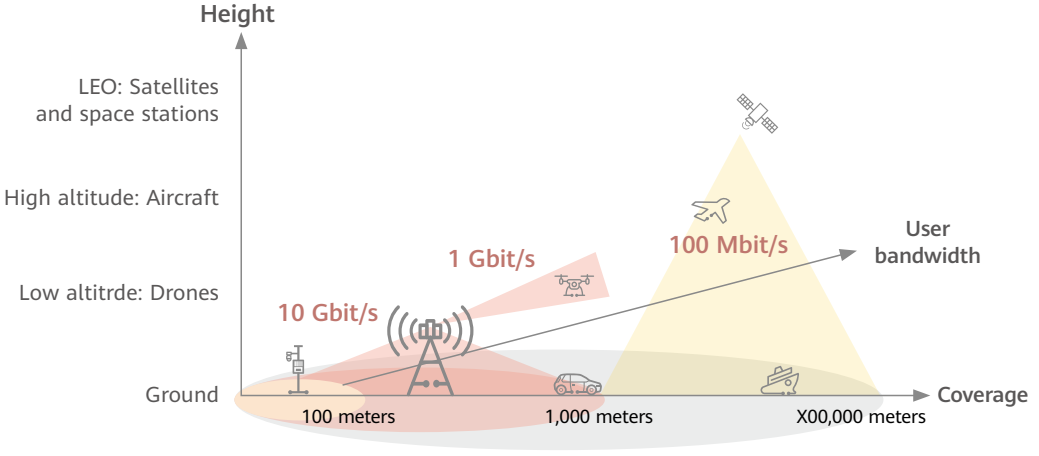
Huawei predicts that by 2030, the penetration rate of personal cloud disks in home cloud storage will reach 35%, that of home cloud computer services will reach 17%, and that of home healthcare using privacy-protection 3D radar optical sensing will reach 8% globally. Moreover, the penetration rate of home guard and security cameras will reach 24% in China and 15% globally. With the intergenerational development of global fixed optical fiber networks, F5G-A will become the mainstream by 2030, the number of global fiber broadband users will reach 1.6 billion, the penetration rate of gigabit or higher home broadband will reach 60%, the penetration rate of F5G-A 10 Gbps home broadband will reach 25%, the penetration rate of FTTR for Home fibers will reach 31%, and the penetration rate of FTTR for SME broadband will reach 41%.

If networks are to meet the needs of these new technologies and provide a consistent experience across our three spaces (home, office, and self-driving cars), we will need to build new network capabilities that deliver the high bandwidth, high availability, and low latency required.

2.3 Space-Air-Ground Cubic Network: Borderless Broadband for Seamless Global Coverage

In the foreseeable future, broadband coverage will extend beyond the ground, encompassing the air and even space. These networks will connect devices at various heights, such as drones and manned aircraft flying at altitudes of less than 1 kilometer, aircraft at altitudes of up to 10 kilometers, and spacecraft in low-earth orbit (LEO), hundreds of kilometers above the earth's surface. A space-air-ground cubic network will consist of small cells with a coverage radius of 100 meters, macro sites with a coverage radius of 1–10 kilometers, and LEO satellite networks with a coverage radius of 300–1000 kilometers, which will provide users with consistent and seamless broadband experiences of 10 Gbit/s, 1 Gbit/s, and 100 Mbit/s, respectively. Broadband will be omnipresent in daily life; the diversification of leisure activities and the growing demand for unmanned operations in intelligent industry and agriculture underscore the necessity of providing broadband everywhere, from land to sea and sky. (Figure 2 All-domain cubic broadband network)

Figure 2 All-domain cubic broadband network



2.3.1 Terrestrial Networks

Wireless networks have already proven their importance in boosting the digital economy and creating huge socio-economic value. To facilitate the diversified experiences of emerging services, terrestrial networks are constantly evolving toward faster speeds and deeper coverage of indoor spaces.

To support XR, naked-eye 3D display, and other services that require ultra-high network speeds, 5G-A increases the network bandwidth 10-fold. Specifically, the downlink bandwidth is increased from 1 Gbit/s to 10 Gbit/s and the uplink bandwidth from 100 Mbit/s to 1 Gbit/s.

Ultra-broadband spectrum is the basis of this 10 Gbit/s capability. Therefore, equipment is being developed to support multiple frequency bands and broadband. In addition to the nearly 100 MHz of FDD spectrum and 100–200 MHz of TDD spectrum currently allocated for 5G, higher-bandwidth upper 6 GHz (U6G) and mmWave are also introduced to provide 200–400 MHz and up to 800 MHz of spectrum, respectively. Operators in different regions can choose when to deploy sub-6 GHz, U6G, and mmWave based on their service requirements and network construction pace, adding spectrum as the need or opportunity arises. Because sites for installing wireless base stations are often expensive and difficult to acquire, the most economical and efficient way to build a high-speed terrestrial network is to add new frequency bands to existing macro and micro sites. This means that a single piece of equipment may need to support multiple frequency bands, and new technologies will need to be introduced to overcome the limited coverage of U6G and mmWave, in order to maximize the utilization of existing base station site resources.

Wireless networks must provide deep coverage to serve indoor users, as indoor use accounts for 80% of total wireless network use. Digital indoor solutions can provide large enough capacity for indoor scenarios, such as airports, stadiums, and shopping malls, by using technologies such as Distributed Massive Multiple-Input Multiple-Output (Massive MIMO). In addition, technologies such as FDD Massive MIMO, supplemental downlink (SDL), and super uplink (SUL) can be introduced to the sub-6 GHz frequency band to improve the outdoor-to-indoor (O2I) penetration of outdoor macro sites, thereby meeting the experience requirements of most indoor scenarios.

2.3.2 Non-terrestrial Networks (NTNs)

71% of the earth's surface is covered by water. The open ocean is beyond the reach of terrestrial broadband networks, as are uninhabitable or sparsely populated places on land, such as remote mountains and deserts. However, as economic globalization promotes the extraction of natural resources, these places are increasingly visited by people and equipped with IoT devices, underscoring the demand for broadband coverage there. Terrestrial networks cannot meet this demand, but satellites can. LEO satellites are located 300–2000 kilometers above the earth — high enough to provide ultra-wide coverage over unpopulated or sparsely populated places. Therefore, satellite broadband and narrowband communications are gaining popularity. In the past decade, rocket recycling technologies have matured, substantially lowering the cost of putting a satellite into orbit. Some enterprises have already deployed constellations of LEO satellites that provide 100 Mbit/s broadband for home users in areas beyond the coverage of cellular networks. Many other enterprises are planning similar deployments in the near future.

However, due to spectrum constraints and communications disruptions, the peak capacity of an LEO satellite in a satellite network is about 10–20 Gbit/s and the single-user-perceived speed of broadband access is 100–200 Mbit/s. Suppose a global satellite network comprises 10,000 satellites distributed on multiple orbital planes from very low earth orbits (VLEOs) to LEOs, and each satellite maintains links with satellites around it in all directions using over 100 Gbit/s lasercom. Considering at least half of the areas passed over by the satellites are areas where demand for broadband is minimal (e.g., oceans and deserts), the actual effective capacity of the satellite network will be around 100 Tbit/s, and the capacity density will be less than 2.5 Mbit/s/km2 (just a few percent of the capacity density of a common terrestrial 4G network in urban areas).

The 3rd Generation Partnership Project (3GPP) is defining a global mobile communications protocol standard for NTNs. In Release 17, it introduced the first 5G-based transparent payload technical standard. In Release 18, it improved the coverage and performance of IoT-NTN, and completed research on features such as air interface transmission link enhancement. For the upcoming Release 19, 3GPP is studying a network structure of regenerative satellite payload and inter-satellite link technology in order to further improve the performance and efficiency of satellite networks. Release 20 is expected to introduce a new standard for smart handheld NTN broadband terminals.

LEO satellite broadband terminals are becoming smaller. The latest commercially available portable broadband satellite CPE weighs just 1.1 kg, and is small enough to be carried in a backpack. Powered by batteries, the product can meet the typical mobility requirements of individuals, such as use in connected cars, camping trips, and exploration. It is foreseeable that satellite communications will be used as a supplement on the fringes of terrestrial 5G-A networks to meet the narrowband and broadband service requirements of people and things and achieve borderless global coverage.

2.4 Industrial Internet: A New Type of Network for Intelligent Manufacturing as Well as Human-Robot and Robot-Robot Collaboration

The industrial Internet is a new type of infrastructure that deeply integrates ICT into the industrial economy and fully connects people, machines, things, and systems. For industries, this means the birth of a brand-new manufacturing and service system that covers entire industry value chains and paves the way for digitalization, network-based operations, and the intelligent transformation of all industries. The traditional industrial Internet system consists of four key components: industrial control, industrial software, industrial network, and information security. The industrial network is the foundation of the entire system.

Traditional industrial networks are built based on the International Society of Automation 95 (ISA-95) pyramid model. This architecture was introduced more than 20 years ago and is a manufacturing system centered on human management. However, the development of intelligent manufacturing requires a new architecture that will facilitate human-robot and robot-robot collaboration.

The new architecture will be built upon three equal elements – humans, robots, and an intelligent platform (cloud/edge computing). Private industrial communication buses will be replaced by universal industrial networks and open data layers that support real-time data transmission. The intelligent platform will aggregate data collected from humans and robots for real-time analysis and decision making and support effective collaboration between humans and robots.

To support the stable development of the industrial Internet, the network must meet the following requirements:

- Deterministic network latency: Industrial applications like automatic control and motion control pose strict requirements on the latency, jitter, and reliability for network data transmission.
- Network reliability: Control services on industrial sites are typically performed within milliseconds. This requires protection switchover to be completed within sub-seconds.
- Intelligent O&M management: Effective O&M management for industrial networks hinges on achieving zero workload through streamlined processes that minimize the burden to industrial production.

Huawei predicts that the total number of global connections will reach 200 billion by 2030, including about 100 billion wireless (cellular) connections (including passive cellular connections) and about 100 billion wired, Wi-Fi, and short-range connections. In industrial settings, the multitude of connected devices will include not only pressure, photoelectric, and temperature and humidity sensors, but also numerous intelligent cameras, drones, and industrial robots. With the advent of the AI era, Huawei predicts that 20 million industrial robots will enter the cutting-edge smart manufacturing field by 2030. Consequently, industrial networks, currently characterized by a fragmented landscape of different narrowband technologies, will adopt universal broadband technologies.

Universal industrial networks will erase the technical boundaries between consumption, office work, and production. These networks will support multiple types of services using deterministic broadband networks and slicing technologies, such as 5G, Time Sensitive Networking (TSN), IPv6 Enhanced, and industrial optical networks, allowing enterprises to connect any workforce and migrate all consumption, office work, and production elements to the cloud.

Universal industrial networks will enable on-demand data sharing and seamless collaboration between office and production systems within a company, between different companies in the

Direction for exploration: Sharing vehicles for faster, low-carbon transportation

The transportation system of the future will be a multi-layered, efficient, and comprehensive system that integrates multiple modes of transport, including road, rail, air, and water transport. When a centralized transportation management system schedules the vehicles and resources, passengers can be provided with tailored mobility solutions based on their individual needs. This also means that vehicles will be used much more efficiently. The shared vehicle model avoids wasteful, carbon-intensive transportation where a single passenger drives a vehicle to a single destination.

Snapshot from the future: Mobility as a Service available on demand

According to the International Road Transport Union, Mobility as a Service (MaaS) is to put the user at the core of transport services, offering them tailor-made mobility solutions based on their individual needs. MaaS is the integration of various modes of transport into a single mobility service accessible on demand. It combines all possible modes of transport, enabling users to access services through a single application and single purchase. [25]

A key objective of MaaS is to provide integrated and convenient public transport services and develop green transport. MaaS systems aim to integrate local transport (e.g., buses, rail, shared cars, and shared bikes) and intercity transport (e.g., planes, high-speed rail, and long-distance coaches) and provide useful local information about dining, accommodation, shopping, and local tourist attractions. These systems will build on the intelligent scheduling functions of public transport systems, and identify passenger travel models while prioritizing green transport. With online payment functions integrated, MaaS systems can offer travel booking, one-tap itinerary planning, seamless connections between different modes of transport, and one-tap payments. MaaS will improve satisfaction with transport services while also providing green transport options.

Many EU cities are building MaaS showcase projects. Different cities have different levels of integration in terms of facilities, fares, payments, information, communications, management systems, and transport services. Gothenburg, Hanover, Vienna, and Helsinki were the first cities to explore MaaS. These cities have made full use of digital technologies to optimize their transport systems, including buses, shared cars, bicycles, and urban deliveries. This will help them incubate emerging transport service providers and drive urban decarbonization. [26]

MaaS can bring tangible benefits: Individuals can cut their transport costs while enjoying better safety and a better experience. Governments can optimize their investment in transport infrastructure for more sustainable urban management and higher citizen satisfaction. In addition, MaaS will create more opportunities for transport service providers, as they can cut service costs and expand their services. When MaaS is widely deployed, we will see integrated scheduling of transport resources, better shared resources, a user-centric experience, and low-carbon transport.

same industry, and even between the related services of different vertical industries. They will support broadband-based interconnectivity and multi-cloud data sharing of any workload.

Universal industrial networks will also be smarter than ever, facilitating the movement of data in boundary-free and mobile scenarios across industries and across clouds. They will support intent-driven automated network management and AI-based proactive security and privacy protection, ensuring service security and trustworthiness at any workplace.

An enterprise usually has multiple types of services, so a universal industrial network must ensure

the availability, security, and trustworthiness of services. For example, smart healthcare involves services such as remote diagnosis, monitoring & nursing, and remote surgery; a smart grid involves video-based inspection, grid control, and wireless monitoring; and smart manufacturing involves factory environment monitoring, information collection, and operation control. (Table 6 Network requirements of intelligent enterprises)

Table 6 Network requirements of intelligent enterprises

Industry	Service Type	Network Requirements of Services															
		Number of Connections per Enterprise	Service Availability (Requirements per User or per Service)										Security		Trustworthiness		
			Bandwidth per User (Mbit/s)					Latency (ms)									
			B1	B2	B3	B4	B5	T1	T2	T3	T4	T5	S1	S2	M1	M2	M3
			1~10	10~20	20~50	50~100	>100	50~100	20~50	10~20	5~10	<5	Logical Isolation	Physical Isolation	Visible	Manageable	Operable
Smart health	16K remote diagnosis	10					1G										
	Monitoring & nursing	2K															
	Holographic remote surgery	5					10G										
Smart grid	Video-based inspection	-															
	Grid control	-															
	Wireless monitoring	-															
Smart manufacturing	Factory environment monitoring	100															
	Information collection	10K															
	Operation control	1K															

Reference: CAICT, Research Report on Industry SLA Requirements for 5G E2E Network Slicing

Based on the typical bandwidth and latency requirements of each service and forecasts on the number of devices used by enterprises in 2030, we predict that a medium- to large-sized enterprise will require network bandwidth of 100 Gbit/s and the maximum bandwidth per user will reach 10 Gbit/s. Acceptable latency will vary greatly from one use case to another, from as low as 1 ms to as high as 20 ms. In addition, it will be necessary to ensure the security and trustworthiness of industrial networks.

2.5 Computing Power Network: Orienting Towards Machine Cognition and Connecting Intelligent Computing Centers, Massive Amounts of User Data, and Computing Power Services at Multiple Levels

The social value of communications networks is reflected in the services they support. In the past, networks helped establish communications channels between people by providing communications services. Today, with smart devices and the cloud connected to networks, more diverse content services are provided through communications networks.

The networks we use today are designed for human cognition. For example, the frame rate for motion video (typically 30 frames per second [FPS]) is chosen based on the human ability to perceive motion, and the audio data collected is compressed with mechanisms that take advantage of the masking effects of the human cognitive system. For human perception, such encoded audio and video can be considered high quality. However, for use cases that require beyond-human perception, the level of quality may be far from enough. For example, robotic monitoring systems will need to detect anomalies by listening to sounds beyond the human audible frequency range. In addition, the average human response speed upon seeing an event is about 100 ms. Therefore, many applications have been designed based on this latency. However, for certain applications that are beyond human usage, such as emergency stop systems, shorter response time is required.

The Innovative Optical and Wireless Network Global Forum Vision 2030 and Technical Directions states that compared with today's networks that are designed for human cognition, future networks designed for intelligent machines such as XR, machine vision, and self-driving vehicles will have enhanced performance in four dimensions:

- Cognitive capacity: Systems will be able to capture objects in the physical world more finely, precisely, and in a multi-sensory manner. For instance, in manufacturing monitoring systems, motion capture at 120 FPS will detect anomalies that would otherwise be undetectable.
- Response speed: Systems will be able to respond to the status change of a controlled object within 10 ms.
- Scalability in computing: Systems will be able to accommodate varying and uncertain workload while achieving high resource utilization, through methods such as dynamic linear scaling of computing resources.
- Energy efficiency: Energy efficiency can be greatly improved if enterprises eliminate on-premise computing resources and adopt a cloud-based model. Moreover, energy efficiency will be further improved with an event-driven approach where a system is deployed on a serverless computing platform.

Intelligent machines will create more accurate data. For example, network clocks and geolocation stamps can be used for precise modeling of the physical world in a digital twin system. This will lead to a shift in data processing and computing, from today's Internet platform-centric model to a data-centric model, decoupling data, computing, and communications.

The network infrastructure designed for machine cognition should satisfy the following requirements:

- Accommodating the collection and transmission of massive amounts of data, having an ultra-low latency, and supporting a very large number of subscribers.
- Managing publishers' data generation and injection based on the overall condition of the system and the importance of the data.
- Supporting the storage and sharing of data among communications and computing nodes in the network.

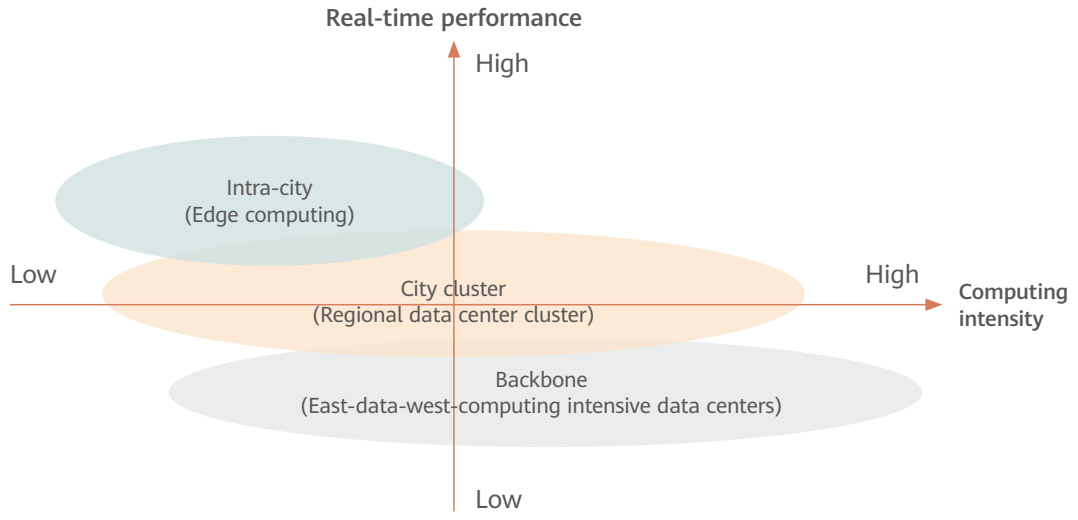
- Supporting precision time and geolocation stamping.
- Providing strong protections for data security, privacy, and integrity.
- Providing a data brokerage between IP and non-IP nodes, with the data brokerage being accessible through multiple networks.

As AI foundation models continue to advance, data centers will witness the number of model parameters increasing to trillions, tens of trillions, or even hundreds of trillions. Such huge numbers will overwhelm a single intelligent computing center. A common intelligent computing center can now host an array of 10,000 to 50,000 Graphical Processing Units (GPUs)/Neural Processing Units (NPUs), while a cutting-edge one usually accommodates up to 60,000 such GPUs/NPUs. As the scale of GPUs/NPUs exceeds 80,000, a single intelligent computing center will face a variety of challenges, ranging from unstable power supply and inefficient heat dissipation to inadequate network bandwidth. These technical bottlenecks make it difficult for a single intelligent computing center to accommodate 100,000+ GPUs/NPUs. As such, distributed computing power collaboration across data centers has emerged as a pressing necessity to keep pace with AI's growing demands for computing power.

In terms of terminals, the computing industry can no longer rely on Moore's law for rapid development given the miniaturization of chips is approaching its physical limits. For example, manufacturing Central Processing Units (CPUs) with more than 128 cores in smart terminals presents economic bottlenecks. Furthermore, the local computing power of devices cannot support the running of ultra-large models due to volume and power consumption constraints. In addition, due to bandwidth costs and latency, cloud data centers may not be able to satisfy the massive amount of time-sensitive service processing required by intelligent systems. That said, the new type of network oriented to machine cognition must allow foundation models to be deployed at the edge for data analytics, processing, inference, and more, without needing to transmit all data to the central cloud.

In the future, the cloud, edge, and devices will be connected, and computing workloads will be apportioned to one of three levels (distributed edge nodes in a city, regional data center clusters that cover multiple cities, or backbone centralized data centers) in real time based on their latency thresholds. In use cases that can tolerate latency of about 20 ms, data may be sent to a

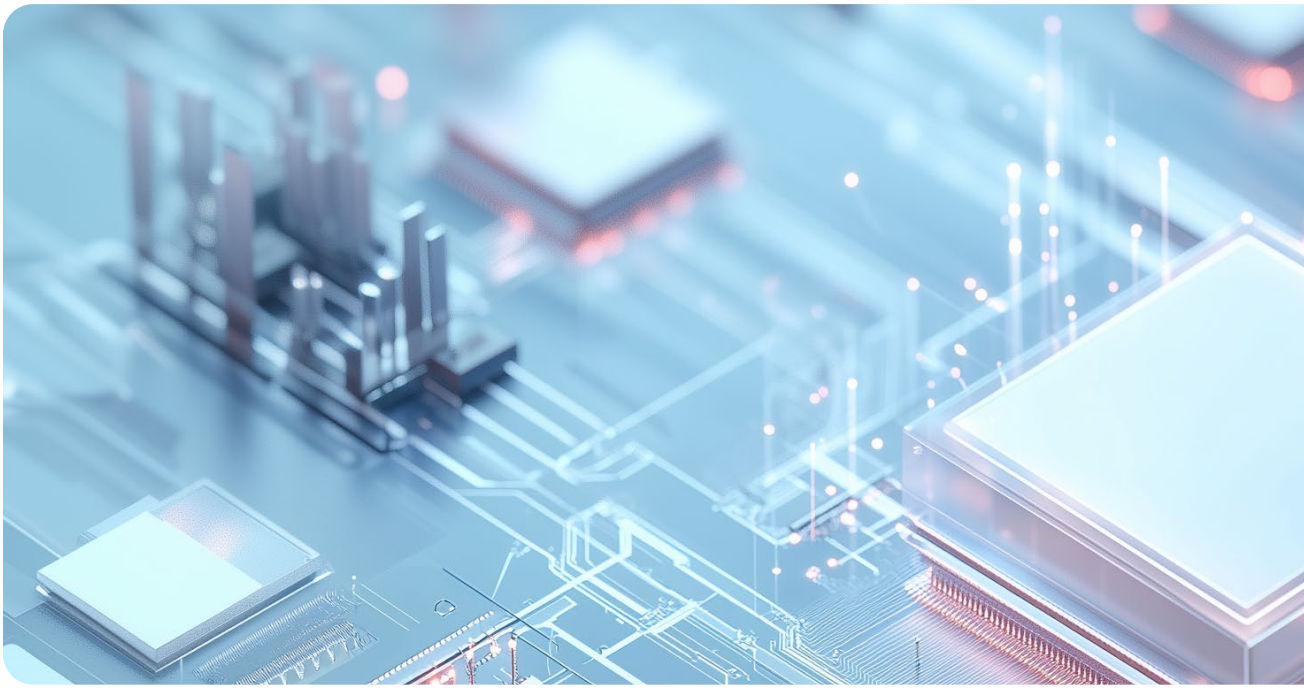
Figure 3 Three levels of computing resources for machine data services



centralized data center. In use cases with lower latency tolerance (from 5 ms to as low as 1 ms), computing will be performed in a regional data center cluster or at the edge. (Figure 3 Three levels of computing resources for machine data services)

Computing efficiency and reliability are correlated with network bandwidth, latency, security, and isolation. Therefore, computing and networks should be coordinated. Major carriers have articulated a new business vision for computing and network convergence services based on a new concept of "computing power network". They aim to connect diverse computing power in the cloud, on the edge, and across devices to implement on-demand scheduling and sharing for efficient computing power services at multiple levels. The computing power network represents a significant shift in network design, from focusing on human cognition to focusing on machine cognition.

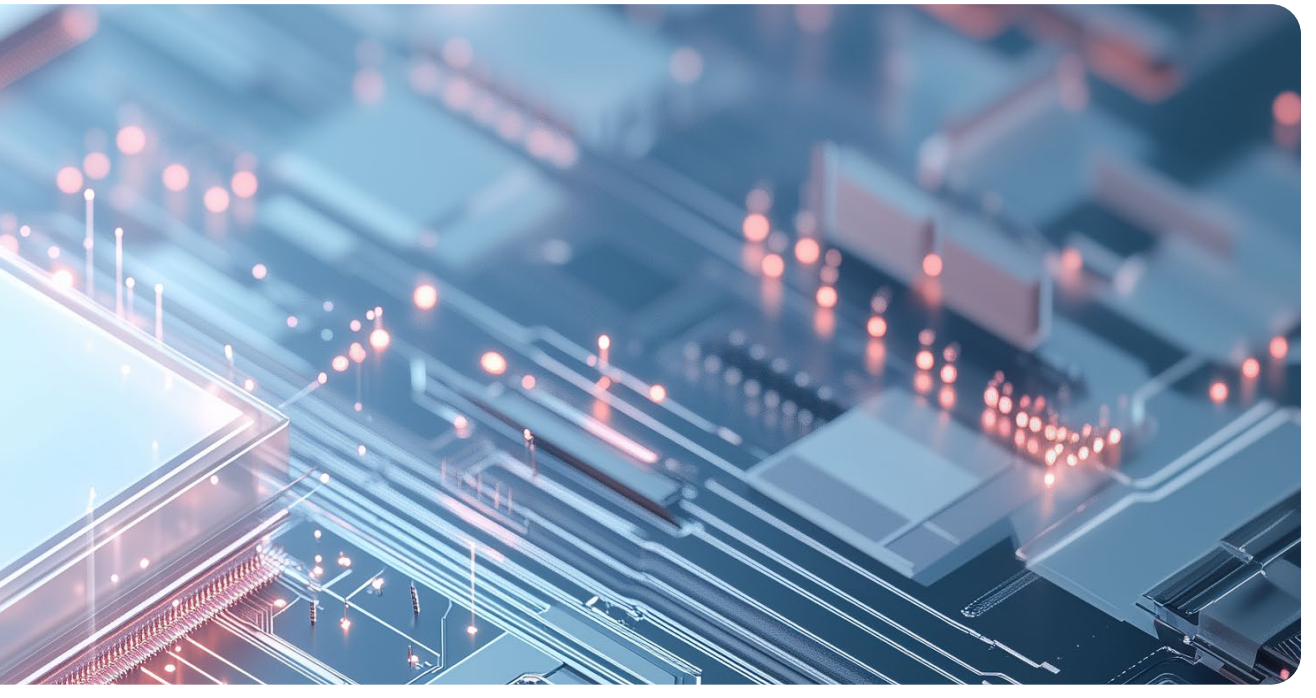
The Chinese government released the Guiding



Opinions on Accelerating the Construction of Collaborative Innovation System of National Integrated Big Data Centers, which states: "With the acceleration of digital transformation and upgrade in various industries, the total volume of data being created by society as a whole is growing explosively, and the requirements for data resource storage, computing, and applications are greatly increasing. Consequently, there is an urgent need to promote an appropriate data center layout, balance between supply and demand, green and centralized development, and interconnectivity. We should build a new computing power network system that integrates data centers, cloud computing, and big data, in order to promote flows and application of data elements and achieve green and quality development of data centers." In addition, the document proposed that "as data centers should be developed on a large scale in a centralized and green manner, network transmission channels between national hubs and nodes should be further streamlined to accelerate the program of 'Eastern Data and Western Computing' and improve cross-region computing power scheduling."

To support proactive development of computing power network standards, the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) has launched the Y.2500 series of computing power network standards, with Y.2501 (Computing Power Network – framework and architecture) as the first standard. This series of standards will be compatible with a raft of computing power network standards developed by the China Communications Standards Association (CCSA). Many carriers have incorporated the computing power network into their 6G and future network research. The computing power network will be a key scenario for communications network evolution over the next 10 years.

Huawei predicts that by 2030, optical connections in a city will be extended to all city scenarios, such as homes, buildings, enterprises, and 5G base stations, to enable 1 ms access to cloud and computing, and every 10,000 people will have four all-optical OTN anchors, among which 100G anchors will account for 25%, and the OTN coverage rate of transmission networks will reach 100% in government agencies, financial institutions, key universities and scientific research institutions, large hospitals, and large industrial enterprises, as well as development zones and industrial parks at the county level or above.



2.6 AN: Unattended Self-evolving Network

Throughout the history of the communications industry, the continuous innovation in crucial technologies has helped to improve the network capabilities of carriers, which in turn stimulates the innovation in a diverse range of services while advancing various industries. Additionally, networks have proposed an objective of achieving Highly Autonomous Networks featuring agile service provisioning, precise user experience assurance, and efficient cross-domain O&M. The significant breakthroughs in GenAI and rapid development of crucial technologies such as the telecom foundation model have created unprecedented opportunities of new traffic, connections, and services in the telecom industry, significantly accelerating the transformation toward network intelligence. In the future, carriers will leverage AI agents and digital assistants to deliver a zero-trouble, zero-wait, and zero-touch service experience to customers and implement self-configuration, self-assuring, and self-optimizing O&M for networks. This will help them promote service innovation to increase revenue while empowering various industries to develop new quality productive forces. Huawei expects that AN Level 4 will be reached by leading carriers by 2025 and by most carriers by 2030.

From the technical perspective, AN Level 4 is characterized by machines functioning as the major O&M entity which is assisted by humans during decision-making in key tasks/scenarios. In the future, machines will be widely used to understand human intents through AI, generate suggestions on network planning and optimization, and use decision-making AI models to complete intelligent decision-making, so as to achieve machine-oriented AN.

TM Forum also characterizes AN Level 4 from the perspective of service value, offering a valuable reference for the industry to systematically evolve towards AN Level 4. (Table 7 AN Level characteristics)

Table 7 AN Level characteristics

Perspective	Dimension	Level Characteristic	
		L3/Machines assisting humans	L4/Humans assisting machines
Customer	Zero-wait	Automated service provisioning	Automated service delivery
	Zero-trouble	Experience awareness and visualization	Experience evaluation and assurance
	Zero-touch	Visualization	Interaction
Network	Self-configuration	Automated configuration delivery	Pre-event simulation Post-event verification
	Self-healing	Precise fault diagnosis	Potential risk prediction and prevention
	Self-optimizing	Single-objective exclusive optimization	Multi-objective collaborative optimization

The telecom foundation model is a key enabling technology for AN Level 4. It employs AI agents and digital assistants to redefine network capabilities in an E2E manner, empowering carriers to achieve the transformation toward network intelligence. AI agents can leverage the intent understanding and logical reasoning capabilities through the telecom foundation model to break down complex problems and find optimal solutions. System capabilities can then be employed to implement scenario-specific autonomy, delivering a new network O&M experience. AI digital assistants can be customized using the natural language understanding capability of the telecom foundation model to complete complex human-machine interaction through different roles, simplifying learning and improving learning efficiency for employees. Additionally, AI digital assistants can offer vast amounts of knowledge and data to employees on demand to enhance O&M efficiency.

The increasing application of AI agents and digital assistants across the entire network production process — involving planning, construction, maintenance, optimization, and operations — will transform communications networks from the following perspectives by 2030:

- O&M mode: creating a new interactive O&M mode based on natural language
- System capabilities: improving system capabilities in terms of awareness, analysis, decision, and execution
- Service processes: designing service flows centering on machines to automate E2E processes and reduce the time to market (TTM) of services
- Integration mode: replacing the conventional application programming interface (API) integration mode with the self-service integration mode through large models to reduce the TTM of services

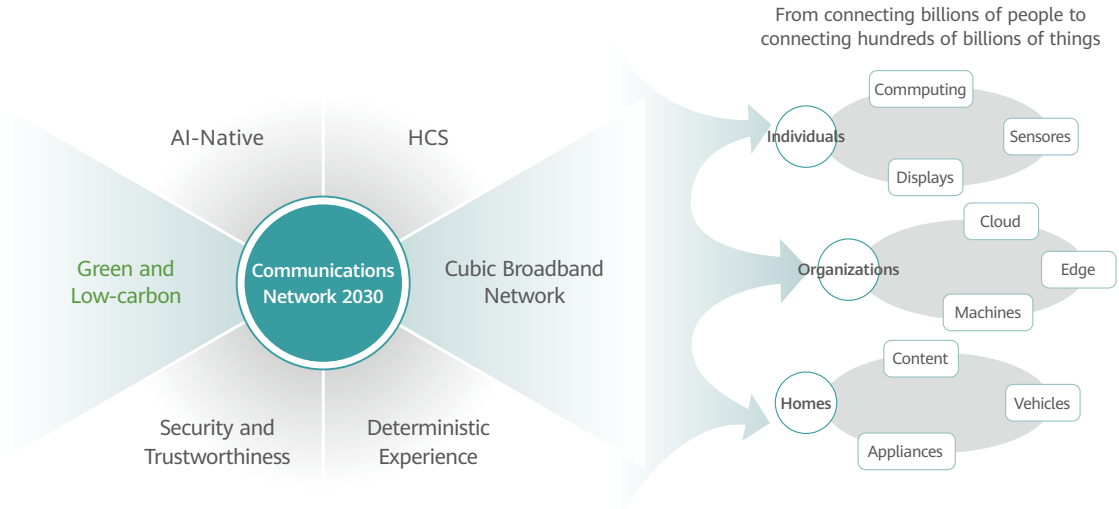
03

Defining Features of Future Networks

3.1 Development Directions of Future Network Technologies

Future networks won't just connect billions of people; they will connect hundreds of billions of things. We envision those connections as being supported by green and cubic broadband networks that are AI-native, secure, trustworthy, and capable of providing deterministic experiences and HCS. (Figure 4 Vision for the communications network of 2030)

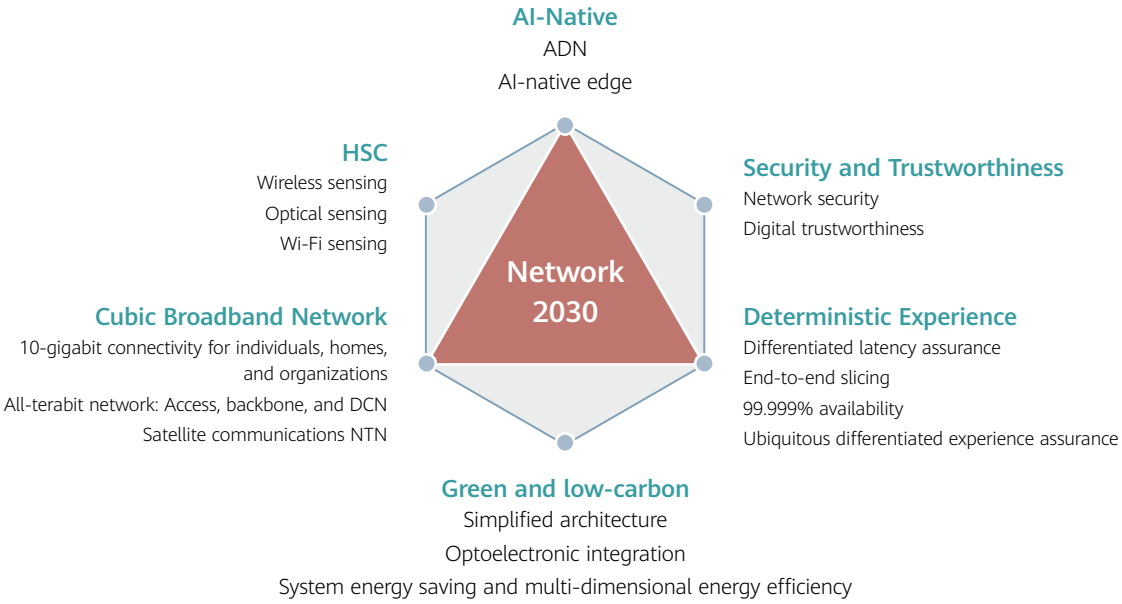
Figure 4 Vision for the communications network of 2030



3.2 Defining Features

The communications networks of 2030 will have six defining features enabled by 17 key technologies, and each key technology will rely on research on multiple technological fronts. (Figure 5 Defining features of the communications network of 2030)

Figure 5 Defining features of the communications networks of 2030



3.2.1 Cubic Broadband Network

The coming decade will see continual improvement in network performance. We will see an evolution from today's gigabit access enabled by 5G, F5G, and Net5G toward 10-gigabit capacity enabled by 6G, F6G, and Net6G. The global penetration rate of gigabit or higher home broadband networks is expected to reach 60%, and that of 10-gigabit home broadband networks is expected to reach 25%. The average monthly network data usage per household is forecast to increase by 8-fold to 1.3 TB. Network ports will be upgraded from 400G to 800G or even 1.6T, and single-fiber capacity will exceed 100T. In terms of coverage, network construction up until now has focused on connectivity on the ground, but in the future, we will see the construction of networks connecting the ground, air, and space.

1) 10-Gigabit Connectivity for Individuals, Homes, and Organizations

Fiber networks are expected to be widely deployed globally over the next 10 years, transforming today's gigabit connections for individuals, homes, and organizations into 10-gigabit connections.

To deliver 10G home broadband, 50G passive optical network (PON) technology will be applied to more than one million ports by 2027 and 200G PON technology will likely be used by 2030 on optical access networks. The coherent detection technology typically used for wavelength division multiplexing (WDM) will be used in the PON field, which will significantly improve receiver sensitivity and support modulation formats with higher spectral rates, such as quadrature phase shift keying (QPSK) and 16-quadrature amplitude modulation (16-QAM), to achieve higher data rates.

To deliver 10-gigabit broadband for individual users, mobile network research has been focusing on full use of sub-100 GHz spectrum combinations and iterative evolution of Massive MIMO, and has produced the following landmark innovations:

- Extremely large antenna array-Massive MIMO (ELAA-MM): This helps overcome the limited coverage of high frequency bands (such as mmWave and 6 GHz) to achieve ubiquitous 10 Gbit/s downlink experience.
- Multi-band serving cell (MBSC): By binding multiple frequency bands into a virtual large carrier to share control channels, control channel resources are conserved, downlink user experience is improved by about 30%, and cell capacity is increased by about 20%.
- Flexible spectrum access (FSA): Uplink resource bottlenecks are alleviated through flexible uplink and downlink slot configuration, achieving 1 Gbit/s uplink experience.

In 2020, 3GPP Release 16 defined two frequency ranges, Frequency Range 1 (FR1) and Frequency Range 2 (FR2), for 5G new radio (NR), covering all spectrum bands for International Mobile Telecommunications (IMT) between 450 MHz and 52.6 GHz. In 2022, Release 17 defined the 6 GHz spectrum as an IMT licensed band for NR (numbered n104). In 2023, Release 18 specified the range of the band n104 (6425–7125 MHz), and its freezing marks the start of 5G-A. Release 19 will cover the major evolution directions of 5G-A, with the first batch of initiated topics covering new services and technologies such as AI for air interface, integrated sensing and communication (ISAC, channel research), and Ambient IoT, and will start the research on the channel model for the 7–24 GHz spectrum.

Regarding the ongoing evolution of Massive MIMO, Release 17 defined FDD CSI enhancement and TDD SRS capacity expansion standard features, and work has been initiated on some important aspects of multi-antenna technology, such as hybrid beamforming (HBF) for the new U6G band, sub-band full duplex, and FDD 64T/128T Massive MIMO, with an eye to inclusion in Release 19.

To make 10-gigabit campus networks possible, more research is needed on all-optical Ethernet technologies for 10GE and 100GE access and

next-generation Wi-Fi technologies that support millimeter-wave and high-density MIMO. Theoretically, Wi-Fi 7 standards should be able to support 10-gigabit user access. With wireless air interface technology approaching Shannon's limit, further evolution of Wi-Fi and mobile technologies will require more spectral resources, which are scarce. This has prompted industry-wide discussions about the feasibility of converging Wi-Fi 8 and 6G.

2) All-Terabit Network: Access, Backbone, and DCN

Taking into account the growing broadband requirements of individuals, homes, and enterprises, as well as in AI training and prediction scenarios, future access network equipment will need to support terabit-level interfaces. Backbone equipment will support 40–100 Tbit/s per slot and data center equipment 400 Tbit/s per slot.

By 2030, there will be broadband networks that can achieve terabit-level transmission speeds in many parts of the networks, from access, backbone, and data center to the Internet. These will mostly serve the world's largest cities – those with populations of 10 million or higher.

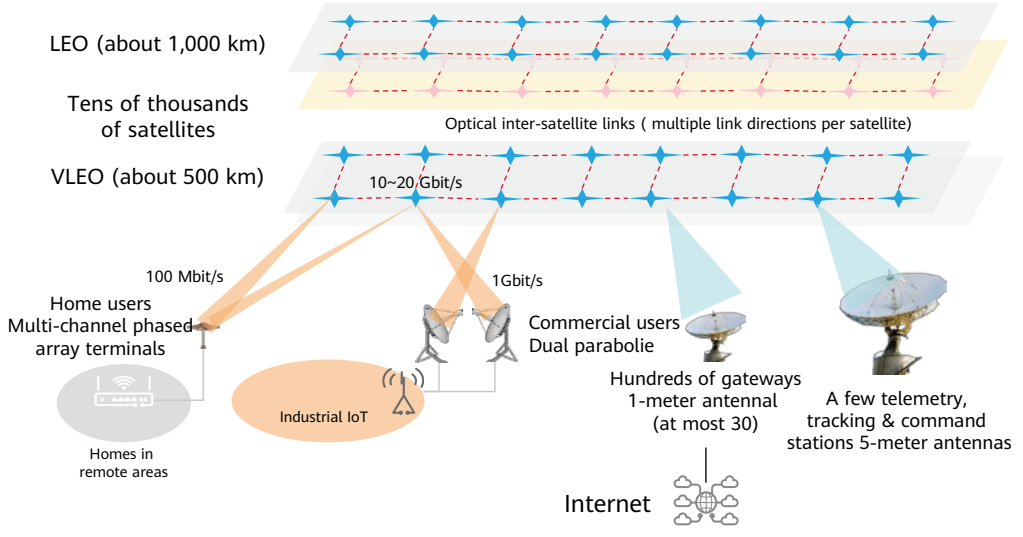
In the terabit era, datacom equipment will need to have Ethernet interface technology that supports speeds of 800 Gbit/s or even 1.6 Tbit/s to meet service development needs. Unlike 200G/400G Ethernet, 800G Ethernet is a nascent technology that has yet to be standardized. From a technical standpoint, there are two routes that will take us to 800G: continuing evolution of existing pluggable optics modules and the adoption of new co-packaged optics (CPO) modules. Both module types will have a place in the future market, but pluggable optics modules with a capacity of over 800G are expected to encounter power and density problems, so CPO modules will likely become the preferred choice.

Moreover, enabling a transmission capacity of more than 100 Tbit/s per system will require technical breakthroughs in WDM equipment, including high-baud-rate electro-optic modulator materials, spatial division multiplexing (SDM) transmission systems and modules, and new optical amplifier technology that goes beyond C band to L band and S band.

3) Satellite Communications NTN: Effective Supplement to Terrestrial Network Coverage

LEO satellite broadband mainly serves homes and enterprises in remote areas, and ships at sea. It can be used for backhaul and combined with cellular networks and wireless local area networks (WLANs) on the ground to provide both broadband and narrowband coverage for villages or enterprises in remote areas. (Figure 6 Satellite communications network)

Figure 6 Satellite communications network



For satellite-terrestrial transmission, broadband CPEs and handheld NTN terminals must be able to efficiently access satellite networks. To make this possible, we need to study new air interface technologies to overcome the deep fading, high latency, ultra-high Doppler shift, and highly dynamic nature of satellite networks. We also need to explore key air interface technologies, such as time-frequency synchronization technology for random access and high-speed handovers over air interfaces, as well as optimized encoding, decoding, waveform, modulation, and multiple access technologies for satellite-terrestrial links. These advancements will enable highly reliable access, efficient multiple access and wireless transmission, and high-speed mobility management.

To improve the coverage capability and network spectral efficiency of satellites, there is a lot of research that needs to be done. Specific areas of research include:

- High-performance multi-antenna beamforming technology and ultra-large-aperture and high-gain multi-antenna technology, to enable spatial multiplexing for ultra-high beam concurrency and high-speed beam switching, in order to support the access of high-performance broadband CPEs and smart handheld terminals likely to be defined in 3GPP Release 20

- Technologies to mitigate interference between the beams of an individual satellite, as well as inter-satellite interference, to improve spectrum multiplexing rates and spectral efficiency

- Unified multi-user scheduling of time, frequency, space, and power domain resources of satellite-terrestrial links in a multi-layer LEO satellite constellation, to fully and efficiently utilize network resources.

- Large-bandwidth satellite-terrestrial lasercom technology, to meet increasing feeding bandwidth requirements, along with solutions for mitigating atmospheric turbulence during laser transmission

Inter-satellite transmission requires satellites at different orbital heights to form multi-layer constellations, with each layer networking through inter-satellite links. Inter-satellite links are established on demand between satellites in the same orbit, at the same layer, and at adjacent layers, forming a cubic space network. Inter-satellite links will use lasercom and terahertz technologies to support a speed higher than 100 Gbit/s. This will require research on adapting industrial products to aerial settings, making phased array antennas more compact, and enabling dynamic inter-satellite lasercom tracking and pointing.

The network management and control domain comprises an operation and control center, network management center, gateway, and converged core network. In order to perform the tasks of satellite network management, user management, and service support, we need to research flexible and efficient dynamic routing protocols between gateways and constellation networks, and hyper-distributed converged core networks that support intelligent switching of multi-layer satellite constellations.

3.2.2 Deterministic Experience

The ability of communications networks to provide deterministic experiences is key to supporting online office and learning, as well as meeting the security and reliability needs of production environments.

1) 20 ms, 5 ms, and 1 ms Latency Assurance for Differentiated Service Requirements

Over the course of this decade, the Internet traffic model will undergo a fundamental shift from today's top-down content traffic generated primarily from online services, retail, and entertainment to bottom-up data traffic from pervasive intelligent applications deployed across various industries. Intelligent machines will generate massive amounts of data, and this data will need to be processed in data centers. This decade will see a push toward the coordinated development of electricity and computing power to enable society-wide green computing power. Therefore, the networks of the future will need to be able to support more centralized operations of data centers. That will entail meeting differentiated latency requirements, with the acceptable latency for backbone, inter-city, and intra-city network services being 20 ms, 5 ms, and 1 ms, respectively.

In addition, networks will need to schedule resources in real time at the network layer based on service attributes in order to make computing power greener and more efficient.

In addition to meeting differentiated latency requirements at the network architecture and system levels, the industry also needs to research E2E deterministic latency.

Computing power and data components are progressively moving edgeward. Wireless access latency now accounts for 30% to 60% of network latency. Reducing wireless access latency has become the focus for enhancing session experience. However, wireless air interface sharing leads to resource sharing by multiple user devices, making it difficult to guarantee real-time performance and high speeds. To solve this problem, multi-carrier aggregation and multi-antenna spatial multiplexing technologies need to be developed to optimize carrier configurations and increase air interface capacity. These technologies, together with differentiated and hierarchical scheduling policies, will improve the bandwidth of services under latency constraints on multi-band carriers and provide deterministic experience for applications.

Furthermore, an intelligent closed-loop experience assurance mechanism needs to established so that intelligent core networks can implement real-time experience awareness and scheduling, ensuring deterministic service experience.

The optical access networks we have today feature PON technology, which is based on time division multiplexing (TDM). PON uses uplink burst to prevent collisions, making it ill-suited to scenarios requiring low latency. Frequency division multiplexing (FDMA) needs to be explored to allow concurrency of multiple optical network terminals (ONTs) and guarantee low latency by addressing fundamental issues.

For wide area networks (WANs), the current best-effort forwarding mechanism needs to be changed, protocols at the Physical (PHY) and Medium Access Control (MAC) layers need to be improved, and new technologies such as time-sensitive networks (TSNs) and deterministic IPs need to be integrated to ensure on-demand, E2E latency.

2) E2E Slicing: Logical Private Networks and Services That Are More Adaptable to Vertical Industries

E2E slicing provides vertical industries with customized private network services that run independently and are isolated from each other. This is a key area we can work on in order to serve vertical industries. E2E slicing is a network virtualization technology with Service Level Agreement (SLA) assurance. Through network slicing, different logical or physical networks can be isolated from the network infrastructure to meet the SLA requirements of different industries and services. Types of slicing include wireless slicing, transport network slicing, and core network slicing. When a carrier provides a slice to a customer, the carrier also provides E2E management and services.

Wireless slicing: It can be further classified into hard slicing and soft slicing. Hard slicing is achieved through resource isolation, such as

through static resource block (RB) reservation and carrier isolation for specific slices. Soft slicing is achieved through resource preemption, such as quality of service (QoS)-based scheduling and dynamic RB reservation. Currently, the bitrates of different network slices can be guaranteed based on priorities. The next step in the development of network slicing is to explore the most appropriate wireless protocols for the PHY, MAC, Radio Link Control (RLC), and Packet Data Convergence Protocol (PDCP) layers. For example, we could have a PHY layer with a low-latency coding scheme for slices that support ultra-reliable low-latency communication (URLLC) services, or a MAC layer with an optimized hybrid automatic repeat request (HARQ) mechanism.

Transport network slicing: This is achieved through physical isolation or logical isolation. Physical layer isolation technologies include optical-layer fine-grain OTN (fgOTN), allowing different services to be carried through different wavelengths or through fgODU within a single wavelength. Flexible Ethernet (FlexE) at the MAC layer is also used to isolate services by scheduling timeslots. Physical layer and logical layer isolation complement each other in terms of technology, providing both deterministic and flexible network capabilities for transport networks. Further research is needed in the industry to explore the integration of technologies such as congestion management mechanisms, latency-oriented scheduling algorithms, and highly reliable redundant links for FlexE, TSN, and deterministic networking (DetNet), as well as PON+OTN/IP E2E slicing capabilities. This can deliver deterministic latency and zero packet loss for physical slicing, as well as low-granularity FlexE interfaces.

Core network slicing: In 5G standalone (SA) architecture, microservices are the smallest modular components of core network functions. In the future, microservices will need to be flexibly orchestrated into different slices based on service requirements, and flexibly deployed in different parts of the network based on differentiated latency and bandwidth requirements.

E2E management and services: 3GPP has defined an E2E network slicing management function (NSMF), which streamlines network slice subnet management functions (NSSMFs) to enable E2E automatic slicing. This can facilitate elastic slice service provisioning and capacity expansion or reduction. Moving towards 2030, the SLA awareness, precision measurement and scheduling of slicing need to be further researched in the industry to achieve automated closed-loop slicing control. In addition, customers in vertical industries must be able to flexibly customize slicing services on demand. More efforts are needed to study how to meet industry customers' Create, Read, Update, and Delete (CRUD) requirements for slices, and how to coordinate the configuration of slices, private networks, and edge services.

3) Higher than 99.999% Availability for Industry Production Control Systems to Enable Enterprises to Migrate All Systems to the Cloud

Traditional enterprise management and production systems are human-centric and built based on the ISA-95 pyramid model, such as enterprise resource planning (ERP), manufacturing execution system (MES), supervisory control and data acquisition (SCADA), and programmable logic controller (PLC) systems. As enterprises become intelligent, these systems will be built on human-thing collaboration, and we will see the wide adoption of a new flattened architecture for cloud, edge, things, and humans.

Currently, enterprises are primarily migrating their ERP and MES systems to the cloud, which do not have real-time requirements and require

the availability of the cloud and network to be just 99.9%. By 2030, however, enterprises will be migrating all of their systems to the cloud, including systems that require higher than 99.9999% availability for the cloud and network (and edge), such as SCADA and PLC.

Moving forward, improving radio access network availability will be a major area of research. 5G can already meet the basic reliability requirements of URLLC scenarios such as ports and mines, in which availability has reached 99.99%. In the future, AI technologies will be introduced to improve the availability of mobile networks to 99.999% by better predicting channel fading characteristics, identifying envelope channel changes, increasing the number of URLLC connections supported by a single unit of spectrum, and enabling intelligent prediction, interference tracking, and E2E collaboration.

A data center is generally limited by the site scale and power supply, preventing computing hardware from being continuously expanded. Distributed training in collaboration with other data centers can effectively break computing bottlenecks, and will be the trend of computing development in the future. The reliability of network connections must reach 99.9999% or higher to ensure efficient and reliable model training and greatly reduce the time and cost caused by data transmission interruption or retransmission.

4) Ubiquitous Differentiated Experience Assurance

In the future, immersive communication, multi-modal communication, cloud gaming, and cloud phones will take off, gaining wide popularity. These services have additional requirements on bandwidth and latency, starkly different from those of common applications such as video and web. As such, an intelligent differentiated experience assurance mechanism is required to provide differentiated assurance based on user types, terminal types, application types, busy/idle hours, and scenarios/areas. This is necessary to meet network requirements of new applications while maximizing network efficiency.

3.2.3 AI-Native

1) ADN: Continuous Evolution Toward AN Level 4+

ADN is an advanced stage of network automation that aims to use an intelligent, simplified network architecture driven by data and knowledge in order to deploy a self-fulfilling, self-healing, self-optimizing, autonomous network. Such a network will be able to provide new services, deliver a superior user experience, implement fully automated O&M, and maximize resource and energy efficiency.

ADN is still in the development phase transiting from Level 2 to Level 3 and possesses partially and conditionally autonomous capabilities. These capabilities allow the system to achieve closed-loop O&M for certain units in particular external environments based on the AI telecom foundation model. ADN will evolve toward Level 4+, achieving closed-loop automation for a diverse range of services throughout the entire network lifecycle in a more complex cross-domain environment. (Table 8 ADN levels)

Table 8 ADN levels

Level	LO: Manual O&M	L1: Assisted O&M	L2: Partial Autonomy	L3: Conditional Autonomy	L4: High Autonomy	L5: Full Autonomy
Service	N/A	Single use case	Single use case	Multiple use cases	Multiple use cases	Any use cases
Execution	Manual	Manual/ Autonomous	Autonomous	Autonomous	Autonomous	Autonomous
Awareness	Manual	Manual	Manual/ Autonomous	Autonomous	Autonomous	Autonomous
Analysis/ Decision	Manual	Manual	Manual	Manual/ Autonomous	Autonomous	Autonomous
Intent/ Experience	Manual	Manual	Manual	Manual	Manual/ Autonomous	Autonomous

Reference: TMF 2020

To support the evolution of ADN toward Level 4+, we need to study the following key technologies:

The management and operation layer: This layer unifies data modeling to decouple data from functions and applications and ensures data consistency across layers. On this layer, the network's digital twin is built to analyze and manipulate the physical network through simulation. In this regard, research should focus on the following technologies:

- Objective-based adaptive decision-making architecture: The traditional function-oriented architecture must evolve toward a multi-objective decision-making architecture to offer system capabilities that are adaptable to complex and unpredictable environments. For instance, channel-, module-, device-, and network-level energy saving requires collaborative optimization algorithms with multiple objectives (including time, space, frequency, and power objectives). These algorithms must take both user throughput

and overall energy saving performance into account. In the future, multi-objective optimization involving network and terminal energy saving needs to be achieved in addition to reassuring multi-user deterministic SLAs.

- Model-driven and data-driven hybrid architecture: The model-driven architecture requires detailed risk analysis and identification of harmful incidents in the design phase. Its advantages include being trustworthy, explainable, and applicable to critical tasks. The first step of the evolution toward ADN is machines using situational awareness and adaptive decision-making capabilities in the data architecture to replace humans in complex and uncertain scenarios. However, this high-performance architecture has suboptimal explainability, relies on the training sample space, and cannot be easily generalized for different NEs or scenarios.

- Semantics-based intent: In ADN, autonomous systems interact with each other through intent-based interfaces in a simplified manner, and differentiated internal implementation processes are shielded from the outside, which enables an out-of-the-box feature. Autonomous systems are decoupled from each other by focusing only on achieving the objectives, regardless of the implementation methods. There are four types of intent: user intent, business intent, service intent, and resource intent.

- Network digital twin: In terms of data awareness, research on high-performance networks should strive for near-zero-error measurement. At the modeling and prediction layer, a high-precision approximate simulation model needs to be constructed for research on how to provide high-performance, SLA-supported simulation that has theoretical guarantee based on network calculus and queuing theory. In terms of control management, the issues of resource allocation and optimization of giant network systems need to be resolved by exploring the theory of fast and slow control structure.

In addition to the advancement of software systems, building Level 4/Level 5 capabilities into ADN also requires that network architecture, protocols, equipment, sites, and deployment solutions be simplified, so as to offset the complexity of network connectivity with a simplified architecture.

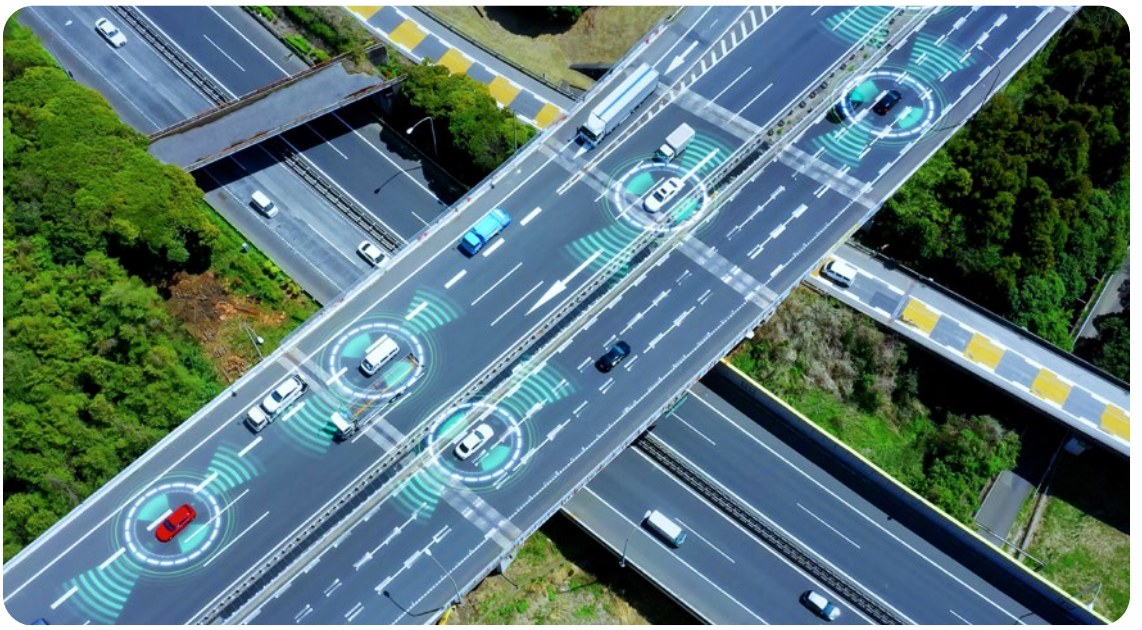
2) AI-Native Edge: Reconstructing the Intelligent Edge with Cloud Native and AI Technologies

Within the architecture of the communications network of 2030, the cloud core network will build an AI-native edge by combining the flexibility and openness afforded by a cloud-native architecture and the service-aware capabilities of AI.

The AI-native edge needs to support AI-based service awareness capabilities. Networks for individual consumers will provide efficient encoding and decoding, optimized transmission, experience assurance, and coordinated scheduling capabilities for full-sensing, holographic communications services. In addition, such networks will provide terminals with computing power offloading for foundation model applications. Private networks for industries can enhance the scheduling framework, and provide service assurance for various industries based on deterministic operating systems. For example, during machine vision processing, the Multi-access Edge Computing (MEC)-based 5G to Business (5GtoB) + AI inference service uses the AI-powered image feature recognition function on the edge to reduce the bandwidth requirements of the backbone network and improve real-time service performance.

The AI-native edge needs to support mesh interconnection and horizontal computing power scheduling. As networks connect to multi-level computing power resource pools, they should be able to sense various resources in order to use computing power efficiently.

To develop computing power awareness, the first thing to do is explore how to measure and model



the computing power requirements of AI services. There are various types of computing chips on a computing power network, such as CPUs, GPUs, application-specific integrated circuits (ASICs), TPUs, and NPUs. The computing power of each type of chips needs to be accurately measured in order to identify the service types to which they can be applied.

Second, computing nodes of a computing power network need to send their computing power resource information, computing power service information, and location information to network nodes. To enable the network to sense multi-dimensional resources and services such as computing power and storage, new computing power routing control and forwarding technologies need to be developed. These could include IPv6 Enhanced-based computing power status advertising, computing power requirement awareness, and computing power routing and forwarding.

Third, in addition to being able to sense computing power, networks should also be able to flexibly adapt to different IoT device scenarios. Huawei predicts that IPv6 adoption must exceed 90% by 2030 to ensure all things that can be connected are connected. It is thus necessary to develop innovative technologies for hierarchical IPv6 address architecture and ultra-large-scale high-speed addressing and forwarding. These technologies should be compatible with both traditional IP networks and lightweight protocols, so as to ensure the global accessibility of data and computing.

As smart home and enterprise advance rapidly, the extension of new capabilities such as sensing, storage, computing, and control based on the FTTR connectivity foundation and AI needs to be continuously studied to further improve users' smart home experiences. For example, AI-based whole-house Wi-Fi optimization, natural language search of photos using FTTR+NAS, family album, voice recognition, and semantic understanding are implemented.

3.2.4 HCS: A New Area Emerging from Communications Technologies

From 1G to 5G, communications and sensing have been independent of each other. For example, a 4G communications system is only responsible for communications, and a radar system is only responsible for functions such as speed measurement, sensing, and imaging. This separation wastes wireless spectrum and hardware resources, and the separation of functions often results in high latency for information processing.

As we approach the 5G-A or 6G era, the communications spectrum will expand to include millimeter wave, terahertz, and visible light. This means the communications spectrum will soon overlap with the spectrum previously reserved for sensing systems. HCS facilitates unified scheduling of communications and sensing resources. Technically, HCS can be broken down into the following three types:

1) Wireless Sensing

HCS is one of the three new use cases proposed for 5G-A, particularly in scenarios such as connected vehicles and drones. With Release 16, precise positioning functionality can already achieve meter-level accuracy for commercial use cases, and future releases are expected to hone this accuracy further, to the centimeter level. As wireless networks move toward higher frequency bands, such as millimeter wave and terahertz, HCS will be applied in areas such as smart cities, weather forecasts, environmental monitoring, and medical imaging.

Wireless HCS technology is still in its infancy and more research is needed to develop foundational theories such as optimal compromise. More research needs to be devoted to electromagnetic wave propagation in complex channel environments; spatial target reflection, scattering, and diffraction modeling; and spatial sparsity sensing modeling. Work needs to be done to improve the performance and energy efficiency of radio frequency (RF) chips and components. There is also a need for further research into extremely large antenna array (ELAA) structure, and efficient distributed cooperative sensing algorithms such as active radar illumination, environmental electromagnetic control, multi-point cooperative transmitting and receiving, target imaging, scene reconstruction, and channel inversion.

2) Wi-Fi Sensing

IEEE 802.11bf defines Wi-Fi sensing standards applicable to indoor, outdoor, in-vehicle, warehouse, and freight yard scenarios, among others. It covers functionalities such as high-precision positioning, posture and gesture recognition, breath detection, emotion recognition, and perimeter security. Moving forward, more research needs to be directed at both the PHY layer (i.e., new signals, waveforms, and sequences) and the MAC layer (e.g., compromise between measurement result feedback and sensing precision for sensing scenarios based on channel state information [CSI] or signal-to-noise ratio [SNR]). Synchronization and coordination between nodes for single-, dual-, and multi-station radar systems is another problem to address. The last issue concerns collaborative sensing across multiple protocols, including 802.11az, 802.11be, and 802.11ay.

3) Optical Sensing

Optical sensing can be divided into fiber-based sensing and laser radar ("lidar") sensing. Fiber-based sensing is more often seen in energy, electricity, government, and transportation sectors where it is used to sense changes in temperature, vibration, and stress to inform fire monitoring and warning, equipment and pipeline fault diagnoses, and environmental and facility stress monitoring. Lidar sensing is more commonly seen in homes and vehicles, providing functions such as spatial environmental sensing, high-precision positioning, and posture or gesture recognition. Currently, fiber-based sensing tends to have a high false alarm rate in complex environments. More research should be directed at reducing the false alarm rate by introducing AI and big data analytics. For lidar sensing, the 3D panoramic modeling algorithm technology needs to be improved to enable multi-radar coordinate system registration based on lidar sensing data.

Huawei predicts that by 2030, 10G Wi-Fi network penetration in enterprises will reach 60%, and F5G private network penetration in medium/large enterprises will reach 42%. In addition, the penetration of 5G private networks in medium/large enterprises will reach 35%. While providing broadband services for enterprises, communications networks will use HCS capabilities to gather static information (e.g., spatial environments, communications blind spots, and obstacles) and dynamic information (e.g., positions, motion tracks, postures, and gestures of people, and the movement of vehicles and objects) to perform data modeling. Coupled with simulation technologies based on the idea of digital twins, the data can help identify and predict changes, empowering numerous industries. HCS represents a new frontier of communications technologies and has huge development potential.

3.2.5 Security and Trustworthiness: A Six-Layer Framework for a New Security Foundation

The networks of the future are on course to be varied, diverse, ubiquitous, and cloud-based, and ToB and ToC services are likely to be converged. These trends will increase network exposure to attacks and further blur traditional network security boundaries. Reactive defense measures such as border isolation and add-on security capabilities will prove insufficient in the face of constantly evolving network attack methods. Therefore, the network security systems of the future must be capable of native, secure, trustworthy, intelligent, and flexible proactive defense.

Security and trustworthiness cover six layers: trustworthiness of components (chips and operating systems), equipment security, connectivity security, management security, federated trustworthiness, and data trustworthiness. Equipment security, connectivity security, and management security fall under network security, while component trustworthiness, data trustworthiness, and federated trustworthiness fall in the trustworthiness realm. The two focus on different aspects but interact in many ways. Ensuring security and trustworthiness requires systemic efforts, involving hierarchical security and trustworthiness technologies such as cross-platform trustworthiness operating systems and chips, endogenous network security, cloud security "brain", multi-intelligent-twin and cross-domain trustworthiness federation, and differential data privacy processing. (Figure 7 Six-layer network security and trustworthiness framework)

1) Component Trustworthiness

Credible data sources are the basis for security and trustworthiness. The Trusted Execution Environment (TEE) at the component (chip and operating system) level is a widely recognized and used solution. Moving forward, chip-level trustworthiness computing technologies will be introduced to network elements (NEs). This will help build a secure and trustworthy running environment for software and hardware based on the underlying NEs, thereby enabling level-by-level verification of chips, operating systems, and applications to ensure data authenticity.

2) Equipment and Connectivity Security

Communications protocols and network equipment can be modified to embed trustworthiness identifiers and password credentials in IPv6 packet headers. Network equipment can verify the authenticity and legitimacy of requests based on

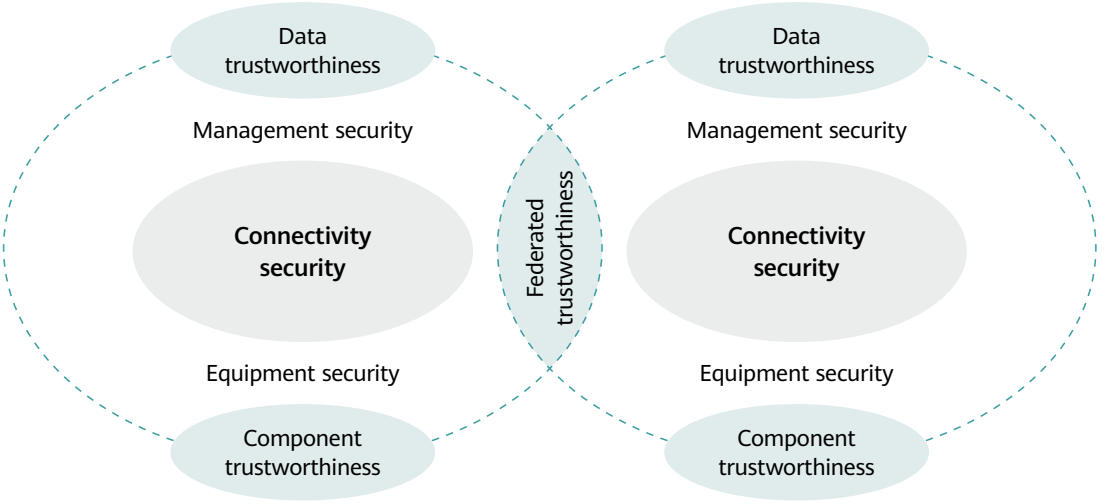


Figure 7 Six-layer network security and trustworthiness framework

identifier authentication, preventing identity theft and spoofing and building fine-grained access authentication and source tracing capabilities.

3) Management Security

First, future networks need to adopt a service-based security architecture that integrates cloud, network, and security, so that security functionalities are provided as components and microservices, and can be centrally orchestrated and agilely deployed. Second, as the user base grows and complexity increases, security policies are growing exponentially to the point where the conventional manual approach to planning and management can no longer keep up. More research is needed on traffic and service self-learning and modeling technologies, model-driven risk prediction and security policy orchestration technologies, and security policy conflict detection and automatic optimization technologies.

4) Federated Trustworthiness

To meet the security and trustworthiness requirements across networks and clouds, blockchain technology will be used to build a trustworthy service system for basic digital resources (including connectivity and computing) for future networks. Distributed accounting, consensus mechanisms, and decentralized key allocation will help ensure the authenticity of resource ownership and mapping relationships and prevent anonymous tampering, illegal hijacking, and other security and trustworthiness issues. The centralized trust model of today's mobile communications network infrastructure results in problems such as excessive permissions of central nodes and single-point authority failures. This kind of infrastructure may pose risks to network security, reliability, and equality, and is not up to the task of serving as a secure and trustworthy foundation for the networks of tomorrow. The infrastructure of next-generation networks must be decentralized, transparent, and auditable, and support trusted identity management.

5) Data Trustworthiness

Networks process user data at user access nodes and service-aware nodes. Therefore, user data passing through the network must be made opaque to the network, so as to ensure user information security. Research should go into technologies that enhance encrypted transmission of user IDs and communications data, as well as pseudonymization and homomorphic encryption technologies that make user information fully invisible to the network.

With the development of quantum technologies, new quantum algorithms that compromise the security of the public key cryptosystem are very likely to emerge with future networks. Quantum computers search for and decompose things much faster than classical computers. A quantum computing breakthrough could render all existing public key cryptographic algorithms useless. Even increasing the parameter length would offer little defense in a post-quantum world. Therefore, networks will need to introduce post-quantum cryptographic algorithms to defend against quantum attacks. Similarly, quantum computers will also reduce the security of symmetric cryptographic algorithms. Symmetric cryptographic algorithms will need to be hardened to support the encryption and decryption of data when throughput and concurrency are high.

3.2.6 Green and Low-carbon

The escalating global energy crisis and climate crisis are driving the demand for low-carbon development around the world.

Several leading operators in Europe have set themselves the clear goal of reducing their carbon emissions by 45% to 55% by 2030, compared with 2020 levels. They have also raised specific requirements for their equipment suppliers to reduce carbon emissions at the organization and product levels. Equipment vendors can meet these requirements and help operators achieve low-carbon goals by moving toward simplified architecture, optoelectronic integration, and system energy saving and multi-dimensional energy efficiency.

1) Simplified Architecture: Low Carbon Realized by Simplifying Foundation, Cloud, and Computing Networks

Traditional networks are divided by technical specialty, resulting in the fragmentation of O&M services. This model is increasingly difficult to adapt to the development of automated and intelligent networks. In the future, networks need to be reconstructed based on the nature of the services they carry, building a simplified three-layer network architecture consisting of foundation, cloud, and computing networks.

The foundation network is used for connectivity at the equipment port level. It uses optical fibers to achieve one hop to computing and provides access (wired/wireless), switching, and core networks from end to end, based on the 100% fiber-to-site optical foundation that supports optical cross-connect (OXC) or ROADM. The foundation network provides high-bandwidth, low-latency, and high-reliability broadband services, and enables green, low-carbon networks with simplified O&M based on all-in-one full-spectrum antennas, fully converged core networks, and simplified protocols. The cloud network is used for connectivity between the cloud and devices at the tenant level, and is overlaid on the foundation network using E2E slicing technology. It enables agile and open virtual networks that provide SLA assurance, and uses a network for multiple purposes to increase network utilization and save network energy.

The computing network is used for connecting data and computing power at the service level and providing computing power routing services and trustworthiness assurance for data processing.

It is constructed based on distributed and open protocols. Through flexible scheduling of data, the computing network enables green, centralized multi-level computing power infrastructure that has a reasonable layout.

The three networks are interdependent. The computing network depends on the cloud network to enable agile building of virtual pipes and open interfaces that can be provisioned on demand, so as to provide real-time, elastic connections between data and computing power. The computing network also needs the support of the foundation network to enable its most important features: low latency and high bandwidth.

2) Optoelectronic Integration: Profoundly Changing the Architecture and Energy Efficiency of Communications Network Equipment

In the communications network industry, optical technologies have traditionally been relatively independent from other specialized technologies such as wireless communications and datacom. However, as networks develop toward higher bitrates, higher frequencies, and greater energy efficiency, traditional electronic technologies will soon encounter sustainable development bottlenecks, such as in distance and power consumption. The solution to this is optoelectronic integration.

In the next decade, the development of new products, such as optical input/output chips and CPO, will improve electronic components' high-speed processing capabilities and reduce their power consumption. Coherent optical technologies will be applied to extend the transmission distance of high-speed ports on datacom equipment. New types of antennas that directly connect to optical fibers will be used to reduce the weight and power consumption of base stations. Microwave communications will be superseded by laser communications to support high-speed data transmission between LEO satellites. To meet the communications requirements of underwater mobile devices, wireless coverage will be replaced by visible light which achieves higher penetration than radio waves. Due to its higher transmittance, far infrared light technology will be used to detect brain waves more accurately. To train ultra-large AI models, optical switching is used to connect switches and servers in data centers, and all-optical cross-connections are used between data centers.

Optoelectronic integration is the way forward for structured improvement of equipment energy efficiency. CPO chips based on optical buses are expected to be in commercial use by 2025. Some academic institutions are researching optical cell switching technology that could potentially replace electrical switching networks. Equipment-level optoelectronic integrated products using optical buses and optical cell switching technology are expected to be developed by 2030. Further into the future, chip-level products that combine optical computing, optical random access memory (RAM) cores, and general-purpose computing cores will also emerge.

Optoelectronic integration technology at the network, equipment, and chip levels can continuously improve the energy efficiency of communications equipment, and meet the green network objective of increasing network capacity without increasing energy consumption.

3) System Energy Saving and Multi-dimensional Energy Efficiency: Building Energy-Efficient High-Performance Networks

Energy efficiency has always been crucial to high-quality network construction. As networks evolve, service scenarios are becoming more complex and service types are diversifying, necessitating scenario-specific energy saving strategies and differentiated service assurance. Meanwhile, networks and equipment systems are also becoming increasingly complex, driving the need for a highly energy-efficient E2E system that covers everything from software to hardware, from main

equipment to auxiliary devices, and from single-point optimization to overall optimization. Such a system would be expected to improve energy efficiency from three perspectives: energy flow, service flow, and control flow.

Energy flow: By reducing energy losses throughout the energy chain from supply to transmission to use, energy efficiency can be significantly improved. Energy supply efficiency can be improved by optimizing power supply architecture (such as adopting a modular power supply architecture) and simplifying the topology. Line transmission losses can be reduced by decreasing the types of voltages and the number of voltage conversion classes. Energy utilization efficiency of loads can be boosted through partitioned power supply and dynamic shutdown. In addition to iteratively optimizing the conversion efficiency of each individual link, the overall conversion efficiency of the energy chain can be maximized through cross-link collaboration.

Service flow: Resources will need to be accurately allocated on demand and elastically scaled in real time based on the mapping between services, resources, and energy consumption. This is essential to meet the experience assurance requirements of a diverse range of services and adapt to the dynamic changes of services in time and space. In addition, the depth, speed, precision, and flexibility of equipment shutdown will need to be improved to approach the goal of "0 Bit 0 Watt".

Control flow: A complete energy efficiency optimization, control, and evaluation mechanism is required for dealing with dynamically-changing complex systems. By implementing hierarchical autonomy at the network, site, equipment, and chip levels, we can improve the energy efficiency of each level. We can also maximize the energy efficiency of an entire system through cross-layer collaboration such as software-hardware synergy and device-pipe-chip synergy, as well as vertical integration. Any evaluation of efficiency needs to take into account both service volume (such as traffic and coverage area) and service quality (such as user-perceived speed, latency, and reliability). To strike a balance among service volume, service quality, and energy consumption, we need to develop intelligent multi-objective comprehensive optimization capabilities at the network and equipment levels. These capabilities are important for building energy-efficient high-performance networks.

3.3 Summary and Technology Outlook

By 2030, we will be living in a multi-network and multi-cloud world. Billions of people and hundreds of billions of things will be connected to an intelligent world of hyperreal experiences where multiple clouds coexist, including public, industry, and telecom clouds. Connections will be supported by cubic networks consisting of 10-gigabit personalized home networks, 10-gigabit industrial campus networks, 10-gigabit individual networks, and global satellite networks.

In future communications networks, energy efficiency will be continuously improved through optoelectronic integration at the network, equipment, and chip levels in the foundation network. The cloud network will use E2E virtual slicing to connect the breakpoints of specialized networks on top of the foundation network, so as to provide differentiated capabilities guaranteed by SLAs for different tenants. The computing network will provide high dynamic connectivity between data and computing power through innovation in IP network protocols, meeting the requirements of intelligent services. Green, low-carbon networks will be enabled by a three-layer simplified network architecture and three-layer optoelectronic integration.

Future communications networks will support deterministic service experiences critical to the intelligent transformation of industries. Users will be connected to multi-level computing resources: 1 ms latency will be guaranteed for data transmission within cities, 5 ms latency within city clusters, and 20 ms latency through backbone networks. The networks will also provide greater than 99.999% availability, and develop secure, trustworthy network capabilities to support the migration of all systems to the cloud across industries.

Future communications networks will support AI-native. By combining NE status data with AI and innovating in algorithms, the networks will approach the theoretical limit and turn non-determinism into determinism, improving network performance. With the combination of network O&M data and AI, big data analytics, and closed-loop optimization, the networks' automation and all-scenario service capabilities will be comprehensively improved. With the AI-native edge, the networks will also be able to sense diversified service requirements of various industries, thereby improving service experiences.

Future communications networks will support HCS. Wireless, optical, and other multimodal sensing technologies will allow networks to collect environmental data and combine it with digital twin technology to provide industries with the brand-new service capabilities enabled by HCS.

Over 20 years ago, IP technology started reshaping the forwarding architecture of communications networks. Over 10 years ago, cloud technology began to profoundly influence the network management control architecture. Over the next 10 years, AI will be embedded into all layers of the network architecture, driving the networks to evolve toward advanced intelligent twins. To support the development of intelligent networks in the future, networks' computing capabilities will be enhanced, and optoelectronic integration will be adopted to enable green, low-carbon communications networks.

In conclusion, the architecture of the communications network of 2030 will evolve towards cubic broadband networks, deterministic experience, AI-native, HCS, security and trustworthiness, and green and low-carbon networks.

04

4 Recommendations

William Gibson, famous science fiction novelist and author of Neuromancer, once said, "The future is already here. It's just not evenly distributed yet." AR, the key technology for integrating the virtual and real worlds, was invented by the Royal Navy 60 years ago, and used for the sighting devices of fighter aircraft. Later, MIT established in the 1980s the Media Lab, which is dedicated to changing the way humans interact with computers and delivering personalized digital experiences.

Communications technology and computing technology share the same origin. Less than five years after IBM launched its first personal computer in 1981, the world's first router was invented. Compared with computers, the main distinguishing features of communications equipment are enhanced optical and wireless functions, and network protocol interfaces.

Cloud, AI, and optical, the three key technologies influencing the development of future communications networks, are also reshaping the computing industry. While we may be more familiar with cloud and AI, optical technologies have also been profoundly influencing the computing industry over the past decade. Currently, the industry is focusing on two research areas of optical computing. One is replacing electronic components with optical components to develop optoelectronic integrated computers. The other is using optical parallel processing to build an optical neural network which will increase computing power by 100 times while consuming very little power. The

application of optical technologies in computing will also play a part in realizing a green, low-carbon network architecture.

Currently, we cannot find an accurate keyword to represent the target network. 6G/F6G/Net6G may be the keyword based on the improvement of network capabilities from ubiquitous gigabit networks to 10-gigabit cubic networks (5G-A/F5G-A/Net5.5G). Industrial Internet may be the keyword based on the shift of network application scenarios from consumer Internet to industrial Internet. At the same time, computing power network may be the keyword based on the shift in the nature of services from human-oriented cognition to machine-oriented cognition that supports massive amounts of user data and multi-level computing power services.

In addition, optical network may be the keyword based on the evolution of the underlying technology from electronic technologies to optical technologies. The cognitive network or digital twin network may be the keyword based on the improvement of network intelligence from L3 to L4+ ADN.

The next decade in communications networks will open up huge space for imagination while also bringing an abundance of uncertainties. All players in the industry need to work together to explore new technology directions and jointly make the vision for the communications network of 2030 a reality.

Appendix A: Acronyms and Abbreviations

Abbreviation/Acronym	Full Spelling
3GPP	3rd Generation Partnership Project
5G NR	5G New Radio
5G SA	5G Standalone
5GtoB	5G to Business
ADS	Advanced Driving System
ADSL	Asymmetric Digital Subscriber Line
AI	Artificial Intelligence
AMR	Automated Mobile Robot
ADN	Autonomous Driving Network
AN	Autonomous Networks
API	Application Programming Interface
AR	Augmented Reality
B2B	Business to Business
CAICT	China Academy of Information and Communications Technology
CCSA	China Communications Standards Association
CPO	Co-Packaged Optics
CPU	Central Processing Unit
CRUD	Create, Read, Update, Delete
CSI/SNR	Channel State Information/Signal-to-Noise Ratio
DCNN	Deep Convolutional Neural Network
DetNet	Deterministic Networking
E2E	End to End
ERP	Enterprise Resource Planning
ELAA	Extremely Large Antenna Array

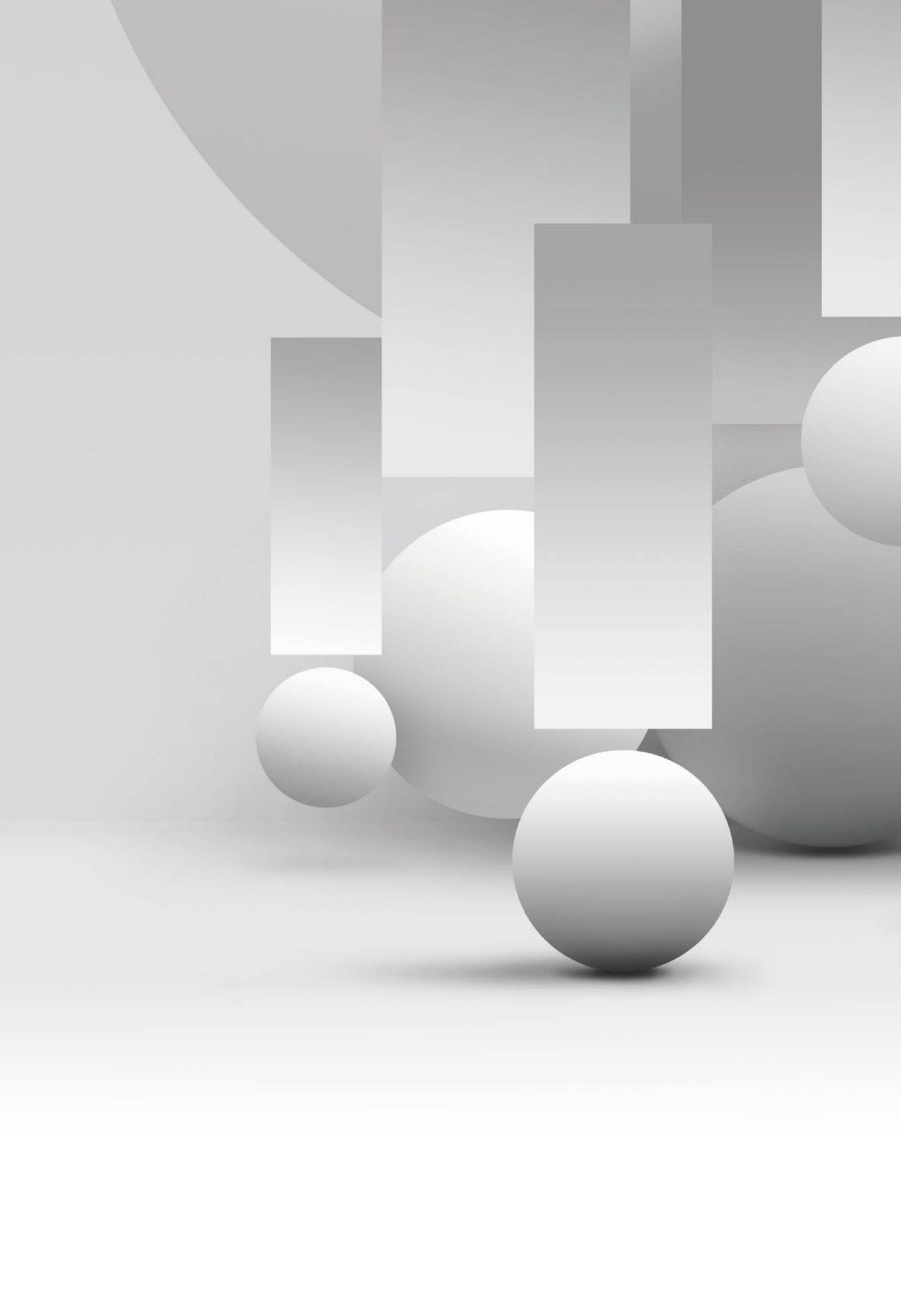
Abbreviation/Acronym	Full Spelling
ELAA-MM	Extremely Large Antenna Array-Massive MIMO
F5G	5th Generation Fixed Network
FDMA	Frequency Division Multiple Access
fgOTN	Fine-grain OTN
FlexE	Flexible Ethernet
FOV	Field Of View
FPS	Frames Per Second
FR1/FR2	Frequency Range_1/Frequency Range_2
FSD	Full Self-Driving
FTTR	Fiber To The Room
FSA	Flexible Spectrum Access
GenAI	Generative AI
GPU	Graphical Processing Unit
GSMA	GSM Association
HCS	Harmonized Communication and Sensing
HBF	Hybrid Beamforming
IMT	International Mobile Telecommunications
IoT	Internet of Things
ISA-95	International Society of Automation 95
ISAC	Integrated Sensing and Communication
ITU-T	International Telecommunication Union-Telecommunication Standardization Sector
LEO	Low-Earth Orbit
MAC	Media Access Control
Massive MIMO	Massive Multiple-Input Multiple-Output
MEC	Multi-access Edge Computing
MES	Manufacturing Execution System

Abbreviation/Acronym	Full Spelling
MIIT	Ministry of Industry and Information Technology
MR	Mixed Reality
MTP	Motion-to-Photon
MBSC	Multi-Band Serving Cell
NPU	Neural Processing Unit
NSB	US National Science Board
NSMF	Network Slice Management Function
NSSMF	Network Slice Subnet Management Function
NTN	Non-terrestrial Network
ONT	Optical Network Terminal
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
PLC	Programmable Logic Controller
PON	Passive Optical Network
PPD	Pixel Per Degree
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
OXC	Optical Cross-Connect
O2I	Outdoor-to-Indoor
RAM	Random Access Memory
RB	Resource Block
RLC	Radio Link Control
ROADM	Reconfigurable Optical Add/Drop Multiplexer

Abbreviation/Acronym	Full Spelling
RTT	Round-Trip Time
RF	Radio Frequency
SCADA	Supervisory Control And Data Acquisition
SDM	Spatial Division Multiplexing
SLA	Service Level Agreement
SLM	Spatial Light Modulator
SDL	Supplemental Downlink
SUL	Super Uplink
TDM	Time Division Multiplexing
TSN	Time Sensitive Networking
TTM	Time to Market
URLLC	Ultra-Reliable Low-Latency Communication
U6G	Upper 6 GHz
VR	Virtual Reality
WDM	Wavelength Division Multiplexing
WAN	Wide Area Network
Wi-Fi 6	Wireless Fidelity 6
WLAN	Wireless Local Area Network
XR	eXtended Reality

Appendix B: Notes on Version Updates in 2024

Huawei makes continuous efforts to explore the intelligent world through in-depth exchanges with renowned scholars, valued customers, and key partners in the industry. The intelligent world is rapidly evolving, with new technologies and scenarios emerging faster than ever, leading to dramatic changes in industry-related parameters. In response, Huawei systematically updated its *Communications Network 2030* released in 2021 to envision the scenarios and trends in 2030 and adjust related forecast data.



HUAWEI

Version 2024

Data Storage 2030

Building a Fully Connected, Intelligent World

Foreword

Human history is a story of data storage and transmission, from the first oracle bone inscriptions 3,500 years ago, to the advent of papermaking 2,100 years ago, and the emergence of digital storage just more than 60 years ago. We have never stopped in our quest to make knowledge more accessible and data storage more efficient.

In the next decade, the rapid development of 5G/6G, AI, big data, and cloud computing give us the potential to produce yottabytes of data every year, as innovative storage technologies usher in a new era of civilization. The data-centered infrastructure which is efficient, green, and secure will drive us to better understand and explore the smarter world.

01

Future Data Storage Scenarios

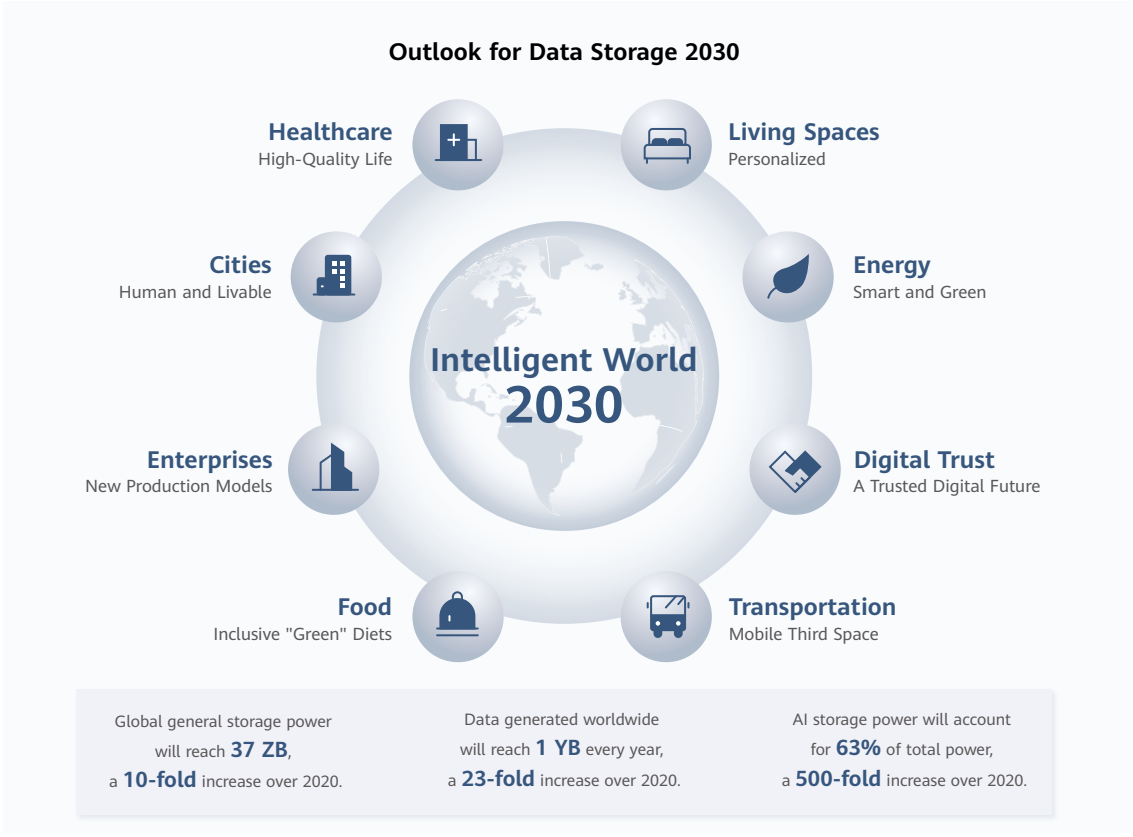


Figure 1-1 Future Data Storage Scenarios

A decade ago, humanity generated just a few zettabytes of data every year, and mobile Internet, cloud computing, and big data were still in their infancy. Today, these technologies are profoundly changing our world, and new technologies such as AI, blockchain, 5G/6G, AR/VR, and the metaverse are driving human society into a new stage of an intelligent world.

By 2030, we will be producing yottabytes[1] of data every year. Compared to 2020, annual data growth will be twenty-three times higher, general storage power in use will be ten times higher, and AI storage power will have increased by a factor of five hundred[2]. The digital and physical worlds will be seamlessly converged, allowing people and machines to interact perceptually and emotionally. AI will be ubiquitous and will serve as scientists' microscopes and telescopes, enhancing our understanding of everything from the tiniest quarks to vast cosmological phenomena. Industries which already make extensive use of digital technology will use AI to become more intelligent.

In the next decade, digital technologies will help us move to an intelligent world – a process of the same epochal significance as the age of discovery, the industrial revolution, and the space age.

1.1 Digital Technologies Move Human Society from Informatization to Digitalization

Healthcare: Digitalizing Health, Improving Quality of Life

Over the past decade, the health of humanity as a whole has improved markedly. According to the World Health Organization's (WHO) World Health Statistics 2021, global life expectancy at birth has increased from 66.8 years in 2000 to 73.3 years in 2019. The pace of population aging is accelerating worldwide. Projections indicate that 16.5% of the global population will be 60 years old or over by 2030. This is expected to drive a surge in demand for healthcare services[3]. According to the WHO's 2019 findings, spending on health is growing faster than the rest of the global economy, accounting for 10% of global gross domestic product (GDP). The WHO also predicts that by 2030, there will be a global shortfall of 5.7 million nurses and 10 million health workers in total. At the same time, we are seeing wide disparities in the global distribution of medical resources – disparities that become especially clear when viewed in terms of population growth.

Looking to the future, new methods of reducing healthcare costs, diversifying healthcare resources and services, and creating new prevention and treatment methods are desperately needed to increase quality of life and make medical treatment more accessible and affordable for all. Many innovative solutions are emerging that may find application within the next ten years. Real-time health monitoring and health data modeling can help weave disease prevention into the fabric of our daily lives. This shifts the paradigm governing our healthcare system from treatment to prevention and covers the following scenarios:

Building a knowledge graph for better health management

The growing popularity of wearables and portable monitoring devices combined with advances in technologies such as the Internet, IoT, and AI will make personal health data modeling a realistic prospect in the near future[4]. User-specific knowledge graphs can be built based on data, including health indicators, medical diagnoses, and treatment results. They can compare and analyze these knowledge graphs to formulate personalized health solutions. We can take intervention measures which include guidance on nutrition, exercise, and sleep, as well as mental health support to incrementally improve our lifestyles. For example, a digital health company built a knowledge graph to examine relationships between diet and disease. The company used the knowledge graph to help individuals improve their sleep quality and manage their weight. The health management survey conducted by the company showed that the participants recorded an average 35 minutes more sleep daily, and a total body weight roughly 1.5 kg lighter across the year, which translated into lower probability of disease.

Making infectious disease prediction more accurate

New digital technologies, such as natural language processing, can also broaden the amount of data that can be used for epidemiological management. These technologies allow public health institutions to collect and analyze news articles, reports, and search engine indexes to track major public health events around the world. These institutions can extract valid information from the collected data, build new scientific models, and conduct intelligent analyses of the data so that they can respond to incidents faster and more effectively. For example, a technology company has used natural language processing and machine learning to gather data from hundreds of thousands of public sources, including statements from official public health organizations, digital media outlets, global airline ticketing agencies, as well as livestock health reports and population demographics, to analyze the spread of disease 24 hours a day.

More accurate drug trials shifting treatment from "one-size-fits-all" to "bespoke"

AI can help doctors develop personalized treatment plans by analyzing thousands of pathology reports and treatment plans, and determining which would be most appropriate for each patient. One research institute in Singapore has even created an AI-powered pharmaceutical platform that optimizes medication dosages. The platform can quickly analyze a patient's clinical data, provide the patient with a recommended drug dose or combination regimen based on their specific condition, and revise tumor sizes or biomarkers levels based on available data. In addition, doctors can use the data to determine new courses of treatment for patients.

Achieving safe & precise identification of cancer cells with AI

Precision medicine can help the fight against cancer. During traditional radiation therapy, the radiation typically also kills a large number of healthy cells since the targeted area is quite broad. With the help of AI technology, adaptive radiation therapy (ART) systems can automatically identify changes in lesion positioning and more accurately outline the target areas for radiation treatment. This helps focus the radiation on just the cancer cells and reduces damage to healthy tissue. AI is already enabling accurate identification and automatic contouring of target areas for various types of medical imaging, including CT, ultrasound, and MRI. A contouring workload that would once have taken hours can now be completed in less than a minute, and the damage caused by radiation therapy to healthy tissue can be reduced by 30%.

By 2030, we will be able to track our own physical indicators in real time with sensitive biosensor technologies and intelligent devices. We will also be able to build health knowledge graphs to manage our health independently, reducing the reliance on doctors.

Driven by ICT technologies, portable medical devices powered by advanced software and hardware, cloud-edge-device computing, and stable networks will be available in grassroots-level hospitals, communities, and households. These devices will collect medical data in real time and upload the data to the cloud for processing. Thanks to the big data knowledge base and AI scheduling, users will be able to access coordinated telemedicine services and track their health from the comfort of home. Building knowledge graphs on the cloud requires the large-scale deployment of storage power.

Huawei predicts that by 2030, the global general storage power capacity will reach **37 ZB**, a **10-fold** increase over 2020. AI storage capacity will account for **63%** of total capacity, a **500-fold** increase over 2020.

Food: Data-driven Food Production for More Bountiful, Inclusive, and "Green" Diets

Food is a necessity for all, so the UN has made "Zero Hunger" one of its Sustainable Development Goals (SDGs) for 2030[5]. Current estimates show that nearly 690 million people are hungry, and if recent trends continue, the number of people affected by hunger would surpass 840 million by 2030. The agriculture workforce is shrinking: According to the International Labor Organization, the proportion of the world population working in agriculture has dropped from 43.699% in 1991 to 26.757% in 2019. Arable land per capita is decreasing: According to World Bank data, arable land per capita has fallen from 0.323 hectares to 0.184 hectares from 1968 to 2018 – a drop of 43%. Overuse of pesticides is causing severe soil pollution: According to statistics, 64% of global agricultural land (approximately 24.5 million square kilometers) is at risk of pesticide pollution, and 31% is at high risk. Simultaneously, the focus of people's diets around the world is shifting to more nutrition and food safety standards. For example, 13,316 food products in China received some kind of green certification in 2018. This number increased to 14,699 in 2019, up 10.4% YoY. This higher demand for green-certified products results in higher requirements on agricultural conditions and technologies.

As we move towards 2030, technology and data are key to empowering agriculture, helping it overcome traditional growth constraints, increasing food production across the board, and bringing "green" food to every table around the world. Digitalized agriculture will cover the following scenarios:

Using accurate data, not experience, to guide cultivation

As the saying goes, there is "a time to plant, and a time to pluck up that which is planted." Decisions on when to sow, fertilize, and use pesticides are highly informed by personal experience. However, this leaves a lot of room for uncertainty, and whether any given year yields a good harvest is still ultimately up to fate.

ICT technologies bolster agriculture with accurate data based on analysis of soil moisture, ambient temperature, crop conditions, terrain, climate models, and pests. Precise controls are provided for optimally paired soil and crops. With maize, for instance, data-powered adaptive sowing can increase crop yield by 300 to 600 kilograms per hectare of land.

Taking a digitalized, factory-like approach to protect agricultural production from environmental conditions

One typical example of industrialized agriculture is indoor, vertical farms that use data to build standardized growth environments without needing to consider geographical constraints. In vertical farms, farmers are able to artificially create ideal environments for their crops by precisely controlling light, temperature, water, and nutrient delivery based on the needs of each crop at every step of the process. Vertical farms don't require pesticides or soil, and reduce agricultural water waste. They are not affected by climate, providing consistent and ideal conditions for fresh produce. These smart agricultural models are globally replicable, as the ICT control system and data model used in one vertical farm can be used anywhere else to achieve almost identical results. Recent pilot programs for vertical farms have found that, if harvested every 16 days, a 7,000 square meters area can yield a staggering 900,000 kilograms of vegetables every year.

By 2030, ICT technology will enable us to connect key agricultural production factors, such as farmland, farm tools, and crops, and collect and utilize data on attributes like climate, soil, and crops. New plant modes like vertical farms and precise data analysis will lay the way for yield-boosting agricultural processes.

Huawei predicts that by 2030, the data generated worldwide will reach **1 YB** every year, a **23-fold** increase over 2020. With the wider application of data in agriculture, we will build a more resilient and green food system. The total amount of data generated by global agriculture will reach **4 ZB** each year, a 23-fold increase over the amount in 2020.

Living Spaces: Whole-House Intelligence Enables Personalized Spaces

As demand for personalized home experiences continues to rise, ICT-enabled smart home technology is gaining popularity. A survey found that about 80% of millennials and 69.2% of baby boomers are interested in smart home technologies[6]. In the UK, 80% of consumers are now aware of smart home technology and it is second only to mobile payments in consumer awareness of a basket of tech trends. Interoperability has risen as one of the most important buying considerations. Interest in smart living spaces that offer enhanced convenience and safety is also on the rise.

Digitalization and data enable future home experiences, of which scenarios are as follows:

Digital cataloguing and automated delivery for offsite storage

New communities will deliver comprehensive services to residents, powered by the Internet of Things (IoT), Gigabit fiber networks, and other new advanced infrastructure. Services such as virtual community events and smart pet management will bring residents and their communities more closely together. Groundbreaking new design concepts will also start changing the way our homes look at the household level.

One potential solution to the overwhelming number of possessions that now fill households is offsite storage. Some proposed solutions include digitalization and cataloguing of all household items, with technologies like 3D scanning, and then storage in local shared warehouses. This would mean when you decide to go to a party, you can flick through a 3D hologram menu to pick out the dress and accessories that you want, and, at the touch of a button, have those items delivered to your door, either by robot or through the building's internal delivery system.

Whole-house intelligence that understands usage and creates intuitive experiences

Smart home systems collect data from a wide range of smart home appliances and sensors, over highly-reliable, high-speed networks that reach every corner of your home. They use AI engines to determine what is happening and run appropriate applications. The AI engines, in turn, make informed decisions on how to configure your home appliances, which could be taken independently or in collaboration with other systems, to meet your needs in real time. When implemented properly, smart home systems deliver immersive, personalized, and intelligent experiences that evolve as your usage needs change. The variety of smart home appliances we will see in the coming years is expected to explode. They will work together to intelligently anticipate and meet your needs in different situations. For example, a sleep support solution could easily be created for the bedroom by designing a system that automatically adjusts the softness of your mattress and pillow to suit your body and sleeping habits, and changes the bedroom lighting to stimulate the production of melatonin – the hormone that helps you fall asleep. Bedroom speakers could play music to relax you, and air conditioners could keep track of temperature, humidity, and oxygen levels.

In 2030, your home may be full of smart appliances that bring a new level of interactivity to your lifestyle and entertainment. The building you live in may be supported by a great variety of smart control systems, and smart functions may be more widely available in your local community. However, none of this will be possible without connections that deliver high bandwidth and extremely low latency.

Huawei predicts that by 2030, 25% of homes will have access to 10 gigabit fiber broadband. The number of global smart home households will soar to 1.8 billion and the annual data volume will be a staggering **23 ZB**.

Transportation:
Smart, Low-Carbon Transport Powered by Data Opens up the Mobile Third Space

Travel by private cars is an important part of modern life. In 2020, the vehicle-miles traveled across the US totaled 2.83 trillion miles. In Europe, vehicles travel more than 12,000 kilometers a year on average. Transport systems now face many challenges: Traffic jams are becoming increasingly common and transportation accounts for 26% of global carbon emissions[7].

All of the key elements (vehicles, traffic lights, pedestrians, etc.) need to be connected using ICT technologies and are infused with big data to support decision-making, so that each phase of a journey can be more intelligent and less carbon-intensive. Digitalized transportation will cover the following scenarios:

Self-driving vehicles in the fast lane

Self-driving vehicles are achieving higher levels of automation, from L2 and L3 to L4 and L5. Buses, taxis, low-speed logistics, and industrial transport (logistics and mining) will be the first commercial applications of autonomous driving.

Low-speed public roads: Self-driving vehicles have delivered positive results in fields such as logistics and distribution, cleaning and disinfection, and patrolling. Unmanned vehicles for logistics and distribution can successfully drive at low speeds on roads with less complicated conditions. This means they can provide safe unmanned delivery services on public roads. Low-speed unmanned vehicles have provided valuable support during the fight against COVID-19, especially in the transportation and distribution of medical supplies, cleaning and disinfection, patrolling, and checking temperatures.

High-speed semi-closed roads: Heavy trucks drivers are expensive, and they frequently breach rules by overloading their vehicles and working overtime. So autonomous driving of heavy trucks would quickly help industries cut costs and work more efficiently, making this a compelling business case. According to a Deloitte report on smart logistics in China, technologies like unmanned trucks and artificial intelligence will mature in a decade or so, and will be widely used in warehousing, transportation, distribution, and last mile delivery.

Special non-public roads: Autonomous driving demonstrates its commercial value with high levels of safety and efficiency in environments like mines and ports. While working autonomously, many mechanical vehicles, such as mining trucks, excavators, and bulldozers can work together. In the event of a fault or danger, the commander can remotely pilot the vehicle to a safe area from the control center. At the Yangshan Port in Shanghai, 5G-powered L4 smart-driving heavy trucks can drive at speeds of up to 80 km/h, and the distance between vehicles can be shortened to 15 meters. Thanks to the centimeter-level precision of the BeiDou GPS system, the vehicles can come to a stop within 15 seconds of a command, with an error of only 3 centimeters. The use of autonomous vehicles has brought a 10% improvement in vessel loading/unloading times.

Public roads: Robotaxis are an obvious business model for self-driving companies. According to one study, robotaxis could replace 63% of carshare and taxis and 27% of public transport. Autonomous driving technologies will lead to more innovative changes. Cars can become the mobile third space, catering to many different scenarios. This will disrupt the business models of existing industries. Self-driving food trucks may become the standard of the future, and dinner with friends and family may take on a whole new form: After you book a lunch, a self-driving food truck will pick you up and carry you along whatever scenic route you choose. You can enjoy the views while dining and chatting, all within a private space. This model would eliminate the need to visit restaurants and ensure privacy during the meal.

Urban air mobility

In the future, airspace will become an important resource for urban transportation. An efficient air-based urban transportation network will greatly free up roads, reduce travel times, and improve the efficiency of logistics and emergency services.

Air emergency rescue systems: Over the past decade (2010–2020), skyscrapers have sprung up in many cities, creating new safety hazards. Firefighting and emergency medical services in skyscrapers will be a new challenge for cities. Air emergency rescue offers a new solution to these challenges, allowing firefighters and medical personnel to better protect lives and property by quickly reaching higher floors to put out fires or assist people.

Air metro/air taxis: Convenient and efficient transportation is one of the core needs of urban residents. eVTOL will prove to be an effective tool to improve the urban transport experience. Four-seat aircraft of multiple companies are now capable of reaching cruise mileage of about 100 km. Pilot projects have begun for air passenger transport services. In 2019, a Chinese company launched the world's first urban air mobility (UAM) service in Zhejiang, cutting road trips that normally took 40 minutes to a five-minute air hop.

UAM scenarios require a fast and stable space-air-ground integrated network and positioning system, cost-effective and reliable visual sensors and lidar, secure and stable automatic flight algorithms, and an efficient, real-time command and dispatch platform.

Future transport will be a multi-dimensional innovative system. The shift to electric, autonomous, shared, and connected vehicles will create an intelligent, convenient, low-carbon transport experience to reshape the transport experience, create innovative mobility services, enable more efficient sharing of transport resources, alleviate traffic congestion, and reduce the environmental pollution caused by traffic. This is how we will resolve the conflict between the surging demand for transport and the urgent need to decarbonize.

Huawei predicts that by 2030, electric vehicles will account for 82% of all global vehicles sold, and the penetration rate of new self-driving vehicles in China will reach 30%. The vehicle-level storage capacity will exceed **500 PB.**

Cities: New Digital Infrastructure Makes Cities More Human and Livable

Rapid advances in new technologies, such as 5G, cloud, AI, blockchain, and intelligent sensing are opening up more possibilities for smart cities, which represent the best places to create new applications for these technologies. In 2020, nearly 1,000 exploratory smart city projects were underway worldwide. In 2020, this spending totaled nearly US$124 billion, an increase of 18.9% over 2019[8].

Advancing digital transformation has become one of the key pathways to sustainable development for the world's leading cities. Digitalization and data support the following future scenarios:

Nanosensors track the pulse of the city

Digital cities depend on data, which comes from a wide array of sensors scattered throughout the city. While there are various types of sensors in use currently, one particularly cost-effective and disruptive technology — nanosensors — is expected to drive the next revolution. As such, the MIT Technology Review listed Sensing City as one of the 10 Breakthrough Technologies 2018.

Graphene gas nanosensors are ultra-sensitive to odors. An American university has created a novel type of nano-coating using graphene, and when the coating is applied as a nanofilm on gas

sensors, it delivers a 100-fold increase in molecular response compared to the best available sensors that use carbon-based materials. In the future, these sensors will be able to accurately identify hazardous, toxic, or explosive gases in the air, thereby greatly enhancing cities' capability to detect dangerous substances.

Nanocrack-based acoustic sensors are able to recognize specific frequencies of sound, which perform far better than conventional microphones at separating out sounds in a given frequency range. For example, when these nanosensors are placed on the surface of a violin, they can accurately record every note of a tune and "translate" it so that a connected device can accurately recreate an electronic version of the tune. When this kind of sensor is worn on the wrist, it can accurately monitor a person's heartbeat. Breakthroughs in this technology will greatly enhance acoustic monitoring in urban infrastructure.

All-optical information exchange, 10 gigabit interconnection

The digital transformation of cities poses challenges to massive flows of information, however, 10 gigabit interconnected all-optical cities can unleash tremendous value and growth potential. In April 2021, Shanghai became the world's first all-optical smart city. Owing to its F5G optical network, the city is able to deliver stable connections with latency below 1 millisecond anywhere in the urban area. The deployment of this high-speed optical network has laid a solid foundation for Shanghai's future digital transformation. The future architecture of an all-optical city will consist of four parts:

All-optical network access: All network connections will be optical, including homes, commercial buildings, enterprises, and 5G base stations. The all-optical transmission network will be extended into edge environments like large enterprises, commercial buildings, and 5G base stations. This will enable the digital transformation of many different industries, and support the development of F5G+X and 5G2B industrial applications.

All-optical anchors: Connections originating in home broadband, enterprise broadband, 5G networks, or data centers will be routed and transmitted through all-optical networks; optical networks will support multiple different fixed access technologies and provide one-hop connections to the cloud.

All-optical switching: One-hop access to services through urban optical networks. All-optical cross-connect technologies are used to build multi-layer optical networks that support one-hop access to services; high-speed inter-cloud transmission; and high synergy between cloud and optical networks.

Fully automated O&M: Real-time sensing of network status with proactive, preventive O&M. This supports elastic network resources, and automated service provisioning, resource allocation, and O&M.

Intelligent hubs cut out the human factor from urban management

With the breaking down of data barriers, AI evolves from partial intelligence to all-scenario intelligence, creating a new public governance subject. Future cities need a powerful smart central platform that aggregates massive data from all corners of the city, on the other hand, the platform transforms data into an advanced city governance capability, benefiting thousands of industries and greatly improving city governance efficiency and user service experience. Early-stage explorations by Toyota: In Toyota's plan for the city of the future, each house, building, and vehicle will be equipped with sensors. Data from these sensors will then be aggregated into a city's intelligent hub. Then AI will be used to analyze people's surroundings and then guarantee the safety of pedestrians and drivers by keeping them separated.

Proactive, precise provision of government services

Machine recognition technology makes contactless services possible. Today, in most of China's developed provinces, citizens do not need to go to government offices to access government services. They can now access them directly through their smartphones. Over the next decade, the digitalization of government services will be taken to the next level.

In the future, as governments aggregate more massive data and their AI technologies mature, they will be able to deliver government services in a more precise and proactive way; manage their municipalities more efficiently; and improve their service experience. Let's look at smart care for the elderly as an example. Communities in Shanghai have installed smart water meters for elderly people who live alone and agree to the installation. If the total water used within a 12-hour period falls below 0.01 cubic meters, the meter will send an alarm to the central network and community workers. These workers will then visit the elderly person in question to check whether everything is normal. Such attentive care demonstrates a real human touch and concern for the elderly.

ICT technologies like 5G, optical networks, AI, cloud, blockchain, and intelligent sensing will all be rolled out rapidly over the next decade. Cities will soon welcome a period of 10 gigabit connectivity, with 10 gigabit wireless services becoming available to organizations, homes, and individuals.

The application of these ICT technologies in cities will significantly boost our ability to make use of limited resources, to manage our cities efficiently, and to give citizens a positive experience. ICT technologies will help cities achieve their sustainable development goals, and make our cities more human and livable.

Huawei predicts that by 2030, cities contribute to **96%** of the data, and **42%** of the data emanates from monitoring, scheduling, and management of infrastructure linked to cities.

Enterprises:
Digital Factory Reshapes Production Models and Enhances Enterprise Resilience

Over the next decade, population ageing will lead to a huge worldwide labor shortage. According to a report published by the UN, the global population aged 65 and over is projected to exceed 12% of the total population by 2030, while the global population aged under 25 will decrease from 41% in 2020 to 39% in 2030. By 2030, we can expect a deficit of 85.2 million workers around the world. Take manufacturing for example. By 2030, this sector is estimated to face a global labor shortage of 7.9 million workers, leading to an unrealized output of US$607.14 billion[9].

Consumer demand is set to become much more diverse, which will profoundly change production models, forcing businesses to innovate. For example, as the "singles economy" gains traction, companies can rapidly adjust their products by targeting solo dining, small home appliances, and mini karaoke booths. In addition, companies need to take the initiative and stimulate demand through emotional appeal, and rapidly produce combinatorial designs for product appearance, images, and implications. For example, they can customize limited-edition products or launch co-branded products within the shortest time possible.

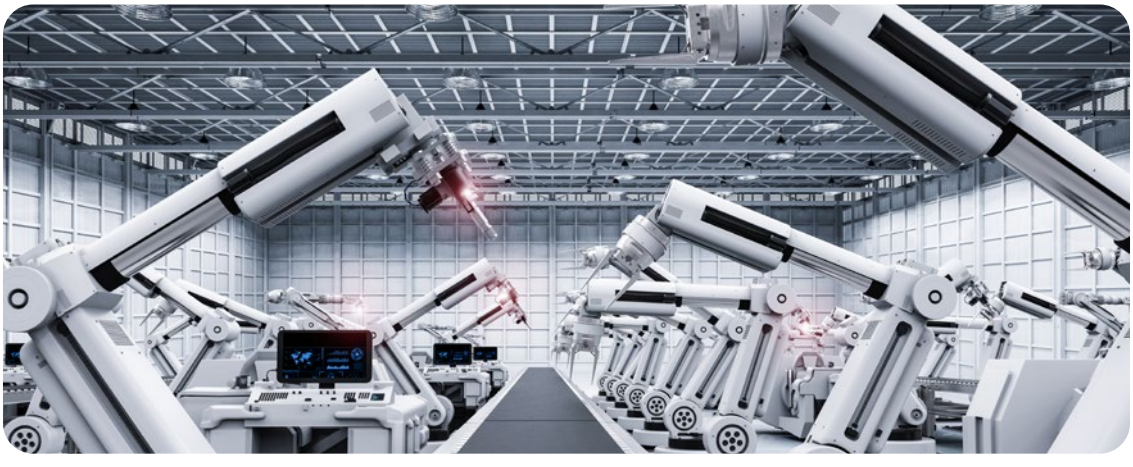
What's more, black swan events pose new challenges to production continuity. Due to the pandemic, it is estimated that global GDP suffered US$3.94 trillion in lost economic output in 2020. The top risk to enterprise growth is supply chain disruptions. Therefore, determining how to use and protect data, reshape production modes, and enhance supply chain resilience are now vital challenges that companies must take very seriously.

Collaborative robots

More and more enterprises are subject to labor shortages, which requires new forms of productivity come in. Collaborative robots are a type of industrial robot. They were initially designed to meet the customized and flexible manufacturing requirements of small- and medium-sized enterprises. To combat labor shortages, they are now an essential complement. Collaborative robots are suitable for jobs that people are unwilling to do, such as highly repetitive work like sorting and packaging. Collaborative robots have several unique advantages:

Safer: Collaborative robots are compact and intelligent, and their sophisticated sensors enable them to stop in an instant. They can work closely together with human workers on the production line to get the work done.

Faster and more flexible deployment: Collaborative robots feature user-friendly programming, such



as programming by demonstration, natural language processing, and visual guidance. They can be placed in new positions at any time, and programming and commissioning can be completed rapidly, so they can start working very quickly.

Lower total cost of ownership (TCO) and shorter payback period: The price and annual maintenance cost of collaborative robots are significantly lower than those of traditional industrial robots. According to China's Forward Industry Research Institute, the average selling price of collaborative robots has halved over the past several years.

Collaborative robots are currently most widely used in the manufacturing of computers, communications equipment, consumer electronics products, and automobiles. They are also starting to be used in the medical industry for analysis and testing, liberating medical professionals from repetitive and time-consuming procedures (e.g., urinalysis) and reducing the risk of infection among medical workers by taking care of tasks like throat swabs.

Autonomous mobile robot

Autonomous mobile robots (AMRs) are a key enabler to help the manufacturing industry become flexible and intelligent. They will reshape the production, warehousing, and logistics processes. AMRs have rich environmental awareness. They feature dynamic route planning, flexible obstacle avoidance, and global positioning. The AMRs used in industrial manufacturing and logistics are mainly powered by the simultaneous localization and mapping (SLAM) technology, laser navigation, visual navigation, and satellite positioning to enable autonomous navigation. AMRs make automated and unmanned logistics possible. This includes unmanned operations for sorting, transporting, and storing goods.

Digital simulation and flexible manufacturing

To respond to changing market conditions and set themselves apart in the face of fierce competition, companies must take the initiative and embrace new production models. That's why an increasing number of companies are looking to concepts like flexible manufacturing. Flexible manufacturing relies on ICT technologies. Simulation, modeling, VR, and other ICT technologies can be used to simulate the entire new manufacturing process. This will reduce the cost of new product development and design, and support more accurate adjustment costs and capacity planning. In addition, the intelligent task scheduling system schedules the production tasks and allocates production materials and tools based on known features such as the factory's production capacity, order complexity, and delivery deadlines. Flexible manufacturing uses ICT technologies such as visual programming, natural language interaction, and action capture to help factories reprogram

and define equipment quickly and easily. This will help companies promptly meet consumer demand for flexible manufacturing. Flexible logistics uses ICT technology to effectively manage warehousing and logistics, which prevents omissions and other errors in the shipment process. Take furniture producers as an example. With large-scale customization, every board, decorative strip, and handle may need its own identification code or radio frequency identification (RFID) tag to facilitate automated packing and loading, and to support traceability throughout the whole transportation and distribution process. This is the intelligent customized production that centered on consumers.

Resilient and intelligent supply chains that help enterprises respond to crises

More and more companies regard building a resilient and intelligent supply chain as one of their most important strategic priorities. Supply chain visualization uses ICT technology to collect, transmit, store, and analyze upstream and downstream orders, logistics, inventories, and other related information on the supply chain, and graphically display the information. Such visualization can effectively improve the transparency and controllability of the whole supply chain and thus greatly reduce supply chain risks. Supply chain visualization supports the tracking of materials and equipment in upstream activities. Logistics information is displayed in real time, including information on packing, goods logged in, goods logged out, and inspections; goods can even be traced throughout the production process. This enables companies to detect and rapidly respond to any logistics emergency by promptly adjusting logistics routes to ensure the timely and safe delivery of goods.

A remote monitoring system monitors the environment in warehouses in real time. This system uses various sensors to graphically display operations and maintenance (O&M) information such as temperature, humidity, dust, and smoke. This allows the timely detection of any signs of fire or water leakage, which enables prompt

intervention and prevents material losses. Goods can be tracked in real time as they are logged in to and out of warehouses. With the movement of goods, IoT, RFID, and QR code technologies are used to automatically identify and register goods, and the warehousing status data of goods can be accessed remotely in real time.

By 2030, digital technologies will be transforming companies. Technologies such as AI, sensors, IoT, cloud computing, 5G, and AR/VR are poised to become new drivers of productivity. They will help make up for labor shortages, so that companies can seize new business opportunities and expand their possibilities.

In the future, product design, process design, equipment functions, logistics, and distribution will all be reshaped to become more flexible and serve new people-centric production models. Powered by digitalization, supply chains will be visualized and expand into supply networks. This will enable companies to become more resilient than ever and more capable of responding to volatile markets.

Huawei predicts that by 2030, the digital transformation of enterprises will further promote the application of data services in enterprises. In addition, data services are forecast to account for **87%** of enterprises' application expenditures, while AI computing will account for **7%** of a company's total IT investment.

Energy: Data Helps Build Energy-Efficient, Low-Carbon Data Centers

Climate change is a global challenge, and many countries have come together to tackle it. At the UN Climate Conference (COP 21) in 2015, parties to the Paris Agreement agreed to intensify efforts to limit global warming to well below 2°C, preferably to 1.5°C, compared to pre-industrial levels, and set the goal of reaching net zero CO2 emissions globally around 2050. In other words, by the midpoint of this century, the CO2 emitted by human activities needs to be matched by the CO2 deliberately taken out of the atmosphere[10]. At the 75th UN General Assembly in September 2020, China pledged to peak its carbon emissions by 2030 and achieve carbon neutrality by 2060. Concerted efforts are needed to combat climate change and drive the transformation of the global energy in three areas: energy supply, consumption, and carbon fixation.

With the increasing complexity of energy networks and the increasing digitalization of the energy sector, ICT technologies have become an important part of decarbonization solutions. The key questions for global warming now are: How can we further increase the share of renewables in the energy mix? How can we adapt to the new energy mix? And how can we fully harness the power of ICT technologies? Smarter green energy drives sustainable economic development and supports the following scenarios:

Offshore wind, a promising energy source for the future

In 2020, the worldwide energy installed capacity from renewable sources increased by 280 gigawatts (GW) or 45%. Of this, 114 GW was contributed by wind, an increase of more than 90%. Some European countries are actively exploring offshore power generation. In 2020, the installed offshore wind capacity in the UK and Germany exceeded 18 GW, accounting for 51% of the world's offshore wind capacity. Offshore wind energy still provides only 0.3% of the electricity globally, and there is huge room for expansion.

Wind conditions at sea are better than on land, with wind speeds typically 25% higher than on coastal land and less turbulence, resulting in a dominant and stable wind direction. The capacity of offshore wind turbines can be 3 to 4 times greater than that of inland wind turbines. There are fewer calm periods at sea, so offshore wind turbines can generate power for 3,000 hours a year, which makes for more efficient use of generator capacity. The technology advances have led to a significant reduction in the cost of offshore wind installations, and the offshore

power generation cost in 2040 is expected to be 60% lower than in 2019. The Global Wind Energy Council (GWEC) forecasts that global offshore wind capacity will increase from 29.1 GW today to 234 GW by 2030. Annual installations of offshore wind capacity are expected to grow at 31.5% per year over the next five years. This is a boom time for offshore wind power.

Floating PV, the latest trend in solar PV

According to the Snapshot of Global PV Markets 2021 by the IEA, the total installed capacity of photovoltaics at the end of 2020 was 760.4 GW. In 2020, solar PV accounted for approximately 42% of the total power generation from all new renewable energy sources. Large inland PV power plants are the most common form of PV installation, but there are a number of problems associated with inland solar farms: land acquisition, high costs, and poor performance under high temperatures. Floating PV (FPV) is a new direction for solar PV. Compared with land-based PV (LBPV) systems, installation of FPV systems on water saves land for agricultural use. The lack of obstacles on the surface of the water means less shading loss and less dust. In addition, the natural cooling potential of the body of water may enhance PV performance. In 2020, a research team from Utrecht University in the Netherlands simulated an FPV system on the North Sea. They found that the apparent temperature at sea was much lower than on land. The apparent temperature difference was nearly twice that, at 9.36°C. This study found that an FPV system performs 12.96% better on average on an annual basis than an LBPV system.

As the technologies mature, rapid growth is anticipated in FPV. On July 14, 2021, Singapore's Sembcorp Industries unveiled a floating solar farm deployed on the Tengeh Reservoir. With 122,000 solar panels spanning across 45 hectares (equivalent to about 45 football fields), the 60 megawatt-peak (MWp) solar farm is one of the world's largest inland floating PV systems. According to Rethink Energy, the global FPV market capacity will exceed 60 GW by 2030. Globally, the estimated potential capacity is 400 GW, meaning that FPV could double the current global installed capacity of solar PV. The floating solar market is set to accelerate as the technologies mature, opening up new opportunities for scaling up global renewables.

Low-carbon data centers and sites

According to the IEA, since 2010, the number of Internet users has doubled, global Internet traffic has increased by 12 times, and the electricity consumed by data centers and transmission networks has increased significantly. The global electricity demand from data centers was about 200 terawatt-hours (TWh), accounting for about 0.8% of global electricity demand. Data networks consumed approximately 250 TWh in 2019, accounting for approximately 1% of global electricity consumption, with mobile networks making up two-thirds of this figure. Data center power consumption in China alone is expected to exceed 400 billion kWh in 2030, accounting for 3.7% of the country's total power consumption. If data center power usage effectiveness (PUE) improves by 0.1, the result will be 25 billion kWh of power saved and 10 million fewer tons of carbon emissions. If all data centers use green power, carbon emissions will be reduced by 320 million tons each year. Green power and PUE optimization are key measures for low-carbon data centers.

In addition to applying renewable energy and free cooling, AI is another effective way to make data centers more efficient and save energy. Sensors in data centers collect data such as temperature, power levels, pump speed, power consumption rate, and settings, which are analyzed using AI. Then, the data center operations and control thresholds are adjusted accordingly, reducing costs and increasing efficiency. AI is used in data center cooling to reduce the energy used for cooling by 40%. According to Datacenter Dynamics, the Boden Type Data Center (BTDC), an experimental data center built in Sweden with funding from the EU's Horizon 2020 programme, has achieved a PUE level of 1.01 by using AI algorithms to achieve synergy between the cooling system and computing loads, server fans, and temperatures, in addition to environmental cooling. As AI technology matures, with green electricity and free cooling, data centers and communication networks will be more efficient and reach zero-carbon goals eventually.

By 2030, the global carbon emissions need to be reduced by half. For production sites, new energy such as wind and PV can fuel new deployment modes, while on the consumer side, electrification can help achieve electricity substitution goals. ICT not only helps itself but also other industries to reduce carbon missions.

Huawei predicts that by 2030, the power consumption of data centers in China will account for 3.7% of the total China's power consumption (incl. **25%–32%** consumed by storage systems). Green energy will play an important role in slashing emissions.

Digital Trust: Data Security Applications Shape a Trusted Future

Driven by digital transformation, interactions between organizations, between organizations and customers, and within organizations, are migrating to the digital world ever more quickly. Valuable digital assets are generated during these processes. Digital trust is a complex system that covers a range of areas, including privacy, security, identity, transparency, data integrity and governance, and compliance[11]. New technologies such as blockchain, privacy-enhancing technology, and AI and new rules will help shape a trusted digital future.

Smart contracts on the blockchain

Digital assets bring unprecedented quick access and convenience to organizations and individuals with potential risks of theft and misappropriation. Blockchain-based smart contracts contain terms expressed in a digital form on a blockchain, and the recording and processing of these terms are completed on the blockchain. Blockchain technology allows information to be recorded and distributed, which ensures that the entire process, from contract storage and access to performance, is transparent, traceable, and tamperproofing. Smart contracts have huge market potential in logistics, e-commerce, finance, insurance, and other sectors. According to Capgemini Consulting, smart contracts may help US consumers save US$480 to US$960 per mortgage loan, and enable banks to cut costs in the range of US$3 billion to US$11 billion annually by lowering operational costs in the US and European markets. Consumers in the US and EU could save US$45 to US$90 per year on their motor insurance premiums, and insurers would reduce the cost of settling claims by US$21 billion a year globally.

New mechanisms for collecting personal information online

More and more laws and regulations concerning the over-collection of data have been passed in recent years. In the context of big data, a fair digital strategy would contain optimized

mechanisms that balance the privacy of individuals and the interests of data users creating value with consumer data. The level of control data subjects has over their own personal information will be further enhanced while preserving the conventional approach of obtaining informed consent. In 2021, China promulgated its first Personal Information Protection Law. This law emphasizes multiple basic principles for protecting personal information, including openness, transparency, knowledge of purpose, and minimization. In the future, regulatory frameworks will be further refined so that users will have more knowledge and control over the ways in which their data is collected and used, and the associated risks.

The General Data Protection Regulation (GDPR) is currently the most stringent privacy and data security law in the world. It was drafted by the EU and took effect on May 25, 2018. In 2020, the US published its Federal Data Strategy 2020 Action Plan, which includes the goals of protecting data integrity, conveying data authenticity, and ensuring data storage security. On May 27, 2020, Japan passed the Act on Improving Transparency and Fairness of Digital Platforms, which was designed to regulate specific digital platforms and enforce obligations to the public on those platforms. These regulations represent a global trend towards antitrust action in the data domain. As antitrust laws are further modernized and adopted, data users and third-party companies will be granted more data rights against industry giants. This will help develop a healthy digital trust ecosystem, and prevent large platforms from committing digital security violations or engaging in other behaviors that compromise fair competition, such as illegally obtaining, abusing, and trading personal data.

Blockchain, AI, and other technologies will be the foundation of a digital, trustworthy, world, as they provide better personal privacy and asset protection, can accurately highlight disinformation, and mitigate fraud or data theft risks. Further, privacy-enhancing computation ensure data shared among multiple parties is encrypted without risk of private information leakage.

Huawei believes that by 2030 half of all computing environments will use privacy-enhancing computation, and 85% of enterprises will use the blockchain. While privacy-enhancing computation, blockchain, and IPFS technologies all offer clear advantages in security, it will cause a huge increase in encrypted and distributed ledger data, generating **17 ZB** of new data each year. **Additionally, over 80% of enterprises are expected to deploy multi-layer ransomware protection systems which cover the storage systems.**

1.2 The Digital Economy Leads Humankind into the Yottabyte Era

By 2030, the digital economy will account for 60% of the global economy, and data will become the basis for industry digitalization.

Today, technologies and industries are transforming, and the digital economy is thriving, changing the way people live and work and influencing economic and social development, global governance, and civilization. According to the 2020 Global Digital Economy report by China Academy of Information and Communications Technology, the global digital economy was worth US$31.8 trillion in 2019, accounting for about 36% of global GDP. The digital economy has maintained rapid growth, developing significantly along the way. The added value of the digital economy has reached CNY35.8 trillion, accounting for 36.2% of GDP and contributing 67.7% to GDP growth.

By 2030, the global digital economy will account for 60% of the global economy. The digitalization of conventional industries is speeding up. By 2030, the output value of the digital industry will reach 9%, a catalyst for economic growth. The digitalization of conventional industries requires digital tools to become more Internet-based, intelligent, and automated, to expand customer scope, to reduce costs, and to improve efficiency.

By 2030, 45% of industries will be digitalized, enabling us to better understand the world and promote AI and smart manufacturing.

Technology allows us to better observe, monitor, track, and process human, social, and earth activities, creating data that enables us to understand and describe the world more accurately than ever. Data and machine learning technologies are driving the development of AI, which will allow for increased automation of services, processes, and communication. By providing customized products based on customer preferences, AI will take efficiency and productivity to a new level.

As the Data Volume Increases from 175 ZB to 1,003 ZB, We Enter the Yottabyte Era

According to IDC and Huawei GIV team, the amount of data generated globally every year increases rapidly with the development of digitalization, from 2 ZB in 2020 to 175 ZB in 2025. In 2030, this figure is expected to reach 1,003 ZB, marking the start of the yottabyte era (1 yottabyte = 1000 zettabytes).

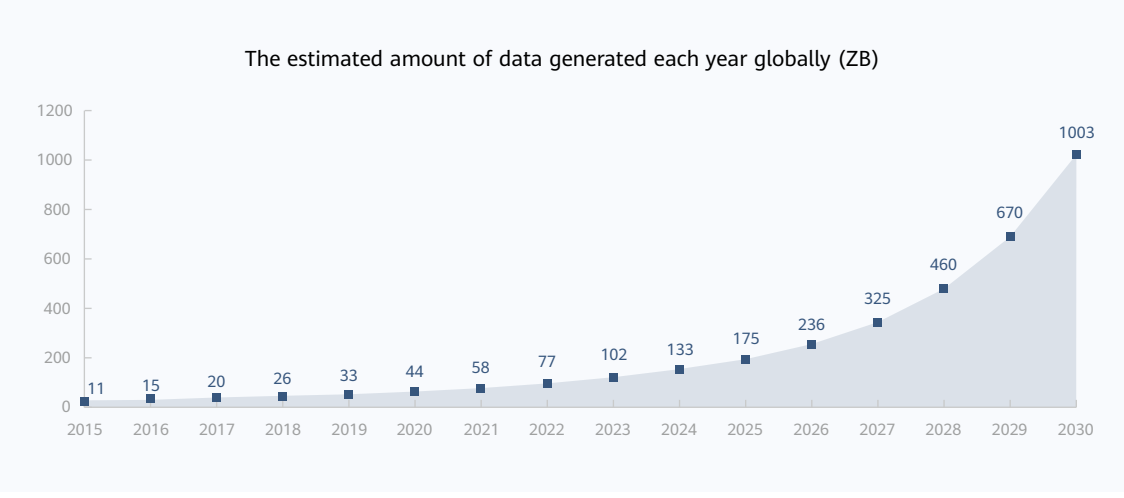


Figure 1-2 Prediction of the total amount of new global data generated each year

Diverse Data Applications Generate Many Different Types of Data

As the wave of digitalization sweeps across industries, data applications are becoming increasingly diverse. In addition to conventional database applications, new applications such as distributed databases, big data, and high-performance computing (HPC) are emerging. On average, an enterprise now has more than 100 types of data applications.

Evolving digital and mobile technologies have significantly changed how enterprises interact with customers. Internet applications, such as mobile apps, have become the most effective platform for driving customer purchases, and they have led to rapid service growth. The resulting surge in the volume of structured data has made core system workloads less predictable and more volatile. To cope with this, enterprises need to have core systems with flexible resource scalability, so that they can quickly expand resources during peak times and release idle resources during off-peak times to avoid waste. In addition, the multi-read and multi-write capability is becoming a mainstay feature in core systems as it ensures high system reliability.

Unstructured data is becoming a key enterprise asset because it contains a lot of valuable information and comes in different formats, such as text, images, videos, and audio files. **By 2030, 1 yottabyte (YB) of data is expected to be generated globally each year, and more than 80% of it will be unstructured.** Unstructured data is widely used in enterprises. 56% of enterprises use AI for at least one business function to analyze and process unstructured data in many different scenarios. The increasingly improved enterprise data governance capabilities enable data-driven service growth. Enterprises have also started to use unstructured data in systems that make decisions about production. Examples include online real-time credit approvals in the finance industry and pathological analysis in healthcare. **By 2030, it is estimated that 80% of unstructured data will be used to support production-related decisions.**

AI Promotes Data Awakening and Transforms Data Tiering

Large AI models are having an unprecedented impact on our daily lives, propelling us towards a more intelligent world and ushering in an era of data awakening. Data, as one of the three elements of AI, determines how far AI can go,

so the different types of data and how they are stored and accessed are very important.

First, the amount of hot data has skyrocketed. Statistics show that in China the new data stored in 2023 represented only 2.9% of the total data generated that year. A massive amount of data was discarded at the source and not stored. As AI continues to evolve, the volume of hot data and its importance are also on the rise. More and more data is being stored, and this can be used as a real-time and valuable input for AI. **It is estimated that by 2030, all hot data will be stored on SSDs.**

Second, the value of warm and cold data is being reexamined, and these types of data are gradually becoming hotter. Warm and cold data refers to information that is not frequently accessed, such as backup and archived data, so it has traditionally been seen as being less valuable. However, since AI requires extensive data for training, warm and cold data is becoming more important. By incorporating this data into the training process, we can enhance the accuracy and generalization of large AI models while also unlocking the value of previously overlooked data. Additionally, warm and cold data that needs quick access is termed active archived data. **It is projected that by 2030, more than 60% of enterprises will need to be able to access active archived data at least once a day.**

The Surge of Cloud and Internet Data Necessitates Changes to Data Architecture

In recent years, the rapid development of cloud computing and Internet technologies has had a significant impact on various industries. The demand for data storage in the cloud and Internet fields is growing at an unprecedented rate. Statistics show that about two-thirds of enterprise-level SSDs are delivered to cloud and Internet vendors.

To cope with the explosive rise in the volume of data and rapid changes in service requirements, cloud and Internet vendors, such as Google Cloud, are promoting the use of diskless reference architectures. A diskless reference architecture allows local disks of servers to be remotely deployed to form a brand-new architecture consisting of diskless servers and remote storage pools. It decouples compute resources from storage resources and enables flexible resource sharing, which greatly improves resource utilization and scalability while simplifying O&M and reducing energy consumption. With its flexibility and efficient storage resource management capabilities, it is expected to provide strong support for the sustainable development of the cloud and Internet industries and become the mainstream architecture. **Estimates suggest that by 2030, more than 80% of cloud and Internet enterprises will use the diskless reference architecture.**

70% of Data Generated on the Cloud, Edge, and Devices Is Concentrated, Resulting in Intensive Large-Scale Data Centers

The differences between data generated at the edge, endpoints, and core data centers are as follows:

Endpoints refer to devices on the edge of the network, including PCs, mobile phones, industrial sensors, automobiles, and wearables. By 2030, more than 75% of endpoint data will be processed by AI in real time.

Edge refers to servers and devices that process enterprise-level loads. Instead of being located in core data centers, these servers and devices are placed in server rooms, workplaces, or wireless base stations of branch offices to facilitate data processing with reduced network latency. By 2030, more than 80% of edge data will be processed by AI in real time.

Core data centers refer to large-scale data centers, including enterprise data centers, IDCs, and the cloud data centers of public cloud vendors. **By 2030, more than 90% of data at core data centers will be processed by AI in real time.**

By 2030, Endpoints Will Be the Primary Source of New Data. However, the Proportion of Data Generated by Edge and Data Centers Will Also Increase in the Future

As the number of endpoints continues to increase, endpoints will remain the main source of new data by 2030. It is predicted that the new data they generate will increase by 14 times, accounting for 52% of the total new data by 2030. This is due to the dramatic growth of smart vehicles, wearable devices, and industrial IoT.

There will also be a significant increase in edge devices by 2030. 5G MEC, CDN, remote and branch offices (ROBO), and high-tech media processor applications will become universal and home digital processing centers will begin to scale. In the future, every family will have a digital processing center that connects all home digital or intelligent endpoints, such as mobile phones, wearable devices, and smart home appliances like refrigerators, to store and process data and assist with everyday routines. By 2030, the data generated at the edge will have increased by 22 times, accounting for 21% of the total data generated.

The cloud is a key node for data aggregation, processing, backup, replication, and transfer. As each operation generates new data, data operations in the data center have an amplification effect, which is enhanced as more data is aggregated in the data center in the future. By 2030, the data generated by data centers will have increased by 18 times, accounting for 27% of the total new data.

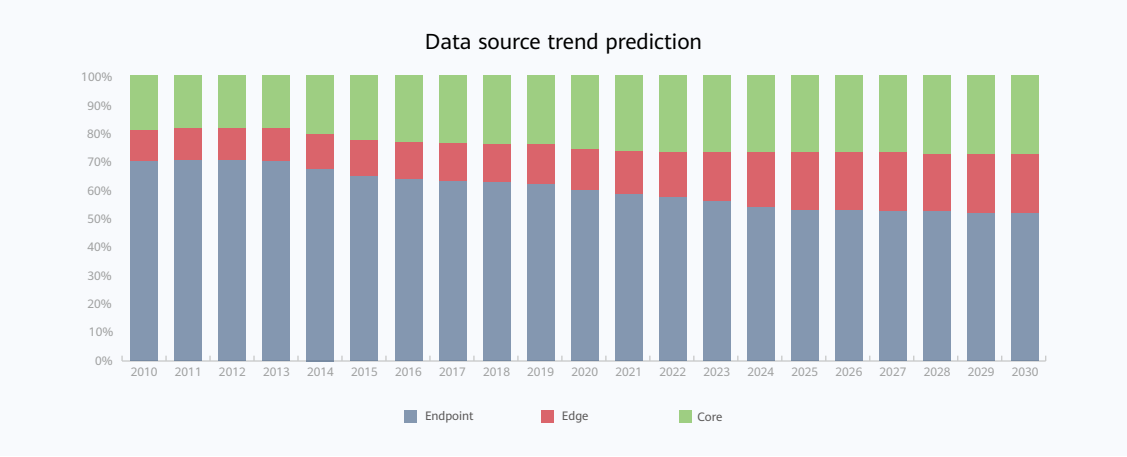


Figure 1-3 Data source trend prediction

A large proportion of data generated on endpoints will be stored in data centers through application systems and backup systems by 2030. The optimization of network and bandwidth make it more convenient and secure to store data, such as web disks, photos, account information, and applications, in data centers. Take the application account as an example. You can use one account to log in to the system via different clients and access the same view and services based on the account and its status stored in the data center.

About 65% of stored data will be stored in data centers by 2030. Data will be periodically backed up to data centers rather than being stored in endpoints. With more applications requiring real-time processing and low latency, data storage scenarios at the edge will become diversified. This includes intelligent driving training endpoints, real-time edge stream processing, 5GMEC, and VR/AR edge centers. The proportion of data processed at the edge will reach 10% by 2030.

Scattered data will be concentrated to data centers, allowing us to more easily mine data value and lay a solid foundation for digitalization.

02

Vision and Key Features of Data Storage by 2030

Over the next 10 years, the compound annual growth rate (CAGR) of data is expected to reach close to 40%, and data types are diversified. A single storage media cannot meet diversified data storage requirements. Diversified storage media are required to cope with challenges such as high storage costs, high power consumption, and poor durability. Diversified mass data promotes the development of diversified advanced media and media applications. Intelligent data reduction and joint data coding technologies will increase storage capacity density by several times.

The fast-growing data volume is contradicted by the slow-growing data processing capability, and the data storage capacity and data development are severely unbalanced. The classic CPU-centric architecture cannot meet the requirements of mass data storage and processing. Therefore, the entire architecture needs to be reconstructed with data as the core[12]. The new architecture supports storage-compute decoupling at the macro level, and computing in-memory at the micro level. The high-throughput, ultra-low latency, and high-scalability interconnection bus break the resource boundary and form a resource pool of processors, memory, and storage, supplementing computing through storage and improving data processing efficiency by several times.

The mounting data transfer requirements and increasingly severe data gravity have formed a fundamental contradiction limiting the value of data. The intelligent data fabric supports cross-region intelligent and efficient data flow, breaks space constraints, achieves what you see is what you get (WYSIWYG), and improves data flow efficiency by one hundredfold. Intrinsic data resilience separates

data use rights, management rights, and ownership rights, promoting trusted data flow[13]. A secure and reliable data application environment must be built through proactive defense to ensure data privacy and improve trusted data transfer efficiency by thousands of times.

Complex storage systems cannot meet the intelligent data service requirements of emerging multi-cloud applications. Therefore, data service logic needs to be decoupled from data intelligence. Future data storage will have new data awareness and understanding capabilities, supporting the rapid growth of data services in thousands of industries.

The continuous storage energy consumption increase still lags behind the global low-carbon development targets, placing new requirements on the green and low-carbon capabilities of storage. New energy-saving materials, transmission of data with optical signals instead of electronic signals, and dynamic energy-saving technologies promote chip energy saving. New liquid cooling heat dissipation and intelligent system control technologies decrease energy consumption across the entire system. System-level multi-dimensional and intelligent resource control technologies will reduce emissions throughout the data lifecycle, improving energy consumption efficiency by several times in the future and supporting the sustainable development of the future data industry.

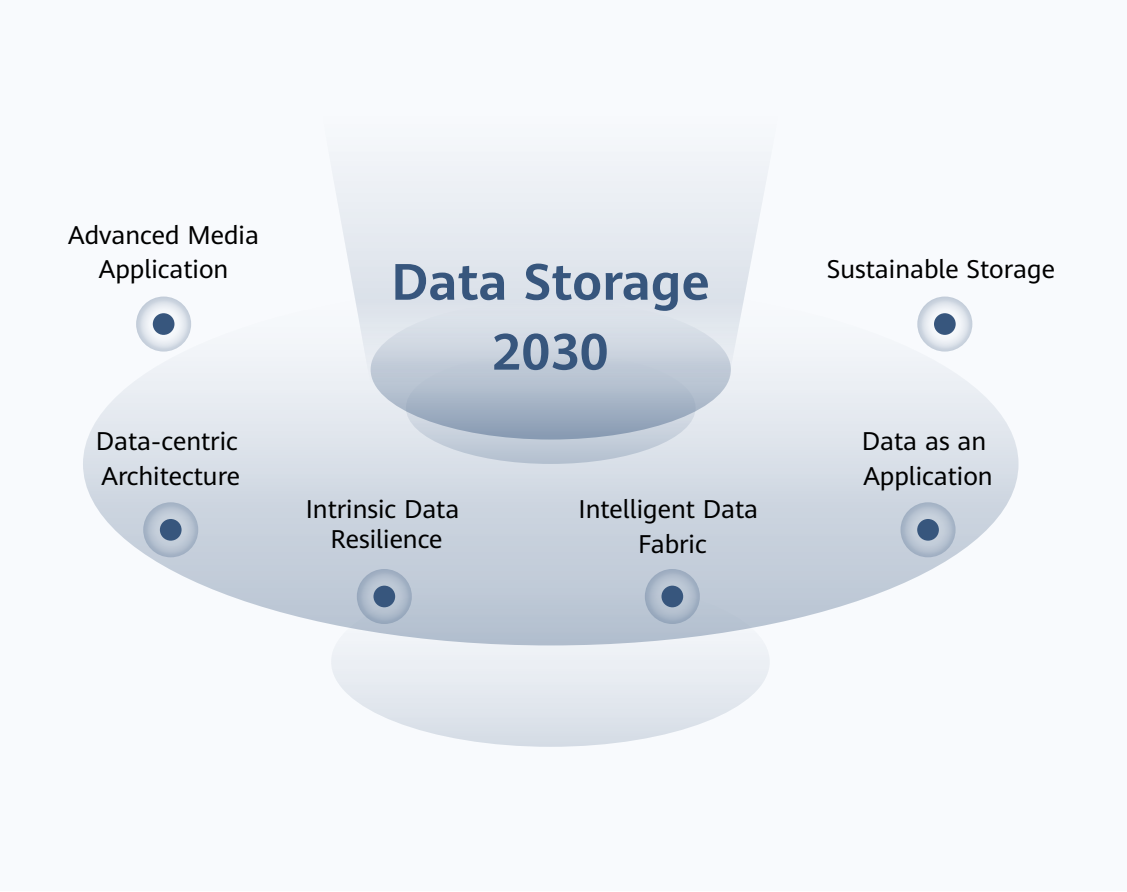


Figure 2-1 Six key features of data storage by 2030

To sum up, future storage will have the following six key features: advanced media application, data-centric architecture, intrinsic data resilience, intelligent data fabric, data as an application, and sustainable storage.

2.1 Advanced Media Application

The evolution of large AI models to being multi-modal is gradually awakening data. More and more video and image data will be saved for training. It is estimated that from 2030, 1 YB of data will be generated each year. The volume of data used for large AI model training is expected to increase more than one thousandfold to 400 EB. Nearly 50 ZB of valuable data will need to be stored every year, a 23-fold increase compared with 2020. Storage media must provide large capacity with cost-effectiveness and low energy consumption, featuring high reliability, high scalability, durability and high security. In addition, storage devices must have data computing and analysis capabilities to obtain data faster.

Different media have their own advantages and disadvantages. Therefore, multiple media need to be combined to cope with challenges. The evolution trend of different media suggests that the media capacity density will increase by 10 times by 2030. However, compared with the 23-fold increase of data volume, the growth of media capacity density still lags far behind. Media application innovation is required to fill the gap.

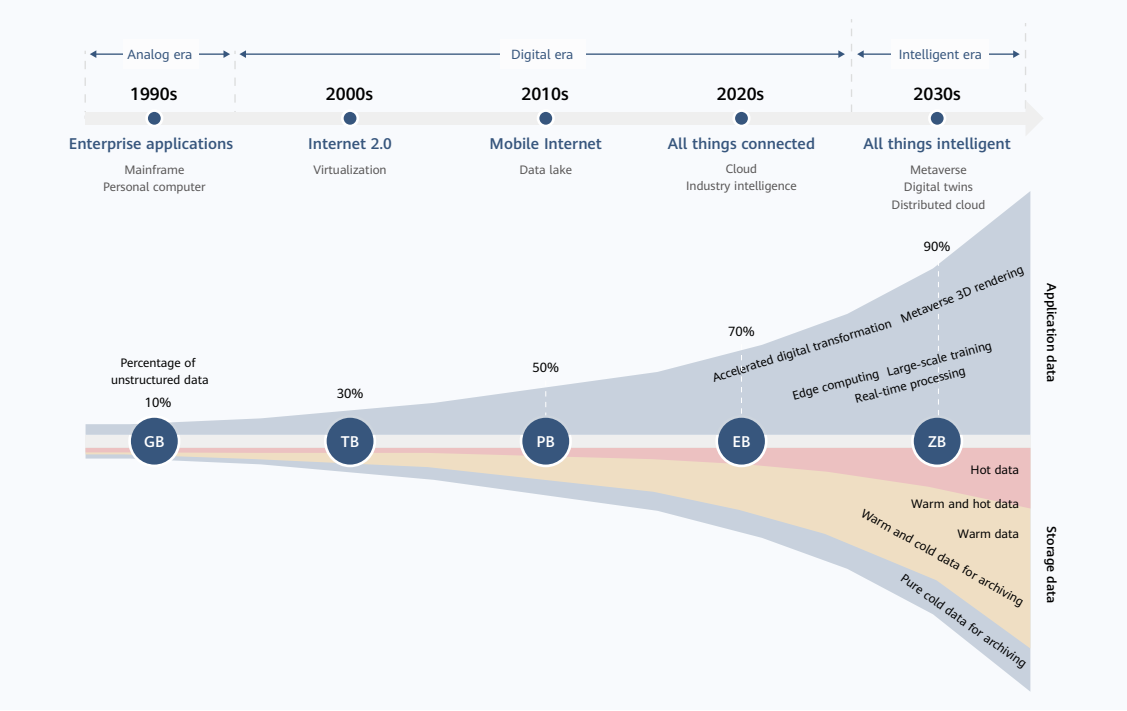


Figure 2-2 Data volume growth trend

Data can be classified into hot, warm, and cold data based on the access frequency. Different data is stored on different storage media.

Hot data: accounts for around 30% of total data. Of this, real-time data processing of AIoT, edge computing, robotics, and autonomous driving requires data access in nanoseconds. Such data is considered extremely hot, accounts for 1.5% of the total storage capacity, and requires memory media with very high performance. Online transaction services such as banking and e-commerce, and industrial manufacturing services such as EDA also require frequent real-time data access. Such data is categorized as common hot data, will increase by more than 35 times, and requires high performance flash storage media.

Warm data: Data-intensive services, such as HPDA, need to analyze a large amount of data but do not have high requirements for access frequency and real-time performance. Such data accounts for 60% of the total data and is expected to increase by more than 25 times by 2030. Apart from the need for large capacity media, it is also cost-sensitive with high sensitivity to power consumption, and requires cost-effective storage media.

Cold data: Historical documents, national archives, and other data that needs to be stored for a long time as required by laws and regulations are seldom accessed and therefore categorized as cold data. Such data accounts for around 10% of the total and is expected to increase by nearly 20 times by 2030. The long-term storage of such data requires high storage reliability and long-life storage media.

To train large AI models, increasing amounts of cold data will be activated to become warm data. Traditionally, data is stored in 3 tiers—hot, warm, and cold—in a ratio of 2:3:5. Now, cold and warm data will be merged into a combined tier. Data will be stored in two tiers—a hot tier and a warm/cold-combined tier—in a ratio of 3:7. The proportion of warm data as a percentage of all data will exceed 60% and sit somewhere close to 70%. The activated value of cold data will prompt data awakening.

Advanced Media Technology

Diversified data drives diversified storage media, requiring greater competitiveness in different application fields. Memory media for storing extremely hot data will be mainly DRAM supplemented by SCM, and memory tiering will become a new form. All hot data media will be NAND Flash, and Flash technology will evolve towards high density and low latency. In warm and cold data storage media, magnetic tapes are expected to continue to evolve towards high density and high concurrency, while optical disks will move towards larger capacity, higher concurrency, and longer service life.

1.Hot Data Media Technology

Memory plays a very important role in the computer system architecture for caching programs and data. With the development of data-intensive applications, the amount of data to be processed increases from GB-level to TB-level, driving memory media to provide larger capacity, lower power consumption, and higher concurrency.

(1)The memory architecture will become multi-layered.

DRAM currently dominates memory media. Due to the limited space for improving the capacity density of processes for chips below 20 nm, the 10 nm-chip process will keep developing for 10 years. With the increasing requirements of big datasets for large memory, the development of new media technologies such as SCM promotes the multi-layer memory architecture and gradually complements DRAM.

(2)SCM will continue to explore new scenarios.

SCMs that are based on new materials and structures have a performance comparable to DRAM and have the novel feature of data persistence. Computing in-memory (CIM) implemented by SCM has been used to supplement DRAM in certain fields, achieving a good acceleration effect. In the future, the new ecosystem centering on SCM will be enriched. Various SCM media with persistence capabilities allow fast, high-performance access to hot data. As for processors in existing storage systems, a large amount of time is typically consumed in I/O waiting. Innovative in-memory persistent storage subsystems are likely to resolve this problem in the future.

(3)Continuous evolution of NAND Flash in the 3D stacking accelerates the replacement of HDDs.

Compared with HDDs, SSDs have obvious advantages in terms of performance, power consumption, and capacity. In the consumer market, HDDs have largely been replaced by SSDs, while in the enterprise market, the replacement of HDDs by SSDs is expected to accelerate.

SSD development involves increasing the number of stacking layers, thereby increasing the storage capacity per unit silicon wafer area and reducing the cost per unit storage space. However, as the number of stacking layers increases, the depth-width ratio (the ratio of the hole depth to the hole diameter) of stacking memory holes increases, which brings greater challenges to the etching and deposition processes and limits the continuous increase of the number of stacking layers. To further improve the storage density and increase the effective area percentage of NAND arrays, the three-dimensional (3D) architecture of stacking peripheral CMOS circuits and NAND arrays will become the mainstream in the future.

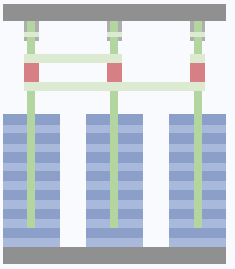


Figure 2-3 Working principle of the 3D NAND

Through stacking and 3D architectures, the capacity density per chip area is expected to increase by 10 times in 2030 compared with 2021. However, due to factors such as technology complexity and process yield rate, the cost of SSD in 2030 will not see a 10-fold drop in costs compared with 2021. Furthermore, due to impact of processes, interference of internal cabling, and an increase in density, the bottom-layer bit error rate of SSDs may further deteriorate, which poses a new challenge to an error correction algorithm with a low bit error rate, a low latency, and a high throughput.

2.Warm Data Media Technology

The evolution trend of SSDs and HDDs suggests that HDDs will still be cost-effective in 2030. As a result, HDDs will remain the dominant media in warm data storage scenarios requiring cost-effectiveness[14].

HDD technology improvements center around density improvement. Because the magnetic recording of HDDs can only be attached to the surface of the substrate, the density can only be improved by increasing the number of disks and enhancing the magnetic density. Restricted by the HDD form and superparamagnetism, the capacity density of HDDs is close to the limit. Therefore, short-term HDD density improvement will evolve towards breaking through form and superparamagnetism restrictions, such as ultra-thick HDDs and energy-assisted magnetic recording (EAMR) (HAMR and MAMR) [15]. Long-term technology evolution includes breakthroughs in magnetic recording technologies and materials, such as skyrmion and magneto-optical and magneto-electrical combination technology and materials.

3.Cold Data Storage Media Technology

By 2030, cold data storage media will still be mainly magnetic tapes and optical disks. Featuring high reliability, long service life, and low requirements on storage environments, optical disks are more suitable for ultra long-term storage of cold data. Magnetic tapes are mainly used for medium- and long-term storage of cold data. In the data-driven intelligent era, data becomes hotter, resulting in two new requirements for cold data storage media: low cost and fast read speed.

(1)Magnetic Media Technology

Magnetic tape (tape for short) recording implements data storage through moving tapes. Tape capacity is usually expanded via space folding. For example, the media recording area of LTO-9 is 100 times that of HDDs in the same period. Currently, the capacity density of tapes is about 1/100 of that of HDDs. In the future, the capacity of tapes is expected to exceed 100 times that of HDDs by leveraging magnetic domain miniature, high-precision servo control, and ultra-low bit error rate (BER) magnetic channel coding technologies.

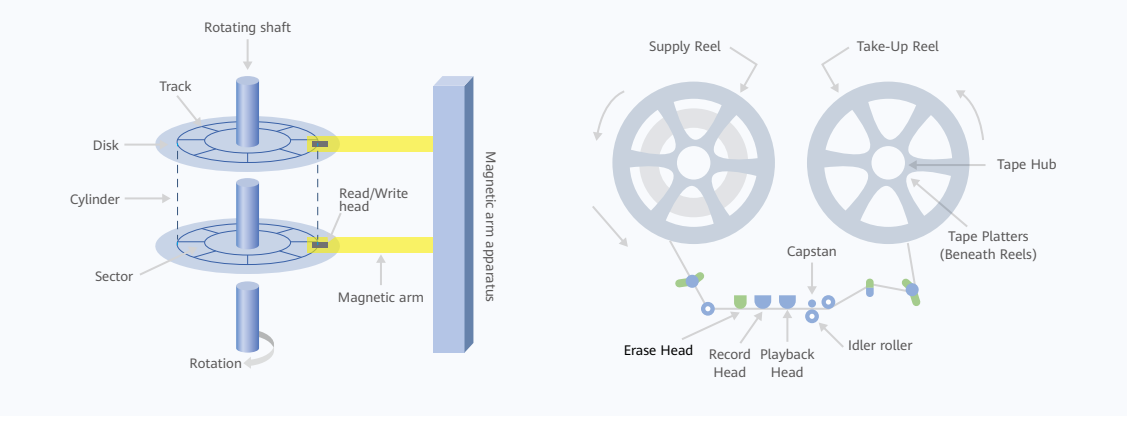


Figure 2-4 Working principle of magnetic storage (disk and tape)

Tape-based linear motion enables more heads to concurrently read and write data. Currently, the concurrent bandwidth of 32 heads in LTO-9 is more than twice that of HDDs. In the future, the bandwidth will be 10 times that of HDDs. The working principle of tapes suggests that they have innate advantages in sequential read/write. However, during random read/write, the head positioning time increases along with the capacity, affecting real-time data access[16]. In the future, data access will be faster for devices with high bandwidth, and data layout and scheduling algorithms can further improve real-time data access performance. The material suggests that their service life is significantly affected by the environmental temperature. When the temperature ranges from 35°C to 40°C, the service life of tapes decreases dramatically, increasing the risk of data loss. In the future, new magnetic materials, manufacturing processes, and efficient environment control technologies need to be further explored to prolong the service life of tapes.

(2)Optical Storage Media Technology

In the future, optical storage media will have larger capacity and cost less. Currently, Blu-ray storage is the mainstream optical storage media. Blu-ray was initially used in the consumer field, but its capacity is only 500 GB/disk, and the throughput of a single optical head is only over 40 MB/s[17]. In the future, optical storage will make breakthroughs in technologies such as super-resolution, multi-level, multi-dimensional, mirror ultra-multi-layer, and body material to increase the capacity of optical storage to 300 to 700 TB/disk and the throughput to 100 MB/s. In 20 years, the capacity of a single optical disk is expected to reach 100 TB.

Due to the requirement for long service life in cold data storage, another major future challenge for optical storage is how to ensure that data in optical storage media can be securely and accurately read after thousands of years[18].

Super-resolution optical storage technology: Optical storage records information by using a laser beam to physically and chemically change the recording material. Reducing the wavelength and increasing the numerical aperture can reduce the size of the laser spot and improve the recording density of optical storage. However, the wavelength and aperture are limited by the diffraction limit. In the future, the diffraction limit is expected to be exceeded through the use of multi-beam superposition interference. This will help further improve the recording density and increase the capacity of optical disks.

Multi-dimensional/multi-level optical storage technology: Unlike single-dimensional optical storage that can record only a single bit, multi-dimensional optical storage can record multi-bit information. The technology that is currently under research is five-dimensional optical storage. Five-dimensional optical recording stores data in five different dimensions: three spatial dimensions of storage media, polarization dimension, and intensity dimension. In the future, five-dimensional optical storage is expected to resolve the spatial interference of optical signals and develop towards six or more dimensions, further improving the capacity density of optical disks.

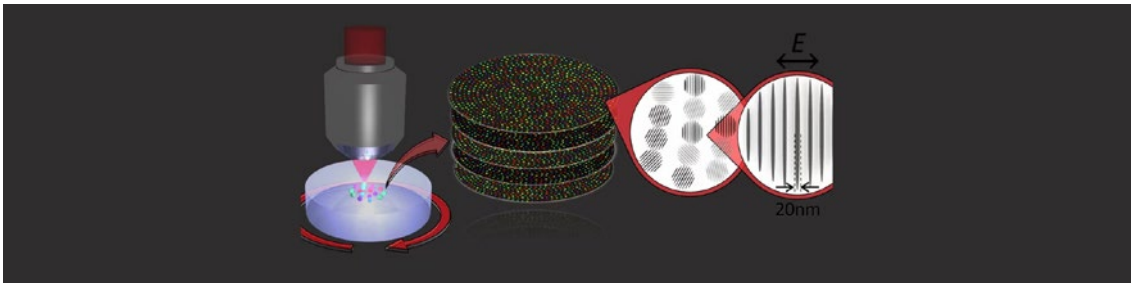


Figure 2-5 Working principle of optical storage

Multi-layer/Body material optical storage technology: The storage density of optical disks can be improved by superimposing the number of disk layers. For example, Blu-ray storage is commercially available with six layers. In the future, the industry is expected to solve the issue of inter-layer optical interference, allow dozens and even hundreds of layers to be achieved. Holographic optical recording uses phase change body materials to record information at different layers and angles inside the storage media. By combining multi-layer and body material recording technologies, optical storage can reach even higher densities resulting in a storage capacity of more than 100 TB/disk.

Servo drive technology: An optical disk drive includes a laser and an optoelectronic modulation device. Currently, the femtosecond laser and optoelectronic modulation device used in multi-dimensional optical storage are costly. With the development of the femtosecond laser industry, further breakthroughs in high-frequency and high-voltage optical circuit technology are expected to be made in the future to reduce the cost of gemstone-level crystals so that they can be put into large-scale commercial use in the optical storage industry. Limited by the write principle of optical storage, the read/write bandwidth of a single laser is only dozens of megabytes per second. In the future, the high-precision servo control technology is expected to implement parallel reads/writes of multiple optical channels and improve the throughput.

Media Application Innovation

1.Media Process Technology

The semiconductor manufacturing process and physical limits of media structure mean that the integration of media such as SSDs and DRAMs cannot be continuously improved. In the future, wafer-level innovation, chiplet-level innovation, and interface and protocol innovation can further improve media density and service life, reduce media power consumption, and enhance media reliability.

Wafer-level innovation: The die-on-board (DOB) technology can integrate storage chips into circuit boards to provide higher density and better performance. The Wafer-Scale technology directly uses wafers of multiple NAND dies without cutting or packaging the wafers, achieving higher density, faster speed, and higher reliability. The Wafer-Scale technology is currently immature. Problems such as manufacturing of ultra-large chips, function management and monitoring of chips, cross-chip connection, chip heat dissipation, and reliability management need to be solved. In the future, advanced process technologies, innovative chip design methods, and intelligent test methods are expected to be used to achieve higher capacity and better durability while maintaining the advantages of high density and low power consumption.

Chiplet-level innovation: Chiplets can integrate different functional modules into separately packaged chips to achieve better flexibility and scalability, higher performance, and higher power efficiency. Currently, the Chiplet technology still faces many technical challenges, such as inter-chip communication and synchronization, cache consistency, and transmission rate matching.
In the future, technologies such as intelligent control algorithms, efficient chip cache consistency protocols, encapsulation processors inside storage media, heterogeneous processors, and accelerators are expected to encapsulate computing chips and media chips together to build a Chiplet media that integrates storage and computing, achieving high performance, low power consumption, and easy scalability.

Interface and protocol innovation: As media become diversified, data transmission between multiple media interfaces has a large protocol conversion overhead, which can be greatly improved in terms of performance, security, and

universality. Zone Namespace (ZNS) is a high-speed storage protocol used for flash storage devices. It supports efficient space management based on smaller data blocks, alleviates the performance imbalance of SSDs, and improves the performance of garbage collection and data migration of SSDs. Currently, issues such as compatibility and application migration need to be resolved. Plogs are used for persistent data storage management. They can transmit and process data between different storage systems across multiple storage media. The Plog protocol uses the automatic retransmission and self-healing mechanisms to ensure data consistency, reliability, and integrity, improving data transmission and access efficiency. In the future, with the continuous development of diversified media technologies, new high-performance interfaces and protocols need to be defined to further improve compatibility and data access efficiency.

2.New Data Coding

Data coding technologies include compress encoding for reducing data volume (Sayood, 2017), error correction coding for preventing data errors, and erasure coding for preventing data loss (Peterson & Weldon, 1972). They are the core technologies that achieve large storage space and long storage period. In the future, storage systems that integrate diversified media for mass data call for breakthroughs in data coding technologies through intelligent data compression, joint coding, and intelligent data classification to improve effective storage capacity, save energy, and ensure long-term reliability.

$$H(p) + D(p \parallel q) \leq E_p l(X) < H(p) + D(p \parallel q) + 1$$

The theorem shows that the average code length $E_p l(X)$ after data compression is equal to the information entropy $H(p)$ of data plus the extra length $D(p||q)$ caused by model inaccuracy in the best case, and the code length in the worst case is 1 larger than that in the best case.

Figure 2-6 Lossless data compression theory

Intelligent data compression: Data compression is the process of using short-bit data to represent information based on a specific coding mechanism. In data storage, lossy compression coding and lossless compression coding coexist. The current lossy coding cannot break the classical rate-distortion theory. In the future, semantics extraction and compression technologies need to be explored, rate-distortion functions need to be extended, and a new theoretical system needs to be established to make technical breakthroughs in lossy compression. Mainstream lossless compression methods in the industry use LZ and entropy coding as the core, and the effect of compressing unstructured data is poor. The compression method based on statistics and dynamic prediction models can effectively improve the reduction ratio of unstructured data. However, the models depend on data and expert experience and develop slowly. AI-based prediction models can surpass expert-designed predictors through automatic extraction of data features and self-learning of models. The existing AI-based compression algorithms face the problems of poor generalization capability and high computing power consumption. In the future, transfer learning, meta-learning, and large model technologies are expected to improve the model generalization capability and algorithm efficiency, improving the reduction rate by several times in the storage system.

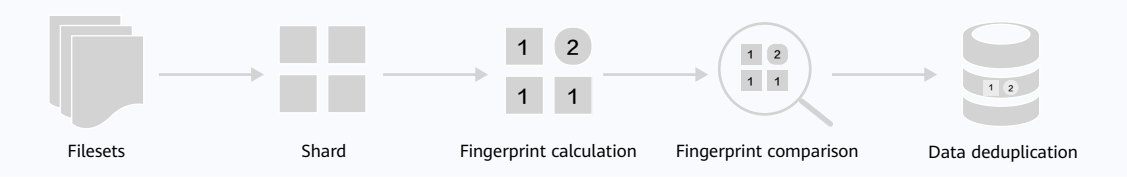


Figure 2-7 Basic principles of data deduplication

Data deduplication: The deduplication technology deletes duplicate data blocks by identifying the content at the data block level. With the emergence of processor technology and new storage media, the deduplication technology is gradually shifting from offline to online processing. The data deduplication granularity is shrinking and changing from file-level deduplication to recent byte-level similarity-based deduplication, creating challenges for the computing power and I/O throughput of the system. In terms of deduplication of diversified mass data, the deduplication ratio in high-dimensional data scenarios is several orders of magnitude lower than that in structured data scenarios. With the development of semantics deduplication technologies in the future, the storage efficiency of unstructured data is expected to be fundamentally improved.

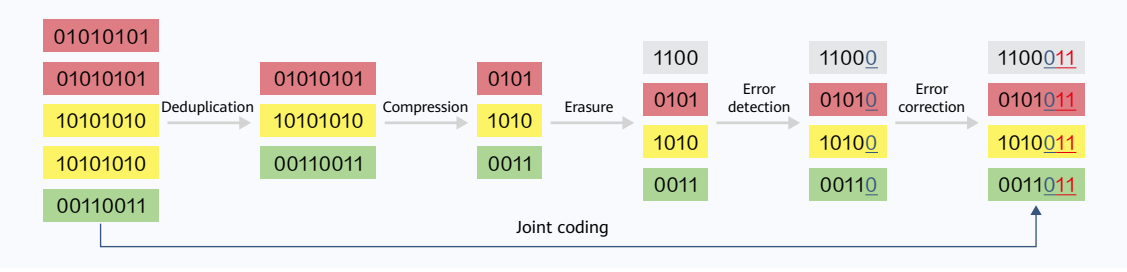


Figure 2-8 Data joint coding

Data joint coding: Shannon's separation theory (Shannon, 1948) proves that separate design of source coding and channel coding can achieve optimal system performance when the code length tends to be infinite. If the code length is limited, combining source coding and channel coding may achieve gains (Jiang & Bruck, 2008). In the future, joint coding can be designed to achieve higher-density storage, simplify the system, and reduce power consumption.

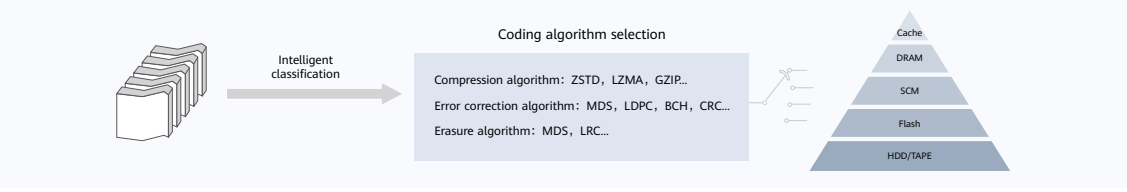


Figure 2-9 Intelligent classification

Intelligent data tiering and classification: Storage is a system with diversified and hierarchical media. The reliability, latency, bandwidth, and cost of different storage media vary greatly and, therefore, a matching data coding algorithm (Kim, Gupta, Urgaonkar, Berman, & Sivasubramaniam, 2011) (compression, error correction, or erasure coding) must be selected. In the future, breakthroughs in intelligent data classification technologies are required to optimally match different data coding technologies with different media, improve data density and reliability, and reduce latency.

2.2 Data-Centric Architecture

Driven by new data-intensive applications such as big data, AI, HPC, and IoT, the data volume is increasing explosively with a compound annual growth rate (CAGR) of nearly 40%. Hot data will account for more than 30% of the total data volume. With the slowdown of Moore's Law and Dennard Scaling, the annual growth of CPU performance has reduced to 3.5%. The fast-growing data volume and the slow-growing data processing capability present a challenge to the data industry, as storage capacity and data development are severely unbalanced.

In a typical CPU-centric data center architecture, uneven distribution of services in space and time lead to low utilization of local storage resources. The idle rate of local memory and storage exceeds 50%[19]. In addition, data movement and repeated data format conversion consume a large amount of CPU resources, resulting in a low data processing efficiency.

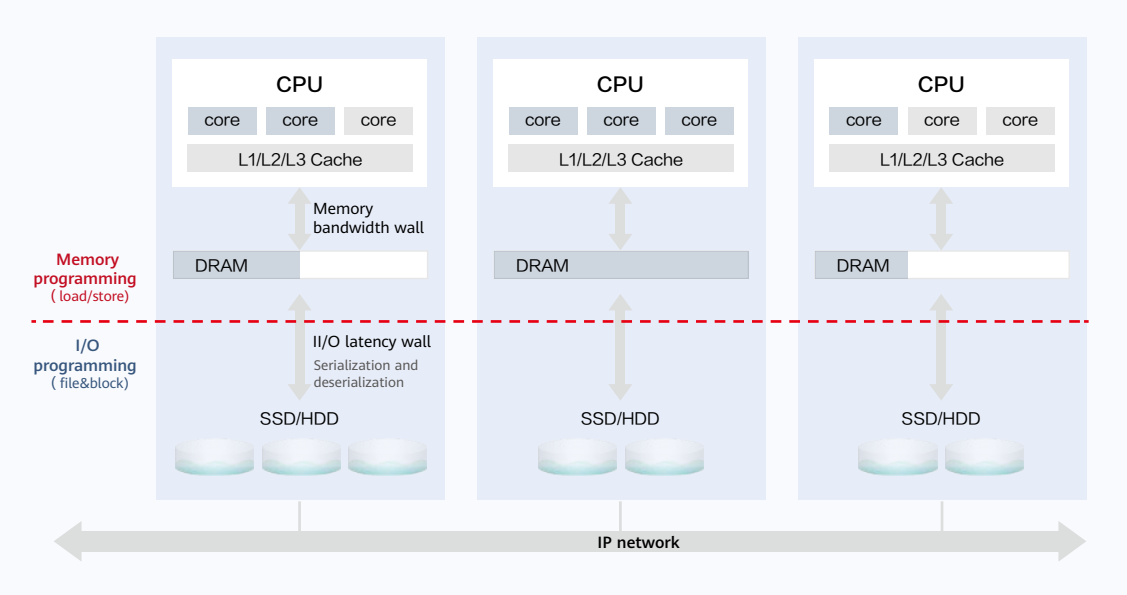


Figure 2-10 CPU-centric architecture

Storage and computing resources cannot be efficiently used in the existing data center architecture. To improve data processing efficiency and storage resource utilization, the future data center architecture needs to shift from CPU-centric to data-centric, including:

1.Storage-compute decoupling at the macro level: Computing and storage resources are independently deployed and interconnected through high-throughput[20] data buses for unified memory semantics access, implementing decoupling and flexible scheduling of computing and storage resources and maximizing resource utilization.

2.Computing in-memory at the micro level: Data is processed nearby with data as the core, reducing unnecessary data movement. Dedicated data processing computing power is deployed at the edge of data generation, on the data flow networks, and in the data storage system. The convergence of network, storage and computing improves the data processing efficiency.

3.Highly scalable cluster storage: The scale-out and scale-up capabilities are increased by a factor of several dozens. Cluster storage can be scaled out from dozens of controllers to hundreds of controllers and EB-level capacity is available. Hundreds of xPUs can be scaled up to thousands of xPUs, achieving acceleration via near-storage data processing.

Storage-Compute Decoupling

Storage-compute decoupling will no longer be limited to the decoupling of CPUs from SSDs and HDDs but will instead completely break the boundaries of various storage and computing hardware resources and builds them into independent hardware resource pools (such as CPU pools, DPU pools, memory pools, and flash storage pools) for elastic expansion and flexible sharing of hardware. The storage-compute decoupling architecture has three features: storage resource pooling, global memory semantics access, and high-throughput peer-to-peer interconnection bus.

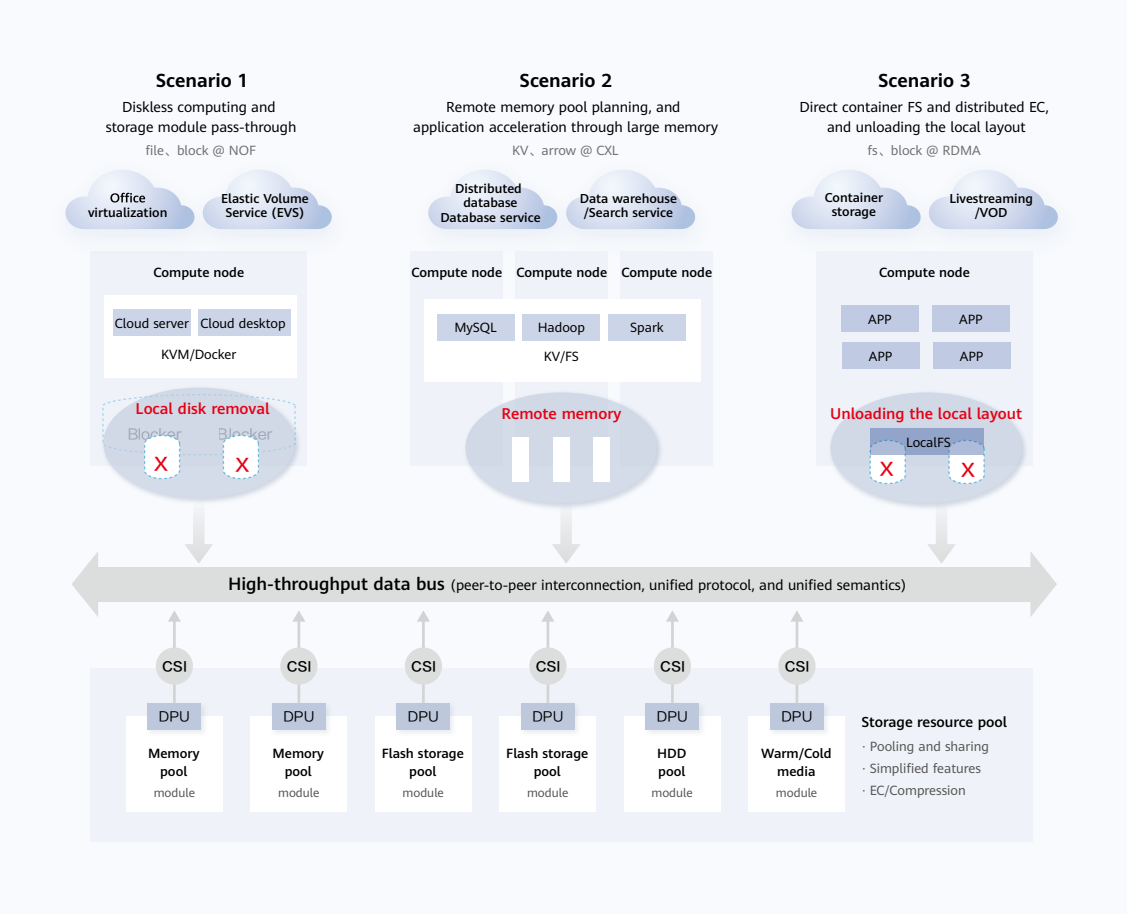


Figure 2-11 Storage-compute decoupling architecture

1.Storage Resource Pooling

The new storage-compute decoupling architecture remotely deploys local disks of servers to form diskless servers and remote storage pools. In addition, the remote memory pool is used to expand local memory, implementing true storage and compute decoupling and greatly improving storage resource utilization. In service scenarios, virtual disks with different performance and capacity and pooled memory space can be selected based on application requirements. First, storage resource pooling prevents space waste caused by local storage space over-configuration. Second, resource pooling prevents data from flowing across buses and devices, reducing data movement, improving performance, and reducing power consumption. Finally, if a server is faulty or replaced, data migration is not required.

The NVMe over RDMA network technology implements SSD pooling and provides consistent

local access performance for remote access to SSDs. In the future, new memory networks (such as CXL and Unified Bus), intelligent tiering of memory media, and unified addressing technologies are expected to be used to implement memory pooling, expand the memory capacity by 10 times, and reduce the cost of large memories for applications.

2.Global Memory Semantics Access

Traditional applications access data via file, object, and block interfaces. The I/O stack protocols are complex, and the application I/O overhead exceeds 30%. Memory semantics and memory data format are used for data access, achieving zero I/O stack overhead, zero format conversion overhead, and zero data flow overhead. Currently, memory semantics data access still faces challenges from the application data access interface ecosystem and memory semantics network standardization. In the future, a unified memory semantics standard protocol is expected to be formed for memory semantics data interworking and further improvement of data access efficiency.

3.High-Throughput Data Bus

The traditional interconnection bus is CPU-centric. The CPU becomes the system bottleneck, and the system cannot be expanded on a large scale. The protocol types are different from each other, and the protocol conversion is repeated, reducing the system efficiency. Different devices have different communication semantics, and the data format conversion is repeated, causing extra overheads. The high-throughput data bus needs to be defined to support peer-to-peer communication among devices, eliminate protocol conversion, and simplify data access. The high-throughput data bus has the following features:

(1)Peer-to-peer connection: The CPU-centric structure is broken. CPUs, DPUs, and storage devices are interconnected in peer-to-peer mode. Data access does not pass through CPUs. Heterogeneous and diversified data processing devices directly access data in peer-to-peer mode, improving data migration efficiency.

(2)Unified protocol: Different communication requirements are abstracted in devices, cabinets, and data centers. Unified basic protocol functions are formulated and unified access protocols are used between processors and storage devices and among different storage devices.

(3)Unified semantics: Different access requirements are abstracted into unified access semantics, and data can be shared and accessed across systems and devices of different types.

(4)High throughput: The bandwidth of a single SSD will evolve to 25 GB/s, and that of the memory will support 100 GB/s, with a latency of less than 50 ns. New buses are used to interconnect SSDs, memories, and processors, as well as extend to inter-rack interconnection. In addition, new buses need to meet the requirements of high bandwidth for large-block data transmission and low latency for small-block data transmission. In the future, buses need to support TB/s-level bandwidth and 10 ns-level latency.

Computing In-Memory (CIM)

In the data-centric processing paradigm, data processing has shifted from general computing to professional data processing, and from data migration to processors to near-data computing power deployment. Data is processed with the most appropriate computing power near data, and data is processed nearby at the edge of data generation, during data movement, and in the data storage system. As a data carrier, data storage not only provides data access services, but also near-data processing acceleration services. There are three main methods for near-data processing: diversified storage and computing convergence, convergence of data storage and networks, and convergence of data processing and networks.

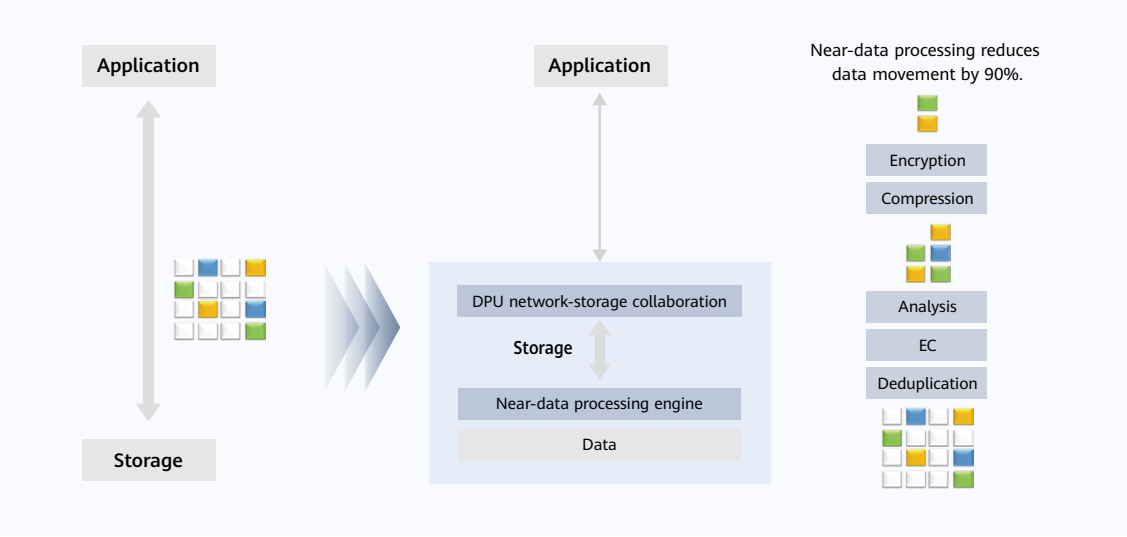


Figure 2-12 Working principle of storage and compute convergence

1.Diversified Storage and Computing Convergence

Storage and computing convergence is a technology used to offload host processing to memories to reduce data migration and network latency, overcome bandwidth bottlenecks, and improve data processing efficiency. Storage and computing convergence includes storage and computing integration (SCI) and computing in-memory (CIM)[21].

SCI integrates the instruction computing unit and operator unit on the storage component for data preprocessing, for example, adding a solidified data preprocessing unit (such as a compression engine or an encoding engine) to an SSD or a memory to accelerate data processing, or, integrating a large-capacity memory into the processor to reduce data access and, in turn, improve data processing efficiency. In the future, how to define efficient forward-compatible instruction sets and new operator abstraction in the aforementioned scenario is still a challenge that needs to be resolved but it is expected that implementation of efficient data processing in common scenarios through common instruction set research and customized operators will be achieved.

CIM uses a non-Von Neumann architecture and integrates storage units and computing logic to break the boundary between computing and storage and transfer a small amount of data during data processing. Compared with the traditional Von Neumann architecture, CIM improves energy efficiency by more than 10 times. Due to the limitations of current storage media, there are still great challenges to overcome in digital-to-analog conversion efficiency, computing precision, and scale. In the future, breakthroughs are expected to be made by improving storage media and discovering new media materials.

2.Collaboration of Data Storage and Networks

Convergence of data storage and networks senses the storage semantics to offload data storage services and schedule data flows, improving data access performance and accelerating data application services. Currently, it can be potentially applied in offloading storage access protocol (file protocol, object protocol, KV, etc.), accelerating storage I/Os (data passthrough and I/O zero copy), and offloading data layout (such as index). Storage services can be flexibly offloaded to intelligent network interface cards (NICs). However, this faces challenges in terms of programming friendliness and operational efficiency. In the future, efficient storage operators

are expected to be defined to achieve greater flexibility and higher performance.

3.Collaboration of Data Processing and Networks

By collaborating with the network, data processing in hosts by general-purpose processors is offloaded to dedicated data processors, such as security data processing (such as SHA256 and lattice-based cryptography), data compression (ZSTD, LZ, and CDC), data protection (EC), and data analysis (Scan, Filter, and Merge). Dedicated data processors represented by DPUs feature lower costs, lower power consumption, plug-and-play, and swap-and-play. They accelerate data processing during data flow, release the computing power of general processors, and improve the performance of big data, HPC, and databases by multiple times.

Cluster Storage

In a large-scale AI computing center, neither a single storage node nor scalability to any fewer than 100 nodes can meet the compute cluster's requirements for hundreds of PBs of capacity, retrieval from hundreds of billions of files, and hundreds of TB/s of bandwidth. By 2030, we expect that a storage cluster will be scalable to more than 500 nodes. In addition, more and more data reads and writes are being offloaded to xPUs to achieve near-data processing and improve efficiency. Storage can support an increasing number of xPUs. In the future, thousands of xPUs will be able to work concurrently and will be elastically scalable. The cluster storage capacity is expected to increase one hundredfold to hundreds of PBs, and this will meet the demands for both high performance and large capacity.

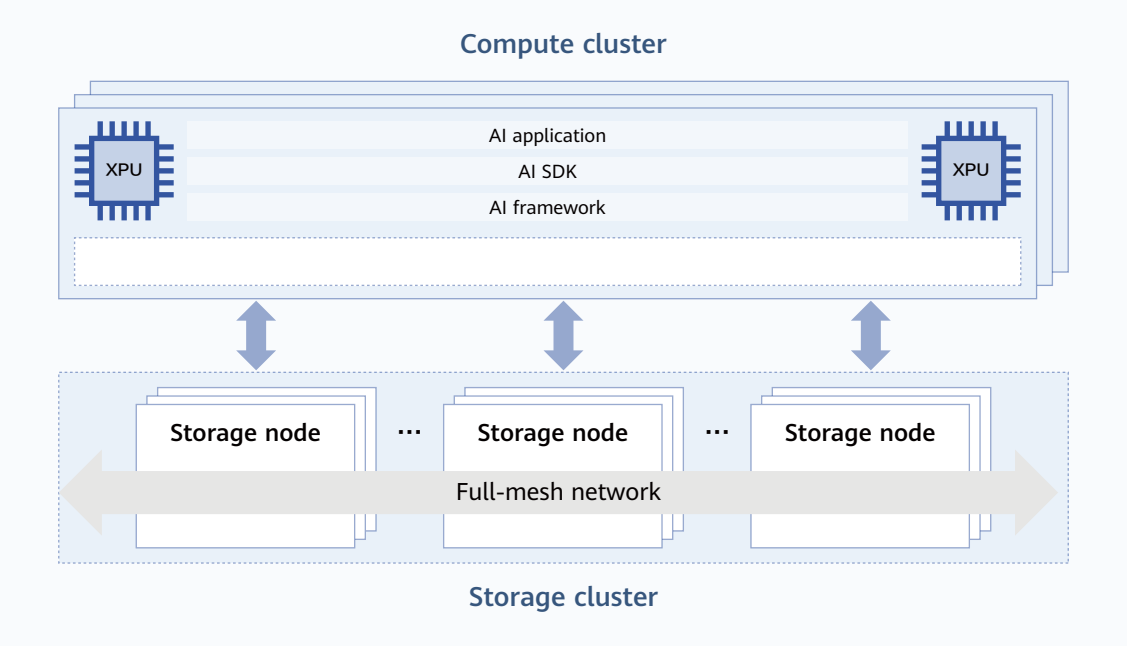


Figure 2-13 Cluster storage

2.3 Intrinsic Data Resilience

As a new production factor, data is becoming increasingly valuable, increasing the attack surface and attack intensity. The current border-based passive defense system cannot meet future data security requirements [22], and data privacy protection requirements are mounting during data value release. Privacy computing centering on "usable but invisible" data converts and releases data value while fully protecting data and privacy. Data transfer is necessary for releasing data value. The replicability, sharing, and unlimited supply of data means that protecting data property rights, use permissions, and control rights during data transfer is the primary issue the current data infrastructure must resolve.

In the future, data infrastructure will feature intrinsic data resilience. As such, we need to make continuous breakthroughs in technologies such as proactive data protection, zero data copy, and zero-trust storage.

Proactive Data Protection

Research on data security attack and defense situation shows that the current passive defense security system cannot effectively defend against virus attacks such as ransomware. A proactive data protection security system needs to be built from multiple technical directions, such as data security situational awareness, data timeline travel, native anti-tampering, and multi-dimensional linkage response.

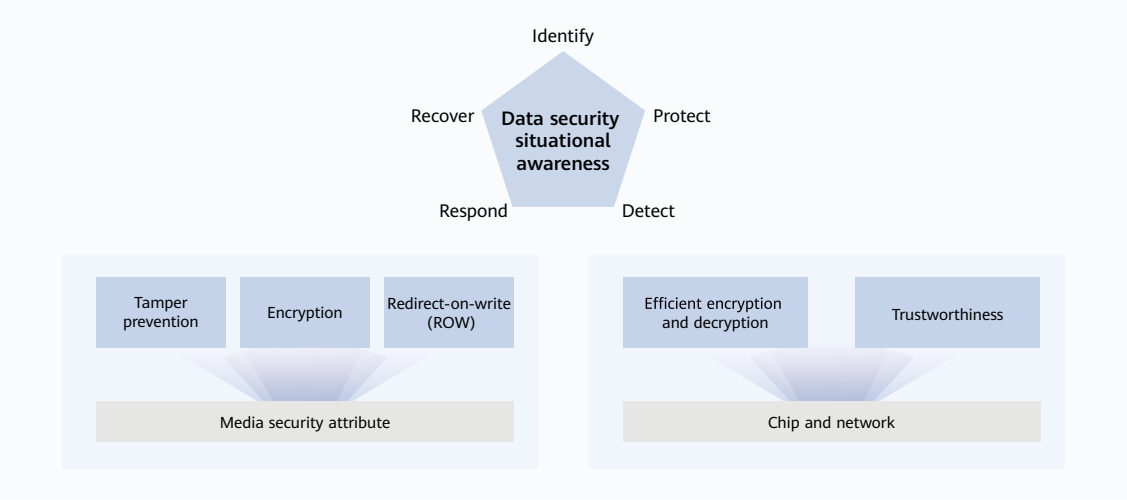


Figure 2-14 Proactive data protection

Data security situational awareness: The data security situational awareness technology collects the data access behavior, data information entropy, internal data correlation, and data distribution within a certain period of time, and dynamically measures and evaluates data security risks and threats based on the big data analysis technology to support subsequent self-defense decision-making and actions. Currently, the industry is facing challenges such as low accuracy and efficiency of threat detection and situational awareness capability and an insufficient dynamic threat evaluation capability. Detection accuracy and performance are expected to gradually improve in the future. The ability to detect unknown data threats through research on the sampling theory of mass data, converged processing of heterogeneous data, and activity

identification based on incomplete information are also expected to be enhanced.

Data timeline travel: If data is damaged due to internal and external attacks, the data infrastructure must be able to quickly restore the damaged data to any historical point in time to achieve zero data loss. In addition, to implement attack source tracing, the data infrastructure must have the finest-grained data replay capability to support the adjustment and optimization of data security policies. The industry is currently facing the challenges of quickly and accurately locating the time point of damaged data and automatically tracing behavior. In the future, technologies such as I/O-level data recovery and root cause analysis based on cause-effect reasoning are expected to be used to retrace the data timeline.

Native anti-tampering: Currently, the data anti-tampering capability is mainly implemented by the system-level data access control technology. However, due to the large attack surface of the system, it is difficult to effectively ensure data anti-tampering. In the future, the system-level data access control technology and the physical anti-tampering attribute of media are expected to be combined to implement the physical anti-tampering capability.

Multi-dimensional linkage response technology: The multi-dimensional linkage response technology requires cross-device collaboration among network devices, security devices, endpoint detection and response (EDR) devices, and storage devices to implement multi-dimensional closed-loop threat processing and prevent threats from spreading. The main challenges in the industry currently lie in autonomous decision-making and response technologies, that is, how to develop intelligent response policies to provide customers with convenient and effective alternative solutions. Future technical breakthroughs in areas such as AI security analysis, causal analysis, and inference are expected to make autonomous decision-making and response more intelligent, thus achieving real quick and accurate response.

Data Zero Copy

The value release process of data elements is divided into three phases. The first phase focuses on supporting business system operation and promoting digital transformation and intelligent business decision-making. The second phase allows external enablement of data flows to aggregate and integrate high-quality data from different sources in new businesses and scenarios, achieving win-win and multi-win results. The efficiency between data sharing and data access control during this phase needs to be improved. Technologies such as cryptography-based access control, data self-protection, efficient and transparent audit, and efficient network encrypted transmission can be used to implement efficient data flow and use while ensuring data sovereignty and security. The third phase is borderless zero copy, which eliminates data silos to the maximum extent. The zero data copy access technology is used to break data boundaries and implement data sharing.

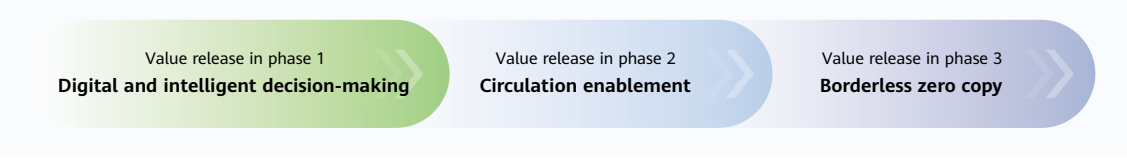
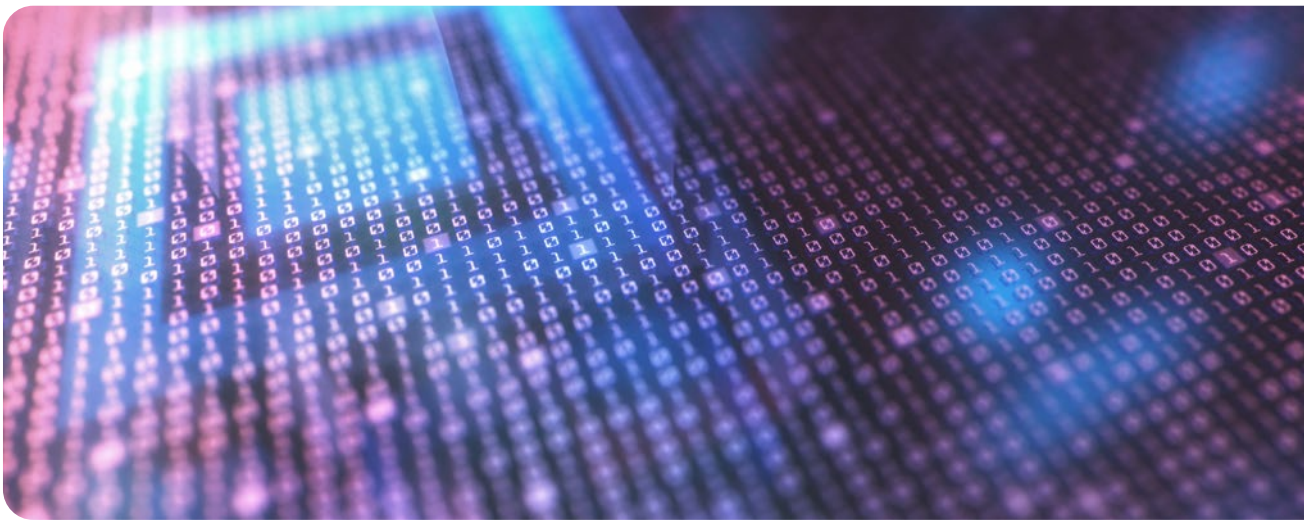


Figure 2-15 Data value release model



Cryptography-based access control: Cryptography is used to ensure data confidentiality and thus users who do not comply with access control policies cannot decrypt data. Typical attribute-based encryption (ABE) solutions support complete access control policies of any logic which, compared with traditional one-to-one public key encryption, offer a one-to-many setup to slash communication overheads and encryption/decryption calculation overheads for key nodes. Future technologies for encryption must be able to implement policy judgment and randomized processing on a ciphertext that is leaving a trusted domain, so as to prevent any data that does not comply with predetermined access control policies from leaving the trusted domain.

Self-protecting data: Data security has evolved from system-centric management and control to data-centric full-lifecycle security protection, with cryptographic computing used for data privacy. Self-protecting data refers to a type of technologies that make data "usable but invisible". However, there are problems of associated information privacy leakage, while the data use scope, use mode, validity period, and access permission are difficult to restrict. In the future, the "data capsule" solution is expected to encapsulate access policies, use control policies, and encrypted data to ensure that data owners can control data and implement secure data transfer.

Efficient and transparent auditing: The mainstream technology for trusted data auditing uses the blockchain. However, the blockchain technology has problems such as high overheads, low consensus algorithm efficiency, and data storage redundancy. Efficient and transparent auditing must be used to build a certificate anti-tampering auditing solution, to implement more efficient trusted data storage and meet the read/write latency requirement in the actual production process.

Zero data copy access: Due to the differences between application data models, most applications are siloed with independent data copies. In the future, application data models are expected to be deployed at the data storage layer and automatically generated based on the same data to eliminate data silos. In addition, technologies such as fine-grained access control and trusted network transmission based on chip certification are used to implement efficient data access across trusted domains.

Zero-Trust Storage

Zero-trust storage is an extension of the zero-trust model, aiming to solve storage security problems such as data leakage, integrity damage, and data availability damage. In zero-trust storage, all data read and write are considered unverified. Based on the minimum authorization principle, access subjects, data, and data operation implement minimum-granularity data access control through continuous verification and dynamic authorization. To achieve zero-trust storage, we need to make breakthroughs in data storage and usage environment security and full-path data security encryption.

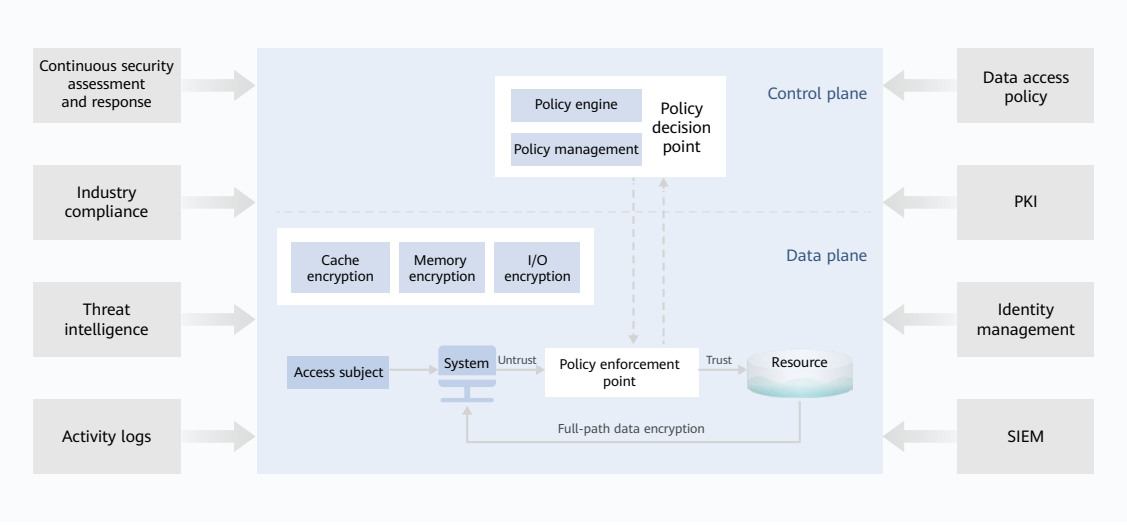


Figure 2-16 Zero-trust storage

Mandatory data access control: Based on the minimum authorization principle, fine-grained data access control uses the mapping among data access subject characteristics, data attributes, and fine-grained data processing to ensure that the minimum-granularity dataset can be accessed and used only by subjects under specific conditions. In the future, the design of access control policies will become increasingly complex due to the huge number of authorized entities and data, complexity of data processing, and uncontrollable conditions, and an improper access control policy can cause major security risks. To cope with this challenge, technologies such as formal verification, automatic policy generation, and compliance auditing will ensure policy consistency and correctness, and solve issues such as those involving large-scale formal verification performance, automatic policy generation mechanism, and complex rule matching.

Full-path data encryption: In the current border-based data security system, the full-path data is not safe, which may cause data leakage. Therefore, we need to consider encrypting the full-path data processing from memory, storage I/O, network I/O, and cache, and sharing native data security capabilities through unified key management.

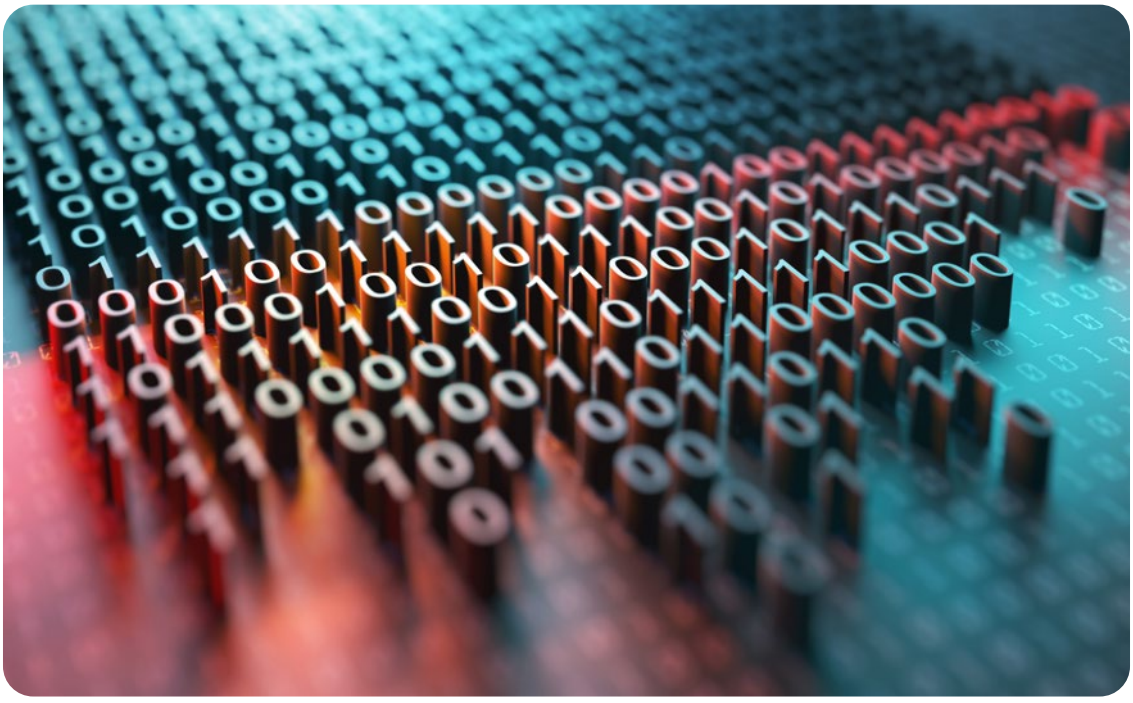
Privacy computing: To ensure data privacy and security during computing, secure data computing technologies have emerged, including Federated Learning for AI[23], hardware-level Trusted Execution Environment, secure multi-party

computation using cryptographic algorithms, and zero knowledge proof.

1.Trusted execution environment (TEE): The main challenge of implementing the hardware isolation technology for sensitive data processing is that the completeness of the hardware security isolation mechanism cannot be proved by data, which may cause security vulnerabilities. However, compared with the cryptography technology, TEE has little impact on performance. In the future, TEE-based privacy computing will become a common industry requirement. It is estimated that this technology will be used in more than 50% of data processing scenarios by 2030.

2.Cryptography-based homomorphic encryption and secure multi-party computing technology, whose security can be proved mathematically, have become the industry's most ideal privacy computing technology. However, the main challenge is that its performance is more than 10,000 times lower than that of general computing and needs to be greatly improved to meet application requirements. With the maturity of approximate computing, homomorphic encryption and secure multi-party computing are applied in specific fields such as health data sharing. In the future, the hardware acceleration-based homomorphic encryption and secure multi-party computing technology will be widely put into commercial use in high-security application scenarios in industries such as finance and healthcare.

3.Multi-party computing is based on secret sharing among multiple parties. If cryptographic methods such as zero-knowledge proof are used to implement multi-party computing, the performance overhead is high. With its mathematically proven security, using TEE to implement secret sharing among multiple parties greatly improves multi-party computing performance, showing potential for widespread application in the future.



2.4 Intelligent Data Fabric

The continuous development of digital technologies has led to a large number of requirements for cross-domain data flow, posing higher requirements on data availability and quality. However, geographical barriers and difficulties in data governance restrict the free flow of data and finally result in data gravity. Data fabric is to coordinate distributed data sources automatically and dynamically, provide integrated and reliable data across data platforms, and support the use of a wide range of different applications[24]. Based on technologies such as AI and knowledge graph, intelligent data fabric continuously identifies and coordinates correlations between available data points to help customers mine value. For networks that transmit data among the edge, data center, and cloud, intelligent data fabric enables continuous analytics on already-existing, discoverable, and inferable metadata assets to integrate cross-platform data and provide efficient data flow and processing for applications. To better implement intelligent data fabric, continuous breakthroughs are needed in technical directions such as cross-domain data collaboration, automatic data orchestration, and efficient and fast storage networks to eliminate the data gravity problem.

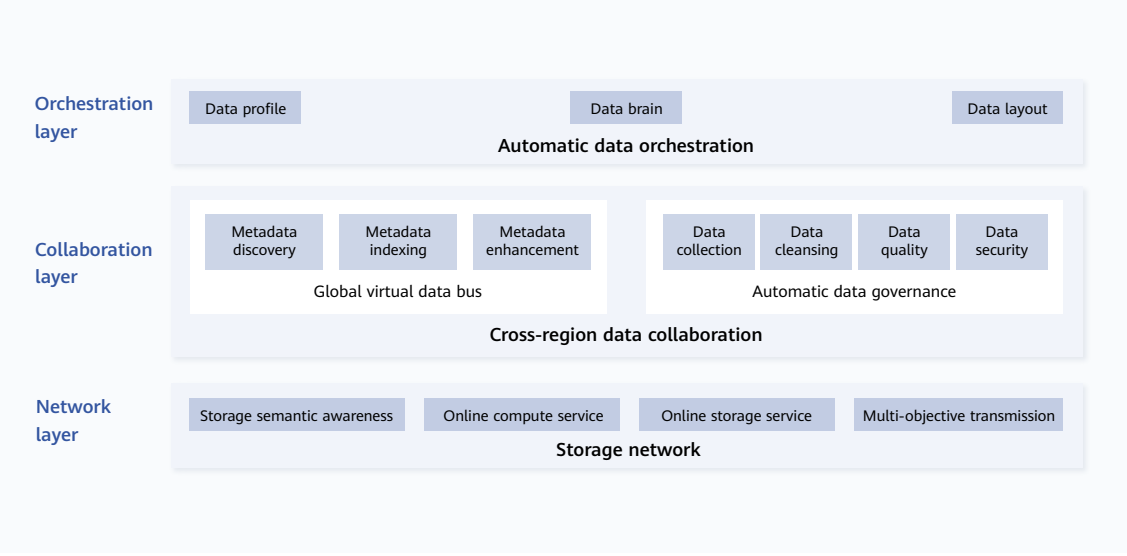


Figure 2-17 Intelligent data fabric framework

Automatic Data Orchestration

The present data content is inadequate at detecting network status, and application intents are unable to transmit to networks effectively. As a result, this unequal distribution of data and networks causes access delays and low network utilization. To address this issue, it is necessary to develop data profiles and data brains that achieve optimal data placement without compromising on service performance.

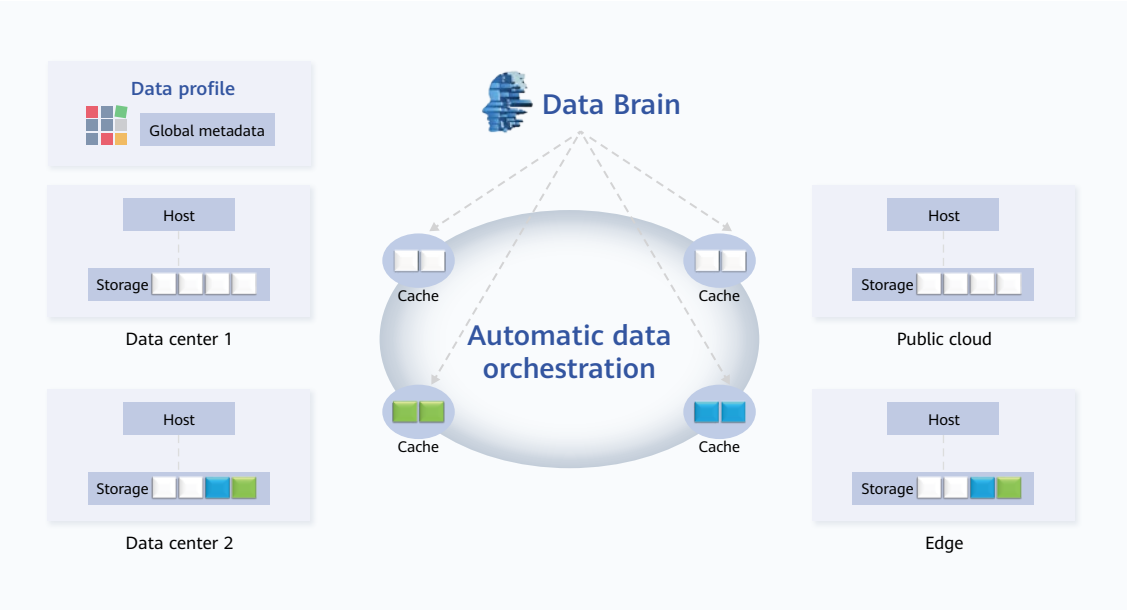


Figure 2-18 Automatic data orchestration framework

Data profiling: This technology senses application characteristics via the storage network status, spatiotemporal information of data blocks, and application labels. The current service awareness lacks granularity and precision. However, there is potential for the use of advanced technologies like deep graph neural networks and cause learning to create multi-dimensional data profiles which take into account aspects such as data gravity, data volume, data activity, network bandwidth, and latency. These technologies help achieve accurate service awareness.

Data brain: Current data orchestration is plagued by scattered data, dimension explosion, a lack of standardization, high technical requirements on developers, and unawareness of customer applications, posing high requirements on data flow management and control in multi-cloud scenarios. The universal data orchestration platform cannot meet the requirements of multiple parties. New technologies such as intent-driven API, machine learning, and big data analysis can help generate the optimal data layout policy based on industry applications, and provide powerful security management and audit capabilities that realize full-process automatic data orchestration.

Data layout: This technology places data to the optimal location based on service policies, so that users can access nearby data by content name and obtain the best experience at the minimum cost. For example, cross-region backup mode can help store cold data in a region with lower operation costs. The current data layout has problems such as poor data sharing between services, long tail data access, low hit ratio of data cache, and high network bandwidth usage. Breakthroughs in separating service logic from data logic, data network coding, and data prefetch and eviction algorithms are expected to implement adaptive data cache and nearby read/write cache acceleration. This cost-efficient and application-unaware measure will improve data access while simplifying data retrieval and utilization.

Cross-Region Data Collaboration

Typically, enterprises store data in multi-region data centers or using multiple heterogeneous cloud providers to provide unified compute-storage services for lower costs and better infrastructure capabilities. The distribution of assets, software, and applications across multiple data centers or clouds necessitates cross-region data collaboration and integration. Such trends can be divided into the following two directions:

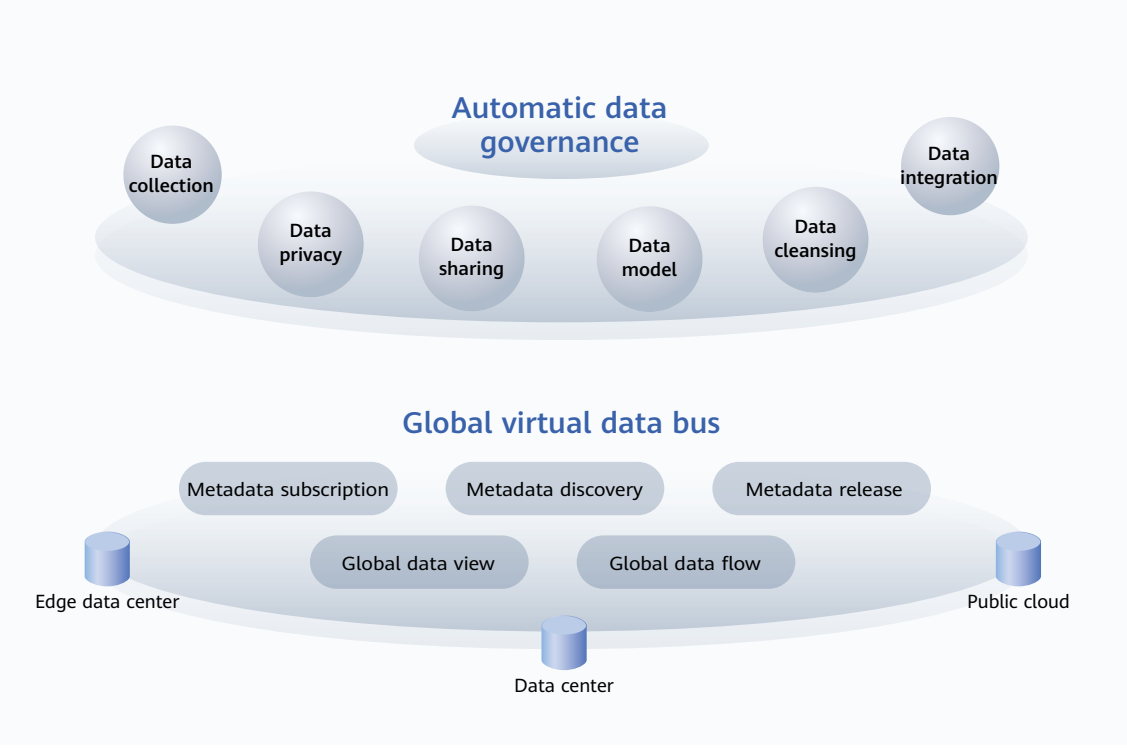


Figure 2-19 Cross-region data collaboration framework

Global virtual data bus: In public clouds and enterprise data centers, data is managed in different regions, and thus a large number of data silos exist. In the future, the release, discovery, and subscription of metadata can help realize efficient on-demand interconnection and build a global virtual data bus. Such virtual data bus must have a unified data namespace and transparent data flow capability to provide cross-cloud global data space and secure, efficient, and easy-to-use data networks.

Automatic data governance: To enhance the efficiency of the process from data collection to data value mining, data from different sources and of different types requires interconnection and efficient collaboration. This can be implemented through unified and standardized data models and systems that provide automatic processing capabilities and integrate basic functions such as data collection, cleaning, integration, quality improvement, and security assurance. The current data governance technology is not mature. Breakthroughs need to be made in heterogeneous data integration, data lineage management, and data classification and grading to develop unified, efficient, and intelligent data cube services to effectively improve data quality and availability.

Storage Network

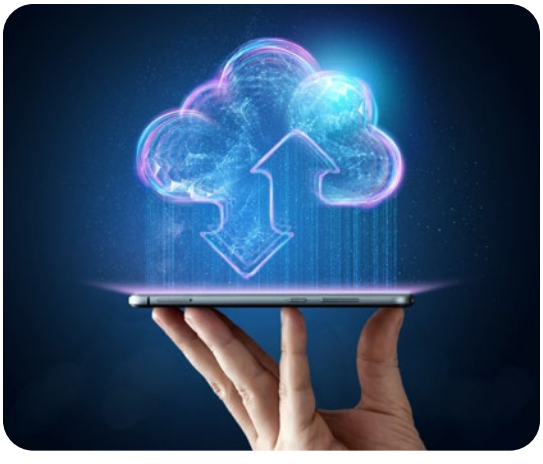
As the amount of data generated continues to increase, data silos will become even more common. It will hence be crucial to facilitate cross-region data flow to ensure better data management. However, the long network latency and low system efficiency of data access severely hinder the development of data applications. Therefore, an efficient and fast storage network needs to be constructed to implement data access that is insensible to applications and regions. Based on current storage trends, future storage networks must provide the following capabilities:

Storage semantic awareness: Traditional networks are aware of only network semantics, such as IP addresses and TCP/UDP port numbers, and treat all network packets in the same way. Future intelligent data network can further sense storage semantics, for example, identify and prioritize packets based on storage semantics to implement policy-based forwarding, identify association between packets to schedule co-flow, and route to fit storage I/O semantics. These features will enable differentiated processing of storage packets and optimized utilization of limited network resources, which will facilitate frequent data exchange among different nodes.

Online compute service: Intelligent data networks will expand network computing capabilities beyond solely packet forwarding and routing capability common in traditional networks. The abstract operator is used to design a Turing-complete instruction set to develop an efficient data processing engine that can be carried by a network forwarding device or device-side NIC. A network forwarding device implements associated-channel processing of data, and performs computing processing such as data encryption, decryption, compression, redundancy removal, and verification during data migration, thereby realizing real-time parallel processing of computing and transmission. A device-side NIC implements near-data computing to save data migration bandwidth and deliver a low-latency service. In addition, it also carries the protocol conversion of interfaces such as Smart Data Accelerator Interface (SDXI)/ NVMe to provide hardware-based data flow capabilities.

Online storage service: While current networks generally transfer data packets, future networks will use the forwarding and processing capabilities of a large number of data packets. This provides diversified associated-channel storage services, such as distributed locks, metadata cache, and transaction concurrency control, to achieve sub-RTT service response time and greatly improve data access efficiency.

Multi-objective transmission: In terms of network control protocols, traditional TCP/IP networks are designed for network survivability and deliver either high throughput or low latency, but this is not a compromise with new-gen technologies such as RDMA over WAN, F6G, and all-optical networks. In terms of network routing protocols, traditional networks are designed for a single objective, whereas modern storage networks handle both real-time database query requests (low latency) and large file transfer requests (high throughput), with a host of multiple objectives, such as the shortest path, maximum network utilization, and load balancing achieved. Therefore, future networks are expected to run on multi-objective protocols to meet diversified data services.



2.5 Data as an Application

Digital infrastructure in 2030 will be closely related to our everyday lives. Consider how digital twins, metaverse, and ChatGPT are made possible on today's networks, but hindered by limitations in intelligent data processing of emerging multi-cloud applications. Future systems must decouple data logic from data intelligence. The data infrastructure faces three challenges: (1) Scattered data causes silos and impact sharing. (2) Data mining consumes huge resources due to multiple modeling, training, and reasoning processes, which is unsustainable. (3) Complex data management of mass applications affects the efficiency of data pre-processing, severely restricting the development of applications.

Data as an application will have benefits in data awareness and understanding and new data services, and support the projected hundred-fold growth in data services in numerous industries. Digital storage is developing towards the ubiquitous, diversified, and cognitive storage trends.

Ubiquitous: Future storage will be miniaturized, portable, green, and intelligent and offer advantages in power consumption, density, and processing. It will be available in new forms (computational, brain-like, and biological DNA storage), and the portable features will enable faster large-scale commercial availability. In the short term, portable storage will improve data transfer speeds on the device side, edge, data center, or cloud. In the medium and long term, its building-block design will form reliable, secure, and O&M-free storage that can implement real-time data sharing, interaction, and processing.

Diversified: Traditional applications are mainly graphics and images in data format. However, the emerging applications such as brain computer interfaces, bionics, and AI will drive the diversification of data formats and create new data paradigms like vectors, tensors, and retrieval-augmented generation (RAG). From the perspective of data semantics, autonomous driving, drones, and robots will generate a large amount of data with composite semantics.

Cognitive storage: Currently, storage devices only provide the data storage function with multiple access layers, failing to offer ultimate application experience. This will be improved in future, smarter storage that is characterized with cognitive capabilities. It offers advantages in automatic processing and analysis, adaptive modeling, domain knowledge acquiring, and optimized data processing capabilities through learning[25].

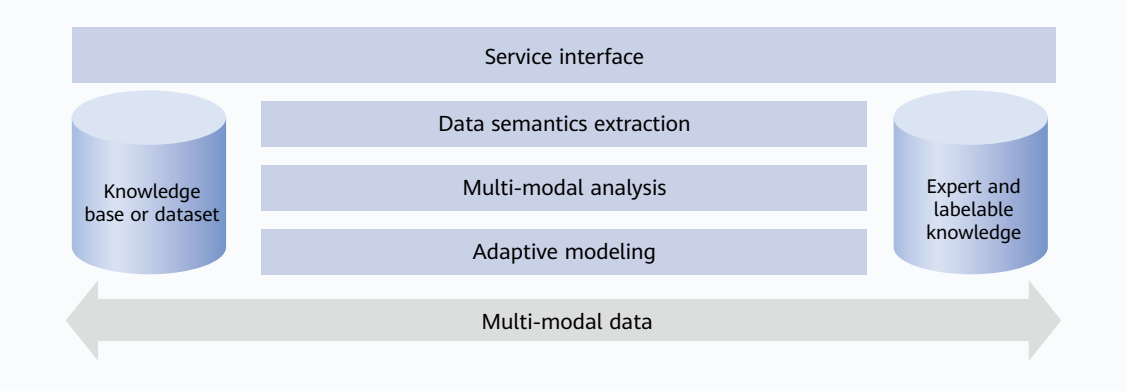


Figure 2-20 Framework of data as an application

The future of data as an application technology will follow four major trends:

Service Interface for Content Consumption

Current levels of storage devices provide basic data interfaces such as block, file, and object interfaces, which can further connect to Table format (databases), DataSet vector (training and inference), and asset interfaces (transaction applications). Next-gen data services and APIs, however, will go provide more advanced functionality, better performance, and secure access to data, and support advanced applications to leverage the higher power of data.

Such interfaces are aimed at helping engineering workloads. By simplifying pagination, data streaming, and event-driven architecture, the advances in custom queries, filtering, and programming provide features tailored to specific cases.

In most cases theses interfaces will be integrated with NLP technologies to provide services like ChatGPT, in which users can interact with AI to ask questions and receive relevant answers.

For business decision-makers, the data interfaces can use predictive analytics to provide insights and predictions based on historical data. This helps users identify patterns and trends in data that may not be visible through simple analysis. Instead, advanced visualization tech can help the users obtain a more comprehensive understanding of system data.

In summary, future storage will evolve from simple data access to content consumption.

Data Semantics Extraction

Data semantics applies smart AI techniques to extract target information related to service objectives from data sets. In doing so, the original data can be compressed and the system efficiency can be improved.

Current semantic extraction is developed on natural language, knowledge graphs, and deep neural networks, but is limited by interpretability, accuracy (inference from raw data), and scale of the deep neural network theory. Current techniques run poor generalization and limited deployment, requiring multiple trainings. High availability is based on intact semantics independent of software/hardware and cross-platforms differences, whereas portability requires a semantic scheme for complete data definition and description to drive the standardization and industrialization of data services. In the future, the NLP and pre-trained model technologies will catalyze new breakthroughs in semantics extraction and lossless semantic inference.

Multi-Modal Data Analysis

There is a growing trend of integrated multi-modal, supported with optimized sensory tech that streamlines the collection and processing of structured and unstructured data [26]. Consider the workloads involved in autonomous vehicles. Such systems must simultaneously aggregate and process multiple data sources from roads, traffic, in-vehicle sensor, and surrounding environments, realizing situational awareness from which informed decisions can be made.

At the same time, the multi-modal data processing must standardize and process data from different sources so that the data can be exchanged and shared between different applications. Data convergence in the future may have the following modes:

1.Multi-modal data convergence: Data of multiple types such as voice, image, and sensor data are converged and analyzed to obtain enough reliable information, solving the main problems common in a single data source.

2.Multi-layer convergence: Converged data from different layers is, such as bottom-layer sensory data and top-layer semantic information, converged and analyzed to

improve accuracy and depth of data analysis.

3.Multi-source data convergence: Data from multiple sources, such as social media, IoT, and enterprise internal systems, is converged and analyzed to improve data integrity and scope and to discover associations and relationships.

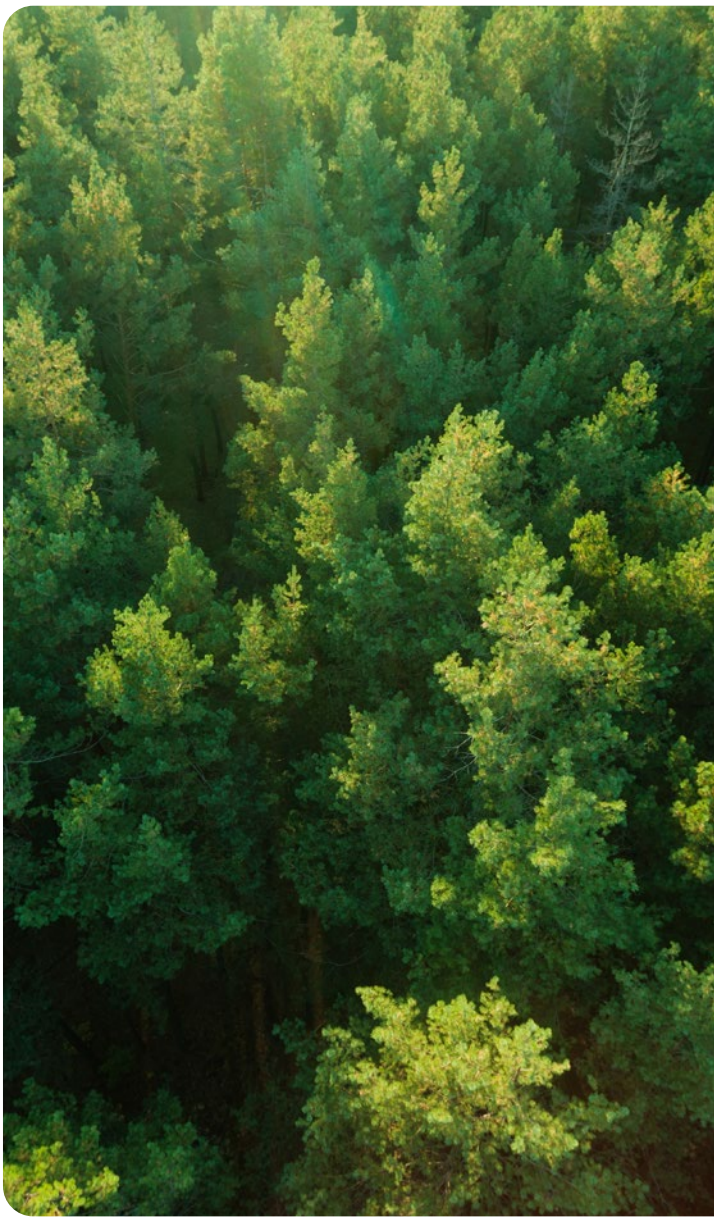
Current multi-modal data convergence and analysis is based on rule, feature, and semantic convergence algorithms, in coordination with machine and deep learning, computer vision, natural language processing, and sensor technologies. In the future, the multi-modal data convergence analysis will solve the problem of strong dependence on data homogeneous distribution and closed domains. Through spatial transformation, self-supervised learning technology, and AI-Generated Content (AIGC), the multi-modal data convergence analysis can implement cross-modal learning and automatically learn the semantic alignment relationship between modals to improve the precision of modal convergence.

Adaptive Data Modeling

Data Adaptive Modeling is an approach that identifies and learns patterns and structures from data as it is collected, and generates corresponding prediction models. Current models are hindered by inconsistent sampling sets and data drifts caused by differences in the application environment and training models. Data drifts mean that the old models cannot adapt to new environment and need to be retrained. In addition, the data adaptive model needs to quickly adapt to new environments and scenarios, for quick response and efficient prediction.

Currently, adaptive modeling is developed on neural networks and machine learning technologies, requiring dedicated network structures and features. In the future, breakthroughs in incremental and transfer learning, domain adaptation methods, along with Generative Adversarial Network (GAN) must enable adaptive modeling to suit a wider scope in complex and changeable deployments [27].

2.6 Sustainable Storage



It is estimated that by 2030, 4% to 6% of the global annual electricity output will be required to read global data once a month. The carbon dioxide generated from this must be absorbed by global trees in seven days. Therefore, in order to build a sustainable data infrastructure, we must reduce the energy consumption per bit of data read/write.

Based on the classic Von Neumann architecture, the energy consumption of data transmission between storage and computing units accounts for 60% to 90% of the total energy consumption of the IT system. The energy consumption problem of data-intensive applications is particularly prominent. However, data-centric architecture will solve the problem of high power consumption during data transmission.

In the future, technologies such as low-power-consumption media and transmission of data with optical signals instead of electronic signals will reduce energy consumption. Energy-saving technologies involving storage systems, entire devices, and environment-related solutions will further reduce carbon dioxide and improve energy consumption efficiency. They reduce energy consumption in terms of chips, media, and networks, achieving optimal energy efficiency per bit and minimum carbon emission.

Storage System-Level Energy Saving

These technologies detect the running status of computing, storage, and network devices to identify hot and cold data characteristics and service load rules used to construct a system optimization model. The model can be used to adjust software and hardware working status parameters to achieve optimal energy consumption of the entire system. The system-level energy saving technologies are as follows:

1.Intelligent power consumption optimization for hardware

Historical data analysis based on big data and AI reveals key factors that affect energy consumption, so as to obtain PUE prediction and energy saving benefit models. Optimization algorithms are employed to obtain optimization parameter groups and predict the optimization policy and total energy consumption of devices (such as CPUs, disks, networks, fans, and cooling pumps), to slash energy consumption of the entire system. Current solutions produce insufficient model generalization and samples, and non-real-

time performance, requiring much manual intervention and energy. In addition, AI models have poor explainability, leading to high operation security risks[28]. Future technologies such as modularization of models will be built on expert experience, probabilistic modeling will use fewer samples, online training/inference will be more efficient, and domain adaptation will reduce manual intervention workloads, which in turn improve model explainability and slash energy consumption.

2.Energy saving in data tiering

One issue is to reduce electric energy consumed by devices such as servers, storage devices, and network devices in non-working hours. Hot and cold data tiering stores data in magnetic-optical-electric hybrid media based on the data usage frequency, to effectively reduce energy consumption and balance performance and costs. Current data tiering policies and capacity planning are based on manual experience, wasting a lot of resources, and issues such as those related to I/O access modeling, data layout, and prefetch must be resolved. There is a need for AI-based refined models that can ensure performance and minimize energy consumption caused by data access.

3.Heat dissipation technology of storage devices

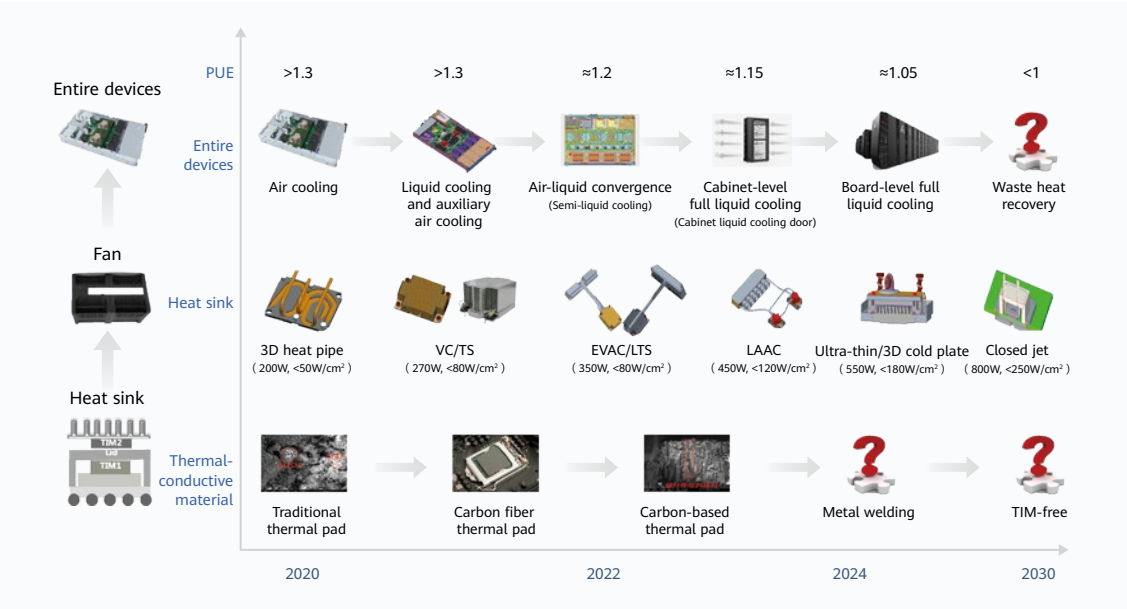


Figure 2-21 Heat dissipation technology of storage devices

Typically, cooling systems account for 30% to 60% of the total power consumption in a data center, making it a key area to slash carbon emissions, specifically by improving the entire-device heat dissipation system and reducing its power consumption[29]. Consider the air-cooled heat dissipation technology. It runs 5–10 COP but with huge noise generation, which has become an insufficient design for most environments. Future breakthroughs in technologies such as heat conduction-free materials on chips, closed jet impingement cooling designs, new cooling technologies, and entire-device waste heat recovery will improve heat dissipation efficiency and reduce carbon emissions across all hardware layers. However, issues must be addressed such as those related to the zero-thermal resistance welding, non-aqueous working substance and liquid-cooling material with high specific heat capacity and corrosion resistance, and electric machine conversion efficiency.

4.Resource awareness coordinated scheduling technology

Clean energy (e.g. solar, wind, water, geothermal, biological, and nuclear) does not produce any greenhouse gases. To reduce carbon emissions, large data centers are intensively deployed in Western China which is home to rich clean energy while small-scale ultra-low-latency edge data centers are deployed in Eastern China to support local services and reduce data migration. Data placement policies and cross-DC scheduling engines are needed to dynamically detect the location, status, availability, and heterogeneity of computing, network, and storage resources as well as regional resource pricing and carbon emission standards in real time, to achieve cross-DC data allocation. This achievement combines with intelligent data collaboration to build a global unified framework. This facilitates data extraction, analysis, and aggregation across DCs to achieve optimal computing, data

Data Transmission Energy Efficiency Improvement

Current network communication devices account for about 15% of the total energy consumed in a data center. Driven by new applications (AI and big data analysis), data centers will require higher transmission bandwidth, leading to higher energy consumption. As the transmission speed reaches 400G and even 800G, power consumption will become a bottleneck for improvement of network bandwidth. It is estimated that the electricity expenditure will account for about 95% of the annual operating expense of data centers in 2030, and networks will account for 20% of the total power consumption of data centers. Therefore, optimizing the energy efficiency of communications networks is a priority.

Mainstream data center network solutions use the optical-electrical-optical conversion process and electrical signal processing, which are the common areas of excess power consumption, making them the obvious starting points for energy savings. One solution is optical switching, which directly maps optical signals to outputs, requiring no optical-to-electrical conversion, to provide 10 TB-scale bandwidth, ns-scale latency, and TB-scale performance per watt. Current optical switching is based on the time switching technology, with the optical path switching latency as high as dozens of milliseconds. The optical-electrical hybrid technology can be used to build a high-throughput network, and breakthroughs will see nanosecond-level optical switching technology and high-speed switching algorithm, to achieve an all-optical data center network with low power consumption.

Chip-Level Energy Saving Technologies

Chips account for most of the energy consumption in current storage systems, making it critical to reduce the energy consumption of chips. As chip components are increasingly integrated, heat dissipation per unit volume is increasing. However, the limited heat dissipation speed of materials restricts chip performance. How to increase the chip computing power and control the chip energy consumption becomes a big challenge. Technologies such as heterogeneous and diversified computing power integration and on-chip dynamic intelligent energy efficiency management can realize both high computing power and low power consumption.

Chip-level energy saving technologies have the following research directions:

1.Low power-consumption raw material

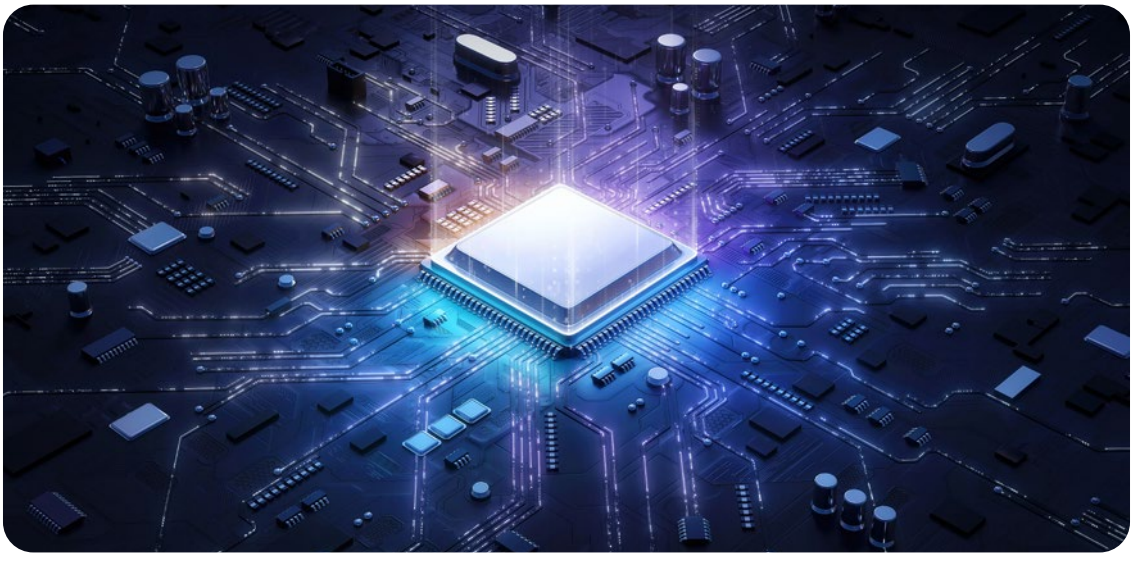
The chip integration density is expected to keep increasing with progress made as follows: emerging chip materials such as cold source structures, oxide materials, and carbon-based nanomaterials; packaging technologies such as 3D packaging and wafer level chip packaging; and low-power consumption technologies such as complementary field-effect transistors (CFETs).

2.High-density and low-power consumption processes

As chip components are becoming smaller and smaller, the energy consumption decreases accordingly. However, this classic physical law is no longer applicable on a nanometer/angstrom scale. In the future, DTCO\STCO technologies are expected to find an optimal chip design and lithography process, to ensure future solutions can integrate 100 billion transistors.

3.Chip energy consumption management

Technologies for on-chip energy consumption management can reduce energy consumption by controlling the chip voltage and clock frequency. Currently, the chip voltage and clock frequency are controlled by chips and maximized according to module requirements, leading to huge energy waste. The on-chip energy consumption



management technology can control the voltage and the clock frequency of the core sub-modules at the core level, to ensure proportional changes in energy consumption and computing power. Future-oriented AI and sensor technologies can be used to implement power prediction, power capping, and component power consumption control, achieving the optimal energy efficiency ratio at the component level.

4.Digital processing specialization

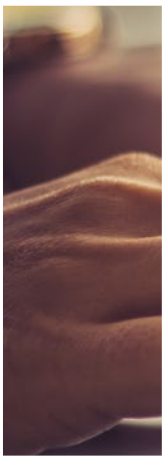
As Moore's Law slows down, the performance improvement of a single CPU is a bottleneck, while annual growth rate of computing power is less than 50%, and the gap between supply and demand is widening. As Dennard Scaling comes to an end, using multiple cores to improve computing power greatly increases energy consumption. The conventional general-purpose processor architecture cannot meet the development needs of diversified applications, requiring specific, tailored architecture designs to meet different computing power needs and achieve low system power consumption[30].

The current field-specific architecture provides diversified computing power through efficient parallel forms, hierarchical memory structures, hybrid precision, and field-specific programming languages. Due to differences in system architectures, instruction sets, and programming models, diversified computing power faces challenges such as difficult cross-platform program running and high programming complexity, fueling the need for technical breakthroughs in unified instruction sets, heterogeneous resource abstraction, efficient resource scheduling, and heterogeneous programming models. This will be the foundation for large-scale multi-system heterogeneous software platforms that integrate compilers, programming languages, acceleration libraries, and development tools.

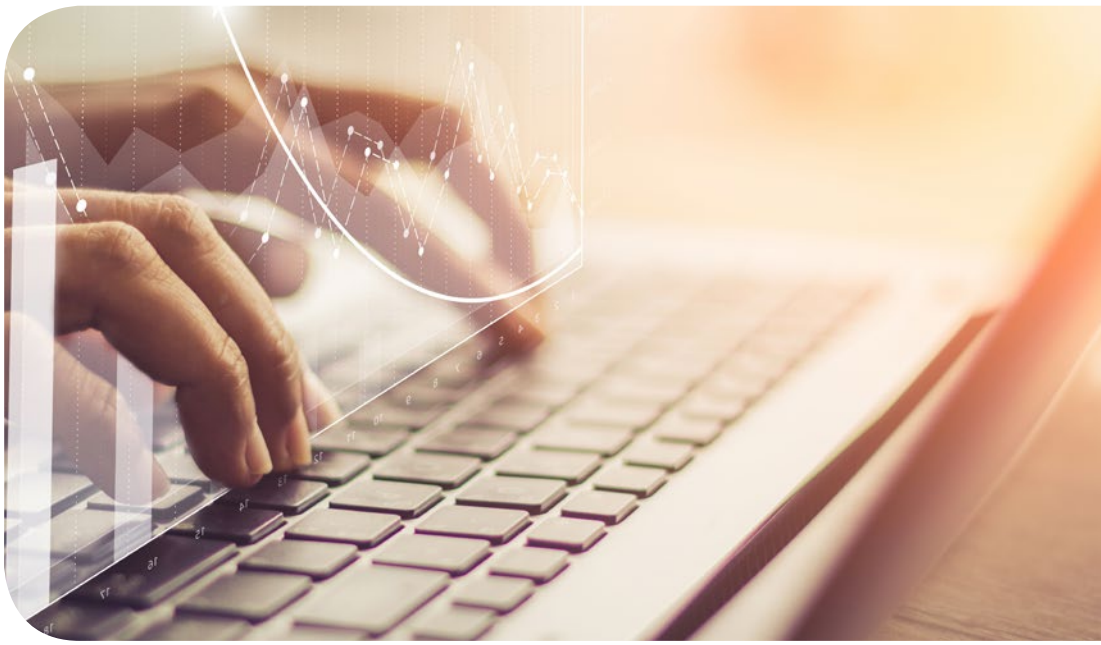
Green and Intensive Storage Standards

Data centers in China consumed about 270 billion kWh of electricity in 2022, representing an increase of 25% compared with 2021 and accounting for 3.1% of the total power consumption of the whole society. It is estimated that the data center energy consumption will double by 2030, resulting in high carbon emissions and heavy pollution. The storage industry urgently needs to gradually improve green and intensive storage regulations and standards in line with the national "carbon peak and carbon neutrality" strategy.

Current green storage standards for data centers cover energy efficiency simulation models, energy saving technologies, LCA carbon emission evaluation, carbon emission reduction and low carbon emissions, and recycling, but to date there is no unified standards for the storage industry. This needs to be addressed in the future to cover key indicators such as the carbon footprint throughout the data lifecycle, chip control interface, data transmission power consumption, and energy efficiency, carbon emission intensity, and renewable energy utilization of storage devices. Such achievements in energy consumption will create the benchmark from which a comprehensive evaluation system will be formed for a green, low-carbon storage industry.



03



Data Storage 2030 Initiative

Data storage is the cornerstone of digital infrastructure that supports the globalization of the digital economy. The industry will experience YB-scales of data by 2030, requiring collaboration to make breakthroughs to build better data storage. We hope that industry collaboration and innovation focus on the following:

1.Diversified media, innovative applications, and improved capacity density and per-bit energy efficiency.

2.Breakthroughs in system architecture that expand beyond the traditional von Neumann architecture, promotion of the construction of the data-centric system architecture, high-throughput P2P interconnection bus, and unified standards and protocols, as well as new-gen infrastructure.

3.Storage power development that boosts computing in storage, establishes a storage power indicator-based model based on the whole process of data processing, and improves efficiency.

4.Zero-trust storage systems that separate property, use, and control rights, establish unified standards for data gravity indexes, and streamline mobility of trusted data flows.

5.Green and intensive standard systems that feature optimal energy efficiency per bit and carbon emission must be built on sustainable IT. This will contribute to the shift from environmental-centered energy efficiency to IT-oriented and sustainable energy savings.

Let's work toward the new era of digital infrastructure together.

Appendix A: References

[1] Seagate and IDC, Data Age 2025, May 2020

[2] Gartner, "Forecast: Hard-Disk Drives, Worldwide, 2020-2026", 2022. https://www.gartner.com/document/4014430

[3] World Health Organization. World health statistics 2021: monitoring health for the SDGs, sustainable development goals. 2021. https://apps.who.int/iris/handle/10665/342703

[4] Deloitte China, Digital health whitepaper, 2021

[5] 2030 Sustainable Development Goals in China, SDG China, http://sdgcn.org/sdg2.html

[6] realtor.com, Smart Home Technologies Reshape Real Estate Preferences in 2020, https://www.realtor.com/research/smart-home-tech-2020/

[7] World Economic Forum, Raising Ambitions: A new roadmap for the automotive circular economy, 2022, https://www3.weforum.org/docs/WEF_Raising_Ambitions_2020.pdf

[8] IDC, Worldwide Smart Cities Spending Guide, 2021

[9] Korn Ferry, Future of Work---The Global Talent Crunch, 2018, https://www.kornferry.com/content/dam/kornferry/docs/pdfs/KF-Future-of-Work-Talent-CrunchReport.pdf

[10]United Nations Environment Programme, Emissions Gap Report 2020, 2020, https://www.unep.org/emissions-gap-report-2020

[11]Abbosh O., Bissell K., Reinventing the Internet to Secure the Digital Economy, 2019, https://www.accenture.com/_acnmedia/thought-leadership-assets/pdf/accenture-securing-the-digital-economy-reinventing-the-internet-for-trust.pdf

[12]Hennessy, John L. and Patterson, David A., Computer Architecture, Fifth Edition: A Quantitative Approach, Morgan Kaufmann Publishers Inc., 2011

[13]China Academy of Information and Communications Technology (CAICT), Data Elements White Paper, 2022

[14]Gartner, HDD and SSD market forecast, 2021

[15]Yang S, Zhang J. Current Progress of Magnetoresistance Sensors. Chemosensors, 2021

[16]Takeshi H., Hitoshi N. A study on high-density recording with particulate tape media for data storage systems, Synthesiology, 2017

[17]SONY, Optical disc archive generation 2 white paper, 2016

[18]Yuan X., Zhao M., Guo X., Li Y., Gan Z. and Ruan H., Optical tape for high capacity three-dimensional optical data storage, Chinese Optics Letters, 2020

[19]Shu Jiwu, Outlook for a new storage-compute decoupling architecture technology, Communications of the China Computer Federation, 2022

[20]Fan Dongrui, Ye Xiaochun, Bao Yungang, Sunninghui, The road to self-developed high-throughput computer of China, HPC Development Strategy of China, 2019

[21]Conte T. M., DeBenedictis E. P., Gargini P. A. and Track E., Rebooting Computing: The Road Ahead, IEEE Computer Society Press, 2017

[22]Wu Jiangxing, Cyberspace Endogenous Security Development Paradigm, China Science: Information science, 2022

[23]Yin X. F., Zhu Y. M., Hu J. K., A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions, ACM Computing Surveys, 2022

[24]Gupta A., Key Pillars of a Comprehensive Data Fabric, Gartner, 2021

[25]Microsoft Azure, Knowledge store in Azure Cognitive Search, 2023, https://learn.microsoft.com/zh-cn/azure/search/knowledge-store-concept-intro?tabs=portal

[26]Baltrušaitis T., Ahuja C., Morency L. P., Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

[27]Wilson G., Cook D. J., A Survey of Unsupervised Deep Domain Adaptation, Association for Computing Machinery, 2020

[28]Yu Y., Wu C., Zhao T., OPU: An FPGA-based overlay processor for convolutional neural networks, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2020

[29]State Information Center, Research Report on Green and High-Quality Development of Data Centers, 2022

[30]Hennessy J. L., Patterson D. A., A New Golden Age for Computer Architecture, Communications of the ACM, 2019

Appendix B: Acronyms and Abbreviations

Acronyms and Abbreviations	Full Spelling
AIoT	Artificial Intelligence of Things
CBA	CMOS-bonded Array
CFET	Complementary Field-Effect Transistor
CIM	Computing In-Memory
CMOS	Complementary Metal-Oxide-Semiconductor
CNA	CMOS Near Array
COP	Coefficient of Performance
CUA	CMOS Under Array
CUE	Carbon Use Efficiency
DNA	Deoxyribonucleic Acid
DPU	Data Process Unit
DRAM	Dynamic Random Access Memory
DTCO	Design-Technology Co-Optimization
EDA	Electronic Design Automation
HAMR	Heat-assisted Magnetic Recording
HDD	Hard Disk Drive
HPDA	High Performance Data Analytics
IoT	Internet of Things
LCA	Life Cycle Assessment

Acronyms and Abbreviations	Full Spelling
LTFS	Linear Tape File System
LTO	Linear Tape-Open
MAMR	Microwave Assisted Magnetic Recording
MRAM	Magnetoresistive Random-Access Memory
PB	Petabyte
PLC	Penta-Level Cell
PUE	Power Usage Effectiveness
QLC	Quad-Level Cell
SCI	Storage Compute Integrated
SCM	Storage Class Memory
SSD	Solid-State Drive
STCO	System Technology Co-Optimization
STT-MRAM	Spin-Transfer Torque MRAM
TEE	Trusted Execution Environment
UB	Unified Bus
Wafer Level	Wafer Level
YB	Yottabyte
ZB	Zettabyte
ZB	Zettabyte

Appendix C: Acknowledgment

We express our heartfelt gratitude to the numerous experts from Huawei and over 100 distinguished scholars from various fields who actively engaged in the discussion and shared their insights towards shaping the future of the data storage during the compilation of Data Storage 2030. Your invaluable input is instrumental in identifying the development direction and technical characteristics of the sector. Thank you for your valuable contribution.

(This list is sorted by the initial letter of the scholar's name.)

Bao Yungang (Researcher Fellow, Institute of Computing Technology, Chinese Academy of Sciences)

Cui Heming (Associate Professor, the University of Hong Kong)

Chen Mingyu (Researcher Fellow, Institute of Computing Technology, Chinese Academy of Sciences)

Feng Dan (Distinguished Professor of the Changjiang Scholars Program, Huazhong University of Science & Technology)

Gu Rong (Distinguished Research Fellow, Nanjing University)

Guo Minyi (Professor, IEEE Fellow, Shanghai Jiao Tong University, Member of Academia Europaea)

Huang Qin (Professor, Beihang University)

Jiang Dejun (Associate Researcher, Institute of Computing Technology, Chinese Academy of Sciences)

Jin Hai (Distinguished Professor of the Changjiang Scholars Program, IEEE Fellow, Huazhong University of Science & Technology)

Li Yi (Associate Professor, Huazhong University of Science & Technology)

Liu Xianming (Professor, Harbin University of Technology)

Lu Youyou (Associate Professor, Tsinghua University)

Miao Xiangshui (Professor, Huazhong University of Science & Technology)

Ren Kui (Professor, ACM Fellow, IEEE Fellow, Zhejiang University)

Shu Jiwu (Distinguished Professor of the Changjiang Scholars Program, IEEE Fellow, Tsinghua University)

Tang Zhuo (Professor, Hunan University)

Wang Cong (Professor, City University of Hong Kong)

Wang Zeke (ZJU 100 Young Professor, Zhejiang University)

Wang Zhaoguo (Associate Professor, Shanghai Jiao Tong University)

Wu Hequan (Academician of the Chinese Academy of Engineering)

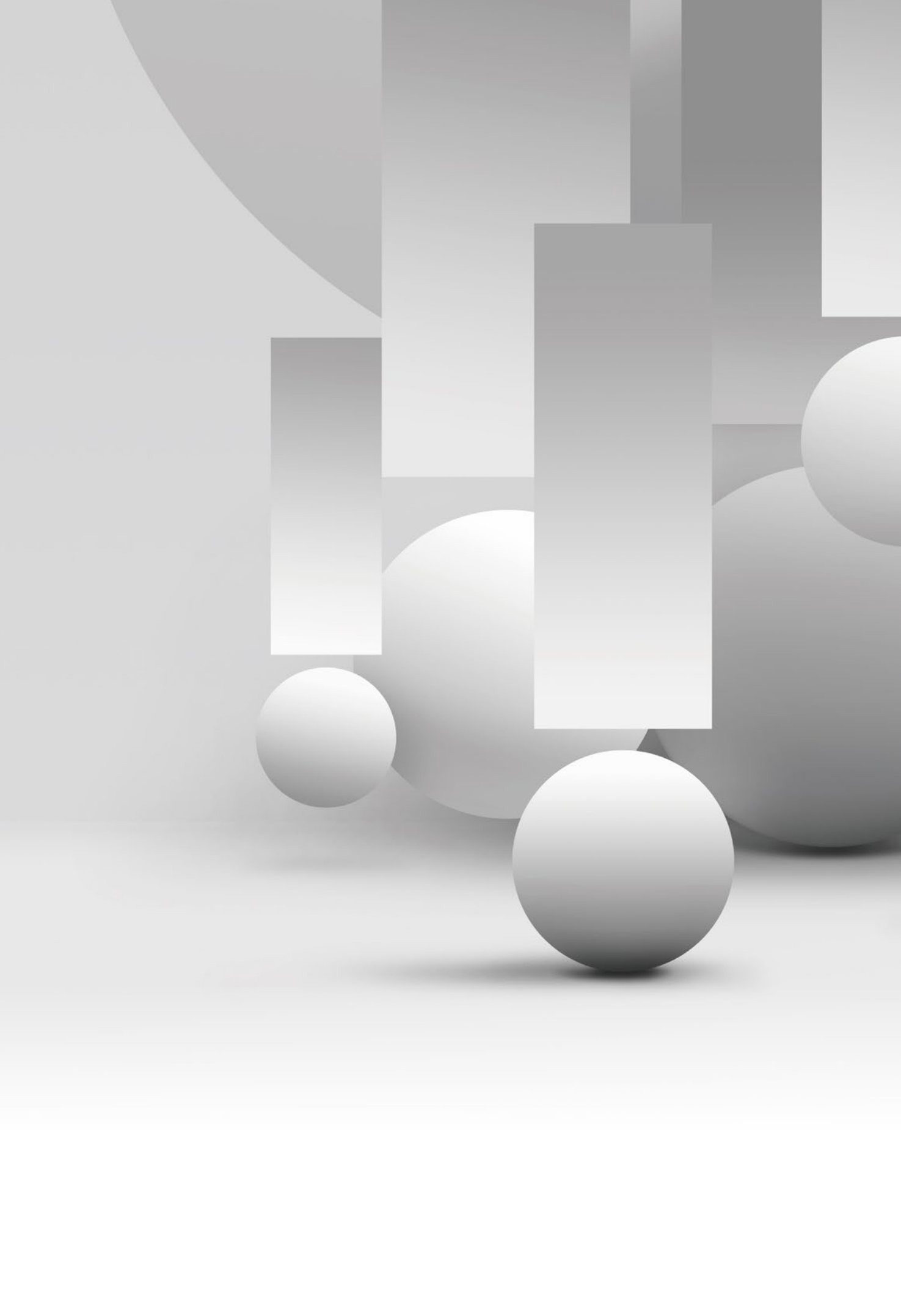
Xie Changsheng (Professor, Huazhong University of Science & Technology)

Zhao Shizhen (Associate Professor, Shanghai Jiao Tong University)

Zhou Ke (Distinguished Professor of the Changjiang Scholars Program, Huazhong University of Science & Technology)

Notes on the update:

Huawei collaborates with industry experts, customers, and partners to explore the intelligent world. The progress towards an intelligent world has accelerated significantly, with new technologies and scenarios emerging constantly, and industry-related parameters changing exponentially. As a result, Huawei has updated the *Intelligent Automotive Solution 2030* report released in 2021, providing insights into the scenarios and trends towards 2030, and adjusting the relevant forecast data.

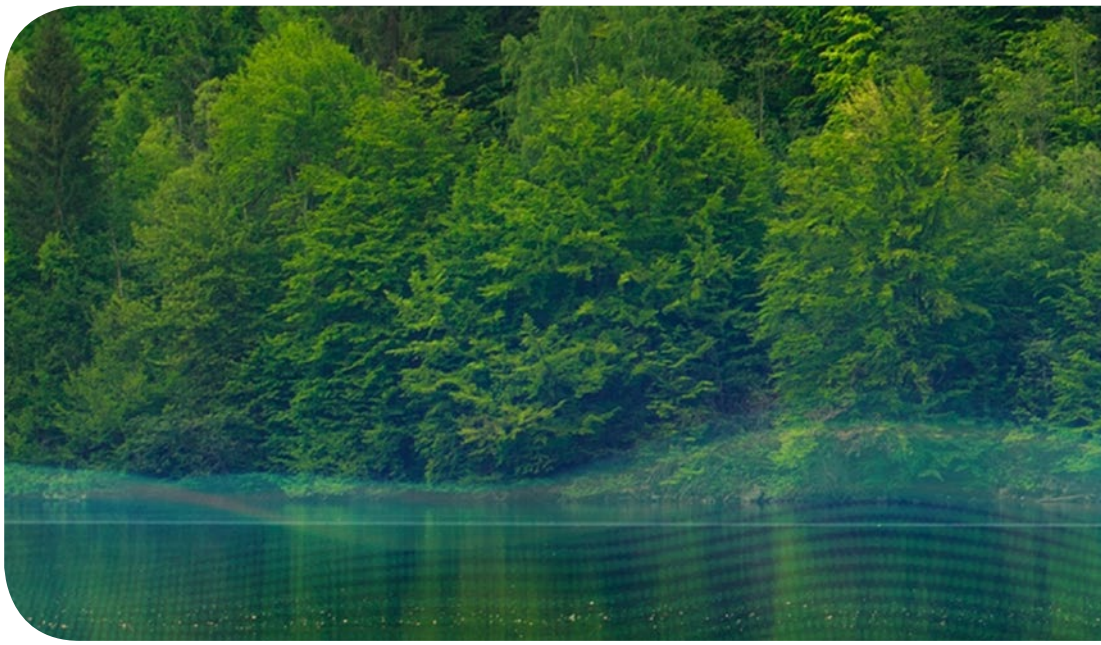


HUAWEI

Version 2024

Digital Power 2030

Building a Fully Connected, Intelligent World



01

From the Paris Agreement to COP28 UAE, the global community is accelerating its journey toward carbon neutrality

COP 28 marked the beginning of the end of the fossil fuel era; more than 150 countries have pledged to cut carbon emissions

Since the 18th century, coal, oil, and electricity have been extensively utilized, each playing crucial roles in the first and second industrial revolutions. These energy sources facilitated the transition from an agricultural society into the industrial economy. As a cornerstone of global economic development, energy has consistently driven social progress, reduced poverty, and improved peoples' livelihoods.

However, human activities have clearly impacted the planet's ecosystem with greenhouse gas (GHG) emissions reaching record highs in recent years. According to the United Nations' Intergovernmental Panel on Climate Change (IPCC), human activities generate approximately 23.7 billion tons of carbon dioxide (CO_2) annually, with around 20 billion tons resulting from fossil fuel combustion. As a result, the amount of CO_2 in the atmosphere now is 27% higher than its average level over the past 650,000 years. The extensive burning of coal during the industrial revolution has resulted in a spike in CO_2 levels, putting our ecosystems at unprecedented risk and contributing to severe ecological and economic imbalances. This has prompted discussions on reducing fossil fuel use to lower GHG emissions.

Fortunately, a clearer consensus has been reached among the scientific community and governments on climate change, and the Paris Agreement, signed in 2015, specifies that our most important goal in the global fight against climate change is to achieve carbon neutrality by the middle of the century. The energy development strategies and practices adopted by major economies around the world have proven that reducing our reliance on fossil fuels is one of the best ways to achieve the carbon reduction goals. This requires countries to step up efforts to develop renewables while simultaneously improving energy efficiency and reducing overall consumption of fossil fuels. Multiple countries have put forward targeted energy reform and GHG control goals. At the COP28 UN Climate Change Conference held in Dubai, the United Arab Emirates at the end of 2023, multiple countries and regions reached a consensus on accelerated actions to reduce GHG emissions by 2030. These actions include transitioning from fossil fuels to renewables such as wind and solar, tripling the installed capacity of renewables and doubling the energy efficiency globally by 2030, and phasing out fossil fuels.

As of the first half of 2024, more than 150 countries had pledged to reduce carbon emissions. For instance, China's National Development and Reform Commission and National Energy Administration released the Energy Production and Consumption Revolution Strategy (2016–2030), which specifies that by 2030, China's new energy demand will predominantly be met by clean energy. The strategy proposes reducing total energy consumption to below 6 billion tons of coal equivalent (TCE), with non-fossil fuel making up about 20% of the total primary energy supply by 2030. China has also pledged to achieve its CO_2 emissions peak by 2030, if not sooner. The EU's 2030 climate and energy framework aims for net GHG emissions reductions of 55% compared to 1990 levels and an increase in renewable energy consumption to 38–40% by 2030. The US government has also pledged to achieve a 50–52% GHG emissions reduction from 2005 levels by 2030, and one of the most important steps to achieve that goal is to require the US grid to obtain 80% of its electricity from emission-free sources by that year as well.

Global sustainable economic growth requires sustainable energy supplies and renewables will become the most important source of energy

Population growth and national industrialization have driven energy demand to unprecedented levels. Since commercial oil drilling began in the 1850s, experts estimate that the world has harvested more than 135 billion tons of crude oil, with that figure increasing every day. Currently, the global annual consumption of primary energy amounts to approximately 14 billion tons of oil equivalent, which consists of more than 85% fossil fuels. This means that fossil fuels will soon dry up. According to BP Energy Outlook, we will run out of global oil, gas, and coal resources in about 54, 49, and 139 years, respectively, if our current extraction and consumption patterns do not change. This underscores the necessity of advancing renewable energy sources to ensure sustainable development.

According to Goal 7, set in the United Nations 2030 Agenda for Sustainable Development, adopted at the seventieth session of the United Nations General Assembly, the following targets are to be achieved by 2030: ensuring universal access to affordable, reliable, and modern energy services; increasing substantially the share of renewables in the global energy mix; doubling the global rate of improvement in energy efficiency; enhancing international cooperation to facilitate access to clean energy research and technology, including renewables, energy efficiency, and advanced and cleaner fossil fuel technology; promoting investment in energy infrastructure and clean energy technology; and expanding infrastructure and upgrading technology for supplying modern and sustainable energy services for all in developing countries, particularly least developed countries, small island developing states, and land-locked developing countries, in accordance with their respective programs of support.

Countries around the world are making the development of renewables an important part of their future energy strategies. Numerous countries have formulated specific strategies, plans, targets, regulations, and policies to support the development of renewables. In 2023, the Indian government released the latest national electricity plan, which unequivocally states that the accumulated installed capacity of renewables will reach 336.6 GW by 2026–2027 and 596.3 GW by 2031–2032. The Vietnamese government predicts that renewables will account for 30.9%–39.2% of the country's power supply by 2030 and 67.5%–71.5% by 2050. The Malaysian government announced a new renewable energy development target, aiming to increase the share of renewables in the country's electricity mix to about 70% by 2050. The United Arab Emirates plans to triple its renewable energy production by 2030, with approximately US$55 billion earmarked for investment in renewables. The Italian government has boosted the country's 2030 renewable capacity goal from 80 GW to 131 GW. The Portuguese government plans to raise the installed capacity of renewables from 27.4 GW to 42.8 GW by 2030. In September 2023, the European Parliament approved a proposal to promote the deployment of renewables. According to the proposal, the share of renewables in the final energy consumption across the EU will increase from 32% to 42.5% by 2030. Additionally, EU countries are urged to strive for a 45% share. We expect that renewables will contribute 65% of global electricity generation by 2030.

Cost-effective wind and solar power will have a share of 70% in renewables by 2030 thanks to rapid development

Fossil fuels continue to dominate the global electricity supply due to their cost competitiveness compared to other energy sources. If we want to transition to a truly decarbonized energy system that primarily relies on renewables, we must ensure that renewables are cheaper than fossil fuels. As an alternative, the global renewable energy industry has emerged as a promising new market in recent decades. Many countries have made wind and solar power generation part of their new energy strategies, and invested significantly in R&D and industrial development in these areas.

Driven by technological innovation, wind and solar power generation is also growing increasingly affordable. Oxford University's Max Roser found that the levelized cost of electricity (LCOE) of utility-scale photovoltaic (PV) plants was US$0.36/kWh in 2009 and that within just one decade the price had declined by 89% to US$0.04/kWh. However, electricity from fossil fuels, especially coal, is not getting cheaper. Coal-burning power plants have a maximum efficiency of 47%, often leaving little room for substantial efficiency improvements. The price of electricity from fossil fuels is also not only based on the cost of technology itself but, to a significant extent, the cost of the fuel. The cost of coal that the power plants burn accounts for around 40% of total costs. This means that even if the cost of constructing a power plant declines, the price of the electricity it generates will not continue to drop until it reaches a certain point. However, each time the cumulative installed capacity doubles, the price of PV modules declines by 20.2%. The LCOE of PV power will continue to drop as new PV module technologies and processes mature.

In addition to these cost benefits, wind and solar power generation is more flexible than traditional fossil fuel power plants. Resource endowments have long influenced domestic energy development and utilization. However, as wind and solar are becoming the preferred new renewable energy sources, they can transcend the limits of resource endowments and produce electricity anywhere as long as their relevant requirements are met. For example, distributed PV has attracted investors from many industries due to its low investment threshold. As wind and solar power generation becomes more affordable and flexible, more users are willing to use distributed PV systems in campuses, industrial complexes, and commercial and industrial (C&I) scenarios, changing how energy is produced and utilized around the world. Offshore wind power is an important type of wind power that occupies zero land space. The power generated from offshore wind turbines is directly delivered to coastal load centers nearby, avoiding the waste that long-distance transmission causes. Because of this, we are currently seeing a shift from onshore wind farms to offshore wind farms. The proportion of distributed PV systems keeps increasing, with C&I distributed PV systems holding a major market share. Floating PV plants have become popular in many regions because they offer larger power generation capacity, no land requirements, and a lower impact on water bodies. According to IRENA, by the end of 2023, the global accumulated installed capacity of wind power and solar power had exceeded 1,000 GW and 1,400 GW, respectively. We expect that the accumulated installed capacity of solar power will approach 6,000 GW by 2030.

02

Technologies will drive clean energy development and expedite green transitions across industries

New energy systems based on power electronics equipment will drive the future transformation of the energy industry

Power electronics technologies play a key role in electricity generation, distribution, transmission, and consumption. As more electricity is generated from renewables such as wind and solar, energy industry transformation efforts will focus on building an energy system that will be centered on electricity, connected to power grids, and based on power electronics equipment. The inclusive interfaces, fast response times, and high conversion efficiency of power electronics devices are already being widely implemented in electric power generation, transmission, and consumption.

Electric power generation: Power generation systems using renewables, such as wind and solar, cannot directly transmit electricity to local grids like conventional electric generators. Power from these renewables first needs to be converted into frequency-adjustable AC using power electronics technologies to meet the grid transmission requirements. For example, PV inverters and wind power converters can adjust voltage waveforms through power electronics switches to enable the transmission of renewable electricity to local grids, making power generation more efficient.

Electric power transmission: Intelligent power electronics equipment can significantly enhance long-distance power transmission performance, optimize power flow distribution, and improve the reliability of power supply. This strengthens the reliability of electrical grids, thereby making power

transmission over large-scale grids more secure and efficient.

Electric power distribution: Large numbers of distributed power supplies, microgrids, and flexible loads are being connected to power distribution networks, increasing the requirement for plug-and-play power supply and the overall amount of reactive power in the transmission lines. Problems such as voltage spikes and harmonic distortion are also becoming increasingly serious. There are limited ways to improve the power quality and supply stability of traditional distribution networks, meaning these networks alone can no longer meet user requirements for high-quality electricity. Power electronics equipment, such as multi-functional power electronics transformers, DC circuit breakers, and DC switches, can instead be used to guarantee the power quality of different load categories and meet tailored electricity needs.

Electric power consumption: The demand for DC power and proactive source-load interactions is increasing due to the application of distributed power supply and energy storage devices, and the emergence of new types of facilities, such as data centers, communications base stations, electric vehicle (EV) charging stations, computers, and LED lights. Switching power supplies and switchgears with high efficiency, high power density, high reliability, and low cost are meeting the increasingly diverse personalized needs of users and the demand for quality assurance of electric power.

Demand for new types of power semiconductor devices is set to skyrocket. Future energy systems will need to optimize renewable energy resources and maximize energy efficiency. Consequently, the standards for energy transmission and control subsystems will rise significantly in terms of safety, efficiency, and intelligence. We will need entirely new electricity transmission and distribution networks designed specifically for renewables, more efficient terminal systems that work better with other subsystems like distributed power supply and energy storage, and more comprehensive service systems that are integrated with information systems. Changes introduced by these new systems will need to be managed, compensated, or controlled by power electronics equipment, which currently rely on silicon-based components to a large extent. However, the reality is that silicon-based components are going to hit a wall soon. The physical properties of silicon mean that there will no longer be a way to further improve performance. Many are already struggling

to further reduce the energy use of silicon-based components. These components simply won't be a good fit for generating, transmitting, consuming, and absorbing clean energy at scale in future energy systems. Third-gen power semiconductor chips and components, based on silicon carbide (SiC), stand out for their high voltage, frequency, temperature, and speed. These SiC components have delivered a huge boost to the reliability, availability, energy density, and energy conversion efficiency of power electronics equipment while simultaneously reducing overall cost and energy loss. SiC components will be widely adopted as they are ideal for sectors with high requirements on energy conversion efficiency, such as electricity generation based on renewables (e.g., solar and wind), ultra high-voltage direct current electricity transmission, new-energy vehicles, rail transportation, industrial power supplies, and home appliances. The high uptake of SiC components in new-energy vehicles, industrial power supplies, and other domains will also help drive down costs. And there will be a new wave of technologies that will enhance the performance and reliability of SiC components. These trends will prime the SiC sector for explosive growth and market development. In 2023, the market value of SiC components reached US$2 billion. McKinsey estimated that the market for SiC components will expand rapidly, reaching US$10 billion to US$14 billion by 2030. It is estimated that by 2030, over 70% of solar inverters will use SiC components. By then, SiC components will most likely be found in more than 80% of the charging infrastructure and EVs, and be widespread in the power systems of communications networks and servers.

Digital technologies will drive intelligent transformation of energy systems, making renewables smarter, safer, and more efficient

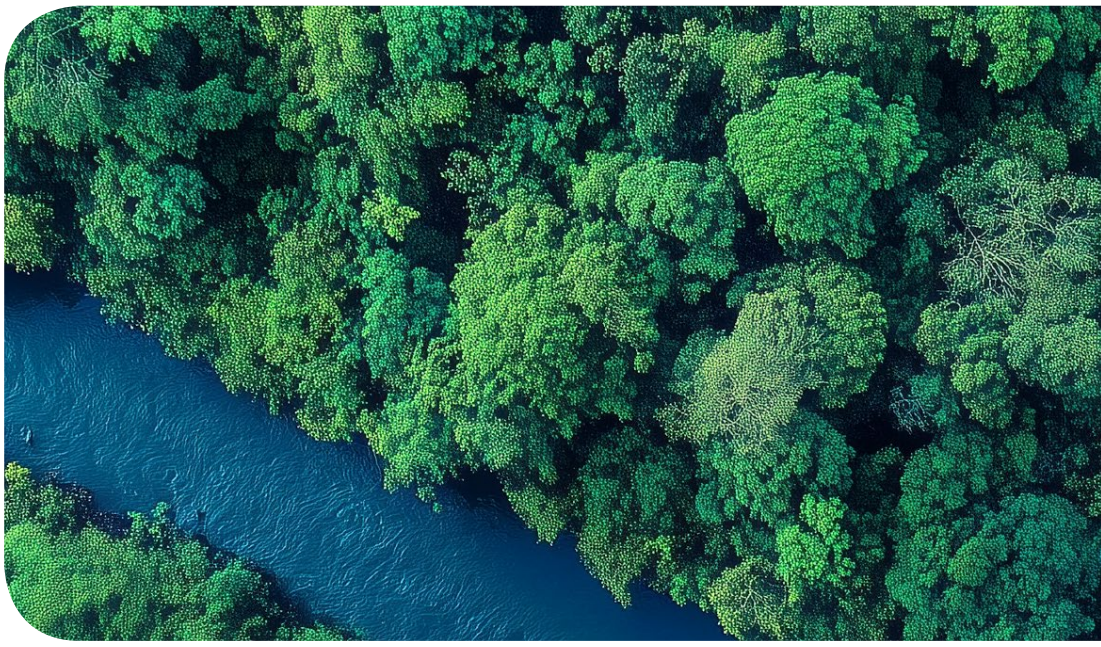
Energy systems will soon become more distributed, thanks to the rapid increase in renewable energy installations (e.g., wind and solar) and the increasing flexibility of applications that support these systems. Energy systems of the future will be decentralized, just like a nebula, with ecosystems made of numerous distributed energy applications. Large power plants, campuses, buildings, households, EVs, and countless other facilities will also develop their own energy systems. These distributed energy systems will not be sustainable if they rely on traditional models. Intelligent connectivity and control powered by digital technologies will make energy systems highly intelligent and connected, which in turn will make them safer and more stable, efficient, affordable, and flexible. They will then be better positioned to reduce carbon emissions and generate clean energy more efficiently.

Advances in emerging technologies, particularly 5G, cloud, AI, big data, and IoT, are ushering all sectors of society into a new digital era. This will be an era where all things can sense, connect, and work intelligently. This vision of ubiquitous connectivity and pervasive intelligence is already becoming a reality. The following new digital technologies are being adopted in the energy sector at an increasing pace and will soon become game changers: Networking: Low-power wide-area networks (LPWANs) are being rapidly commercialized around the world. With wide coverage, low latency, and massive connectivity, 5G is ideal for IoT applications and is permeating a growing array of scenarios that require on-demand, intelligent connections between people, machines, and things. Information processing: Information perception, knowledge representation, and machine learning technologies are advancing rapidly, driving IoT's ability to intelligently process data to levels we have never seen before. IoT virtual platforms, digital twins, and

OSs: Widespread adoption of cloud computing and open-source software is reducing the entry barriers for those who hope to play a part in the energy sector. Cloud computing and open-source software are also boosting the popularity of energy system OSs and digital ecosystems.

As distributed energy systems continue to grow in popularity, users will become prosumers – those who both consume and produce energy. Highly intelligent energy systems can flexibly determine when to switch from generating electricity (when energy prices are high) to storing electricity (when energy prices are low). The systems will leverage generation-grid-load-storage synergy to facilitate energy dispatching and complementation, and they will be able to transfer energy flow to and from each other across time zones and across vast distances. EVs will be able to double as energy storage facilities that feed electricity back into the power grid during peak hours to help meet demand. Data centers will be able to provide heating by reclaiming huge amounts of the residual heat they produce. Intelligent devices for homes will become endpoints that detect, meter, and trade electricity. Distributed energy, energy storage, and the electricity spot market will thrive. There will be an untold number of prosumers that have resources aggregated and controlled through virtual power plants (VPPs), and they will enable energy systems to better respond to demand and provide value-added services. The proportion of PV plants using AI technologies is expected to reach 90% by 2030.

03



The three pillars of new energy infrastructure will become the foundations for the intelligent era

In the next decade, renewables like wind, solar, and hydro power will replace fossil fuels as the primary sources of electricity. The electrification of power consumption is also on the rise. Technologies for EVs, hydrogen energy, energy storage, heat pumps, and thermal energy storage are advancing rapidly. Transportation, heating, and other energy consuming systems are rapidly transitioning from diesel, petrol, natural gas, and coal towards electricity. Energy systems will soon be embedded with more advanced plug-ins, and be supported by an energy cloud operating system (OS) that integrates information flows and energy flows. The connection between electricity production and consumption will enable two-way, Internet-based interactions among various industry players, encompassing everything from energy sources and power grids to load management, energy storage, and consumption.

Transforming energy systems will unlock vast opportunities for innovation in technology, business models, and operations throughout the energy sector. New power system infrastructure involving renewables, like solar power, will present numerous opportunities. Similarly, the electric mobility sector, primarily driven by EVs, and other new digital industry energy infrastructure sectors, particularly ICT energy infrastructure, will also see significant growth.

In terms of energy generation, it is predicted that renewables will generate more than 65% of the global electricity by 2030, the LCOE of PV power will be as low as US$0.01/kWh, and the global installed capacity of PV power will approach 6000 GW. In terms of energy consumption, the electrification rate will reach 30%, the annual energy charged to EVs will exceed 1.1 trillion kWh, and more than 80% of ICT energy infrastructure will be powered by green energy sources.

Energy infrastructure of new power systems will focus on clean energy, and power generation, grid, load, and storage will be preliminarily integrated

1) As grid parity extends from PV to PV+ESS, the PV industry is entering the PV+ESS era

LCOE is a measure of the average net cost of electricity generation for a PV plant throughout its lifecycle. It is used to compare the electricity generation costs of PV plants with other types of plants. Under a full-lifecycle investment model, the LCOE is determined by a plant's upfront investment, operation & maintenance (O&M) expenses, and the operational lifespan of the system. By 2030, the LCOE of PV plants is expected to plummet, possibly even down to US$0.01/kWh.

PV plants are composed of PV modules and balance of system (BOS) components (such as electrical cables and solar inverters). Generally speaking, about 45% of a PV plant's investment goes into its PV modules. Over the next decade, this percentage is expected to decrease by at least 15 points due to advancements in engineering techniques, reduced manufacturing costs, and the continually increasing efficiency of PV modules. This means more investment will go to BOS components and O&M. On top of this, technological innovations will drive up the overall cost competitiveness of PV plants.

With the rapid advancement of battery and system technologies, the levelized cost of storage (LCOS) for ESSs is decreasing, positioning ESSs to play a significant role in electricity regulation resources within power systems. AS grid parity extends from PV to PV+ESS, the PV industry has entered the PV+ESS era. The ESS industry is thriving thanks to the significant benefits ESSs provide in business scenarios such as renewable energy consumption, power grid peak shaving, and time-of-use arbitration. Simultaneously, other technologies such as long-term energy storage are also rapidly emerging. In China, the proportion of installed capacity for pumped storage decreased from 97% in 2016 to 67% in 2023. Meanwhile, the proportions of lithium-ion batteries, sodium-ion batteries, flow batteries, flywheel energy storage, and hydrogen energy storage have been increasing. It is estimated

that by 2030, the annual newly installed capacity of global energy storage will increase from 46 GW to more than 140 GW. As fossil energy generators are gradually phased out, long-term energy storage will play a crucial role in energy regulation for new power systems.

2) Grid-forming technology helps renewables become primary electricity sources

Grid-forming technology has the potential to make power grids more resilient. Due to the significant fluctuations in PV power generation, it can only meet the energy dispatching demands of power grids with the support of regular power supply services such as peak shaving and backup. Consequently, as more electricity generated from wind and solar energy is fed to a power grid, the grid itself will become more vulnerable. For example, the power grid's system inertia may drop, and its ability to regulate frequencies and control system voltage may be compromised. What's more, the characteristics of faults and oscillations on the power grid may change significantly.

Effectively integrating wind power and PV power generators into power grids and harmonizing operations is key to incorporating large amounts of renewable energy into power grids and changing the energy mix. In a power grid, fossil fuel power plants and hydropower plants typically use conventional synchronous generators. These generators employ mechanical structures to provide stable voltage and frequency, facilitating frequency regulation and voltage control. However, as asynchronous generators gradually replace synchronous generators, the fundamental operation of power systems will change. Consequently, renewables-based power systems will need to simulate the technical indicators of synchronous generators to proactively support the grid's frequency and voltage fluctuations. The goal is to enhance the safety and reliability of power grids.

The grid-forming technology combines power electronics, energy storage, and digital technologies to simulate the electromechanical transients of synchronous generators. When connected to power

grids, the grid-forming generators have many of the same external characteristics of synchronous generators, such as inertia, damping, primary frequency regulation, and reactive voltage control. As a result, the grid-forming generators can offer technical specifications that are similar to the synchronous generators used in fossil fuel power plants. The grid-forming generators can proactively support the operations of renewables-based power systems and make them more grid-friendly. This will help renewables become mainstream and provide a solid technical foundation for incorporating them into power grids.

3) Rapidly advancing digital and AI technologies are being widely implemented to develop smart power systems

With the ongoing trends of decarbonization and electrification, power systems are becoming increasingly complex. They now involve trillions of measurement points, tens of thousands of tegawatts of energy for trading, and hundreds of millions of devices. Consequently, there is a growing need for enhanced computing, second-level fast scheduling, and multi-energy comprehensive optimization. In this context, AI has emerged as a crucial technology for energy transformation.

Integrating digital technologies into PV power plants offers a simple, smart, and efficient way of O&M, production management, and asset management, making otherwise "dumb" power plants significantly smarter. AI will play an expert role in enabling PV power plants to achieve autonomous collaboration and optimization. By predicting weather changes and using the smart tracker control algorithm, PV modules, module supports, and solar inverters can collaborate to find the optimal irradiance angle and maximize energy yields. AI can also accurately locate faults, reducing the O&M workload per person from months to minutes, comprehensively improving energy generation efficiency and reconstructing O&M experience to help improve plant productivity and safety. It is estimated that 90% of PV power plants will use AI technologies by 2030.

For power grids, AI algorithms are used to accurately predict the feed-in power and load consumption, improving energy scheduling efficiency. AI models are applied to inspect power transmission lines, improving the operation efficiency 80-fold and significantly reducing the power outage duration.

For energy consumption, AI technologies are adopted in comprehensive energy efficiency management, enhancing the green power application efficiency by more than 15%. In VPPs and electricity trading markets, AI agents can provide optimal solutions for trading entities through swarm intelligence and game intelligence.

4) Energy clouds will intelligently converge energy and information flows to synergize generation, grids, loads, and storage

Energy clouds that converge energy and information flows will function as the OS of the digital power industry. They will direct information flows, regulate energy flows, and spark an energy revolution in which bits can be used to manage watts. In the future, electricity will be the primary energy carrier in energy systems, and digital and power electronics technologies will be leveraged to transform all aspects of power infrastructure, including power generation, transmission, distribution, usage, and storage. Renewable energy will be observably, measurably, controllably, and adjustably enhanced to address the vulnerability of grid integration of renewables and increase renewable energy consumption. Improving the ability to control and regulate extensive terminal systems, like microgrids, integrated energy, and distributed power supply will also enable real-time interaction between power generation units

and users. The data generated by these networks will allow power generation units to learn from and adjust to match user consumption habits, improving resource utilization. This will improve the quality, safety, and stability of electricity systems.

- The physical distribution of energy resources is often the inverse of actual energy demand, but the energy cloud will remove these time and distance limitations from energy flows. Take the situation in China, for example. Northwest and Southwest China have abundant wind, solar, and water resources but low demand for power consumption, while Central, East, and South China have high demand but insufficient energy resources. When renewables are centrally connected to local grids, transmission between these regions is further hindered by high randomness and volatility. On the consumer side, the large numbers of user devices and power supplies, such as EVs and distributed power supplies, result in an increased demand for distribution network resources and increased vulnerability in regional power grids. Grids need stronger zoning and interconnection, simplified system operations, and the ability to provide mutual support. Fault isolation also needs to be strengthened to prevent cascading faults that would cause backbone power grids to break down. An energy cloud can facilitate greater sharing of distribution network resources through technologies like active distribution networks and flexible DC distribution networks. This makes it ideal for scenarios such as microgrids, VPPs, and integrated energy systems. By improving the digitalization of transmission and distribution networks, an energy cloud enhances their flexibility, adaptability, and overall control capabilities.
- An energy cloud makes the balance between energy production and consumption more flexible. The energy cloud will make the connections between electricity production and consumption more resilient by enabling unified management. Synergy between generation, grids, loads, and storage will automate the distribution of integrated energy resources. Regional nodes will be able to be monitored and managed in real time, and regional energy consumption will be equalized to balance production and consumption. In this way, electricity production and consumption are intelligently aligned and collaboratively operated so as to improve resource utilization. For instance, optimization algorithms can ensure that solar PV and wind power generation and ESSs adapt to their respective power markets. These algorithms also consider local weather forecasts and other factors that influence production. Data integration then ensures the optimal combination of power generation. Flexible interconnection and digital control of multiple integrated energy sources will strike a balance between energy supply and demand over larger networks. This will make energy systems more flexible and better able to meet various objectives, such as cost-effectiveness, carbon emission requirements, and comprehensive energy efficiency. It enables the use of a wider range of energy types to meet complementary demands. Flexible conversion and integrated demand responses from multiple energy sources will enhance the flexibility of power systems, making them more capable of incorporating renewable energy.

New-type EV energy infrastructure, exemplified by smart charging networks, is widely applied to mobility services

1) Transportation will be electrified, with EV sales booming

New energy vehicles (NEVs) have grown beyond expectations, leading to an irreversible trend of mobility electrification. By the end of 2023, the number of NEVs on the road in China had exceeded 18 million and that number is set to rise to 180 million by 2034, marking a 10-fold increase over the next decade. By the end of 2023, the number of commercial NEVs on the road in China had reached 2.44 million and it is expected to surpass 22 million, a 9-fold increase in 10 years. Simply put, mobility electrification has become an irreversible trend.

In 2023, EVs consumed 300 billion kWh of electricity for charging. By 2033, this figure is projected to rise eightfold to 2.4 trillion kWh, accounting for 10% of global electricity consumption. The charging network, a vital part of modern mobility infrastructure, is essential for keeping EVs operational and for developing future cities. Charging availability is a major concern for EV owners. Establishing a comprehensive charging network will not only boost EV adoption but also invigorate local industries and ecosystems. Since the charging network benefits from both land and traffic, investments in this area must support its continuous evolution. As the number of EVs increases, the long-term advantages will become more apparent.

The rapid rise of EVs is reshaping the ownership balance, where passenger EV owners are replacing commercial EV owners as the main consumers, accounting for 87% of all types. In light of these changes, the mainstream charging preference shifts from low cost to optimal experience. The existing chargers are still plagued with on-going problems such as poor EV-charger matching, loud charging noise, and major safety concerns: The first-attempt charging success rate of most chargers is still less than 85%; the loud noise of air-cooled chargers brings a bad charging experience to EV owners; more than 50% of thermal runaways still occur during charging or a few hours after charging, which causes buyers to hesitate due to the significant risks involved. Accelerating the construction of charging infrastructure is an important measure to improve user experience and develop the EV industry.

2) Ultra-fast, liquid-cooled, intelligent charging networks are widespread, promoting high-quality and collaborative development of EVs and charging facilities

Comprehensive ultra-fast charging is the future. Technically, third-generation semiconductors like SiC and GaN are now mass-produced for commercial use, enhancing energy efficiency and supporting higher-voltage operations for EVs. The evolution of high-voltage EV architecture allows EVs to support low-current, high-power charging, further promoting ultra-fast charging applications. As one of the core components of EVs, traction batteries have also been upgraded. In these systems, cells determine the charging power. Since 2023, 4C cells have been mass-produced, with prices decreasing to match those of lower-C-rate cells. This price drop motivates automakers to develop ultra-fast charging EV models. Driven by technical and economic benefits, ultra-fast charging will soon be widespread. In 2021, only eight EV models supported ultra-fast charging, but by the end of 2023, that number grew to 113 at an automobile expo in Guangzhou, China. Ultra-fast charging is no longer restricted to high-end EV models, but can also be applied to mid-range and low-end ones, and the number of ultra-fast charging EVs will skyrocket in the future. Ultra-fast charging offers immense value to commercial vehicles, especially where time is money. The time saved translates to lower operational costs and higher revenue. With the trend towards high-voltage and ultra-fast charging, it is estimated that over 60% of EV models on the road will support ultra-fast charging by 2030.

The operation and maintenance of chargers are facing greater challenges due to the emergence of diverse charging scenarios, such as the deployments in tropical regions, coastal areas, and mining sites. At the same time, working conditions are increasingly complex, requiring chargers to work properly in hot, humid, salty, or dusty environments. Traditional chargers adopt air cooling or semi-liquid cooling. As the protection capability is weak, the circuit boards and power components in the charging modules are directly exposed to the external environment. Therefore, the annual failure rate of the modules can be 3%–8% or even higher due to humid air, dust, or heat, reducing the service life of chargers to only 3–5 years. Moreover, fans for the power units and charging modules are vulnerable mechanical components, which require frequent cleaning and maintenance onsite at least four times per year, significantly increasing the O&M costs of charging stations. To overcome these difficulties, the cooling method of chargers is transitioning to fully liquid cooling, covering the charging dispensers, charging modules, and power units. Fully liquid-cooled devices are protected to IP55 or higher, completely isolated from external corrosive materials, and enjoy a longer service life. In addition, during high-current charging, heat generated by charging connector ports is efficiently dissipated by liquid cooling cables, ensuring rapid cooling. Similarly, the heat from power components is managed in real-time by liquid pipes. The system intelligently adjusts the flow rate based on cooling needs, ensuring precise temperature control. The fully liquid cooling architecture offers numerous advantages, including high quality, long service life, wide application scenarios, and easy and cost-effective O&M. The annual failure rate of charging modules can be reduced to less than 5‰ and they can work for one decade or longer in most scenarios.

The lack of advanced digital charging networks results in the isolated management of networks, charging stations, chargers, and EVs. To change this and achieve all intelligence in the future, these isolated parties will be deeply integrated to create the following benefits:

First, EVs and chargers will collaborate more effectively. The head unit in an EV helps the EV communicate with a charger in real time to generate the optimal charging solution and map out the navigation route for the EV owner based on information such as the battery state-of-charge (SOC), EV location, and destination. Technologies like wireless charging, automatic plug-in, and autonomous driving can simplify charging operations, allowing robots to handle the process. After charging, EV owners can pay bills easily and securely using technologies such as blockchain and facial recognition.

Second, the grid-friendly charging networks will support the grid with millisecond-level demand response and high-precision intelligent scheduling. According to the loads and electricity price of the grid, the operation strategy of the charging stations can be dynamically adjusted to balance and optimize the charging demand and power supply.

Third, all-digital O&M of the charging networks will be achieved by using technologies such as cloud management, remote fault diagnosis, and automatic fault recovery. These technologies will detect, locate, and handle charging network faults and exceptions in a timely manner, reducing manual inspection and maintenance, improving the grid connection rate and service quality of the charging networks, and helping operators cut business costs and expand their service range.

3) EVs can collaborate deeply with various energy systems, serving as crucial regulation resources

EVs will become fully involved in interactions with energy systems as important regulators in energy flow control. The large-scale promotion of EVs and renewable energy creates opportunities for EV-grid synergy. There is increasing demand for a large number of flexible power sources on the power generation side and for adjustable load resources on the consumption side. Unlike more common electrical loads such as household appliances, EVs are highly flexible and adjustable. As wireless charging, smart charging, and autonomous driving technologies mature and are widely adopted, EV users will be free to choose when to charge, discharge, and swap batteries, participating in the electricity spot market and ancillary service market based on their individual needs. This will reduce the impact of EV charging on the power grid, provide new resources for the power system to schedule, and avoid a large amount of wasted investment in the power grid and power supply.

The number of EVs worldwide could exceed 150 million by 2030. Ideally, by that time, the energy storage capacity should be 40 times as large as the energy storage capacity installed in 2020, with the potential to serve as an adjustable load and a flexible power source. EVs perform orderly charging as a way to contribute to local peak shaving, bringing about huge economic benefits. In the future, EVs participating in the frequency regulation ancillary service market will have higher value. EVs will be able to take full advantage of their role as flexible loads and perform orderly charging as a way to contribute to user-side applications such as peak shaving, distributed PV charging, demand response, peak staggering ancillary services, and spot market balancing.

Charging infrastructure connects EVs, transportation systems, and mobile lifestyles, as well as diverse energy use scenarios. It is the point of convergence for energy and transportation, in terms of transactions, interaction, behavior, and information. It is one of the important enabling components of the energy cloud.

Large-scale construction of charging networks and the development of technologies such as digitalization, IoT, cloud computing, big data,

and AI bring about multilevel improvements in intelligence: Intelligent charging infrastructure will make charging networks visible, manageable, controllable, and optimizable, significantly reducing O&M costs and increasing efficiency and revenue. As a data interface, chargers can be utilized to build a smart charging network that integrates EVs, chargers, power grids, the Internet, and value-added services. This network will leverage chargers' strengths in terms of scale, integration, data, and connectivity, create multiple new business models, and generate a virtuous cycle of economic and social benefits. Chargers enable charging facility operators to provide data consulting services to support business district construction, real estate development, dealership store planning, second-hand car trading, digital payments, and e-commerce operations as a way to monetize, expand sources of revenue, and improve market operation capabilities. For local governments, chargers can provide data support for urban planning, power dispatching, everyday services, and infrastructure construction, making charging infrastructure an important part of smart cities. It is estimated that by 2030, the annual energy charged to EVs will exceed 1.1 trillion kWh.

Empowering the digital age: green, simple, smart, and reliable energy infrastructure

Consumer data traffic from cellular networks and fixed broadband will grow at a compound annual growth rate of 29% in 2024, increasing from 1.3 million PB in 2018 to 5.8 million PB in 2024. This rapid growth poses significant challenges to existing ICT infrastructure, including data centers, data center interconnection networks, and Internet access networks. To meet these new demands, operators, cloud vendors, and Internet enterprises are upgrading, expanding, and scaling up their data centers. However, data centers consume substantial amounts of electricity to process service loads, resulting in considerable indirect carbon emissions. Building efficient and low-carbon communications networks and data centers is not only an operational necessity for enterprises, but also their civic duty. Leading operators worldwide have made carbon reduction commitments and launched various initiatives. Vodafone and Orange aim to achieve net zero emissions by 2040, while Telefónica has set its target for 2030. Google plans to power all its operations, including data centers and campuses worldwide, entirely with carbon-free energy by 2030. Microsoft has pledged to be carbon negative by 2030, and to remove all the CO2 it has ever emitted, either directly or through electricity use since its founding in 1975, by 2050. Additionally, the municipal government of Beijing mandates that data centers be built with their own distributed renewable energy facilities and be powered 100% by clean energy by 2030. Key players in Europe's cloud infrastructure and data centers have developed a self-regulatory initiative, the Climate Neutral Data Centre Pact.

It is not only crucial to make data centers to be green and low-carbon, but also their significant responsibility as infrastructure to drive the rapid digital and low-carbon transformation of the energy-intensive industries. In the digital economy, the consumption of energy will bring "overlaid returns." Each kilowatt-hour of electricity used in a data center contributes to not only the business value of this data center, but also those of various industries whose applications, including cloud computing, big data, and Internet services, run on these services. Estimates suggest that every ton of standard coal consumed can directly generate CNY11,000 for a data center, contribute CNY888,000 to the added value of the digital industry, and indirectly create a digital market worth CNY3.605 million (excluding parts not directly related to the data center). According to the Global Enabling Sustainability Initiative (GeSI), the ICT sector's carbon emissions will account for 1.97% of global carbon emissions by 2030. However, by enabling

other industries, the ICT sector will help reduce global carbon emissions by 20%, which is 10 times its own emissions. This positive impact is known as the "carbon handprint." Therefore, building green and low-carbon data centers not only promotes the high-quality development of the ICT sector but also enables traditional energy-intensive industries. Through actions such as "migrating to cloud, using digital tools and enabling intelligence," one industry can empower various industries, leading to significant reductions in energy consumption and substantial improvements in productivity and total factor productivity.

We predict that the ICT energy infrastructure will develop in the following directions in the next decade. And more than 80% of the ICT energy infrastructure will be powered by green energy.

1) Green electricity will bring more green computing power

As digitalization progresses globally, the ICT sector has increasingly become energy-intensive. Driven by the carbon reduction targets, the shift toward green energy supply for ICT infrastructure is inevitable. Clean energy sources such as PV, wind, and hydrogen will be integrated into ICT energy infrastructure. Due to their cost-effectiveness and flexibility, more than 80% of power supply systems in ICT infrastructure are expected to incorporate distributed green energy within the next decade. For telecom sites with low power consumption, distributed PV may become the primary power source, enabling zero-carbon telecom networks. Unlike conventional power purchase agreement (PPA) for renewables and renewable energy certificates, data centers will adopt a direct supply of clean energy. This includes building distributed PV plants on data center campuses and rooftops, or building utility-scale PV, wind, and other clean energy plants in nearby areas to directly supply power to data centers. With intelligent control, these distributed energy systems will no longer provide unidirectional power supply but will also participate in ancillary services markets such as power grid peak shaving. This helps smooth out the random and intermittent output of wind power and PV. Consequently, this approach enhances the power supply benefits of ICT infrastructure, maximizes the business value of basic resources, and improves the stability and reliability of the entire energy system.

2) Reliability is the most essential requirement of ICT infrastructure

ICT infrastructure forms the physical backbone for handling vast amounts of data and serves as the core resource for centralized information processing, computing, storage, transmission, exchange, and management. It is vital for the smooth functioning of society and the economy. Consequently, reliability is the lifeline of data centers, yet it often remains the weakest link. Implementing a comprehensive end-to-end assurance mechanism provides the most robust foundation for the reliable and stable operation of infrastructure throughout its lifecycle.

Ensuring the reliable and stable operation of the infrastructure hinges on highly dependable products and professional services. Each infrastructure comprises of tens of millions of components, necessitating an end-to-end full-link assurance mechanism that spans from product inherent reliability to design and O&M by expert teams. Take lithium-ion batteries as an example. During the planning phase, considerations should include remote deployment or separate compartment design with a water fire suppression system for lithium-ion battery rooms. In the construction phase, selecting highly reliable products is crucial. Additionally, strict control over transportation, warehousing, and installation specifications, along with a robust O&M inspection mechanism, is essential to build emergency response capabilities. These comprehensive measures ensure the reliable operation of data centers.

As the power density of ICT infrastructure increases, the time available for emergency handling significantly decreases, presenting greater challenges for maintenance. AI technologies enable risk prediction and data center infrastructure management. AI algorithms can learn from historical and real-time data to predict and identify abnormal patterns. This shift from passive reaction to proactive prevention enhances the reliability of ICT infrastructure through improved O&M.

3) Comprehensive architecture refactoring is making ICT energy infrastructure simple, converged, smart, and efficient

Networks and data centers are becoming larger and more complex. The ongoing pursuit of simplicity is driving the development of ICT energy infrastructure architecture toward greater convergence. Most of today's telecom sites are built indoors, and traditional air conditioners are used for cooling. The overall energy efficiency of these sites is only 60%. Conventional power supply solutions typically use multiple sets of power supply equipment, each supporting a different voltage system, which complicates deployment. We believe that the form of telecom sites will change dramatically in the next decade. What once filled an equipment room can now be squeezed into a cabinet, and what once filled a cabinet can now be mounted on a pole. Sites are becoming simpler and more reliable, with smaller footprints and lower leases. The way in which data centers are built will also change rapidly. Traditional concrete buildings usually take more than 20 months to build, and the building materials are neither environmentally friendly nor recyclable. Prefabricated modular data centers will become increasingly common over the next decade. Prefabrication reduces the use of high-carbon-emission materials, such as concrete, rubber, and rock wool sandwich panels, and dramatically reduces onsite construction and maintenance. This way, a data center housing 1,000 racks can be built in only a few months, meeting the requirements for rapid service rollout. In terms of network and data center power supply solutions, power supply link convergence will become a major trend. Adapting to more renewable energy sources, ensuring compatibility with multiple energy supplies, and being ready for smooth evolution are the directions in which the power supply architecture will evolve. Examples include multi-mode scheduling control and management, modular overlay evolution, and the convergence of different services and devices across multiple scenarios. With this converged

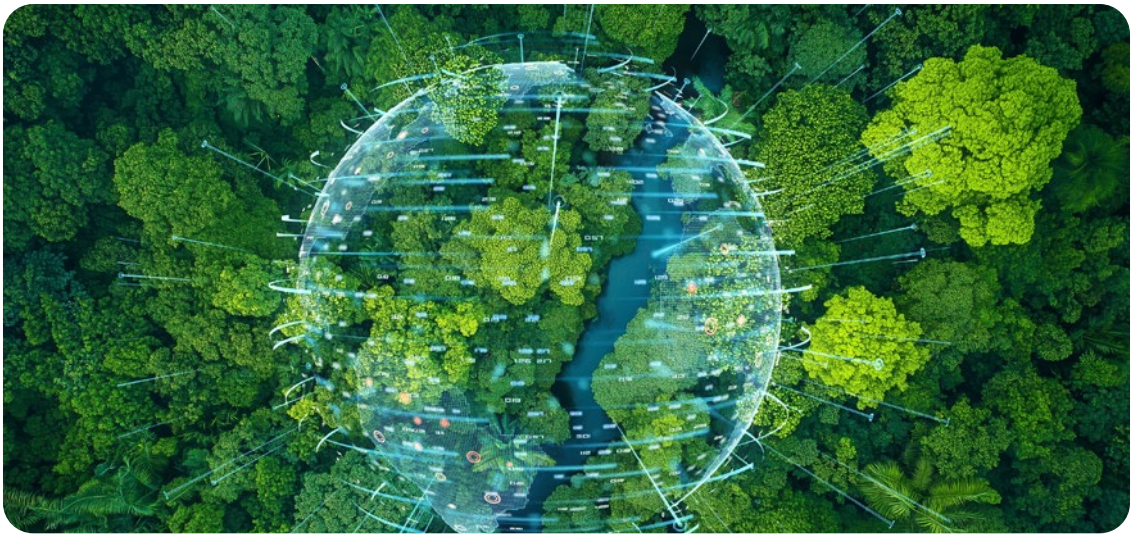
architecture, power supplies and batteries of telecom sites are being integrated into a blade form factor. This approach develops power supply, energy storage, temperature control, and power distribution into individual modules, enabling on-demand evolution to support cross-generational network advancements. All data center power supply links, including transformers, uninterruptible power systems (UPSs), and power distribution, will be converged to reduce the installation footprint. Backup power will rely on lithium-ion batteries, facilitating intelligent collaboration between power generation, storage, and consumption. This reduces the required capacity of the data center UPS, as well as the footprint and construction costs of data centers.

4) DC for AI, AI for DC

With the continuous progress of AI technologies, the operation of data centers is undergoing a revolutionary transformation. AI technologies not only help improve the energy efficiency of data centers and reduce operation costs, but also play a key role in ensuring data center reliability.

In terms of reliability, advanced AI prediction and analysis technologies can predict the service life of key devices such as capacitors and fans in the UPS. In addition, outlier algorithms can identify potential faults of lithium-ion batteries in advance, implementing early fault detection and prevention, which is similar to the concept of "prevention is better than cure" proposed by Bian Que, a famous doctor in ancient China. In terms of energy saving, the AI energy saving algorithms will optimize the cooling system of a data center through real-time analysis and adjustment. Compared with traditional manual optimization, AI algorithms automatically adjust parameters based on real-time weather changes, significantly improving the overall cooling efficiency by an estimated 8%–15%. In terms of simple O&M, AI technologies significantly reduce the workload and difficulty of routine O&M. Traditionally, inspecting the power supply and distribution system requires 6 to 12 onsite meter readings daily. However, with AI technologies, inspecting 2,000 cabinets can be completed in just 5 minutes. Additionally, the AI O&M assistant can monitor devices 24/7, receive real-time device alarms, and provide corresponding solutions. Monthly health reports are automatically generated, offering robust data support for service decision-making.

The application of AI technologies in data centers improves operation efficiency, reduces energy consumption, and enhances data center reliability. With the further development of technologies, we believe that AI will become an indispensable part of data center operations, enabling them to become greener, simpler, and more reliable.



04

Quality and safety will become key challenges for renewables

The extensive use of power electronics devices presents significant challenges to the grid connection and operational safety of renewable energy plants.

As high proportions of renewables and power electronics devices including power supplies, loads, and ESSs with different electrical characteristics, are integrated into existing power systems, the strength of the power grid is significantly reduced. This reduction is due to issues such as voltage instability, frequency instability, power angle instability, and wideband oscillation. Renewable energy devices have poor voltage tolerance capabilities. When a fault occurs, they can provide only 1.1 times the rated current for dynamic voltage support. Traditional thermal power can provide 5–10 times the rated current. In addition, renewable output requires voltage boosts by several stages before being fed into the power grid, and the electrical distance from the grid connection point is 2–3 times that of a common generation unit. With a low short circuit ratio (SCR), a renewable plant may not provide sufficient voltage support for the power grid. Devices such as renewable grid-tied inverters and converters do not have the inertia response capability. As a result, the overall inertia of the power grid is low and the system frequency regulation capability is reduced. The low inertia of renewable generation units also causes the amplitude to decrease for the power angle curve, resulting in power angle instability. The rapid response feature of renewable generation units also causes new problems of wideband oscillations in medium and high frequency bands.

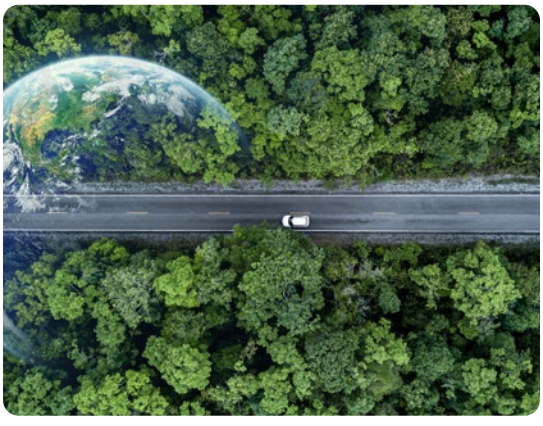
O&M for utility-scale renewable plants are challenging due to their large footprints, high capacities, and numerous devices. For instance, a simple inspection of a 100 MW plant can take at least five person-days. Additionally, most utility-scale projects are situated in remote areas with harsh environments, such as hot deserts with heavy sandstorms, seas with high humidity and salinity, or extremely cold high-altitude plains. These adverse conditions significantly impact the quality and reliability of devices, posing serious threats to the operational safety of the power plant.

For a distributed PV system, more and more devices are deployed in buildings, campuses, and homes, which are closely related to daily production and life. Once an accident occurs in a rooftop PV system, personal and property safety will be seriously threatened. DC arcs have been proved to be a major fire risk in rooftop PV plants. Arcs may occur due to poor contact in PV module weld joints, aging cables, and loose terminal connections. In rooftop PV projects, the DC voltage of PV modules can reach 600–1000 V as long as there is sunlight, even if the equipment is shut down. This poses potential risks to construction workers, O&M personnel, and plant owners. During emergencies such as fires, rescue personnel face significant challenges. They cannot access the rooftop or use water to extinguish the fire due to the high voltage present in the PV array. Consequently, they often have to "let it burn," waiting until all PV modules are burnt before intervening. This approach significantly hampers the rescue process, leading to greater personal and property losses.

Firstly, to cope with these challenges, we should focus on the quality of devices initially. The industry should reach a consensus on the importance of quality and ensure product safety and reliability by using high-quality hardware and software. Secondly, we need to integrate individual technologies such as PV modules, ESS, grid forming, digitalization, and intelligence to develop innovative renewable solutions which shift from grid following to grid forming, achieve intelligent DC safety diagnosis, warning, and protection in system O&M, and ensure personal and asset safety in various scenarios.

Extensive deployment of ESS presents huge challenges such as the risks of carrying massive energy, thermal runaway, and control difficulty in case of fire carrying massive energy, thermal runaway, and control difficulty in case of fire

With the rapid and extensive application of ESSs, numerous serious incidents have occurred in ESS plants, resulting in significant economic losses and casualties. In May 2024, a fire broke out at the world's largest energy storage plant at the time, with a capacity of 250 MWh, in San Diego, California. The fire reignited multiple times and lasted for 11 days, causing severe losses and environmental impacts. According to incomplete statistics, 65 major energy storage fire accidents occurred worldwide from 2019 to 2023. Of these, 27 were utility-scale and C&I energy storage accidents caused by battery quality issues such as process defects and uneven copper foil coating. Twelve were residential energy storage accidents due to battery quality problems like iron shavings falling into the module. Seven were data center accidents caused by water intrusion, management system faults, and battery quality issues. Nineteen were attributed to other quality problems, including improper battery securing methods and insulation faults. Moreover, there are no mature standards and specifications for the design, construction, commissioning, operation, and maintenance of energy storage plants. The management, operation, and maintenance of these plants need to be conducted with greater professionalism. As a new component of the power system, the energy storage industry is still in the exploration phase, and safety has become a significant challenge for its development.



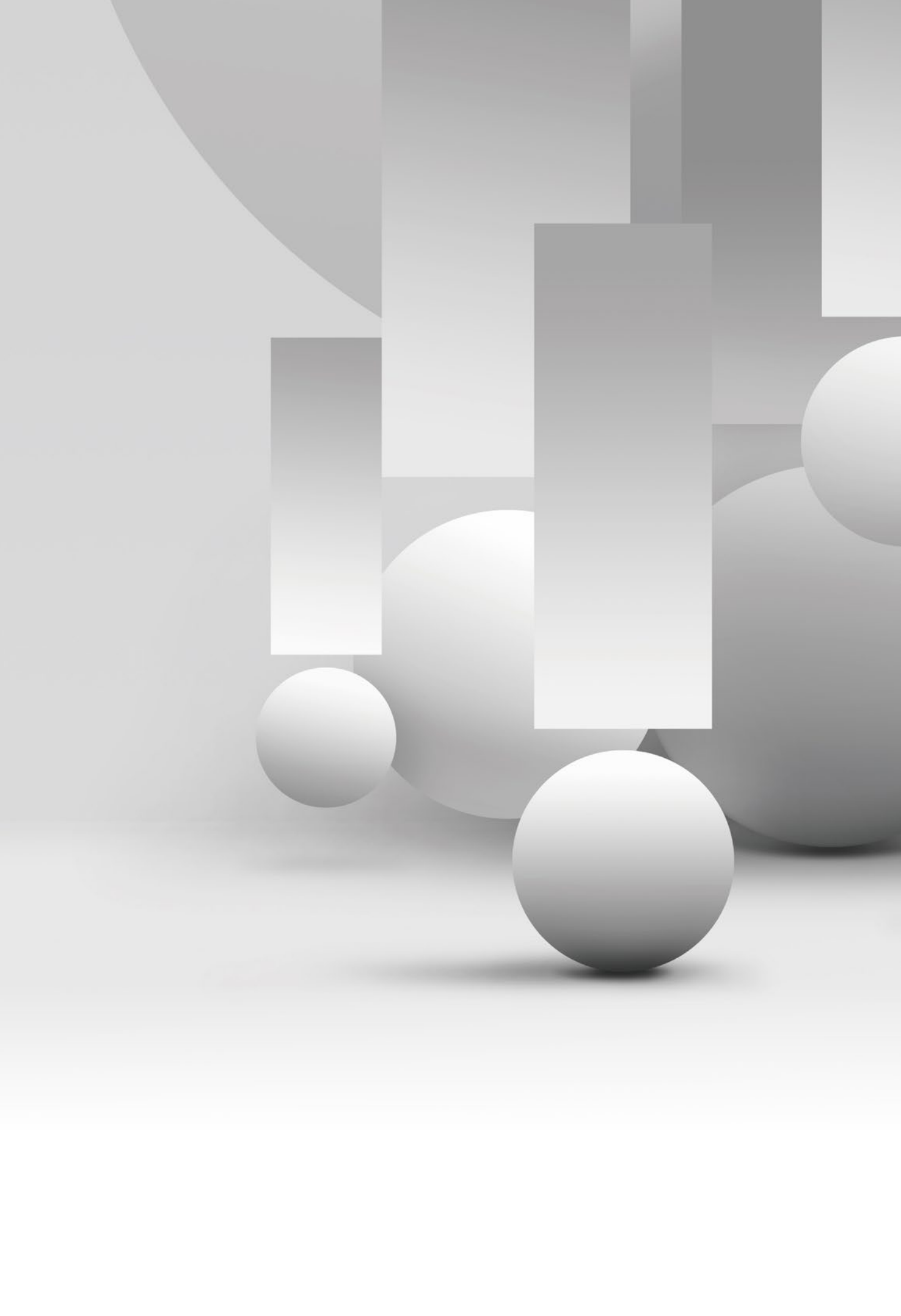
As a kind of regulation resource of an energy system, the ESS is becoming increasingly prevalent. The scale of a lithium-ion battery site is increasing to GWh-level. The large-capacity ESS is developing from 2 MWh per cabinet to 5 MWh+ per cabinet, boasting higher energy densities. Additionally, the adoption of diverse new technologies, such as sodium-ion batteries, flow batteries, and supercapacitor batteries, is on the rise. Consequently, this technological advancement brings about more complex safety risks in energy storage plants.

To mitigate the safety risks associated with energy storage plants, ensuring the high quality of energy storage products is essential. By addressing potential issues from the ESS architecture design, accidents can be prevented before they occur. The manufacturing and usage of key components such as battery cells are important to ensure body safety. Proper heat insulation and flame-retardant materials should be selected. Heat dissipation, smoke exhaust, and fire suppression systems are designed to improve passive safety, while online monitoring, intelligent control, and safety warning technologies are developed to ensure active safety. Safety principles and management should be implemented throughout the entire lifecycle of an energy storage plant, spanning from planning and design, equipment selection, manufacturing, acceptance, transportation, delivery, onsite installation, system commissioning, operation control, repair and maintenance, and plant retirement. Furthermore, relevant safety standards and regulations should be developed and implemented based on industry practices, and consolidated by policies and regulations to promote the high-quality development of the industry.

Conclusion

The energy sector has made remarkable strides, evidenced by the convergence of renewable energy, digital technologies, and power electronics. In the future, electricity-based energy systems will resemble ICT networks, with power grids acting as the backbone, power electronics devices as gateways, and the energy cloud as the operating system. This evolution will transform how energy flows are processed, moved, and stored. We are on the brink of large-scale development and utilization of clean, low-carbon energy. Multi-level energy networks will be widely connected, enabling both active and passive participation from various loads. Collaborative decision-making and operation across multiple service logics will become a reality. Over the next decade, the integration of energy and information flows will deepen, supporting each other and marking a key transition period for comprehensive energy transformation. This shift will shape the energy landscape for the next century. As we enter the intelligent age, technological innovations in information and energy flows are becoming increasingly synchronized. The focus of innovation is shifting from individual devices and scenarios to entire systems and industries. Energy networks are expanding from local to global scales, and operations are transitioning from device-based to cloud-based. Energy systems are becoming more visible, measurable, and controllable on a broader scale. The convergence of energy and information flows is amplifying the value of energy systems, making them more economical, cleaner, and safer to operate. New models of electricity production, transmission, storage, and consumption are emerging, heralding a new era in the energy sector. The integration of energy systems with information and commercial systems is transforming the energy landscape. Energy networks are evolving from standalone entities into critical infrastructure platforms that interconnect with transportation, carbon footprint, and information networks. This collaborative control across industries enhances the scope and method of energy cloud management, extending beyond individual devices and systems.

Technological advancements and energy transformations are mutually reinforcing, profoundly shaping the future of the energy sector. By recognizing major trends, we can better address future challenges and seize current opportunities. In this emerging digital power era, collaboration is key. Building new alliances and exploring innovative ways to collaborate across value chains and ecosystems will drive global energy innovation and development. Together, we can propel energy transformation, creating low-carbon, electrified, digital, and intelligent energy systems. This collective effort will make the world a greener, better place for all.



HUAWEI

Version 2024

Cloud Computing 2030

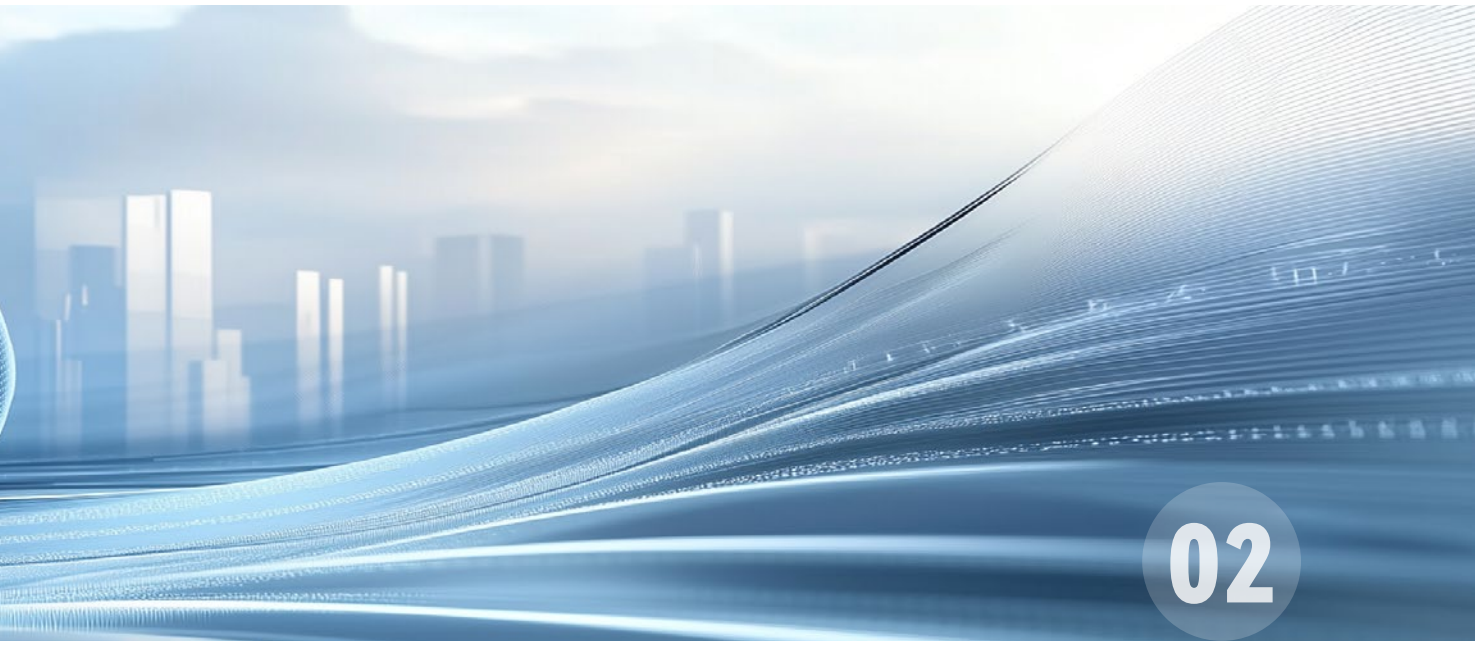
Building a Fully Connected, Intelligent World

Contents

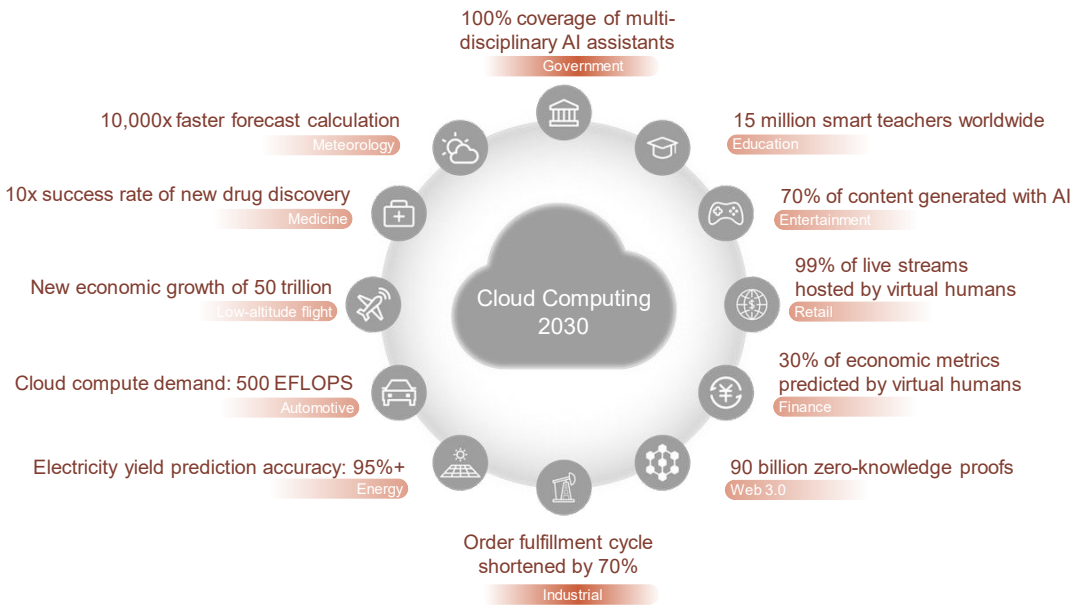
Macrotrends

Since the dawn of the Internet in the late 20th century, we have traversed a remarkable technological landscape. The emergence of cloud computing at the onset of the 21st century and the current surge in artificial intelligence signify not just advancements but a transformation in how we harness human ingenuity. The convergence of cloud and AI has given rise to an intelligent world from smart homes to smart cities, from precision healthcare to personalized education, and from intelligent manufacturing to FinTech innovations. Cloud and AI are the cornerstones of limitless potential for societal development, propelling enterprise innovation and growth to new heights.

By 2030, the ubiquity of cloud computing is anticipated, with an estimated 3 billion smart devices offloading their computational demands to the cloud. Intelligence will be omnipresent, with an expected 1.5 billion enterprise employees benefiting from personalized AI assistants. Furthermore, it is projected that 80% of enterprise applications will be either built anew or rebuilt with AI at their core. The physical world is being reshaped, leading to a data revolution in the 3D space, where the volume of data is expected to exceed current levels by a million-fold. Approximately 500 million individuals will venture into the realm of spatial computing, a fusion of virtual and real worlds.



Future Industry Scenarios



2.1 Pharmaceutical: AI Boosts Drug Design Success 10-fold and Halves Development Time

The development of a new drug is often a lengthy and costly process, spanning over 10 years and requiring an investment of over USD1 billion. However, despite these substantial resources, the success rate of new drug discovery remains disappointingly low, hovering around 10%. This low success rate can be attributed to various factors, including low clinical efficacy, high toxicity, poor druglikeness, insufficient market demand, and unsuccessful product strategy.

Fortunately, AI is here to help. Throughout the four phases of drug development — target discovery, drug screening, lead optimization, and preclinical testing — AI significantly improves efficiency through tasks like AI molecule generation, ADMET property prediction, and molecular dynamics (MD) simulation. Existing data indicates that AI can accelerate drug design by 70% and boost success rates tenfold. The potential impact is immense.

Today, large scientific computing models play a pivotal role in small molecule drug design, spanning diverse innovative drug R&D tasks. These tasks include the development of new antibiotics, antitumor drugs, and drugs targeting the central nervous system, and the discovery of druglike natural products, all of which have yielded impressive outcomes.

Projections suggest that by 2030, AI technology will be integrated into every phase of new drug discovery, potentially reducing the typical 10-year R&D cycle to just five years or less.

2.2 Meteorology: A Data-Driven Earth Decoder Accelerates Weather Forecasting 10,000 Times

According to a Nature Communications report, between 2000 to 2019, extreme events attributable to climate change resulted in costs estimated at approximately USD2.8 trillion, averaging over USD143 billion per year and USD16.3 million per hour. According to a United Nations Environment (UNEP) report, between now and 2050, the cost of adapting to climate change in developing countries may reach between US$280 billion and USD500 billion annually.

Traditional weather forecasting relies on intricate system modeling, incorporating factors like atmospheric circulation, diverse terrains, ocean dynamics, and their complex interactions. The conventional numerical weather prediction (NWP) methods are highly compute-intensive, employing thousands of processor cores for each computation, often spanning dozens of hours. Despite advancements, overall progress in weather forecasting has been gradual. Furthermore, the global weather forecasting services market is predominantly dominated by European and American agencies.

Since 2023, data-driven AI weather models have emerged as a completely different approach to weather forecasting, attracting unprecedented global attention to the research and application of AI technology in this area. In essence, an AI weather model proposes an efficient 3D earth decoder that can better model and predict problems related to Earth science and meteorology.

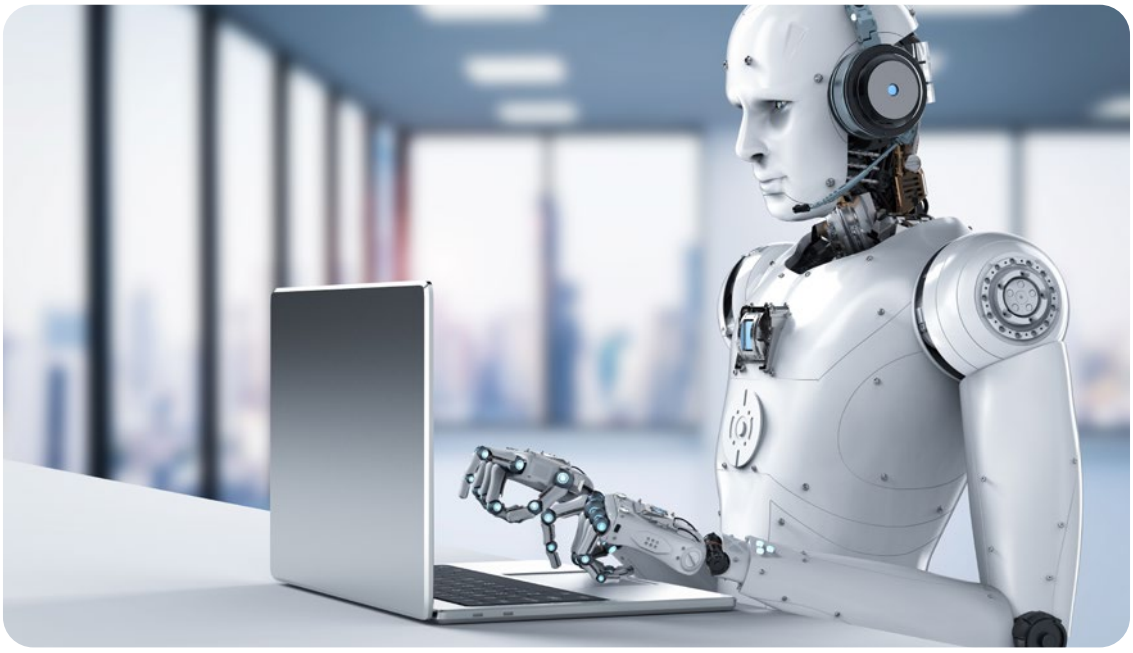
Estimates suggest that by 2030, AI weather forecasting models will be 10,000 faster than traditional NWP systems and 50% more accurate. These advanced models will provide short- and medium-term forecasts for the next 1 to 10 days, mitigating economic losses resulting from extreme weather events, which can reach into the hundreds of billions of dollars.

2.3 Finance: Nowcasting High-Frequency Data for 30% of Economic Indicators

In the financial sector, decisions are made based on external data such as monthly and quarterly economic reports and GDP growth rates. Generally, there is a significant delay before official data like this is published. Nowcasting can use AI models to predict key economic indicators before the official data is published.

Massive amounts of data is collected and processed and AI models are used to improve the sampling frequency from monthly to daily. In this way, economic indicators that you once had to wait over a month to see can now be obtained on the same day or very close to it. By integrating AI algorithms with a real-time, dynamic input-output system, non-linear correlations can be extracted from massive data and factors that always impact the past, present, and future can be discovered. This way, the future states of economic indicators are predicted.

It is estimated that by 2030, nearly 30% of economic indicators and financial benchmark metrics will be nowcasted, with an accuracy of up to 99%.



2.4 Government: Multidisciplinary AI for One-on-One Assistance

Generative AI is fundamentally transforming the way governments operate and serve their public. By 2030, multidisciplinary AI support will be available as both self-developed and externally-provided platforms. These platforms will lighten government workloads by as much as 50% to 70%. With various integrated AI features, such a platform will bring out the best of staff's expertise, responsibility, and availability, working autonomously with people to interact more like a colleague than a mere tool.

Another story worth telling is intelligent assistants, which are built on dedicated government models and constantly evolving their dialog to be more humanlike. **By 2030, each government employee and the enterprises and individuals they serve will have an assistant** for smarter recommendations, easier payments, and better information. These assistants, tailored to each citizen, provide instant, authoritative responses on government affairs and higher administrative efficiency. Citizens stay better informed on policies and regulations, with access to internal guidance and clear explanations, on a single, always-on channel.

By 2030, multidisciplinary AI support will be deployed in more than 95% of governments, and intelligent assistants will be accessible to all served enterprises and individuals.

2.5 Education: Co-Teaching, Co-Learning, and Co-Nurturing with 15 Million Digital-Intelligent Teachers Worldwide

Education is a national priority facing many challenges: In primary education, classroom teaching often involves teachers lecturing to students passively receiving the information. The teaching methods tend to be uniform and untargeted. At home, parents put significant effort into tutoring their children, but they typically lack sufficient guidance and it's difficult for them to focus on interest and innovation. In higher education, university curricula and talent development plans evolve too slowly, so there is often a mismatch between the skills of graduates and industry needs.

The development of artificial intelligence, particularly Artificial Intelligence Generated Content (AIGC), has accelerated the progress of smart education. Large teaching models will reshape education in three main ways:

- Teaching materials can be generated by AI rather than teachers relying on personal experience, and homework can be automatically graded, complete with feedback. These innovations can save teachers over 15 hours per week.
- Adaptive curricula become possible. Study outcomes, including reading and writing abilities and learning attitudes, can improve by up to 90% when students are guided by virtual teachers and supported by regular feedback.
- AIGC can independently conduct certain experiments and generate research papers, reducing the time needed for research by 1/3.

Looking ahead, education will become more open, breaking through the constraints of time, location, and demographics. It will be more sustainable and able to meet the needs of lifelong learning.

By 2030, it is expected that there will be 15 million "digital-intelligent teachers" worldwide working alongside schoolteachers, moving towards a new paradigm of co-teaching, co-learning, and co-nurturing.

2.6 Retail: A Flexible Supply Chain Is Slashing 72% of Inventory Costs. XR and Unmanned Delivery Are Hitting the Mainstream

Global retailers face challenges with perishable goods that can lead to spoilage if they do not sell quickly. With an average daily turnover of USD200,000 per store, just one extra day of unsold inventory across 100 stores can tie up USD20 million in capital.

To solve this, retailers use AI and digital tools to streamline their supply chains, enhancing production, sales, and delivery of goods. AI is really good at guessing how much customers want to buy and figuring out the best prices. This makes their supply chain more flexible and able to work almost on its own. By using AI to look at data, retailers can introduce new products to the market more quickly and handle their stock more effectively. This, in turn, cut down on the costs associated with holding stock. Also, more and more retailers are focusing on making the shopping experience better for customers. They use lots of data and cool tech like XR, unmanned deliveries, and virtual assistants to give customers services that are personalized, fun to use, and understand their needs.

By 2030, it is predicted that AI will be a game-changer for the retailers, affecting up to 95% of customer interactions and having 85% of global retailers investing in it. Full coverage of AI in the retail world could lead to a significant reduction in operating costs for 72% retail enterprises and a revenue increase for 69% retailers. This could potentially add an extra USD611 billion to USD815 billion in overall revenue, a 1.5% to 2% increase.

2.7 Web 3.0: Decentralization in Every Industry, 90 Billion Zero-Knowledge Proofs

Web 3.0 has coexisted with Web 2.0 for a while now, and it is not entirely replacing Web 2.0. Rather, it is evolving and expanding the web's capabilities. Web 3.0 empowers individuals with greater control over their information and reduces reliance on vulnerable centralized servers. It utilizes zero-knowledge proofs to safeguard users against potential privacy breaches and data exploitation. In Web 3.0, there is an enhanced awareness of data and privacy protection. Currently, Web 3.0 remains niche for three main reasons: 1. the poor scalability of existing public blockchains, limiting the scale of upper-layer applications; 2. the lack of industry-changing killer apps; 3. the absence of a development model for Web 3.0 to be integrated with the real economy.

With the approval of BTC/ETH ETFs in Hong Kong (China) and the United States at the beginning of 2024, more real assets will enter the decentralized world. The integration of traditional and decentralized financial systems has been accelerating. Decentralized applications (DApps) will be adopted by all sorts of industries, where they can provide significant automation, transparency, and security for the real economy. Typical use cases include real estate rent collection and property rights contracts, procurement orders, logistics, payments, and settlements throughout the supply chains. DApps have the potential to reduce insurance fraud. In the future, cloud computing performance may be hundreds of times better than today, enabling public blockchains to handle more complex smart contracts and larger transaction volumes. Additionally, cloud computing platforms may integrate quantum computing capabilities to protect data privacy on public blockchains and provide even higher levels of security.

By 2030, it is predicted that the number of global users of digital assets will reach 1 billion, and the Real-World Assets (RWA) tokenization market will reach USD16 trillion. A decentralized finance or game app that impacts our daily lives (with monthly active users exceeding 100 million) may emerge, enhancing the liquidity of these tokenized assets and offering more investment and economic incentive opportunities. By 2030, we expect Web 3.0 applications to generate 90 billion zero-knowledge proofs, creating a computational market space worth tens of billions of dollars.

2.8 Energy: Network-Wide Intelligence of the Energy-based Operating System, Reducing Greenhouse Gas Emissions by 10%

According to the International Renewable Energy Agency, the installed capacity of photovoltaic (PV) energy is expected to reach 5200 GW by 2030, with 68% of the energy coming from renewable sources. The proportion of fluctuating renewable sources, including wind and PV power, is expected to increase to 46%. With the growing power of renewable energy, particularly PV power, the energy mix is undergoing significant changes. This shift represents various challenges for the energy industry in terms of grid connection, operations, and safety. For instance, the intermittent nature of renewable energy affects the stability of the power grid, and the peak load of power demand is becoming increasingly prominent. These challenges significantly impact the safety and stability of the power system.

In the coming years, cloud computing, AI, and 5G will be increasingly used in the energy industry, covering manufacturing, production, and consumption. This will enable the creation of an energy-based operating system that streamlines the entire generation-grid-load-storage-consumption process, leading to network-wide intelligence. By leveraging management systems and AI algorithms to coordinate the running status of each energy node, the overall efficiency of distributed systems can be improved by 5% to 10%. Furthermore, AI prediction models can be applied to new power generation scenarios, such as PV power generation, offshore wind power, and floating PV plants, to improve the accuracy of power generation plans to over 95%. With the help of big data analysis and machine learning algorithms, the power generation prediction error can be reduced to less than 10%, significantly enhancing the grid feed-in capability of renewable energy sources.

By 2030, it is projected that 90% of global PV plants will utilize AI technology, resulting in an annual reduction of 5% to 10% in greenhouse gas emissions.

2.9 Entertainment: AI Creates 70% of Media Content and Unlocks a Personalized Content Market Worth USD500 Billion

Today's media and game sectors face a series of challenges. The speed of content creation and rendering is below expectations, causing huge delays in content distribution. Complex production processes often compromise on reliability, while insufficient collaboration across different areas leaves the potential of teamwork untapped.

However, challenges do come with opportunities. In this case, it's AI, which will transform the production model of media and game content by 2030, with manual video shooting replaced by computer generation. As digitalization is sweeping across industries, AI will dominate some of them, including the media and game sectors. **The maturity of AI supercharges the virtual human industry into a USD38 billion market by 2030.** With AI-generated content (AIGC), reproducing an imaginary world is a breeze, and you can modify your creation whenever you want. Gaming in such a virtual world means unprecedented immersion. In terms of media content, users will no longer just get what they are given, as the popularization of AI makes everyone become content creators. This trend will be accompanied by a revolutionized cloud architecture, industry automation, and new business models.

The AI-infused personalized content market is expected to hit USD500 billion by 2030, and contribute to the monthly traffic growth by 85 billion GB.

2.10 Industrial: 50% Supply Chain Cost Slash and 70% Fulfillment Speed Boost Through the Multi-Agent System

The global value chain is swiftly redefining itself as companies are expanding their reach worldwide. Digital and smart supply networks are leading the wave. However, the growing complexity of these networks requires more costs in procurement, production, logistics, inventory, and quality control, and makes order fulfillment less predictable.

To address this, industrial manufacturers are turning to AI for design, production, and quality control, making processes smarter and more efficient. Robots with AI are becoming more agile, and supply chains are getting smarter too, with AI improving order predictions. At the cutting edge, companies are using agent technology to create flexible, collaborative manufacturing processes with group intelligence. Industry leaders are building networks that drive innovation and efficiency in R&D, production, and supply chain management. The key to this advancement is multi-agent systems technology, which is creating intelligent supply chains. This not only greatly reduces costs but also significantly shortens the time it takes to fulfill orders, making the industry more responsive.

By 2030, agent technology in manufacturing is expected to exceed 80% adoption, slashing supply chain costs by half and reducing order fulfillment times by a staggering 70% with the aid of foundation models.

2.11 Smart, Personalized Human-Vehicle Interaction Powered by 500 EFLOP/s Cloud Computing

Currently, the interaction between humans and vehicles is limited to physical controls, voice instructions, navigation systems, and passive safety technology. Communication and collaboration between vehicles and the external environment are still in the early stages, resulting in insufficient traffic safety and efficiency. This cannot satisfy the increasing demand for intelligent vehicles.

In the future, human-vehicle interaction will be smarter and more tailored. Voice assistants will engage in fluent multi-language conversations with 95% accuracy and 90% emotion recognition, minimizing distractions and keeping drivers alert on the road. AR HUD displays will project real-time traffic and weather information, making information acquisition more intuitive and improving security. Touch and gesture control technologies will be popularized, and face and fingerprint recognition will simplify in-vehicle settings and personalized adjustments. AI will personalize the driving experience with 90% accuracy by adapting to drivers' habits, paving the way for a seamless shift to full autonomy. V2X will enable microsecond-level data sharing between vehicles, infrastructure, and pedestrians, optimizing traffic management and reducing congestion and accident risks.

By 2030, it is estimated that the number of global intelligent vehicles will reach 200 million, and cloud computing power will reach 500 EFLOP/s (exaFLOP/s) to support complex AI models, real-time data processing, and autonomous driving.

2.12 Low-Altitude Economy: A USD50 Trillion Growth Opportunity

The low-altitude economy, a nascent industry driven by general aviation, is rapidly expanding globally. This emerging sector encompasses a wide range of activities, including low-altitude flights, aviation tourism, regional airlines, navigation services, scientific research, agriculture, and emergency response.

Beyond traditional general aviation services, drones have introduced a novel service model, further accelerating the growth of the low-altitude economy. The industry is poised for significant expansion, with projections indicating 50,000 drone manufacturers, 5 million registered unmanned aerial vehicles, and 1 million licensed drone operators in China by 2030. The development and adoption of electric vertical take-off and landing (eVTOL) aircraft and intelligent aviation logistics equipment will also contribute to this growth.

Key enabling technologies for the low-altitude economy include AI chips, multimodal foundation models, and big data platforms. AI plays a crucial role in enhancing aircraft intelligence and optimizing airspace utilization and safety through intelligent traffic management. Low-altitude aircrafts will become an integral part of embodied AI, and low-altitude flight management will increasingly rely on AI-powered digital platforms. In the future, autonomous and self-evolving systems may assist or even replace humans in management and service roles.

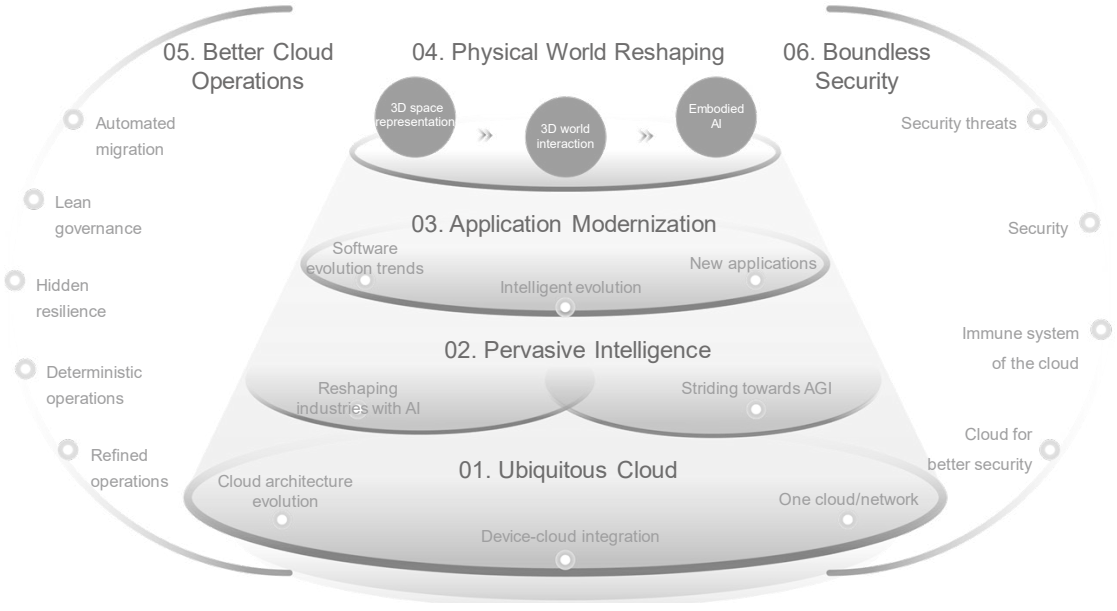
Three primary use cases will drive the growth of the low-altitude economy:

- Green city: Small drones will come into play for sustainable urban development, completing tasks like city inspections.
- Inter-city smart travel: eVTOL aircrafts will be an efficient and convenient option for passengers.
- Large-scale unmanned freight: Large freight drones will address logistics and distribution challenges in remote areas.

The expansion of the low-altitude economy will transform cities from a "2D economy" to a "3D economy", as economic activities extend into the airspace.

By 2030, the global low-altitude economy is projected to contribute USD50 trillion to global economic growth.

Key Technical Features



3.1 Ubiquitous Cloud

3.1.1 Cloud Architecture Evolution: AI Native Architecture

As artificial intelligence (AI) develops rapidly and becomes more widespread, the cloud infrastructure is shifting from a traditional general-purpose computing platform to an AI native architecture. The AI Native architecture should deliver the following technical features to comprehensively improve the performance, efficiency, and environment sustainability of AI applications: diverse compute, peer-to-peer architecture, simplified network, large-scale pooling, proclets, resource flexibility, and low carbon.

1) Proclets: Dynamic Allocation of Process-level Resources Without Compute Unit Segmentation

Cloud resources on VMs or Docker containers require reservation in various ways. However, the granularity of resource allocation is too large. Even serverless applications must run on VMs or containers. Static allocation is inefficient and cannot achieve the best resource utilization.

Proclets categorize resources by memory, CPU, or GPU, and allocate resources at the process level. This eliminates the need for segmentation like function compute units. Proclets use fixed, physical thresholds and combined production, and they are small and fast enough to not be noticed by upper-layer applications.

By 2030, around 20% of compute that is in cloud data centers is expected to be split and assigned as proclets, resulting in significant cost savings of billions of dollars for enterprises.

2) Flexible Compute: The Next Hop of Elastic Compute, Increasing Resource Utilization to 70%

Traditional cloud servers come with fixed specifications and are complex to deploy. This often results in either overprovisioning, which wastes valuable resources, or underprovisioning, which means demand cannot be met during peak hours. Typically, more than 80% of compute resources are allocated, but only a little over 20% are actually utilized.

It is estimated that by 2030, flexible compute will significantly enhance cloud resource utilization, increasing it to 70%.

In the future, flexible compute will incorporate the following key technologies:

Intelligent dynamic overcommitment involves real-time monitoring of resource profiles for individual instances and analyzing their CPU utilization. Resource allocation is then intelligently adjusted to ensure that each instance receives the CPU resources it needs. Dynamic CPU allocation enables zero-latency, user-unaware vertical scaling.

AI can be used to forecast resource requirements by analyzing the resource usage history of applications. It intelligently forecasts service needs and dynamically adjusts compute capacity to ensure that applications always run optimally. The entire process does not require manual intervention. Intelligent horizontal scaling can be automated by recognizing time sequence-based resource needs of large-granularity applications.

Flexible memory enables dynamic memory overcommitment. Unlike traditional methods, that are application-unaware and asynchronous, this technology monitors applications' memory usage and provides synchronous overcommitment. This ensures efficient memory utilization and high application performance, making memory management more intelligent and refined.

3) Diverse Pooling: The Next Step in the Evolution Away from Monolithic Compute

Cloud computing has been evolving from monolithic compute to diverse pooling compute. Traditional servers can only provide the compute of a limited number of CPUs, GPUs, and NPUs, and their ratios are fixed. But AI service requirements are diverse. The same application can have a diverse range of compute requirements, sometimes requiring dozens, hundreds, or even larger number of processors. The cloud architecture needs to be able to flexibly configure diverse compute provided by CPUs, GPUs, and NPUs as required. The selected compute resources can be tightly coupled into a resource pool through an ultra-high-speed network. Resources in the pool are dynamically provisioned

to meet changes in demand. This ushers in a new era of intelligent compute. Tightly coupled resource pools can deliver 10-fold higher computing performance than traditional servers and are suitable for a wider range of scenarios.

4) Peer-to-Peer Architecture: Shifting from Primary/Secondary for Direct Communication

With the rapid growth of the AI and various high-performance computing requirements, the traditional primary/secondary architecture struggles to keep up with to the increasing performance requirements due to its CPU-centric resource bottlenecks and transmission delay. The compute resources in data centers need to be organized and communicate with each other using a decentralized, peer-to-peer architecture. The unleashed compute can reduce the latency as low as microseconds and increase bandwidth sharply. It is estimated that by 2030, 60% of cloud data centers will evolve from a primary/secondary to a peer-to-peer architecture.

5) Simplified Network: An All-in-One Network for Cluster Connectivity Across AZs and Regions

The rapid development of AI applications brings unprecedented challenges.

- As AI applications grow, they require more diverse network capabilities, including scale-out and scale-up, as well as cluster connectivity across AZs and regions. This has led to increasingly complex network architectures.
- Scaling out means retaining redundant resources to handle traffic surges, but over-deployment reduces resource utilization and drives up costs during off-peak hours. In addition, scaling networks out and up, and deploying VPCs independently increases costs and makes O&M harder.
- Traditional network architectures are insufficiently responsive to traffic spikes. Once a cluster's size is determined, it is challenging to expand, limiting flexibility.

To address these challenges, a new simplified network architecture is introduced. The core advantages are as follows:

- Multiple independent networks are integrated so that AI applications, general-purpose compute nodes, and storage resources can share a high-bandwidth network, improving resource utilization, reducing costs, and simplifying management.
- AI applications can run seamlessly across AZs and regions, and the simplified network can scale out during traffic spikes and scale back in when traffic returns to normal.

It is estimated that by 2030, 60% of cloud data centers will adopt the all-in-one simplified network architecture.

6) Core Architecture Transformation: Plug-and-Play, Multi-modal Fusion, Ubiquitous Distribution - Intelligent Cloud-Native Databases

Currently, the core architecture for enterprise services is centered around resources, regions, and loads. Data silos throughout the data lifecycle create problems like inconsistent data across regions, non-uniform security policies, excessive transmission latency, and expensive service development and maintenance.

With the popularization of AI compute power, high-end compute delivered by NPUs and Rack clusters has become more accessible. Against this background, how to use new hardware and new computing power to enable the intelligent transformation of enterprises has become a key issue for data management systems.

In the future, databases will exhibit the following technical characteristics:

Data Access as a Service: Applications do not need to be concerned with the underlying data model that the data is stored in. Serverless, a cloud-native model, empowers databases to

manage, query, and retrieve multiple types of data with the key capabilities. It enables query and storage of heterogeneous data, and it supports Hybrid Transactional/Analytical Processing (HTAP) processing. This reduces the cost of managing applications and using data, accelerating the release of data value.

Intelligent Native Data Management: Intelligent SQL optimization and transformation designed for application developers reduces the barrier to entry for developers. Intelligent Q&A and operations and maintenance (O&M) for next-generation DBAs improve O&M efficiency by 80%. Data management systems have been evolving from relational models to various new models integrating Large Language Model (LLM) and SQL execution. This transition supports real-time inferential and knowledge computing.

Cloud-Native Fully Pooled Architecture: Decoupling and pooling of resources (such as CPUs, memory, and storage) enables elastic scaling, transparent to the applications, in seconds. Decoupled resources can allow you to double the performance you get out of the same amount of compute. This provides performance compatibility for enterprise databases with terabytes or more of data, and offers rapid, smooth scalability for distributed databases.

It is expected that by 2030, we expect to see an all-in-one database management system based on new hardware and cloud-native architecture. We expect to see a system with autonomous data management, intelligent optimization of data processing, and intelligent data security protection. We are truly stepping into the era of data intelligence.

7) Cloud Service Reshaping: From Regional to Global

Traditional cloud services allow geographical flexibility and data localization by region. However, distributed applications are limited in performance optimization and management around the world. By 2030, cloud services will evolve from a regional architecture to a global architecture. With a global design (such as regionless), applications will break through the geographic restrictions and be optimally deployed around the world with SLA guaranteed.

Global data services: By 2030, 80% of applications will use unique IDs for cross-region data access, which improves data management efficiency by hiding cross-region statuses and simplifying the data flow process. Nearby data access around the world enables a subsecond latency. The cross-region data computing performance can get a 5-fold increase and the DR time can be shortened as low as 1 minute.

Global storage services: By 2030, the global storage capacity is expected to increase from 12 ZB to 28 ZB, and the cloud storage capacity to increase from 2 ZB to 5 ZB. Global storage services can be quickly accessed, and the data upload speed will be improved by 50% to 70%.

Global network services: Regional network services are transforming to global network services to handle the challenges of geographical isolation. Network services, such as Direct Connect, VPN, Enterprise Router sharing, and Endpoint, will be globalized to provide low-cost, cross-region VPC communications and cross-region access to gPaaS & AI DaaS services for the same tenant. An application-oriented network model will be provided for simplified management and configurations as well as seamless connections between regions or sites.

Global application distribution: By 2030, the global architecture will greatly simply the development of distributed applications. It is estimated that 80% of applications will be automatically distributed based on the SLAs, and 90% will support global distributed architecture throughout their development lifecycle.

8) Cloud Computing Clusters with ZFLOPS of Compute: No Constraints of Compute, Storage, and Networking

It is estimated that the cloud computing clusters will exceed ZFLOPS of compute by 2030. Cloud data center technologies need to make breakthroughs in the following aspects:

Compute: Physical supernodes need to be evolved into logical supernodes in ultra-large AI acceleration clusters. Elastic cluster compute can then provide efficient training and inference support for multi-modal and trillion-parameter models. A peer-to-peer pooled compute architecture can flexibly adapt to the requirements of various AI applications and intelligently optimize resource configuration.

Storage: A system like this will require petabytes of memory, so a high-performance cache pool and a tiered storage solution are used to improve the storage capability and reduce costs by 75%.

Network: The bus bandwidth can be increased by 30 times with ultra-high bandwidth and high-performance interconnection technologies. A unified protocol is used to connect the AI network and data center network (DCN), ensuring high-speed transmission of a large amount of parameters and gradient information. In addition, a new topology architecture and advanced routing technologies will be used to reduce the number of network hops and provide a deterministic transmission solution to solve long tail latency issues such as P95 or P99

9) Cloud Data Centers: Green and Reliable

1. Low-carbon power supply: Low carbon propels the innovation of cloud data centers. Tomorrow's cloud data centers will feature advances such as liquid cooling, cooling storage, waste heat utilization, and energy storage. By complementing that tech with renewable energy such as solar, wind, and nuclear, these centers slash their energy consumption and carbon emissions. The combination of new energy and energy storage

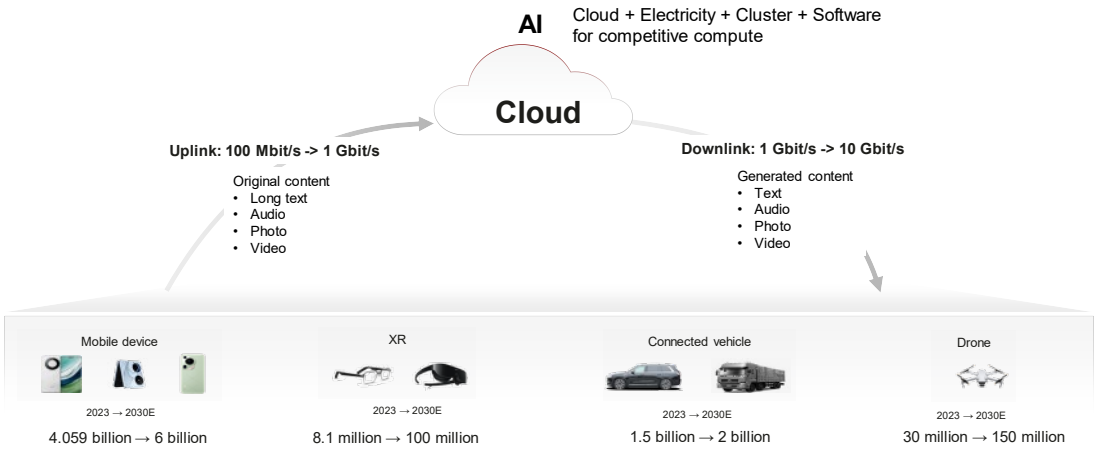
will improve power supply balance and stability of renewable energy systems. It is estimated that by 2030, more than 70% of data centers will be on the cloud, where 100% of power supply is green.

2. AI-powered O&M: Another prominent feature of cloud data centers is O&M intelligent enough to cover timely identification, analysis, prediction, and handling of risks and issues. Smarter O&M makes data centers smarter and more automated. For example, intelligent inspection is 100 times more efficient than manual inspection in identifying risks.

Robots will significantly improve O&M efficiency and quality while reducing the risk of misoperation, especially in labor-intensive and repetitive tasks. 2030 is predicted to see robots take over half of O&M.

3. Streamlined, reliable architecture: Cooling and power supply systems are still the main culprits of data center service interruptions. Evolving centralized cooling to a distributed structure would prevent the failure of a single device from affecting other devices, meaning less impact and higher reliability. This new architecture prevents a single point of failure in the cooling system from crashing the entire data center. The pooled interconnection of power plants makes it possible to simplify the power supply architecture in 2030, where power utilization is more efficient even while removing diesel generators, UPSs, and batteries.

3.1.2 Device-Cloud Integration: 100-Fold Intelligent Computing Enhancement for Devices and Applications Powered by Device-Cloud Synergy



As devices become smarter, applications like XR and autonomous driving demand higher computation and latency lower than 50 ms. Device-cloud synergy offloads processing, boosting on-device intelligence by 100 times and feeding the next wave of new applications.

1) 5G-A Large Uplink: Stimulating Device-cloud Synergy and Promoting Traffic Growth

5G-A technology is set to revolutionize uplink capabilities, with projections indicating a tenfold increase in uplink peak rates to 10 Gbit/s by 2030, complemented by downlink peak rates of 1 Gbit/s. This advancement ensures near-ubiquitous accessibility for cloud applications, exceeding 99% reliability. It facilitates rapid data upload from devices to the cloud, potentially raising monthly network traffic to over 1000 PB.

Concurrently, device-cloud synergy's end-to-end latency could be sub-50 ms.

2) New Computing Architecture: Distributing Computing Tasks Across Devices, Edge, and Cloud

In the age of AI, the demand for processing power outstrips the capabilities of traditional device-centric models. The strain on devices in terms of computing, memory, and power consumption is growing, posing a bottleneck to on-device intelligence. To address this, a novel computing architecture is emerging that distributes processing tasks across devices, edges, and clouds. By 2030, it is anticipated that 70% of computing tasks will be handled on devices, with the remaining 30% offloaded to the cloud. This architecture will enable the cloud to amplify device computing power by 100 times, leveraging the cloud's virtually limitless power and resources like clusters and software. It will facilitate a computing capacity that could scale up to 40 ZFLOPS. Furthermore, the evolution from device-native to device-cloud collaborative OSs will be pivotal. Future applications will inherently support heterogeneous compute across devices and clouds, ensuring secure and reliable data exchange, adaptive task scheduling, and collaborative training and inference capabilities.

3) Explosive Growth of Intelligent Applications on Devices, Vehicles, and Wearables

AI-powered devices and PCs are leading a technological shift, offering end users advanced assistance in document analysis, meeting summarization, graphic and text generation, and AI-aided home education. These capabilities enable daily activities, learning, and professional tasks to be seamlessly managed through cloud-empowered devices with robust compute. The smart vehicle industry is set to accelerate vehicle-cloud synergy and the development of smart road infrastructures. Smart vehicles will diversify their interactions through the cloud, encompassing smart driving, on-vehicle entertainment, and vehicle-road-cloud smart transportation systems. Moreover, advancements in world model technologies, real-time 3D technologies, and zero-latency interactions are propelling the growth of wearables like smart glasses. By 2030, it is projected that the number of AI devices facilitating device-cloud synergy will soar to 3 billion, and that of intelligent networked vehicles hits 300 million.

4) Personal Privacy Protection: Confidential Computing

Personal privacy extends beyond devices to the cloud. Ensuring privacy and confidential computing becomes paramount. Besides the security for devices themselves, the cloud must safeguard against data breaches, ensuring that the computing environment is trustworthy and that data transmitted by devices is anonymized and encrypted. End-to-end security is crucial, creating a secure link between devices and cloud, preventing unauthorized access and ensuring data in transit remains indecipherable. A new, robust security technology stack is in demand, which includes implementing confidential computing, security sandboxes, no privileged access, and data deletion after use. The cloud itself cannot get its hands on any user data neither.

3.1.3 One Cloud

"One Cloud" is the vision of a globally integrated cloud computing platform, unifying data and services across the world. It transcends geographical boundaries, empowering businesses to deploy applications swiftly and conduct global data analytics. This approach not only fuels digital transformation and innovation but also enhances competitiveness in the market.

1) Customers' Perspective: From Siloed IT to Integrated, Tiered Cloud Architecture

By 2030, the corporate landscape will shift from siloed IT systems to a cohesive, tiered cloud architecture. "One Cloud" will streamline

connections between headquarters, branches, and edge facilities, aligning with public cloud services to deliver a seamless customer experience.

One hybrid cloud: Hybrid IT will dominate by 2030, with projections showing 90% of large enterprises and 60% of SMEs adopting this model, a significant rise from 2024's 60% and 30%. This unified platform will harmonize online and offline resources, enabling real-time data sharing and informed decision-making.

One global distributed cloud: The influence of multinational companies will escalate, with an anticipated increase from 100,000 to 140,000 entities by 2030. Global innovation centers are set to more than double, from 800 to 2,000, and the local workforce percentage is expected to grow from 50% to 70%. A unified cloud platform will orchestrate cross-regional services and data, optimizing resource management and service collaboration.

One cloud to connect edge and devices: Edge and on-device computing by 2030 will decentralize data processing further. The global edge device count is expected to surge from 20 billion to 50 billion, with 85% of large enterprises and 50% of SMEs adopting cloud-edge-device architectures. 50% of data will be processed at the edge, and the unified cloud platform will integrate these layers to boost performance, speed, and reliability across the board.

2) Business's Perspective: One Network for High Bandwidth, Low Latency, Massive Connections, and Global Resource Sharing

The evolution of service innovation and network technology has transformed "One Cloud" into a sophisticated global "One Network," enhancing bandwidth, reducing latency, expanding connectivity, and improving security. This advancement empowers businesses to efficiently share global resources and data, driving digital transformation and fostering regional collaboration.

One network for media: By 2030, the online video user base is expected to hit 6 billion. Cloud platforms and real-time streaming will revolutionize sports and entertainment events. With 85% of media companies leveraging cloud and AI for content curation and audience analysis, Huawei integrates high bandwidth and low latency to create one global network for media, offering personalized and interactive experiences.

One network for vehicles: Daily vehicle data generation would range from 0.2 TB to 1 TB by 2030, with a significant rise from 35% to 80% in smart transportation infrastructure in major cities worldwide. The one global cloud, integrated with IoV technologies, autonomous driving, and intelligent transport, utilizes end-to-end encryption to build one global network for vehicles, enhancing mobility intelligence, road safety, and transportation efficiency.

One network for enterprises: The daily data output of global enterprises is set to grow from 125 EB to 500 EB by 2030, with cloud storage capacity expanding from 0.5 ZB to 1.5 ZB and bandwidth demands increasing by 50%. Secure and efficient network technologies would help construct one network for enterprises, enabling efficient global resource sharing and data transmission, and promoting cross-regional collaboration.

One network for cities: The data generated by global cities is anticipated to jump from 0.1 ZB to 1 ZB daily by 2030. 80% of smart cities would employ AI for data analytics, predictive maintenance, and automated management, and 70% of smart city solutions would incorporate edge computing. Integrating intelligent technologies and infrastructures enables one network for cities for efficient digital city management, improves residents' quality of life, and enhances urban operational efficiency.

3.2 Pervasive Intelligence

The vast mountains of knowledge accumulated over the past few decades of the Internet era have been integrated into large language models (LLMs) as tokens, establishing a robust foundation for intelligent services. Looking ahead, these intelligent services are expected to expand from the cloud to the edge and onto end-user devices, permeating every organization and the life of every individual. This evolution will profoundly impact software and applications. By 2030, it is estimated that every enterprise will have at least one custom-developed large AI model, and every employee will be equipped with at least one AI agent. Consequently, every piece of software will be redesigned or refactored using LLMs, and every application will be developed using some form of AI-supported programming tools.

3.2.1 AI Reshapes Industries, Tackling Big Challenges and Driving the Intelligent Economy

Across various industries, AI is being integrated into enterprises' core production systems to address big challenges. It is poised to become a crucial driver of productivity during the current industrial revolution. We anticipate that AI will swiftly transform the workforce and job market, ushering in an intelligent economy.

1) AI Ignites the 4th Industrial Revolution

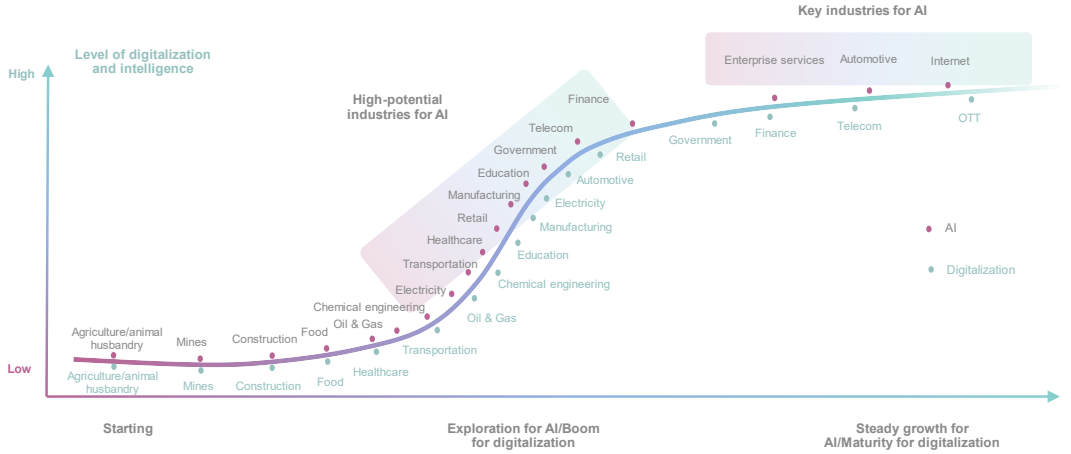
AI is at the heart of the 4th Industrial Revolution, driving innovative business models and fostering an intelligent economy. As the latest general-purpose technology (GPT) following the steam engine, electricity, and information technology, AI has the potential to significantly enhance productivity and reshape the global economy in numerous ways.

Boosting productivity: AI is evolving from AI Copilots, which are intelligent assistants offering information and suggestions, to AI Agents, autonomous agents capable of executing complex tasks independently. The next stage is the AI Workforce, where teams of AI agents collaborate to perform a broader range of complex and creative tasks. This evolution will continuously push the boundaries of automation and intelligence, transforming productivity.

Reshaping the job market: AI is expected to replace some repetitive and routine jobs, such as taxi drivers and graphic designers. However, it will also create new jobs that require skills and creativity, such as AI prompt engineers and AI data scientists. Estimates suggest that by 2030, AI will replace around 400 million jobs, but it will also generate approximately 97 million new jobs.

Increased contribution to economic growth: Statistics indicate that AI currently accounts for approximately 1.6% of the global economy. By 2030, it is projected to contribute an estimated US$2 trillion.

Transforming traditional industries and creating new ones: AI will not only transform traditional industries but also create new ones, such as digital therapeutics, personalized AI companions, and chronic disease management.

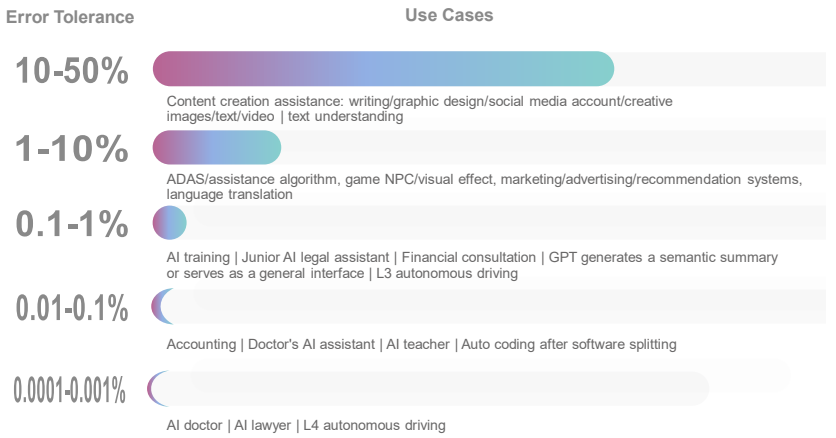


2) AI for Industry: Faster Innovation and Higher Efficiency

AI is set to transform industries by fueling innovation and boosting efficiency.

- **Fueling innovation:** AI supports creative jobs by automating the creative process and offering creative ideas. By using AI to drive automation in tasks such as communication, document collaboration, and interpersonal interaction, employees can free up more time and energy to focus on more creative tasks. This goes beyond accelerating innovation in areas such as customer relationship management, marketing, and sales. It will boost innovative power across entire creative industries.
- **Boosting productivity:** Trained on massive amounts of real-world data, AI has reached or even surpassed human-level performance in many domains. This is particularly true in knowledge-intensive tasks, such as software engineering and development, where AI is significantly boosting productivity.

These advancements, however, do not mean that AI will soon replace humans in all fields of work. The speed and likelihood of AI replacing humans depend on the error tolerance of specific jobs. Jobs that are highly error-tolerant are likely to be the first ones to be displaced. Specifically, creative jobs such as writing, graphic design, and creative copywriting are likely to be first ones to be impacted, followed by jobs that require higher precision, such as education and training and financial consultation. Jobs like doctors and lawyers, which require the highest level of precision, are likely to be the last ones to be displaced.



By 2030, it is estimated that 75% of AI's contribution to global GDP will originate from creative and knowledge-intensive industries.

3) AI for Science: AI Will Transform Scientific Research, with AI Supporting Over 50% of Scientific Computing Tasks

Traditionally, scientific computing has depended mathematics and numerical methods to solve partial differential equations (PDEs). However, computations can be slow or even fail due to limited computing power or the curse of dimensionality. Today, scientists are increasingly turning to AI for solutions. The key is to use data-driven methods, such as machine learning and deep learning, to discover patterns in vast datasets. Looking forward, we anticipate three major shifts in scientific computing:

- **From traditional methods to intelligent scientific research facilities on the cloud:** Scientific research has long relied on traditional labs and theory-driven methods. In the future, it will come to rely more on intelligent facilities on the cloud, including high-performance computing, large, interdisciplinary scientific computing models, and AI assistants. This will enable more efficient scientific research processes and accelerate new discoveries. With on-demand, theoretically unlimited computing power, the cloud solves long-standing challenges for scientific computing.
- **From manual to autonomous experiments:** With advances in AI4S, self-driving labs will become the new norm. AI-powered robots and intelligent lab environments will enhance efficiency, minimize human errors, and enable scientists to concentrate on more complex issues that demand genuine human ingenuity.
- **From independent research to large-scale interdisciplinary collaboration:** Scientific research is shifting from the independent work of individuals or small teams to large-scale, cross-disciplinary collaboration. The large-scale deployment of technologies such as AI agents and privacy computing allows us to build secure, collaborative environments for scientific research. This helps to eliminate data silos while protecting the data ownership of different parties.

By 2030, it is estimated that AI will be utilized in over 50% of scientific computing tasks, potentially accelerating computational speeds by up to 10,000 times in fields such as meteorology and pharmaceuticals.

3.2.2 Striding Towards AGI

Looking ahead, AI is steadily progressing towards artificial general intelligence (AGI). We anticipate rapid advancements in architecture optimization, enhancement of multimodal capabilities, adoption of next-generation architectures, and manual alignment. Generative AI will continue to revolutionize productivity in the media industry, while discriminative AI will maintain dominance in over 50% of all AI use cases. Additionally, AI agents are expected to evolve into super apps, seamlessly connecting enterprise applications.

1. Intelligence as a Service for All Scenarios

As a new general-purpose technology, AI in all its forms - not just chatbots, AI plug-ins, and AI engines that enhance traditional applications, but also AI-native software and hardware - can be offered as intelligent services to augment human perception, cognition, and decision-making. By 2030, Intelligence as a Service will be ubiquitous.

- **All-scenario applications:** In terms of vertical industries, AI will be more widely adopted by the Internet, government, telecom, finance, manufacturing, energy, education, healthcare, transportation, and retail. In terms of horizontal functional domains, AI will see wider adoption in R&D, production, supply, sales, services, operations, and maintenance. We estimate that by 2030, over 90% of companies all over the world will have adopted AI technology in some shape or form.

- **All-modality:** In AI, modality refers to the type of input and output data. As AI applications shift from making predictions to generating things, AI modalities are extending beyond natural language, speech, images, and videos to even more diverse forms, for example, the sense of touch, taste, and smell for humans, and infrared, inertial navigation, and remote sensing data for machine perception. Enhanced data modality support will significantly boost AI's ability to create digital worlds and transform the physical world.

- **Cloud-edge-device synergy:** AI models will be deployed across the cloud, edge, and end-user devices. Smaller models distilled from larger models will be deployed to personal computers in the hundreds of millions, mobile phones in the billions, and IoT devices in the hundreds of billions. Along with new, AI-native devices, they will make Intelligence as a Service available everywhere, facilitating people's life and work everywhere.

- **Multi-size deployment:** 1B to 3B models will be deployed on end-user devices such as mobile phones and tablets, 6B to 7B models on personal computers, and 10B, 100B, or even larger ones on large servers and clouds. Models of various sizes can meet diverse AI needs. Estimates show that by 2030, models smaller than 10B will account for over 95% of all models developed and deployed, instead of the 38% today.

- **Inclusive services:** The marginal cost of Intelligence as a Service is near zero. Today, AI models' inference cost has dropped to less than CNY10 per million tokens. Comparing this with the cost of a typical free search, which is approximately US$ 2 cents, it is safe to say that Intelligence as a Service is becoming more and more inclusive.

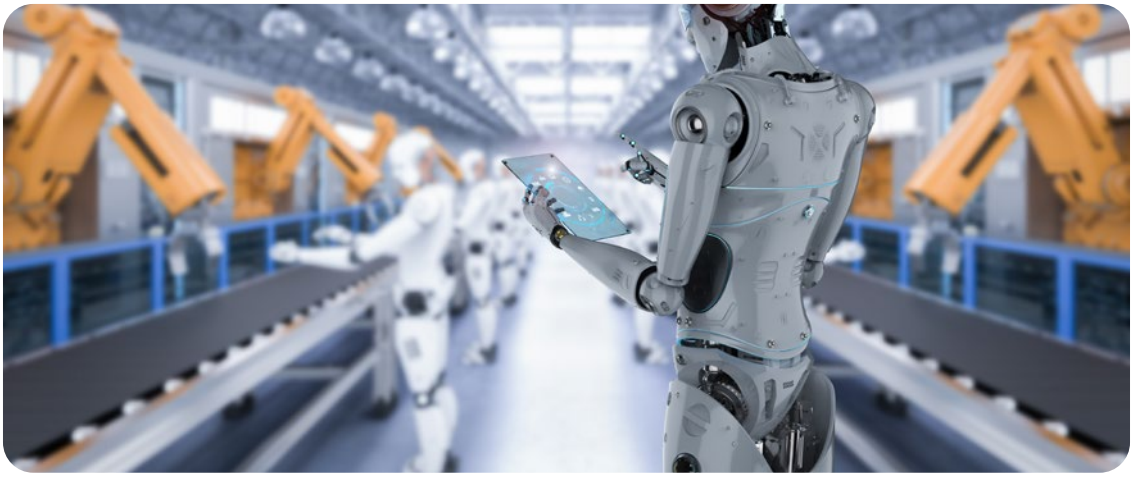
- **Data Intelligence:** From Data-Centric to Knowledge-Centric Enterprises currently use data primarily for business intelligence, process optimization, and control - mainly analyzing structured data. However, in the era of large AI models, current data platforms fall short in data cleansing and knowledge extraction. Building a knowledge-centric data foundation is now crucial. Extracting knowledge from vast datasets helps enterprises consolidate business logic. When combined with industry expertise, this knowledge can be integrated into IT systems, providing companies with quick, relevant, and actionable insights. For instance, in human-computer interaction, it can support efficient, accurate, and intelligent business decision-making. Additionally, it supplies high-quality datasets for AI model training, fine-tuning, retrieval-augmented generation (RAG), and

prompt engineering, enhancing both training and inference processes. By 2030, the digital economy is projected to constitute over 60% of global GDP. As a key production element, data will drive enterprise innovation and growth. Intelligent data operations will form the cornerstone of industrial intelligence.

2. More Than Transformer: The Rise of Hybrid Architectures in AI

Convolutional neural network (CNN) and recurrent neural network (RNN) were once the mainstream architectures for deep learning. Later, the Transformer architecture, which supports highly parallel processing by leveraging the attention mechanism and abandoning the recurrence mechanism, emerged as the leading choice for implementing the Scaling Law. We identify four key technologies and trends that are crucial in our pursuit of AGI.

- **More than Transformer:** Transformer models have advantages in handling tasks that rely on context learning and complex reasoning, while pure Mamba models have better performance in long-sequence training and inference. A hybrid architecture that combines the strengths of both models has already been applied to various tasks, including HD image generation, point cloud analysis, and time series forecasting. This makes it a promising candidate for developing future foundation models. Additionally, these models must continuously increase intelligence per FLOPS as well as intelligence per bit.
- **The shift from multimodal to all-modal, and from language-centric to native multimodal:** The results from existing multimodal models have proven that native multimodal training methods can effectively improve model performance. Currently, data processing predominantly uses "Any to Text" and "Text to Any" approaches. In the future, we may see the rise of native multimodal data processing methods in the form of "Any to Any." By 2030, it is anticipated that unified methods for tokenizing multimodal inputs for models will have emerged, facilitating a comprehensive understanding of the world. Concurrently, unified methods for tokenizing multimodal outputs will also be developed, ensuring AI-generated content more closely mirrors reality.
- **Next-gen neural networks improve model adaptability:** Mimicking natural neural networks, spiking neural networks (SNNs) have demonstrated unique advantages in processing time series data such as voice and video. Liquid neural networks (LNNs) feature adaptive weights, unlike the fixed weights typical of traditional models. LNNs can dynamically adjust their weights based on input data, resulting in a smaller and more interpretable neural network architecture. By 2030, it is estimated that new network architectures, such as SNNs and LNNs, will surpass current mainstream models in terms of performance and cost-efficiency in specific domains.



- **Addressing hallucination and explainability issues through manual alignment:** In tasks that prioritize high precision, AI hallucination can be a significant problem. However, in creative tasks, it may be viewed as imaginativeness that inspires human creativity. AI alignment is about aligning AI with human preferences, goals, values, and ethical principles. Future advancements in AI alignment techniques and capabilities will make AI more explainable and reliable. Specific AI alignment methods include learning from feedback, learning under distribution shifts, alignment assurance, and AI governance. As engineering practices improve, models' ability to handle data contamination and misleading prompts will also significantly improve.

Cloud services have a significant role to play in fueling the development of AI. During model training, the cloud can provide a reliable, large-scale network and essential SRE capabilities, which are crucial for improving model floating-point operations utilization (MFU). Currently, the average MFU is between 30% and 50%. Cloud technologies and optimizations could increase this to 60% to 70%, significantly cutting the training compute cost. Cloud services are also equally important during the inference phase. By increasing the batch size, cloud services can control the overall latency and ensure the efficiency of the inference process. Also, since most of the latency during inference is due to computation rather than data transmission, the powerful compute capacity of the cloud is expected to reduce inference latency to near-zero. Additionally, in a multi-device application scenario, cloud services can consolidate data from different devices to provide rich context, ensuring a consistent user experience.

By 2030, we anticipate that the number of parameters of large AI models will match the synaptic connections in the human brain, that is, between 100 trillion and 1,000 trillion. Models of this scale will require much larger training clusters on the cloud, scaling from 100,000 xPUs to millions of xPUs. In the meantime, the energy requirements for data centers to support the training and inference of these massive models

will surge, increasing from tens of megawatts to hundreds of megawatts.

3. Discriminative AI Continues to Create Value for Enterprises

Discriminative AI models have advanced significantly over the past two decades. They are predominantly used to make predictions based on input data and their training. In the field of computer vision, this means to carefully analyze specific images and determine their categories (image classification), or to identify and locate objects in images (object detection). When applied to structured data, discriminative AI focuses on parsing input data and predicting the target values (regression on structured data).

Although the emergence of generative AI has opened up endless new possibilities, discriminative AI models remain valuable. Their potential lies in the following key areas:

- **Unified discriminative AI models:** Most of today's discriminative AI models are small, task-specific models. For each downstream task (such as image classification, object detection, and structured regression), often a dedicated model needs to be developed from scratch, or highly customized development is required. In the future, large, pre-trained foundation models may be developed for discriminative AI. Such models are expected to achieve the desired performance level without fine-tuning. For more performance-demanding tasks, they will only need minimum supervised fine-tuning (SFT).

- **All-modal pre-trained model:** Compared with a pre-trained, single-modal discriminative AI model, a future all-modal model will support multiple data modalities, such as images, videos, structured data, point clouds, remote sensing, and audio, and integrate different modalities to enrich information. This allows it to significantly improve performance in downstream tasks.

- **Collaboration with generative models:** We see huge opportunities in combining discriminative and generative models. This synergy not only bridges the gap between generation and discrimination but also propels advancements in multimodal fusion, reinforcement learning, and adaptive systems. Such progress will drive AI technology towards more intelligent phases, ultimately leading to AGI.

Due to their reliable performance, small footprints, and high computational efficiency, discriminative AI models are widely adopted in enterprise scenarios. They excel in making accurate, highly generalized predictions and classifications based on multimodal inputs, such as camera-captured images, audio signals, and specialized sensor data. By 2030, it is estimated that discriminative AI will still account for over 50% of all AI use cases and nearly 70% of all enterprise applications.

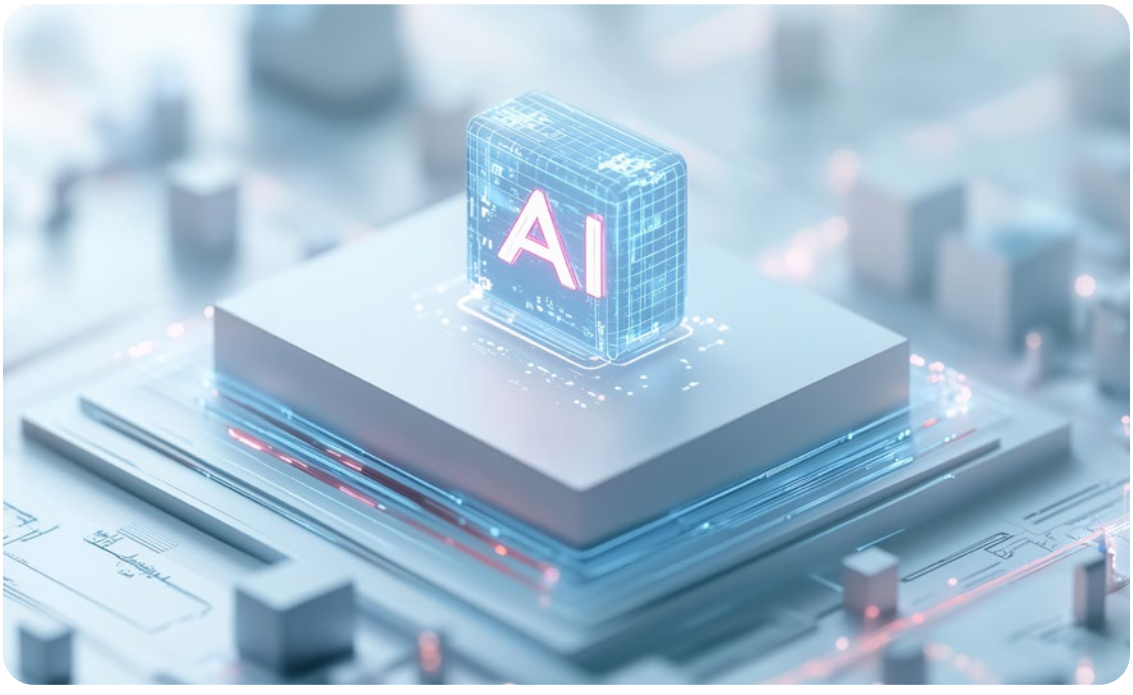
4. AIGC Transforms Media, Offering New Digital Experiences

- Traditional media content is primarily generated by professionals with cameras. Media platforms support 1-to-N unidirectional content distribution only. It is estimated that by 2030, more than 70% of media content will be AI-generated or generated with the support of AI. This will enable more personalized content creation and dissemination. Key features include:

- **Personalized content creation:** AIGC will enable faster creation of personalized content, which will be delivered to the audience by matching the user profile, ushering an era of N-to-M personalized content distribution.

- **Real-time interaction between tens of thousands or even millions of viewers:** Traditionally, online spaces can only serve a limited number of users, often in the hundreds, due to server performance limitations. Now, high-speed cloud interconnects and distributed platforms allow for real-time interaction between tens of thousands or even millions of viewers.

- **Compute-network converged media platform:** Traditional media networks rely solely on a network of caches for content distribution, while generative media platforms integrate caching,



computing, and distribution. Through a synergy between the cloud, edge, network, devices, and chips, along with OS optimization both in the cloud and on end-user devices, we can build a network with zero latency and no stalls, delivering the ultimate viewer experience.

- **Fully digitalized generative media engine:** Deep integration of CG and AI technologies, used along with 3D content segmentation and controllable generation, enable large-scale, distributed, real-time content rendering. AI technology and digitalized devices together enable all-digital content collection, editing, and dissemination.

5. AI Agents Become the New Paradigm for Application Interaction, and the New Enterprise Super App

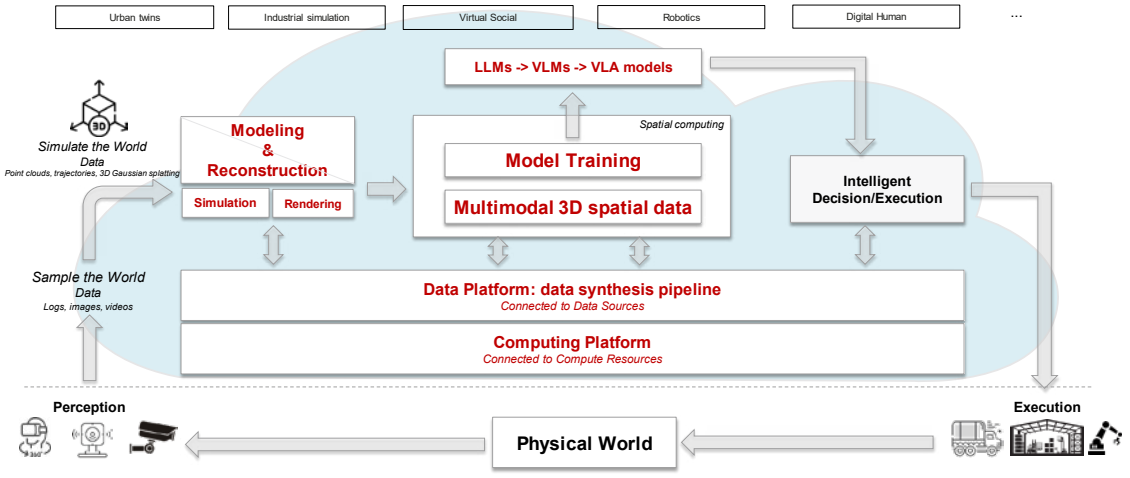
AI agents integrate four key elements of productivity: expert knowledge, data, models, and computing power. They enable intelligent interaction between users and AI-powered applications and systems. Traditionally, users interact with enterprise applications through a software interface. Now, they interact with AI agents, which autonomously execute complex tasks without being given specific instructions on how to navigate these tasks. AI agents give each employee a dedicated, intelligent assistant with enhanced capabilities, potentially making them super apps within enterprises. By 2030, AI agents will impact two-thirds of the world's jobs, and have the potential to replace 30% of the work hours across the globe. By then, over 40 million people may need to be reskilled or change jobs, and 1.5 billion company employees worldwide will have their own intelligent assistants.

Key features of future AI agents include:

- **Autonomous actions based on the preset goals:** Different from today's LLMs, which lack the ability to translate specific goals into actions, AI agents evaluate their goals, develop plans, and take appropriate actions to accomplish those goals. All of these are done autonomously.
- **Persistent memory and intelligent state tracking:** AI agents can be equipped with long-term memory or the ability to track past states. Their knowledge can be accumulated and used as the basis for subsequent decision-making and actions. All these enable more intelligent AI systems.
- **Interaction with the environment:** AI agents can perceive and understand their environment, regardless of whether it is a digital world or robotic system. In the future, AI agents will be able to interact with the physical world through sensors or other physical components.
- **Long-term learning and accumulation:** AI agents can autonomously learn from their interactions with new environments and in dealing with new situations. They can continuously optimize their knowledge systems and update their skills.
- **Multi-domain task handling:** AI agents have the potential to become general-purpose, multi-tasking AI systems that seamlessly integrate multiple skills, such as language processing, logical reasoning, perception and understanding, control and manipulation. They will assist humans in tackling a wide range of complex problems.

3.3 Transforming the Physical World

By 2030, intelligent technologies will have profoundly transformed our physical world. This fundamental transformation encompasses the entire process, from perception to computation, and ultimately to action.



New technologies such as XR devices, eye tracking, gesture recognition, and voice interaction will enable more natural and more efficient interaction (visual, voice, and action) between humans and the 3D digital world, ushering in the new era of spatial computing. Estimates show that by 2030, 60 million XR devices will be shipped annually, and around 500 million people will spend an average of 5 hours per day in a world of spatial computing that joins the physical and virtual worlds.

The transformation in perception is underpinned by wide interoperability between assets and models, enabling us to build global-scale digital twins. Digital twin and generative AI, combined together, will enable people to use these new spatial devices more effectively, generating more diverse, higher-precision 3D spaces by accurately extracting the features of the real world. The data of the 3D digital world, combined with data synthesis technology, will lay the foundation for spatial intelligence and embodied intelligence. The 3D digitization of the real world and the integration of embodied AI models into robots will enable seamless interaction and in-depth integration between the digital and physical worlds.

3.3.1 Representation of 3D Spaces: Integrating AI and CG to Accelerate Information Exchange in a 3D Digital World

The digital representation of 3D spaces has evolved from manual processing of mesh geometry, materials, and illumination to new methods that combine photo-shooting and AI generation. With convenient collection of spatial information using various cameras and innovative representation techniques, key details about real-world spaces—such as illumination, colors, materials, and spatial depths—can now be quickly and accurately gathered at low costs. Additionally, spatial-temporal data can be seamlessly overlaid

onto 3D models, accelerating the comprehensive, standardized representation of complex scenes by hundreds of times.

We estimate that by 2030, city-level 3D reconstruction solutions will cover areas up to 10,000 square kilometers, incorporating real-time city events across all seasons. Such solutions will enable real-time city-level simulations, serving as digital training grounds for L4 autonomous vehicles, drones, and robots.

3.3.2 3D World Interaction: New Paradigm of Spatial Computing, Million-Time Increase in 3D Training Data

New, AI-powered interaction devices are becoming increasingly popular. Unlike traditional cameras and LiDARs, these devices are connected to cloud-based computing power and multimodal AI models, enabling them to capture and collect information about the physical world on a much larger scale. This means computer vision technology is now shifting from "sampling the world" to "simulating the world." Training data is changing from text, images, and videos to fine-grained 3D spatial simulation data across all modalities. Data from the physical world is preprocessed using data engineering pipelines and then synthesized, providing 3D spatial data for training vision-language-action (VLA) models that power spatial intelligence and embodied intelligence. The size of this training data is 106 times larger than the currently available training data for LLMs, which is approximately 13 trillion tokens.

VLA models, pre-trained using real-world data preprocessed using AI, rendering, and simulation techniques, offer enhanced spatial perception and AI capabilities. They are driving new computing paradigms and empowering more end-user devices, creating a data flywheel that will keep reinforcing itself well into the future.

We estimate that by 2030, under the new paradigm of spatial computing, human interaction will demand 1 zettaFLOPS computing power, while data preprocessing and training of VLA models that power spatial intelligence will demand 100 zettaFLOPS computing power.

3.3.3 Embodied Intelligence: Human-like Robots Are Seeing Wider Adoption, Super-human Robots Are Taking Shape

Embodied intelligence, or embodied AI, refers to the integration of AI and robots that enables robots to understand their surroundings and themselves and interact with the physical world to perform designated tasks. It adds the element of embodiment on top of spatial intelligence, enabling the ability to act and interact with the real world.

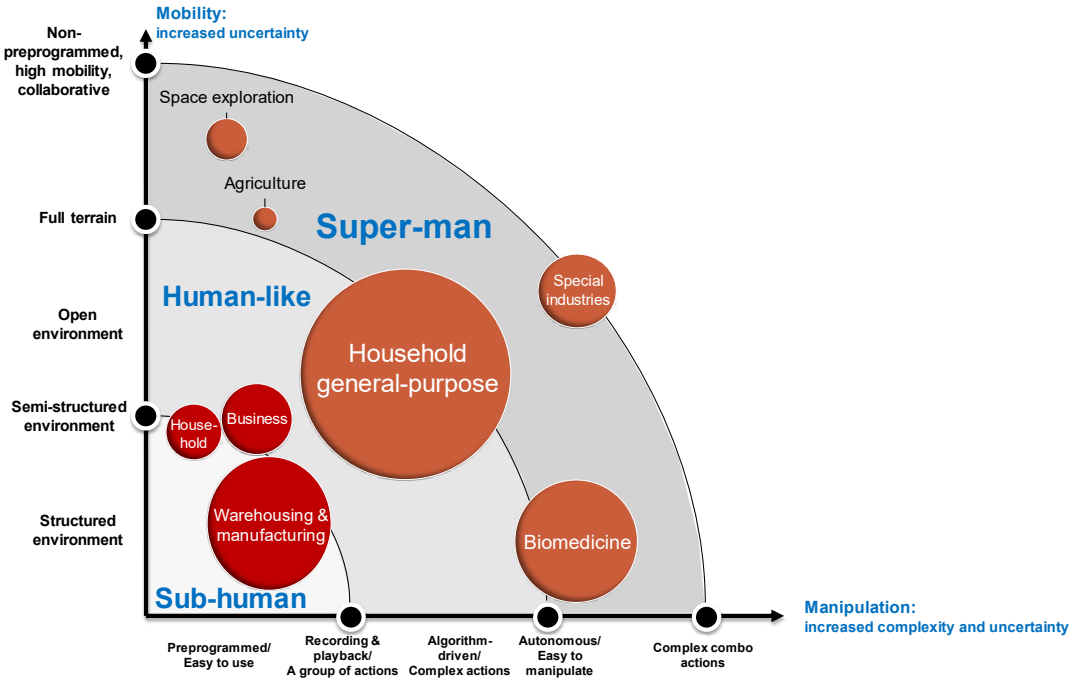
With embodied AI, robots are expected to evolve from sub-human, to human-like, and on to super-human. Sub-human robots primarily handle routine and repetitive tasks. Human-like robots are more adaptable and capable of performing more complex tasks, while super-human robots are supposed to surpass humans in all aspects, working in extreme environments and managing very complex tasks. Rapid advances in LLMs and VLAs are driving exploration of general-purpose, human-like robots. Intelligence lies at the core of robots, while data is key to intelligence. Key technologies include:

A big brain on the cloud and smaller brains on the edge: Together, they power super intelligent robots that may eventually surpass human-level intelligence.

Cloud-based simulation and data synthesis: Data for embodied intelligence is scarce. This means synthetic data is key to training the brains that power embodied intelligence.

Online learning and a closed loop of data: Knowledge is not intelligence. True intelligence can only be learned through interaction with the real world.

We estimate that by 2030, general-purpose service robots, general-purpose factory robots, and general-purpose household robots will see widespread adoption. Powered by embodied AI, their shipments are expected to reach 30 to 50 million units annually. By then, early forms of super-human robots may have emerged, capable of performing tasks that are beyond human capabilities, such as space exploration.



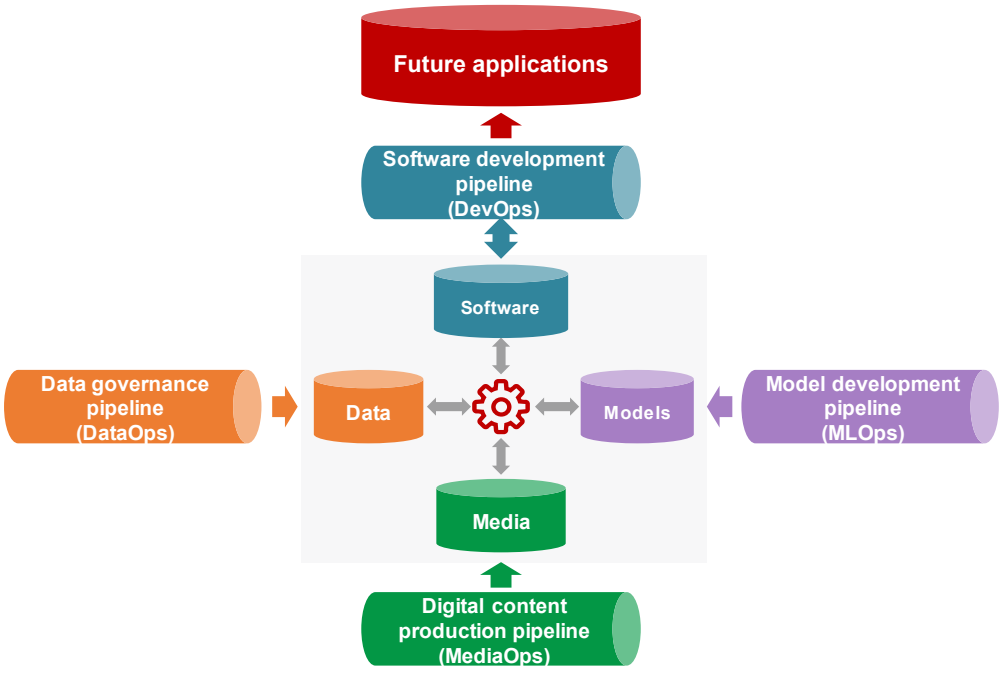
3.4 Application Modernization

According to Gartner, the digital economy will be made up of 500 million new applications in 2025, equivalent to all of the applications of the past 40 years combined. Traditional applications do not have the development, scalability, resource utilization, or O&M efficiency that are needed to adapt to this economy. In response, applications need to modernize, as can be seen by the growing trends in microservices, serverless, DevOps, and low-code development.

Future applications need an intelligent platform that integrates data governance, model development, digital content production, and software development, as envisioned in Software 3.0. They will combine data, models, digital media content, knowledge, code, and services to evolve from Cloud Native to AI Native for human-machine collaboration, continuous learning, growth, high autonomy, and swarm intelligence.

Such applications have an increasingly complex structure and will foster an ecosystem that

dynamically combines and superposes various software elements in both the social (use and management) and physical (generation) spaces, which will be integrated with the information space for a new business model by 2030.



3.4.1 Trends

1) From Code to Integrated Elements, AI Reconstructs 80% of Applications

Future applications are not just about code. They blend code, data, AI models, and digital content for more diversified and intelligent experience.

These composite intelligent systems will comprise system engineering mindsets, foundation model services, hybrid models and architectures, as well as personalized storage, retrievers, generators, and external tools. Yet their development remains economical, thanks to intelligent workflows, job and service orchestration, and component assembly in a converged cloud and AI native framework.

Future applications will integrate traditional code with data for processing, with AI and large language models for intelligence, and with digital content (such as text, images, videos, and interactive elements) for experience.

Their quality will not be just measured quantitatively, but by trustworthiness. These metrics include data privacy, output randomness, result explainability, and legal compliance.

All applications are worth rebuilding with AI foundation models, and 80% of them are expected to achieve this by 2030.

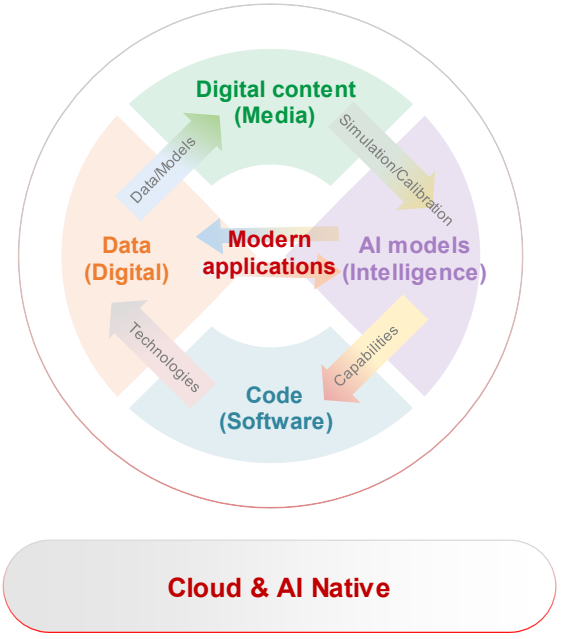
2) From Conventional to AI-powered Convergence Pipeline for 100x Efficiency

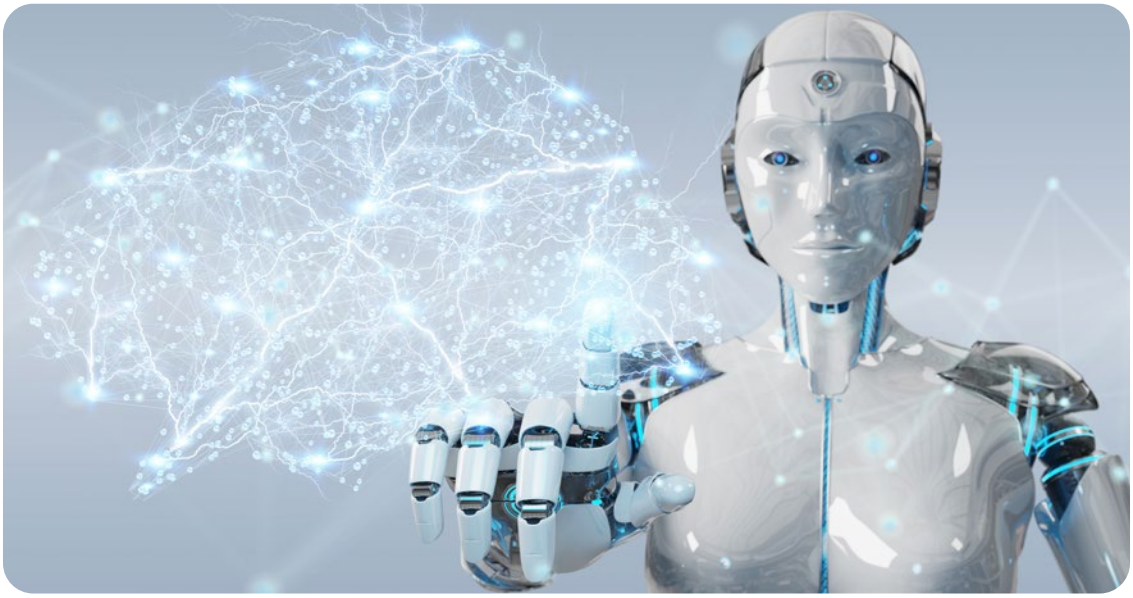
AI drives software engineering into a new era characterized by:

- **Directed programming and development:** Software is now independently developed on the loop instead of in the loop. This intelligent shift involves automated project management, risk control, team collaboration, and analysis. This journey of collaboration between people and highly autonomous agents is not only a technological rejuvenation, but also a fundamental reshaping of R&D roles.

- Intelligent pipelines: DevOps, DataOps, MediaOps, and MLOps remold services and tools to embrace new software engineering concepts, methodologies, and practices.

1. DevOps: This evolution has three phases.





Intelligent assistance (2023–2025): AI enhancement technologies, including R&D models, model fine-tuning, retrieval-augmented generation (RAG), and prompt engineering, are incorporated into existing processes and tools to boost efficiency.

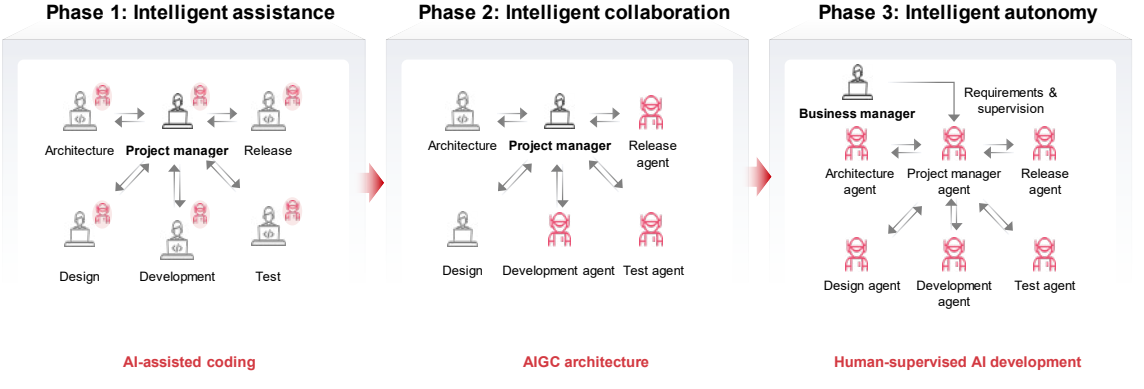
Intelligent collaboration (2026–2028): Professional R&D assistants empowered by AI agents, task decision-making, and tool ecosystems will work with humans on complex tasks to unlock productivity.

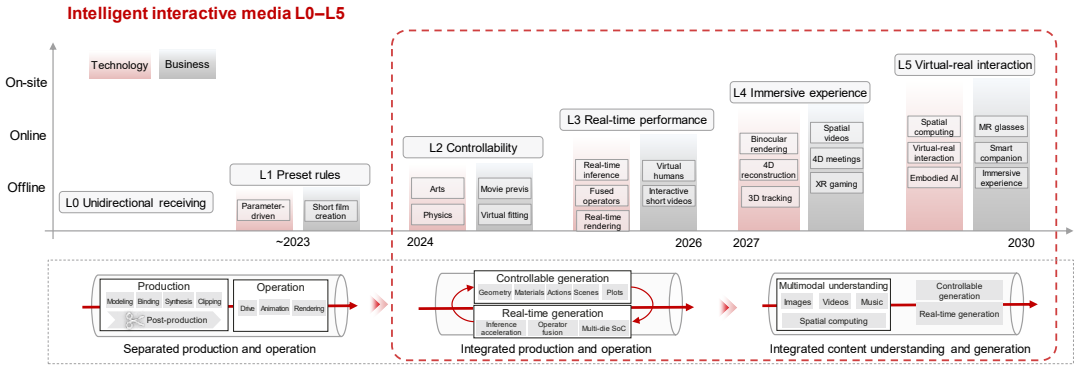
Intelligent autonomy (2029–2030): "Intelligent independent developers", fueled by artificial general intelligence (AGI), will complete R&D from end to end with little human intervention, improving productivity by 10 to 100 times.

Today, DevSecOps is essential for the secure use and governance of open-source software. The agile and automated philosophy of DevOps will merge with industries to form "DevIndustryOps", enhancing competitiveness and market responsiveness.

2. DataOps: Data collection, processing, and analysis are boosted by in-depth AI convergence. Augmented analytics combines GenAI with business intelligence, data science, machine learning, anomaly detection, and action assistance to provide new human-machine interactions in natural languages, automating decision-making with insights, code, and data.

3. MLOps: This key practice streamlines the development, deployment, and monitoring





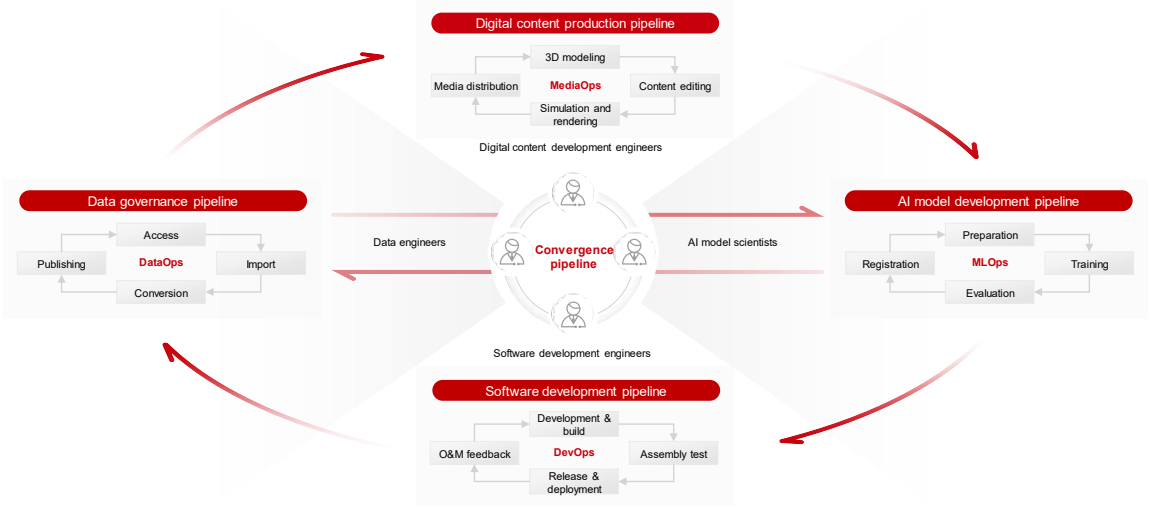
of machine learning models for iteration and processing. It will evolve and deploy AI models more quickly and widely.

4. MediaOps: The media industry is redefined with interactive short videos, scenario-aware ads, six degrees of freedom (6DOF) e-commerce, and immersive movies. These intelligent technologies profoundly change the paradigm of short videos, movies, and games from production and distribution to real-time generation and operation. It is estimated that by 2030, "AI+CG" will help 100 million digital content producers create up to 1 trillion hours of interactive media content.

- **Multi-modal software pipeline**

Each pipeline converges with human wisdom and machine efficiency to make software development more efficient, flexible, and innovative. Combined, this automated, omniscient platform integrates DevOps, DataOps, MediaOps, and MLOps to eliminate silos and bolster efficiency.

By 2030, the convergence pipeline will be widely used with any skill level for software development and innovation, increasing efficiency by 10 or even 100 times.



3) From Led by Humans to Driven by Humans + Compute + Data

Although traditional development has been simplified by emerging IDE tools and high-level programming languages, it still relies on human effort for tasks like requirement analysis, coding, project management, and team collaboration and must be measured by manpower.

Future software will involve more elements such as data, models, and digital content, which complicates the cost structure. Integrating, using, and supervising these elements still require human participation, but compute and data resources play an increasingly important role. By 2030, software costs will not be measured with just manpower, but with a mixture of elements including compute, data sets, and digital copyright.

3.4.2 Intelligent Evolution

Applications utilize cutting-edge AI to optimize, predict, and make decisions autonomously, mimicking human-like thinking and learning. In today's highly competitive market, intelligent applications not only personalize services at low cost, but also promote high-quality decision-making.

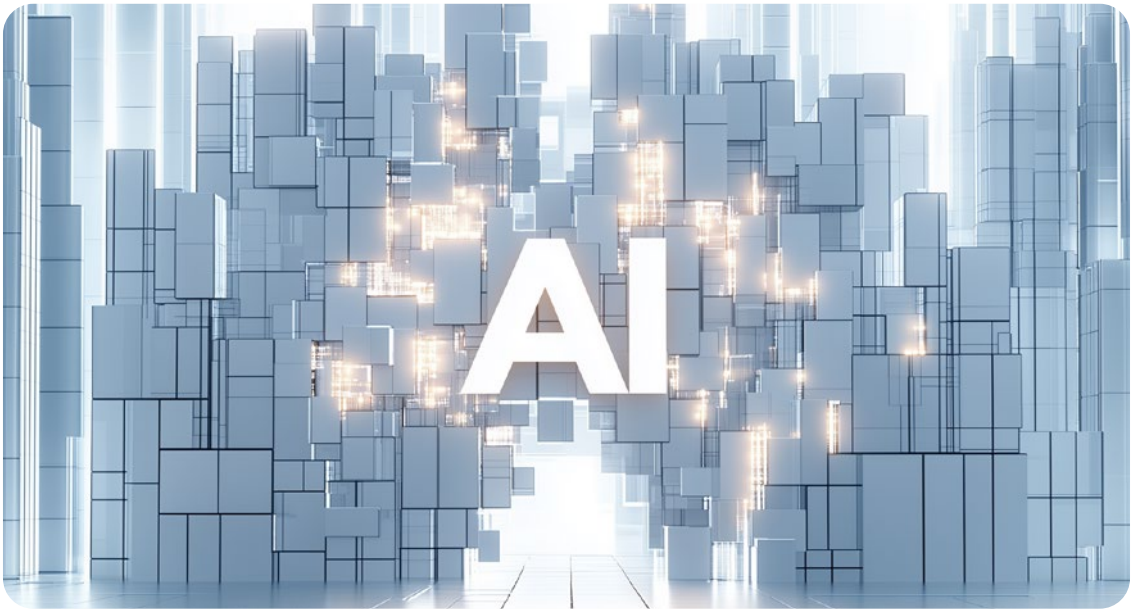
1) From Man-Machine Dialogue to Real Communication

Natural language processing (NLP) takes man-machine dialogue to the next level, where intelligent applications can comprehend complex languages and contexts and provide accurate information and services. Advancements in sentiment analysis allow these applications to recognize and react well to user emotions, interactively comforting or encouraging users based on their tone and language.

By 2030, NLP technology will humanize applications with a speech recognition accuracy of up to 98%. Customer service with sentiment analysis will reduce complaints by 30% to 50%.

2) From Traditional Operations to Intelligent Processes

Intelligent service processes are a key innovation in enterprise operations and management. By 2030, predictive analytics and pattern recognition



will be necessary for efficient operations and quality decision-making. Enterprises will use machine learning and big data analytics for insights into market trends, consumer behavior, and potential risks.

In this transformation, personalized AI assistants will analyze staff workloads and recommend custom improvements. These assistants will take on cumbersome tasks to free employees to focus on more strategic work.

In the future, every employee will have their own AI partner that will always be by their side to help them work more efficiently. This transformation will not only boost efficiency but reduce human error by over 50%. Significant decision-making in enterprises will take only half the time it used to, enabling quick responses to market changes.

3.4.3 New Applications

1) Web3 Applications: Digital Attestations and Decentralized Architecture for Trustworthiness

Web3 establishes a trustworthy foundation that is unavailable in Web2 for the digital world. Its decentralized architecture records all transactions and data in immutable ledgers, eliminating the risks of manipulation and fraud. Here, users own and control data of Web3 applications and profit by participating in the ecosystem. They are no longer passive content consumers, but creators and beneficiaries. Web3 applications can effectively supplement traditional ones for:

- Content creation and copyright protection: Non-fungible tokens (NFTs) and smart contracts help creators own and distribute earnings.
- Supply chain management: A decentralized record system improves transparency and traceability and reduces fraud.
- Financial services: Decentralized finance (DeFi) allows users to manage assets and gain profits without traditional bank middlemen.

Smart contracts and zero-knowledge proofs are crucial for Web3 applications.

Smart contracts provide code functions and the conditions for verifying and executing contracts, helping Web3 decentralized applications (DApps) execute and manage various functions and services. Smart contracts will be deeply integrated with AI to process more complex logic and decisions via autonomous learning and optimization. Bloomberg predicts that the smart contract platform market will soar to USD15–25 trillion by 2030.

Zero-knowledge proofs enable off-chain data and accounts to get verified and obtain trust in the Web3 ecosystem. This can be used for diverse scenarios such as asset proofs, anonymous voting and payment, and transaction privacy protection. With the approval of BTC/ETH ETFs in Hong Kong (China) and the United States at the beginning of 2024, more real assets will enter into the decentralized world, and more traditional and decentralized financial systems will be integrated. The current 400 million people using digital assets around the world are projected to reach 1 billion by

2030. By then, 16 trillion real-world assets will be traded and circulated. All mortgages and pledges that are now physically proven will be electronically verified and evaluated for their actual asset value. By 2030, Web3 applications will execute 90 billion zero-knowledge proofs with USD10 billion compute resources.

2) Quantum Applications: Better for Finance, Physics, Energy, and Biology

A historic "quantum advantage" is expected to happen by 2030 with practical algorithms and powerful computers. Although quantum algorithms like Grover and Shor have shown theoretical potential, they still rely on quantum computers with enough qubits and robust error correction. By 2030, qubit error correction will be precise to 99.999% to 99.9999%, and the number of qubits will grow from 1,000+ to 10,000+. These technological advancements will make it possible for quantum computing to be applied in various industries, including finance, physics, energy, and biology.

By 2030, quantum computing as a service (QCaaS) will become a new industry standard that empowers enterprises and research institutes to easily access quantum computing resources through a cloud platform without investing in or maintaining expensive hardware. Quantum computing infrastructure, algorithm/operator platforms, and application development platforms will mature on the cloud. Presently, quantum computing hardware is developing towards superconducting, ion trap, optical quantum, cold atoms, and spin quantum computing, technologies that have the potential to emerge as the major directors of wide application and commercialization by 2030. At that point, quantum computing will not only be present in scientific research but in industry innovation and transformation.

By 2030, the global application modernization market will be fueled by enterprise digitalization, legacy system reconstruction, cost optimization, and experience enhancement to reach USD380 billion, with a CAGR of 17%.

3.5 Better Cloud Operations

Cloud vendors are continuously refining their enterprise cloud concepts and frameworks based on real-world experience. They are constantly improving their understanding of cloud products and their abilities to combine them in different ways to create new product offerings. Moving forward, we can expect to see AI integrated to deliver advanced capabilities like automated migration, lean governance, hidden resilience, deterministic operations, and refined FinOps. This allows enterprises to fully leverage cloud computing and achieve their business goals faster.

3.5.1 Automated Migration: E2E Automation, 10 Times Faster Than Before

Enterprises are rapidly migrating their services to the cloud, driving the development of migration tools. These tools streamline the migration process by tracking status updates, analyzing services, designing architectures, implementing the migration itself, and verifying the data afterwards. By doing so, they help mitigate the risks and costs throughout the cloud migration. However, enterprises are still facing many migration challenges, such as identifying application dependencies, evaluating technical feasibility, selecting the most appropriate cloud resource specifications.

In the future, cloud vendors will incorporate AI into migration tools to analyze the code, configuration files, and runtime logs of the source application system. They will make efforts to identify application architectures and component dependencies. The target end's technical architecture will be automatically designed based on design principles and best practices. Cloud vendors will use AI to analyze historical performance data of the application system and forecast workloads. They will use it for automated performance testing and capacity planning. AI can help more accurately forecast resource requirements. The most cost-effective cloud resource specifications will be automatically recommended based on the cloud service pricing model. As migration pipeline technologies continue to improve, it is possible that cloud migration will be fully automated. By 2030, cloud migration is expected to be fully automated and unattended. The migration efficiency will be over 10 times higher. Complex application systems will be smoothly migrated from data centers to clouds and between multiple clouds.

3.5.2 Lean Governance: Eliminating 90% of Compliance Risks Early On

A change in quantity can entail a change in quality. Large-scale cloud migration inevitably brings many

challenges in cloud governance. Managing resources and ensuring compliance become exponentially challenging due to the complex dependencies among hundreds of service systems and tens of thousands of cloud resources. Additionally, hundreds of thousands of employees from various business lines and partners require access to these resources. If responsibilities are unclear, you can end up with disorganized permissions configurations. This can result in an exponential increase in data leak channels, making risk control much challenging. To address these challenges, cloud vendors launched the Landing Zone or Governance, Risk, and Compliance (GRC) solution, but there is still significant room for improvement. A survey shows that 76% of enterprises consider cloud governance a significant challenge. As multi-cloud and hybrid cloud environments become more popular, enterprises encounter even more difficulties in cloud governance.

Moving forward, the Landing Zone solution will use zero trust systems to establish comprehensive data perimeters. This ensures that sensitive data can only be accessed when environments, identities, networks, and resources are all trusted. AI will be used to analyze users' historical access and automatically generate and configure strategies to help enforce the principle of least privilege (PoLP) based on user roles and responsibilities. Users will be continuously authenticated and dynamically authorized based on their attributes, the resources they need, and the environments they are in. AI will be deeply integrated into the Landing Zone solution, making it easier to govern people, finances, resources, permissions, and security compliance in a well-architectured, centralized, fined-grained, and intelligent manner. It will help identify most potential risks in advance, based on complex correlation analysis. It can then automatically eliminate these risks early on. Furthermore, a unified platform will probably be used to manage multi-cloud and hybrid cloud environments. AI and automation technologies will be utilized to integrate management interfaces from different cloud vendors to establish consistent cloud governance frameworks, strategies, and processes, simplifying management. By 2030, cloud governance is expected to become more intelligent and streamlined, and over 90% of compliance risks will be eliminated in the early stages.

3.5.3 Hidden Resilience: Zero-Burden, Fully Managed, and Highly Self-Healing

Cloud application resilience refers to the capability of an application system to keep running and quickly recover from faults caused by cloud infrastructure issues, external attacks, external dependency problems, or even regional disasters. Many enterprises enhance application resilience using microservice architectures, containerization, and distributed systems. They implement automated monitoring, fault detection, and recovery systems so that applications can be automatically adjusted and rectified when faults occur. Cloud vendors also provide various tools and services to help enterprises establish more resilient application systems, which further drives the rapid advancement of resilience technologies.

By 2030, application resilience is expected to enter a new era, where it is zero-burden, fully managed, highly self-healing, and completely hidden from view. Nano-level intelligent self-healing technologies will be used to provide micro self-healing for different processes. A global unified sensing system will be used to detect faults and disasters in real time. There is no need to create numerous redundant instances in advance. New instances can be created in seconds when a fault occurs, ensuring seamless fault handling without service interruptions. Tenants no longer need to rely on O&M for resilience. Instead, technologies will be leveraged to implement unattended resilience and O&M. Enterprises will shift their focus from improving application resilience and reliability to using the resilience bastion technology to harden applications through containerization. Traffic control and perception will be incorporated. Steady intelligent monitoring will be used to implement evolutionary traffic consumption and balance. Chaos engineering will no longer be limited to a specific script. Unpredictable chaos engineering will be used to cultivate more creative blue teams.

3.5.4 Deterministic Operations: 80% of Cloud Faults Fixed in Just 10 Minutes

Digital transformation poses severe challenges on O&M. Multi-cloud environments and multiple technology stacks often have to coexist for a long time, which makes O&M more complicated and availability assurance more difficult. It is hard to balance service agility and live network stability, which increases production risks. Various incidents occur all too frequently. Ensuring system stability and reliability has become the lifeline of service development. It is increasingly difficult to establish a highly available service system without breaking the bank. The skills of O&M personnel do not align with the way the future has been shaping up. Traditional O&M capabilities can no longer meet the demanding requirements for stability and reliability of digital services. Enterprises need a more efficient, secure, and reliable O&M quality assurance system.

Deterministic operations continuously develop full-stack technologies and O&M processes for all scenarios. This includes designing and evaluating high-availability architectures, forecasting and preventing risks through proactive O&M, conducting chaos drills, and analyzing end-to-end observations. It also involves recovering from faults in a deterministic manner, dynamically governing risks for large-scale resources, and controlling change risks. These measures greatly improve enterprise cloud O&M, reduce system faults and downtime, and reduce labor costs and operational risks.

In the future, AIOps will be used to create a digital twin world of hundreds of millions of O&M objects on the cloud. All changes to cloud resources will be perceived in real time, and global O&M statuses will be visualized. Foundation models and SRE expertise will be utilized to provide intelligent O&M diagnosis and decision-making for automatic fault decision-making and recoveries. By 2030, enterprises are expected to fully implement deterministic operations for their management systems. This will help them detect 80% of cloud faults within 1 minute, respond within 5 minutes, and recover within 10 minutes. It will automatically track, make decisions for, and handle 80% of O&M tasks.

3.5.5 Refined FinOps: Saving Over $200 Billion USD per Year for Cloud Users

A survey shows that managing cloud expenditures is a major challenge for many enterprises. On average, 27% of public cloud expenditures are unnecessary. 51% of enterprises have established dedicated FinOps teams to cope with this issue. FinOps has become a critical component of many enterprise cloud strategies. By managing and optimizing costs effectively, enterprises can better control cloud expenditures, enhance their ROI, and promote sustainable business growth.

In the future, AI will be deeply integrated into FinOps. AI can accurately forecast costs, identify expenditure spikes and resource waste, and trigger automated repairs and optimization. The entire process, from cost visualization to optimization, will be fully automated. FinOps and AIOps will work closely to combine cloud cost data with operations data. This will provide more comprehensive insights and enable more intelligent decision-making for operations. AI will also make it easier for enterprises, including small- and medium-sized ones, to benefit from FinOps. FinOps will also use GreenOps to help enterprises reduce their carbon footprint and energy consumption while optimizing costs. By 2030, intelligent and refined FinOps is expected to be a must for enterprises to use the cloud. It will save global users over $200 billion USD per year by reducing cloud resource waste.

3.6 Boundless Security

3.6.1 Threats: The Most Frequent and Complex Cyber Attacks Ever

By 2030, cybersecurity will face unprecedented changes resulting from advances in AI and quantum computing, changes in the geopolitical landscape, and the emergence of new services and scenarios. Quantum computing will probably be able to make quick work of the cryptographic methods that are widely used now. There will be an increase in the generative adversarial attacks (such as malicious content, deepfakes, model architecture attacks, and various forms of malware) driven by AI large models. National-level advanced persistent threats (APTs) and professional commercial ransomware will greatly increase the intensity and impacts of network attacks. More countries and companies will need to pay special attention to the security and trustworthiness of supply chain hardware and software. Networks will be more vulnerable with the increase of sensors and controllers of smart devices (such as smart vehicles, smart home devices, and robots). The wide use of digital assets and smart contracts will create severe security challenges for Web 3.0.

It is estimated that by 2030, there will be more professional hacker organizations, highly intelligent attack tools, and new commercialized cyber attack services. The intensity and complexity of cyber attacks will increase by several times, maybe even by as much as 10 fold.

3.6.2 Defense: A Complete, Cloud-oriented, In-depth, Zero-trust System

Traditional security systems only provide reactive or static defenses, depend on border defense, and lack collaboration between products. A future-oriented, intelligent defense system needs to feature proactive defense, data sharing, internal and external consistency, and cloud-network-edge-device-chip collaboration. The architecture needs to be tightly coupled to the services being protected (native security that is deeply integrated into the services). A future-oriented system needs to offer a holistic security view. It needs to be a zero-trust system focused on cloud security, a system able to withstand the fierce network attacks of the future.

1) Zero Trust: Protecting Clouds, Networks, Edges, Devices, and Chips; Changing from Add-on Security to Native, Distributed Security

In the future, a zero trust architecture will protect corporate networks, cloud computing environments, mobile devices, remote offices, remote access, supply chains, IoT devices, applications, data, and more. Companies will deploy more zero trust products than they do today. They will fully integrate different zero-trust solutions, the original cybersecurity architecture, and business applications. It is estimated that more than 95% of companies will implement zero trust policies by 2030. Separate zero trust solutions will be integrated into a zero trust architecture with cloud security at its core.

The zero trust architecture will expand from network border (Zero Trust Network Access [ZTNA] and microsegmentation) to cover chips, devices, identities, data, applications, networks, and infrastructure. It will reshape the network architecture. Zero trust will also expand from external security to cloud native security, which means that security will become an inherent part of the system architecture, not just an additional layer for protection. Security will be shifted left in application development. In the future, security measures will be integrated right from the design stage to ensure synchronous development of security and other functions. In the future, security will be distributed and dynamically scalable based on service and traffic requirement. Security will be an integral part of the services, just like the immune system is an integral part of the human body.

2) Security System: Evolving to Intelligence for Proactive and Automatic Defense

Traditional rule-based security solutions struggle to process massive amounts of data and identify new attack patterns. By 2030, it is expected that 80% of organizations will adopt AI-powered cybersecurity products to address the evolving threat landscape and implement more proactive and intelligent security defenses. The security industry will fully leverage AI's technical advantages to achieve intelligent upgrades, particularly in threat detection, prediction, automatic response, and security decision-making optimization.

Intelligent Threat Detection: AI can identify unknown threats through behavior analysis and extensive threat intelligence, significantly enhancing the detection capabilities of traditional products. AI is primarily utilized in EDR, firewall, and APT products.

Event Prediction: AI introduces new technologies and paradigms for predicting imminent threats. Using techniques such as graph neural networks, malicious behaviors are modeled, and a security foundation model is trained with both benign and malicious samples. This enables the foundation model to deeply understand the internal relationships and rules between attack events, allowing it to predict the types, targets, and methods of future attacks.

Automatic Response and Security Decision-Making: AI-generated threat intelligence helps security operation systems proactively and accurately formulate defense policies

for both current and impending attacks, and automatically respond to them. This proactive approach minimizes security risks by implementing targeted protection measures in advance. Additionally, AI assists security experts in tracking attackers comprehensively, enhancing their situational awareness, and supporting proactive defense actions such as deterrence, trapping, and source tracing.

3) Security Ecosystem: Intelligent Cloud-Based Innovation, Industry Integration, and Win-Win Cooperation

The security industry faces challenges such as fragmentation, market segmentation, lack of systematic architecture for vendors, and insufficient ecosystem channels. In the new wave of security transformation driven by intelligence and cloud technologies, the industry will undergo intelligent upgrades and cloud-based innovations on cloud platforms. This will promote industry integration (resources, technologies, and markets) and foster a win-win, open, innovative, and future-oriented security ecosystem.

In the future, the new security defense system will leverage cloud platforms for technology, intelligence, and market sharing, enabling joint defense and coordinated intelligent operations against security threats. The traditional security ecosystem, which includes security software, hardware, service providers, consulting firms, and training institutions, will adapt to technological trends through cloud-based and intelligent transformations. By 2030, it is estimated that 90% of security vendors will transition to the cloud, embracing it proactively and integrating deeply to address more advanced cyber threats.

3.6.3 Security: The Immune System of the Cloud

1) Large Models: Comprehensive Protection Covering Four Aspects

By 2030, 50% of enterprises are expected to have their own large models. Large model security covers the following aspects: data security, confidential computing, model hardening, and content moderation.

Data security: Protect sensitive data and assets involved when training, using, and storing large models, including data property protection (data source validity, copyright authorization, and open source license), data lineage (upstream and downstream relationships and associations formed during data processing, transfer, and convergence), data encryption, and data cleansing.

Confidential computing: Ensure that enterprises can use sensitive data to train large models, protect data and models, and prevent attacks and damages caused by privileged and internal personnel. The combination of the following ensures security: First, hardware-based security and isolation. Second, zero-trust architecture (ZTA). The authentication service is used to verify the credibility of compute assets. Third, data owners are allowed to collaborate in model training while protecting the confidentiality of dedicated data.

Model hardening: Ensure the security of large models throughout their lifecycle, including development, training, testing, deployment, and management. Model leakage prevention and model encryption are widely used for security hardening. Preventing model leakage aims to protect models, including its parameters and functions, from being replicated or migrated through model technologies such as model theft and model compression. Encrypting models aims to protect model confidentiality, integrity, and availability. The taken measures are as follows: model leakage prevention, that is, replicating or migrating large model capabilities, including parameters and functions, through; model encryption, that is, ensuring with encryption. There are still other measures commonly used, such as model tampering prevention, model backdoor detection, and model copyright protection.

Content moderation: User inputs and large model outputs are checked to ensure that there is no immoderate content, such as the one against social values and privacy laws, or the one including bias, extremism, and discriminatory remarks. Technologies such as text review, image review, video review, AIGC privacy desensitization, AIGC watermarking, and AIGC authentication are used.

2) Cloud-Network-Edge-Device Synergy: A Full-Stack Defense System

It is estimated that by 2030 more than 90% of enterprises will seek for the comprehensive integration of security suppliers. The cloud-network-edge-device synergy will be implemented for the cloud platform by then to create a unified and comprehensive defense system.

Device security: ZTA will be extended to the edge to implement cloud-assisted device, device-cloud synergy, and situational awareness. It is the shift from endpoint detection and response (EDR) to extended detection and response (XDR). XDR driven by AI large models will implement real-time analysis and automatic protection of enterprise terminal devices and IT assets to prevent virus software and network attacks, enabling end-to-end data protection.

Edge security: As edge nodes are distributed, to implement decentralized edge security, there are various challenges, such as increasing border defense risks and physical hardware risks. In the future, applications that are deployed at edges will be more advanced and intelligent. It is vital to create a cloud-centered global security scheduling system based on the synergy of cloud and cloud-assisted edges.

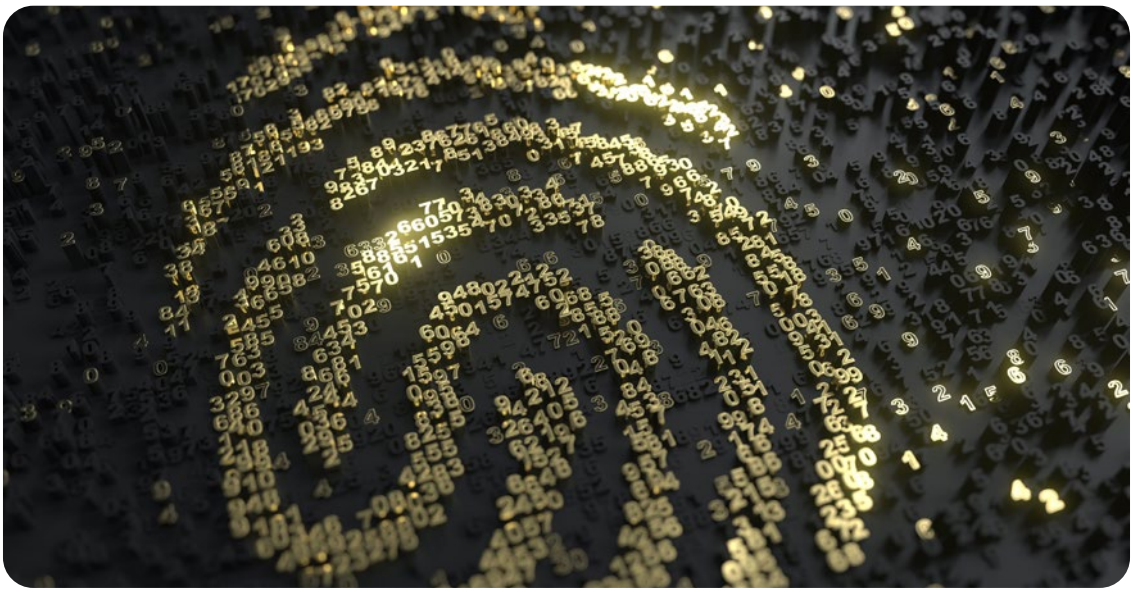
Network security: This is all about comprehensive monitoring and management of network traffic. AI-powered real-time analysis can identify and detect potential threats in the network traffic more efficiently. Based on the cloud and AI algorithms (network traffic models), the analysis precision will be improved greatly, making it possible to quickly identify and defend against more devastating DDoS attacks in the future.

Cloud load security: Apart from baseline defense, the traffic security is combined to monitor and protect the data traffic in the cloud environment, identify potential threats and malicious activities, and ensure data in transit security. As infrastructure as code (IaC) and policy as code (PaC) approaches are used, automated moving target defense (AMTD) for cloud workload security will soon be put into commercial use.

3) Data: Ensuring Encryption, Status Visualization, Transparency, Traceability, and Auditability

By 2030, over 95% of enterprises will adopt data-centric security protection policies.. A confidential computing platform (confidential environment, remote attestation, trusted nodes, and multi-node architecture) is built through automatic data discovery, intelligent classification, full-link encryption, and status visualization. Ensure the security and accuracy of data asset transmission between different systems and platforms. Protection for enterprises can be provided to data assets.

In the future, many new technologies will emerge to ensure the security of core data assets. Hardware-based confidential computing will be widely promoted. Memory encryption will be enhanced based on the traditional data isolation. Zero-knowledge proof will be widely used in financial services, Web3, and blockchain aspects. Secure multi-party computation and homomorphic encryption allows operations on encrypted data, and the operation result is the same as that of the original data after decryption. On the premise of ensuring data privacy, federated learning, zero-knowledge proof, and differential privacy support data interaction and processing. These technologies are expected to mature in the next few years and will be commercially available in 2030.



3.6.4 Cloud for Better Security

1) Network Security: Distributed and Intelligent Network Attack Defense System

With the high bandwidth, abundant computing resources, and hundreds of data centers in more than 200 countries, the cloud platform provides a solid foundation for meeting the increasing computing power requirements. It is estimated that by 2030, the anti-DDoS capability of the cloud platform will reach over 10 Tbit/s, far exceeding the bandwidth of common enterprises' self-built defense systems and several or even dozens of times the defense bandwidth of enterprises. This provides enterprises with more powerful and reliable network security assurance.

The cloud platform uses elastic resources, advanced protection technologies, professional security teams, global data center networks, and automatic response mechanisms to build a defense system that can defend against large-scale network attacks.

2) Data Security: From Enterprise-built to On- and Off-Cloud integrated Protection Policies

Data has become the most valuable asset of enterprises. The security of data is highly valued by countries and enterprises. It is difficult for enterprises to invest in long-term and continuous data security resources, including technologies, talent strategies, and data asset risk management systems.

The cloud platform has a more complete data security defense system (end-to-end encryption, access control, and identity authentication) and more advanced technologies (confidential computing and encryption technologies). In addition, the cloud platform meets the data compliance requirements of governments and administrative organizations and has a more comprehensive data asset governance mechanism, helping enterprises comply with data protection regulations and standards.

Cloud data security technologies will continue to innovate and improve. Confidential computing and cryptographic algorithms will significantly improve the performance of cloud platforms. By 2030, the cost-effectiveness of cloud data security solutions will be more than 50% higher than that

of enterprise-deployed solutions. This will further promote enterprises to use cloud-based data security solutions.

3) Security Operations: Embracing AI-powered Cloud Native SecOps

Cyber security depends 30% on systematic construction and 70% on constant operations. Security operations is a big challenge facing traditional enterprises because of expertise and resource shortages. While a cloud platform makes security operations simple and efficient based on its advanced security tools, professional services, and extensive expertise.

A cloud platform has a wide range of data sources, including run logs of security and cloud services, experience in defense against trillions of attacks, hundreds of millions of pieces of professional security knowledge, tens of billions of malware detection parameters, and more comprehensive, accurate, real-time, and in-depth threat intelligence and indicators. This makes it easier for a cloud platform than a traditional security platform to build threat identification models, provide automated response playbooks, and generate more powerful security models. In the future, the "security brain" powered by large models will help significantly improve security operations, including alarm noise reduction, attack analysis, automatic response, and automatic report generation. By 2030, 70% of enterprises are expected to adopt the AI-powered cloud native SecOps model.

4) Security Innovation: Technological Innovation Is Gradually Migrated to the Cloud

In the future, root technologies such as cryptography, security chips, and certificates will reshape the traditional cyber security industry. The cloud will implement the atomic root of trust based on cryptography and security chips. The cloud will facilitate the innovation, commercial use, and implementation of security technologies.

Cryptographic technologies, as the basis of security, will face new challenges brought by quantum computing threats, international situation changes, and emerging technologies. The cloud can provide a solid foundation for building a comprehensive, efficient, and reliable password protection system based on cloud native encryption services, centralized key management services, more advanced identity authentication mechanism, and more real-time and intelligent intrusion detection and defense capabilities. This will vigorously promote the rapid development of technologies such as post-quantum cryptography, SM series cryptographic algorithms, and hardware-based confidential computing.

Security chips are basic components for trusted computing. The implementation of security chips requires an end-to-end reconstruction of operating systems and applications, which is a complex process that involves in-depth integration and optimization of hardware and software. A cloud platform features software and hardware integration, capability integration, process standardization and automation, as well as good flexibility, security, and economic benefits. These make it an ideal platform for end-to-end reconstruction, making it easier to put security chips into commercial use. By 2030, the availability and cost-effectiveness of security chips will be greatly improved. It is estimated that the market penetration rate will be over 20%. Security chips will be popular in the market by then.

04

Call to Action

The best way to predict the future is to create it. Intelligence is not a mere vision; it is the defining trend of our times. As we stand on the cusp of the intelligent world of 2030, let us join forces to catalyze this transformation. Together, we will navigate the dynamic cloud landscape and harness the full potential of intelligence to reshape industries.

Appendix: Abbreviations and Acronyms

Abbreviation/Acronym	Full Spelling
ADMET	Absorption Distribution Metabolism Excretion Toxicity
AGI	Artificial General Intelligence
AI4S	Artificial Intelligence for Science
AIGC	Artificial Intelligence Generated Content
AMTD	Automated Moving Target Defense
APT	Advanced Persistent Threat
CNN	Convolutional Neural Network
CPU	Central Processing Unit
Dapp	Decentralized Applications
DataOps	Data Operations
DBA	Database Administrator
DCN	Data Center Network
DDOS	Distributed Denial of Service
DeFi	Decentralized Finance
DevIndustryOps	Industry Development Operations
DevOps	Development and Operations
DevSecOps	Development, Security, and Operations
EB	Exabyte
EDR	Endpoint Detection and Response
ER/EP	Enterprise Router/Endpoint
ETH ETF	Ethereum Exchange-Traded Fund
eVTOL	electric Vertical Takeoff and Landing

Abbreviation/Acronym	Full Spelling
FinOps	Finance Operations
GPU	Graphics Processing Unit
GRC	Governance, Risk, Compliance
GreenOps	Green Operations
GW	Gigawatt
HTAP	Hybrid Transactional/Analytical Processing
IaC	Infrastructure as Code
IDC	Internet Data Center
IDE	Integrated Development Environment
LLM	Large Language Model
MediaOps	Media Operations
MFU	Model Floating-point Operations Utilization
MLOps	Machine Learning Operations
MLOps	Machine Learning Operations
NFT	Non-Fungible Token
NLP	Natural Language Processing
NPU	Neural Processing Unit
OS	Operating System
PaC	Policy as Code
PDE	Partial Differential Equation
Proclet	Processlet
RNN	Recurrent Neural Network

Abbreviation/Acronym	Full Spelling
RWA	Real World Assets-tokenization
SecOps	Security Operations
SFT	Supervised Fine-Tuning
SLA	Service Level Agreement
SQL	Structured Query Language
SQL	Structured Query Language
SRE	Site Reliability Engineering
UPS	Uninterruptible Power Supply
V2X	Vehicle-to-Everything
VLA	Vision-Language-Action
VPC	Virtual Private Cloud
VPC	Virtual Private Cloud
VPN	Virtual Private Network
XDR	Extended Detection and Response
XR	Extended Reality
XR	Extended Reality
ZB	Zettabyte
ZTNA	Zero-Trust Network Access

HUAWEI

Version 2024

ICT Services and Software 2030

Building a Fully Connected, Intelligent World

Overview

As the communications industry evolves from 2G to 5G, the ICT service and software industry has also undergone an inter-generational upgrade toward standardization, digitalization, and being tool-based. With the rise of new technologies such as Generative AI (Gen AI) and digital twins, the transition from digitalization to intelligence has become immensely popularized. By 2030, AI will be ubiquitous. We will experience more changes: the intelligent and ubiquitous sensing of infrastructure will become a necessity, large models will make progress toward AGI and think like people do, service models will pivot from majority manual labor to majority machine, and enterprise marketing and enablement methods will become increasingly real-time and agile.

Over the next decade, an intelligent transformation will sweep through various industries, echoing the industrial revolutions in the 20th century. Gen AI will empower machines to learn and think with human-like intelligence, catalyzing a pivotal change of production and redefining the work and life for every enterprise, family, and individual, just as the steam engine and light bulbs did back in their times.

01

Macro Trends and Prospects

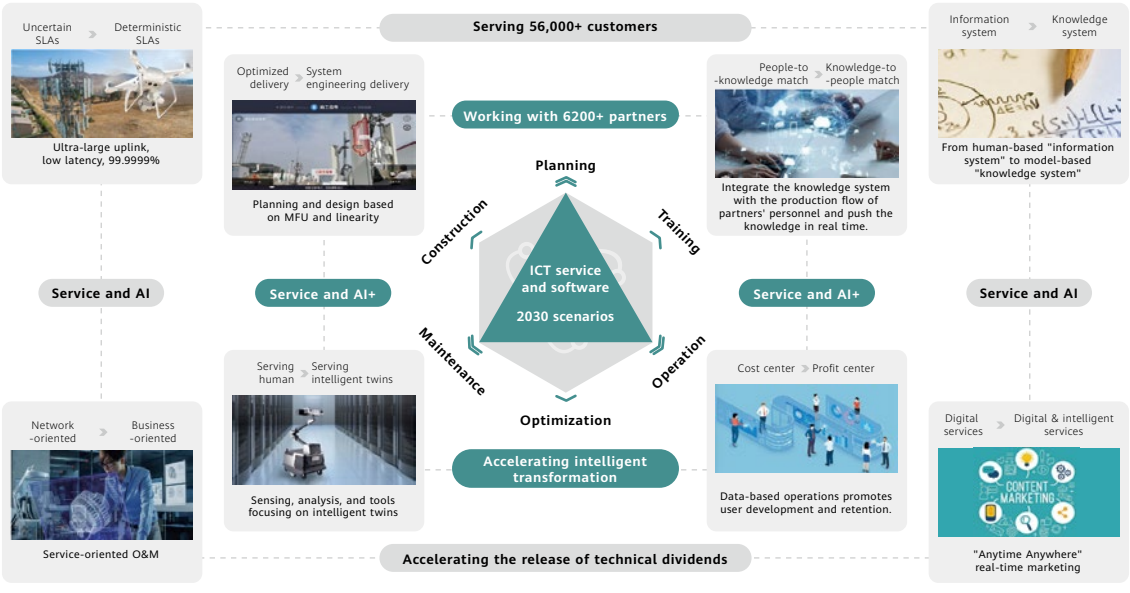
New Technologies, Business, and Models Bring Infinite New Possibilities and Uncertainties Alike

As we rapidly approach an intelligent world, a distinct development trend towards digitalization, intelligence, and low carbonization has emerged. Thanks to certain technologies like naked-eye 3D, AI backpack, autonomous driving oriented to the B2C business, unattended manufacturing factories oriented to B2B business, and mechanical hands and arms for smart mining and ports, forward-thinking businesses are the ones making this happen. With Gen AI as a representative, new technologies such as large models, AI, 5G-A, ultra-large computing clusters, liquid-cooling data centers, digital twins, and agents are making rapid progress. New models of knowledge and data management, AIOps, collaboration between small and large models, as well as AI for Network, and Network for AI have all recently emerged. All these elements work together to drive innovation and the emergence of new, intelligent businesses, sparking new visions and opportunities.

To introduce each new generation of technology and model into the production environment and unleash new productivity, it is essential to achieve continuous evolution alongside the existing business production environment. We must effectively manage complexities and uncertainties, as well as enable orderly evolution of ICT infrastructure throughout its entire lifecycle. It is important to promptly satisfy the new demands for ICT infrastructure driven by new businesses and experiences, foster innovation, and upgrade experiences continuously. The ultimate goal is to maximize investment returns and accelerate the industry's digital transformation. As AI pioneer Fei-Fei Li puts it: "AI is a pervasive technology that will permeate every aspect of our lives and industries like water." As new services, technologies, and models rapidly evolve, the ICT service and software industry will be confronted with unprecedented uncertainties. To unlock business value and

technical dividends, we must harness the power of these innovations while leveraging new technologies to sustain a competitive edge throughout the entire lifecycle. Here are two possible approaches:

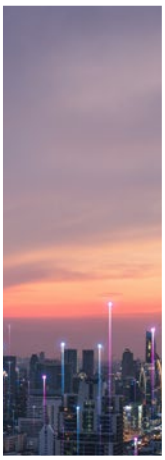
ICT Service and Software 2030 – Future Scenarios: AI+ Changes Service Modes, and + AI Brings New Scenarios



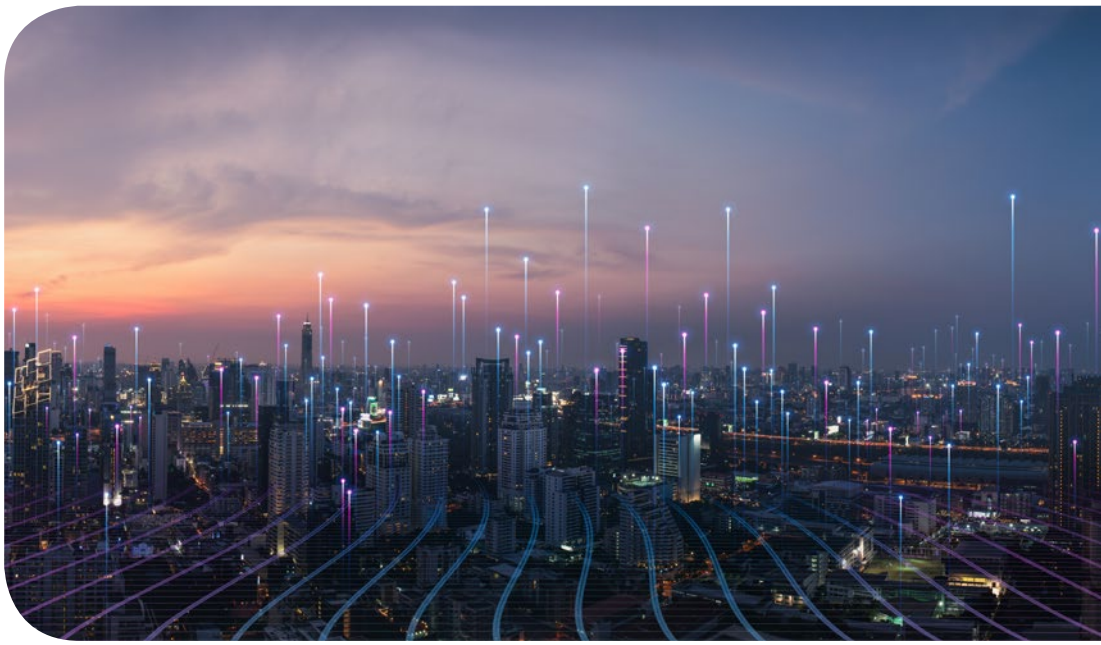
1. Services and AI: Consider what services need to do for better AI development. For example, with the rise of intelligent connectivity of everything, new service level agreements (SLAs) have introduced uncertainties, while network faults are affecting a wider range of services. In this complex landscape, how can we transition from network-centric to business-centric O&M? Data and knowledge management, as a crucial capability for General AI and large models, will redefine the way people learn and are empowered. How do we navigate the future development of ICT talent and services that AI needs?

2. Services and AI+: Consider what AI needs to do for better service development. For example, large models, robots, and embodied AI agents have become an integral part of the future service mode. How do we change the traditional mode of people and platforms through the agent, tool and person mode to improve the efficiency of planning, construction, maintenance, optimization, marketing, and training, as well as reduce costs?

A new horizon is unfolding as we approach 2030. How to utilize definite service capabilities to address numerous uncertain needs of AI and AI+ is a key question that every ICT professional must ponder.



02



Future Scenarios

Planning, Construction, and AI: From Uncertain SLAs to Certain

By 2030, we will have evolved from "connectivity of everything" to "intelligent connectivity of everything". In the digital age, the objects of connectivity are more often people and things in the traditional concept of IoT. According to Gartner and IMT and other organizations' predictions for 2030, AR/VR/MR terminals are expected to occupy 30% of the terminal market. Autonomous driving and unattended industrial manufacturing will also become a reality by 2030. With the help of mechanical arms and hands, unattended factories and mines will become essential for enterprises. On top of that, agents and robots will gradually replace a significant portion of repetitive work currently performed by humans. Huawei predicts that the number of active users for wireless AI agents will reach 6 billion by 2030, By 2030, Huawei predicts that 45% of ICT scenarios will be supported by intelligent agents, and every role will have a dedicated copilot.encompassing not only digital twins in virtual worlds but also embodied AI in the physical world, such as industrial robots, service robots, companion robots, autonomous drones, and self-driving cars. These new service entities will introduce significant uncertainty into future network planning.

We need a network plan based on the new propagation model. Traditional networks, modeled after human interaction, prioritize SLAs that focus on call connectivity rates, call drop rates, mean opinion score (MOS), and call setup delays. The objective is to meet experiential expectations within the bounds of human subjective tolerance, where minor call drops and delays are typically forgivable. However, the network propagation model for intelligent connectivity of things requires providing optimal sensing for machines. As agents increasingly integrate into more B2C lifestyle scenarios and B2B production follows, there is a greater need for deterministic SLAs to ensure ultimate experience in daily life and uninterrupted production. In high-stakes environments like autonomous driving, low-altitude economy, and smart ports, minor setbacks can rapidly snowball into catastrophic failures affecting entire industries,

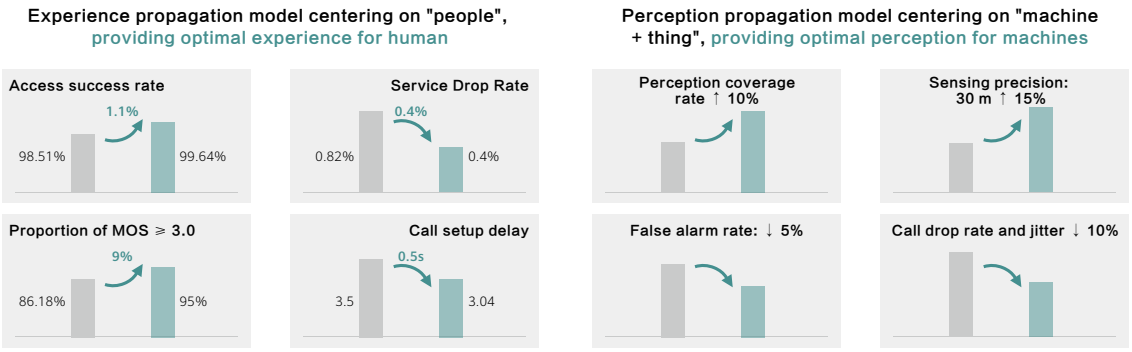
Industry	Business Type	Number of Connections per Enterprise Connections per Enterprise	Requirements of Business for the Network														
			Service Availability (Single User or Single Service)										Security		Trustworthiness		
			Bandwidth Requirement/Single User (Mbit/s)					Service Latency Requirement (ms)									
			B1	B2	B3	B4	B5	T1	T2	T3	T4	T5	S1	S2	M1	M2	M3
			1~10	10~20	20~50	50~100	>100	50~100	20~50	10~20	5~10	<5	Logical Isolation	Physical Isolation	Visibility	Manageability	Operability
Smart Healthcare	16K remote diagnosis and treatment	10					1G										
	Monitoring and Nursing	2,000															
	Holographic Remote Surgery	5					10G										
Smart Grid	Video-based Inspection	-															
	Power Grid Control	-															
	Wireless Monitoring	-															
Smart Manufacturing	Factory Environment	100															
	Information Collection	10,000															
	Operation Control	1,000															

For details, see CAICT 5G E2E Slicing Industry SLA Requirement Research Report.

cities, and global economies. To ensure the high reliability of SLAs, network planning must be compatible with both people-led and machine-led propagation models.

We need a network plan based on the new propagation model. Traditional networks, modeled after human interaction, prioritize SLAs that focus on call connectivity rates, call drop rates, mean opinion score (MOS), and call setup delays. The objective is to meet experiential expectations within the bounds of human subjective tolerance, where minor call drops and delays are typically forgivable. However, the network propagation model for intelligent connectivity of things requires providing optimal sensing for machines. As agents increasingly integrate into more B2C lifestyle scenarios and B2B production follows, there is a greater need for deterministic SLAs to ensure ultimate experience in daily life and uninterrupted production. In high-stakes environments like autonomous driving, low-altitude economy, and smart ports, minor setbacks can rapidly snowball into catastrophic failures affecting entire industries, cities, and global economies. To ensure the high reliability of SLAs, network planning must be compatible with both people-led and machine-led propagation models.

From a business perspective, it is difficult to calculate the return on investment (ROI) for a network that features intelligent connectivity of things, as traditional indicators like packages, DOU, and penetration rates were designed to gauge people-led networks. Each scenario demands a tailored approach, balancing business models and networking needs, necessitating a flexible planning and GTM pace taking into consideration the intelligence level of the corresponding region and city. As a result, refined investment in system integration will have a higher TTM requirement. To gain a market advantage in complex business

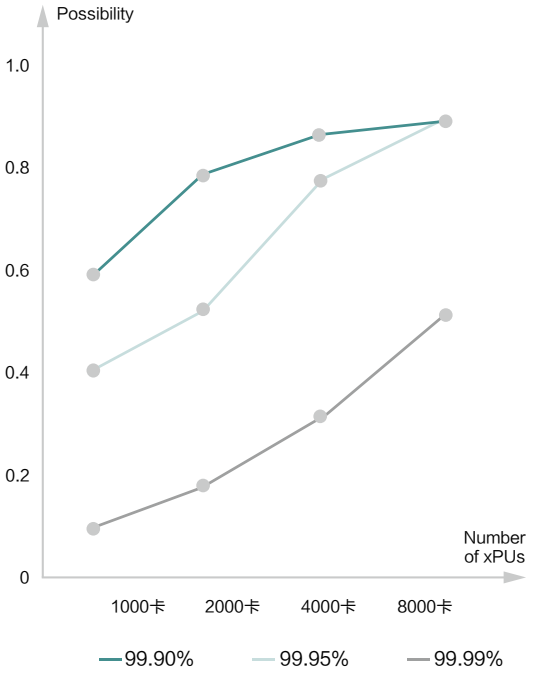


scenarios, it is essential to combine deterministic business scenario SLAs for rapid network upgrade and better ROI. Network planning needs to build real-time network simulation through digital twins, and rapidly define the target network based on the requirements specific to a forward-thinking network construction business scenario. Network planning and design need to simulate the business and network changes in the physical world using digital twins, so as to build human- and machine-led propagation models and network performance simulation and prediction. In addition, planning also needs to be iterated at a small cycle. In this way, we can ensure an accuracy of 99.9% and a 50% shorter TTM.

Planning, Construction, and AI+: From Digital Integration to System Engineering Integration

We are now in the phase of intelligent IT integration. Clusters are scaling up based on scaling laws. Grok 2 and Stargate are now talking about 100,000- and 1,000,000-GPU/NPU clusters. In February 2024, ByteDance unveiled MegaScale, a system featuring 12,288 GPUs, which trained 1,75B models, matching the industry's current peak Google in terms of 10,000-GPU/NPU cluster. ByteDance tried up to 9 optimization methods. In spite of this, ByteDance achieved a Model FLOPs Utilization (MFU) of only 55%, leaving a significant gap to the maximum of 95%. A 1% improvement in MFU will result in over $1 million in cost savings, significantly enhanced performance, and reduced TTM training time. Mason predicts that AI OPEX will rise by 35% compared to traditional computing OPEX, and is expected to increase by over 50% by 2050, driven mainly by water and electricity costs. Intelligent data centers will require AI-driven energy saving based on full-stack DC L1&L2 linkage, as well as high MFU planning.

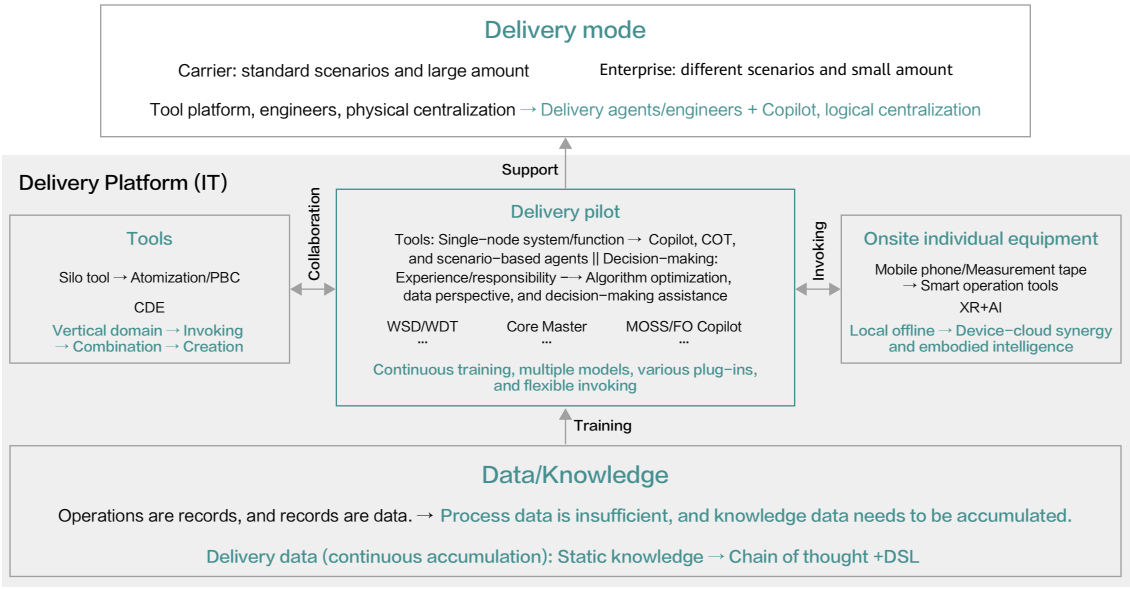
Representative AI Model	Number of Training xPUs
ChatGPT	1000
GPT-4	10,000
Gemini	54,000 TPUs
Grok 2	100,000
AGI	1,000,000 (Stargate project)



AI Business	Computing Power	Network Bandwidth	Network Latency	Bandwidth	Memory Capacity
LLM training	★★★★	★★★★☆	★★	★★☆	★★☆
LLM inference (prefill)	★★★★☆	★	★	★☆	★★★☆
LLM inference (decode)	★☆	★☆	★★★★★	★★★★★	★★★★
Recommendation system training	★★☆	★★★★★	★★★	★★★	★★☆
Recommendation system inference	★	★★☆	★★☆	★★☆	★★★★★

Different AI businesses have unique requirements for building competitive intelligent computing networks. Computing power, network bandwidth, latency, memory bandwidth, and memory capacity needs differ across training and inference scenarios of large language models (LLMs) and small models. The system architecture must be flexible to meet diverse needs. The network provides the foundation for connectivity, allowing for flexible combinations of computing and storage to meet different needs. However, not all five capabilities are needed in a single business scenario at the same time. The network-storage-computing collaborative design, with its characteristics of fully utilizing network to gain extra computing power, utilizing computing to gain more storage power, and utilizing storage to replace computing, is the primary planning direction for future intelligent computing cluster systems. PwC predicts that, by 2030, the potential optimization space for collaborative planning of computing, storage, and networking under the same computational requirements is three times that of the current one. This will make integrated services based on "system engineering" a necessity for meeting future demands for high-complexity MFU and linear scalability.

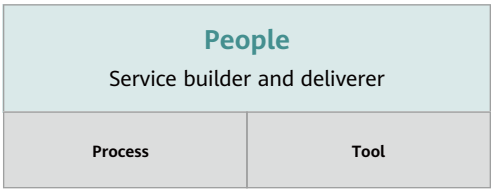
From the delivery mode perspective, future integration services will shift from two-dimensional



to three-dimensional, adding spatial calculation to traditional time (phase and phase effects) and task (outcome and outputs) dimensions to describe services, which involves tracking changes in the system's space over time and tasks using digital twins.

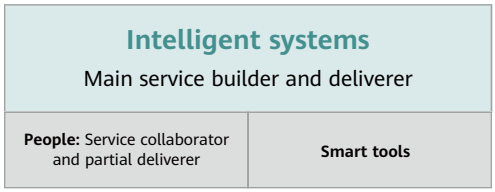
The future delivery model will shift from a traditional, person-plus-tool, physically centralized approach to an agent-plus-copilot, logically centralized model. In the past, delivery was based on the number of sites, (for example, to deliver 3,000 sites, we will need 100 employees and 90 days), and resources were allocated according to the delivery location. The future delivery will be more centralized. Delivery centers will be constructed at the group and province level. Agents and copilots will be used for site surveys, method of procedure (MOP) designs, and original factory configurations. Field work will become more focused and process-oriented. Delivery project managers and technical project managers were crucial in the past, as they possessed project management, key technologies, tool platform capabilities. In the future, models and applications will become the primary builders and deliverers of integrated services. Human-machine collaboration and data-driven approaches will further flatten delivery organizations, resulting in a delivery efficiency improvement of over 50%.

Services focusing on people



- Linear increase of headcount and revenue
- Long service period and small concurrent workload
- High service price, tool-assisted

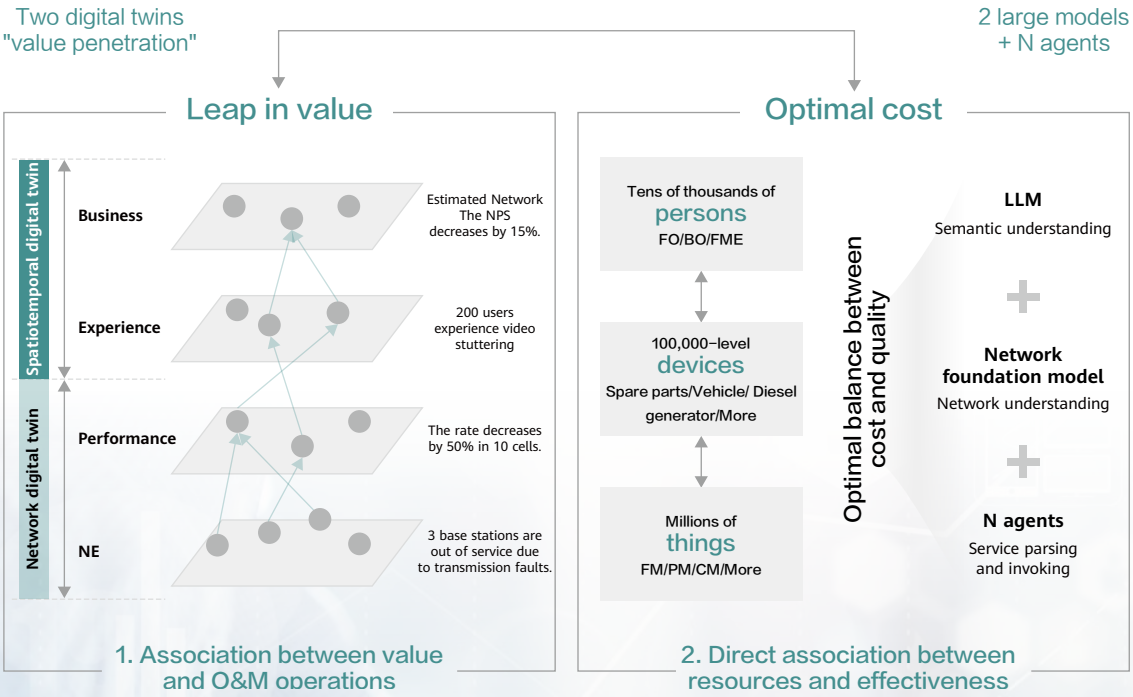
Services focusing on intelligent systems



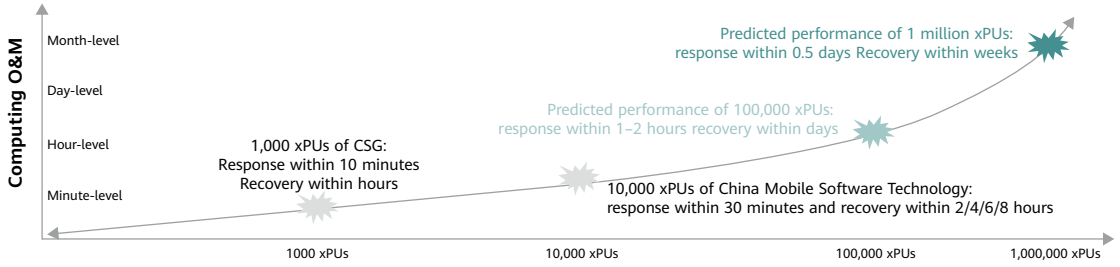
- Use intelligent systems to replace most personnel and upgrade from manual operations to human-machine collaboration.
- Intelligent systems automatically generate processes and data-driven, and multiple concurrent operations are supported.
- Upgrade from functional tools to intelligent tools, improving efficiencyv

Maintenance, Optimization, and AI: From Network-oriented to O&M-oriented

By 2030, as network architecture becomes more complex, O&M objects will include cloud, network, storage, edge, and device. This complexity makes it harder to determine network operations, increasing the need for skilled network change engineers. In recent years, network incidents frequently occur. Compared with five years ago, the proportion of major ICT network faults has increased by 45% in 2024. The primary reason is that as networks become more and more complex, traditional network-based O&M cannot perceive the terminal. It is difficult to connect data and algorithms across the entire stack involving network elements (NEs), performance, experience, and business. With advancements in smart technologies such as Gen AI, large models, and digital twins, the future of O&M will shift from focusing on networks to focusing on business needs.



Computing O&M mode: demarcation efficiency grows exponentially as cluster size increases



Furthermore, as computing power grows according to the scaling law, large-scale model manufacturers now typically use tens of thousands to millions of xPUs. This makes traditional network-based O&M impractical. A report of an Internet vendor on a training cluster involving tens of millions of xPUs shows that over 54 days of training, there were 466 job interruptions, averaging 8 per day. Most interruptions (41%) were caused by software problems, cable issues, and network faults. By 2030, estimated losses from an interruption involving millions of xPUs could exceed CNY100 million under the current maintenance model.

For traditional network O&M, we mainly cope with major and urgent complex network changes from pre-event, in-event, and post-event perspectives.

Pre-event involves maintenance engineers preparing in advance for emergency plans and making corresponding emergency plans to prevent potential impacts on business in the future. In-event involves preparing network change operation scripts in advance and standardizing engineer operation principles. Post-event involves backtracking and summarizing, and iterating on shortcomings in the pre-event and in-event phases, serving as a basis for subsequent case studies. As networks become increasingly complex, this traditional O&M approach finds it difficult to completely avoid major accidents caused by human error. At site X, an engineer mistakenly input an extra '0' in the traffic threshold configuration, leading to a signaling storm that resulted in a two-day disruption of communication services for 30 million users across the province.

Driven by advancements in technologies like Gen AI, digital twins, knowledge graph, and embodied AI, industries are shifting towards replacing traditional network-based O&M with service-based intelligent O&M.By 2030, 30% of leading carriers will deploy digital twin systems using 5G-A as part of their intelligent transformation.

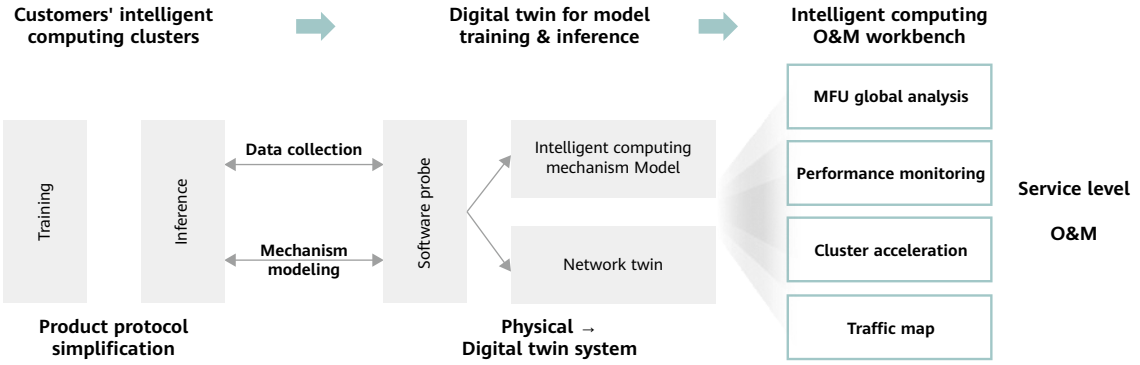
- Start with the end in mind. Simplify computing network protocols based on O&M business needs, providing real-time visibility and access to E2E data.

- Build real-time network digital twins. Create a knowledge graph for O&M to track how network issues affect services. This connects the physical network to the digital world, and links services to sort out key performance indicators (KPIs) and key quality indicators (KQIs), making it possible to visualize and manage the impact of operations and changes on services.

- Develop new applications based on service-level O&M, integrating network and services to enable global visualization and unified management of ICT O&M.

With these capabilities, fault recovery will take hours instead of days, network fault response will take seconds instead of hours, and standby board replacement can be a regular task on a weekly or monthly basis, instead of an emergency operation completed in 4 hours.

Oriented to business O&M:
Streamlining the full stack of product protocols, digital twins, and service transformation



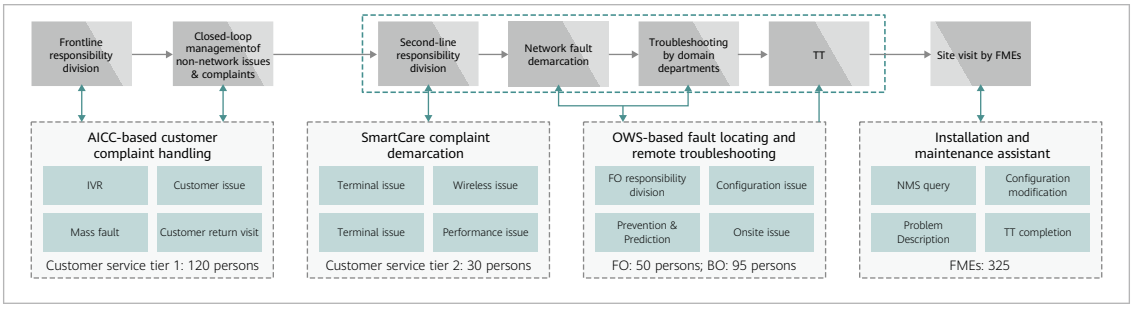
Maintenance, Optimization, and AI+: From Manual to Machine Labor

In the traditional digital landscape, tools and processes were designed with people as the focus. In the intelligent landscape, human-machine collaboration does not require this. There are clear human and machine interfaces. Machines, people, processes, and tools will work together to solve problems.

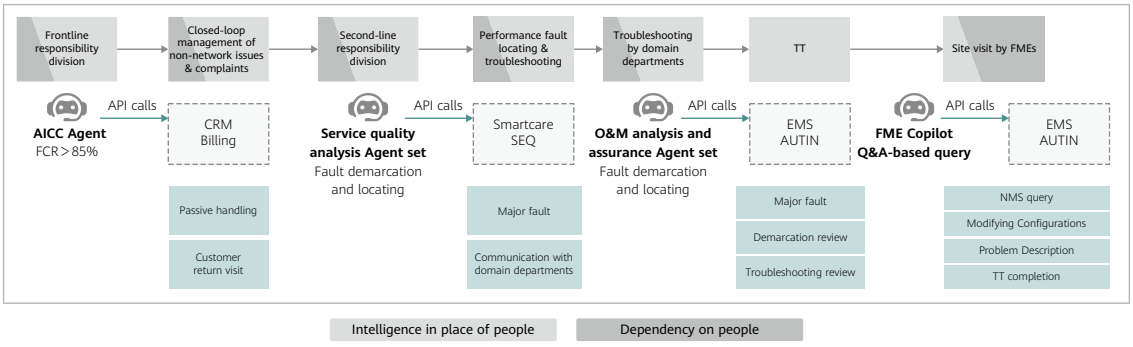
To implement agent-centric O&M, it is generally agreed that three basic technologies need to be built:

1. Computing network O&M large models: Build a computing network O&M large model that understands the O&M mechanism and network protocols based on the basic large model commonly used in the industry, and build role-oriented copilot and scenario-based agents for RE, FO, BO, and FME to reduce repeated manual work. Then, implement real-time dispatch of configuration and commands by incorporating traditional automation large models, and implement unified APIs for the

Digital era: People are the focus. Processes, platforms, and people are interdependent. Processes and small model tools are designed based on the efficiency of manual issue handling.



Intelligent era: Human-machine collaboration is used. People do not need to participate in each phase. Human-machine boundaries are clear. People, processes, and tools are designed by focusing on machine-based issue resolution.



OSS domain, enabling users to use large models easily.

2. Digital twins: The real-time network digital twin system is required to associate services with NEs. Network data in the past was considered black box data, which can be analyzed by layer-by-layer filtering using probes and NMSs. In the future, digital twin systems will use knowledge graphs to collect data quickly at the NE level, aiming to reduce the collection time from 30 minutes to one hour to just a minute.

3. Embodied AI robots: Onsite O&M costs generally account for 60% of the total ICT O&M costs. In the future, embodied AI robots will be deployed in each data center, equipment room, and site. The transformer-based small-sized large model IOS can accurately identify instructions from the NOC/SOC agents, and perform network operations (inspection, live network status awareness, fiber port adjustment, board replacement, etc.) instead of maintenance personnel. This will significantly improve O&M efficiency.

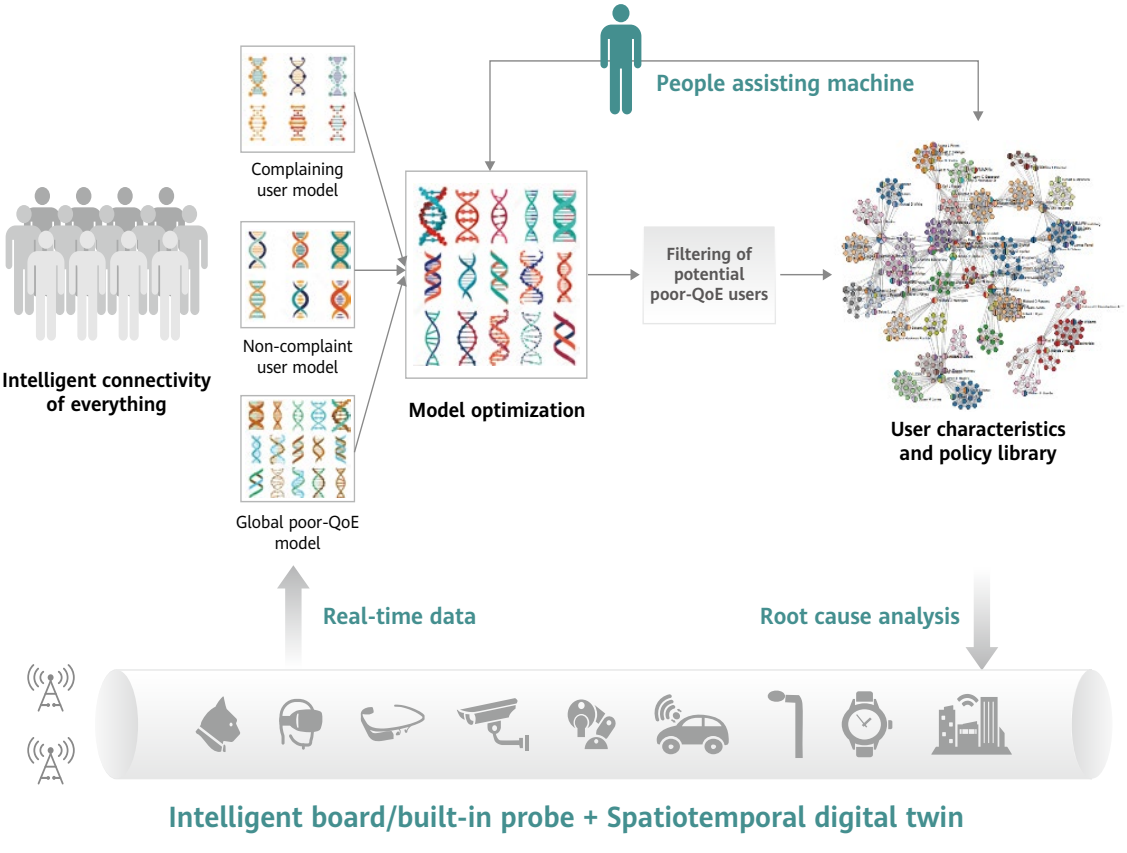
Hardware replacement by a robot hand

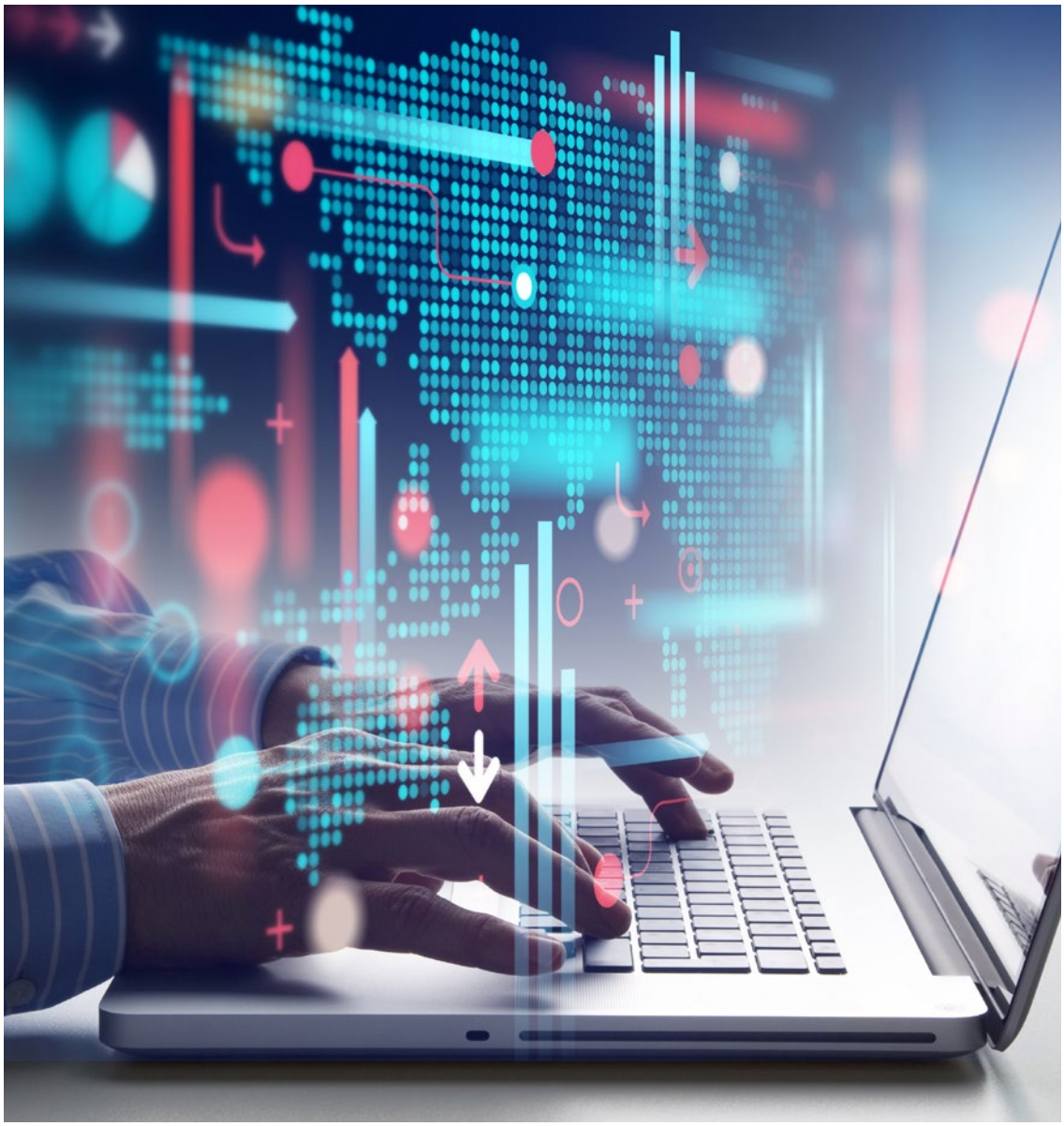
Automatic inspection robot

Optimization and AI: From "People Waiting for Network" to "Network Waiting for People", Fostering the Willingness to Monetize User Experience

By 2030, mobile networks will shift from mainly handling person-to-person and person-to-device communications to mainly handling person-to-agents and agent-to-agent communications. Communications networks will connect not only individuals, but also various sensing, display, and computing resources related to individuals, as well as AI agents. They will connect not only home users, but also related home, car, and content resources. They will also connect machines, edge computing, and cloud resources related to an organization, on top of employees of an organization. In this way, future communications networks can meet various business requirements of an intelligent world.

From an experience assurance perspective, the daily optimization of traditional network is a passive "people-waiting-for-network" approach, where network optimization is a reactive response to customer complaints. It aims to improve product performance by 10-15% based on existing performance, doing the best possible with what's available. In China, an investment of 25,000 person-days is required annually to address a range of routine optimization issues, including collecting complaints about network issues, designing optimization priorities for typical regions/sites, digital road testing, and designing and implementing routine optimization solutions. The "people-waiting-for-network" approach resolves only 35% of network issues, failing to delivery an optimal user experience. A unified large model for network optimization is lacking, causing performance conflicts among multiple optimization solutions. Besides, optimization experience cannot be shared.





By 2030, the network optimization approach will shift to "network-waiting-for-people". We will achieve end-to-end performance sensing using HarmonyOS, smart boards, smart antennas, and fiber iris. We will build a digital twin system (TAZ) based on space-time to precisely predict future traffic, trends, and SLA tendencies for each business category. Using an optimization large model featuring computing & network integration, we will achieve intelligent and autonomous closed-loop of single products in 30% scenarios. For the remaining 70% scenarios, we will adopt a series of methods. First, we will use knowledge graphs and management systems to consolidate core assets, and develop AI driving force based on MR and SEQ performance data generated during daily optimization. We will continuously support model upgrade and iteration, and develop agents and integration twins for VIP assurance and daily optimization. By doing so, we can help network optimization personnel take preventive measures and predict issues based on business changes, quickly generating optimization solutions. Finally, we will invoke traditional small models to perform real-time analysis, decision-making, and closed-loop management of performance experience.

Optimization and AI+: Network Optimization Agent Based on Endogenous Intelligence

According to Joseph Sifakis, a French computer scientist and 2007 Turing Award laureate, the definition of future autonomous network systems in the communication field requires a different approach compared to the general agents based on LLMs. In the communication field, strategies generally need to be formed based on actual network business/network status. Therefore, agents in the field must be able to comprehend network, including its topology, performance, alarms, and events. The high real-time requirements for data in the communication field make it difficult for large models to learn, so digital twins and other technologies are integrated to help. We use real network information to plan and implement solutions, and inject domain expertise into large models through prompts or SFT. This, combined with digital twin models and atomic capabilities for experience, maintenance, and optimization, enables the planning, perception, decision-making, and execution of tasks.

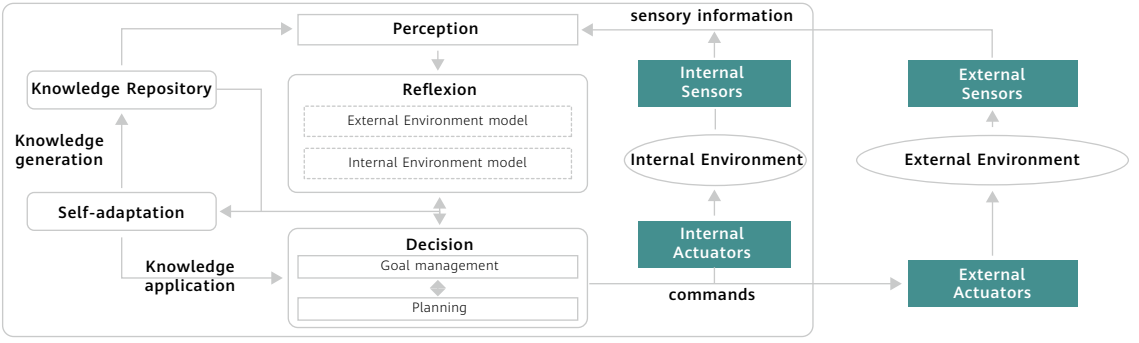
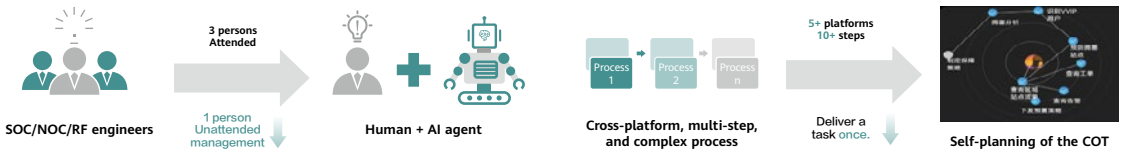


Figure 6: Computational model for cyber physical agent

Take the optimization work pattern in a region as an example. Previously, three network optimization engineers were on standby. Now, leveraging the endogenous intelligence of network optimization agents, unattended network optimization is achieved, requiring only one person to design the agent model.

The network optimization function development process is also simplified. Previously, it involved multiple platforms, over 10 steps, and at least three months. Now, agents design and complete the task in one dispatch with multiple rounds of interactions based on the chain of thought (COT).



Marketing and AI: From Digital Business to Digital and Intelligent Business

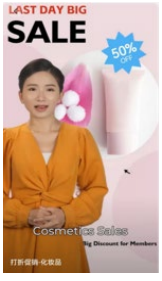
By 2030, the digitalization of marketing businesses transforming into intelligent transformation would have become a consensus. Consulting firms such as Accenture and PwC have identified marketing as the top industry that can be impacted by large models, because there are diverse user demands, numerous human-machine interaction interfaces, and vast opportunities for creative generation and innovation in handling customer needs, from advertising and marketing to billing and sales, as well as customer complaints. This aligns with the three main conditions suitable for current mainstream LLM applications: vast data, creative scenarios, and natural language. The current digital marketing model is no longer sufficient to meet the demands of a future with billions of digital and intelligent humans, autonomous driving cars, and full-line industry intelligence. What do we need in the future?

- Agile innovation: In the digital age, the package design putting people in mind typically takes 3 to 6 months to complete the entire process of market strategy development, resource preparation, package development, and market promotion. By the time a product hits the market, the market has often shifted. In the future, intelligent businesses will need to develop marketing initiatives and package designs based on customer data analysis. Leveraging agile computing and network infrastructure, they can accelerate this process from several months to just a few days. This allows for real-time push of packages at any time, anywhere, in movie theaters, sports stadium VR, and other scenarios.

- Smart creativity: In the digital age, marketing primarily relies on big data user profiling for targeted phone/content pushes. It completely depends on humans for analysis and crafting marketing strategies. By 2030, the efficiency and cost of digital humans in content generation and operations will surpass human capabilities, enabling rapid iteration in response to market changes to meet the latest demands of the masses. Digital humans offer 7x24 online, high-quality interactive solutions to customer issues.

Product overview

Introduction videos for software services, beauty & fashion, 3C electronics, factory workshop, and solutions

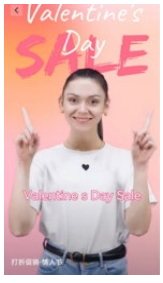


Promotional activities

Voice-over videos on Black Friday, Valentine's Day, Christmas, and other internationalized holiday nodes for voice-over videos.

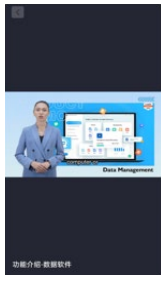
In-feed ads

Google, Facebook advertisements, YouTube, TikTok, Amazon, Instagram platform video



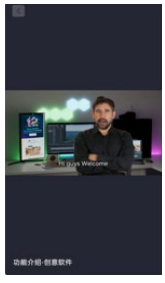
Product recommendation

Video playback for product presentation, details display, comparison and evaluation, purchase guide, etc.



Content marketing

Google SEO, Facebook YouTube video and e-commerce ads



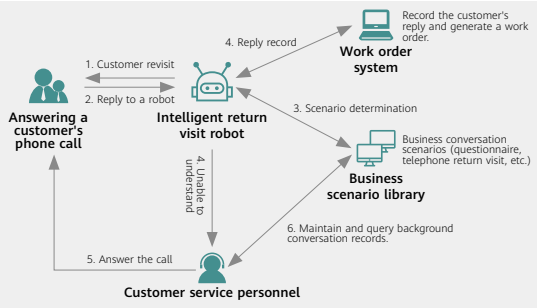
Video tutorial

Customer service, support team content, product operations, Q&A, and explanation videos

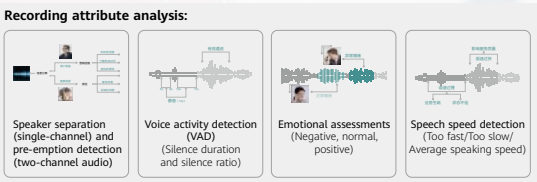
Marketing and AI+: From a Cost Center to a Profit Center

Enterprises have traditionally viewed customer service centers as cost centers. However, with digitalization and intelligence, they are discovering that engaging with complaining customers can attract new users and revenue, making customer service a key revenue driver in the digital landscape.

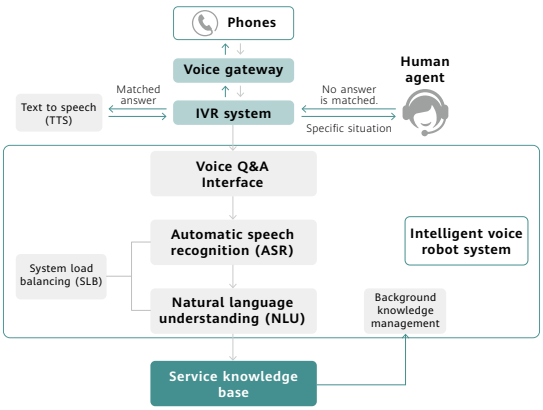
- In the customer cultivation and acquisition phase, digital human intelligent outbound calls are used to replace traditional manual outbound calls. Underpinned by AI and data-based personalized recommendations, the outbound call success rate is improved by three times, and the outbound call efficiency is improved by 300%.



- In the activation and retention phase, digital humans are used to recommend and interpret products in dialog mode, and accurately identify customers' willingness and emotions based on customers' tone, reducing the AHT by 50%.



- In the complaint handling phase, digital humans can handle more than 90% of problems based on the service knowledge base and voice recognition capability. Only 10% of complaints are transferred to manual personnel. This improves the first call right (FCR) from 75% to 90%.

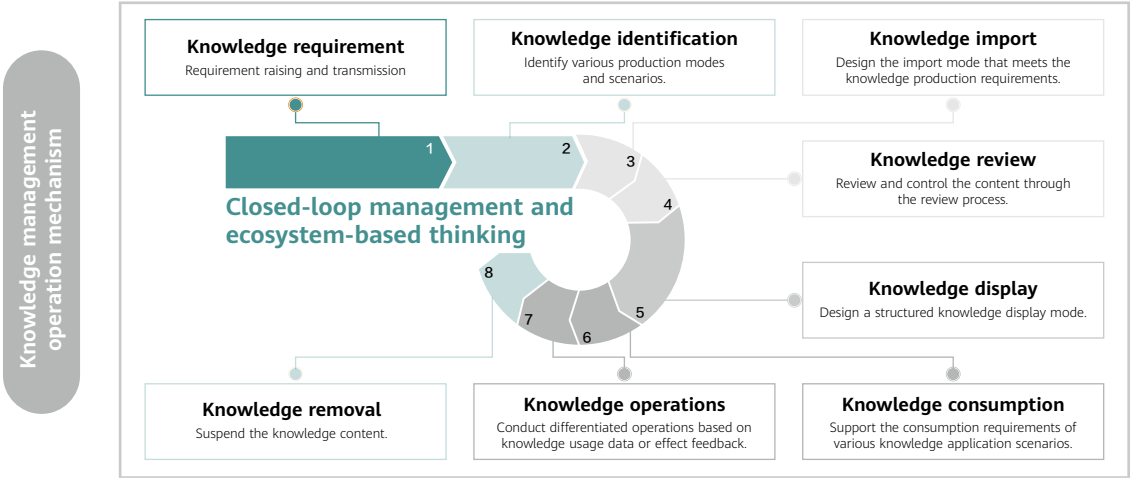


After a customer accesses the intelligent response service, the system implements real-time voice interaction with the customer based on the settings. The background automatically matches related knowledge points through the knowledge base system.

- **The knowledge content can be matched.**
  The system automatically selects voice broadcast or SMS push based on the complexity of the knowledge content.
- **The knowledge content cannot be matched.**
  The system sends messages to the customer, indicating that the related knowledge point cannot be matched and the call is transferred to the agent.

By 2030, digital marketing will keep evolving towards intelligence featuring digital humans powered by large models. Customer service centers will shift from being cost centers to profit centers. Large models will drive a new paradigm for digital business, allowing industries to create more agilely and flexibly while generating ongoing economic benefits.

Enablement and AI: From an Information System to a Knowledge System



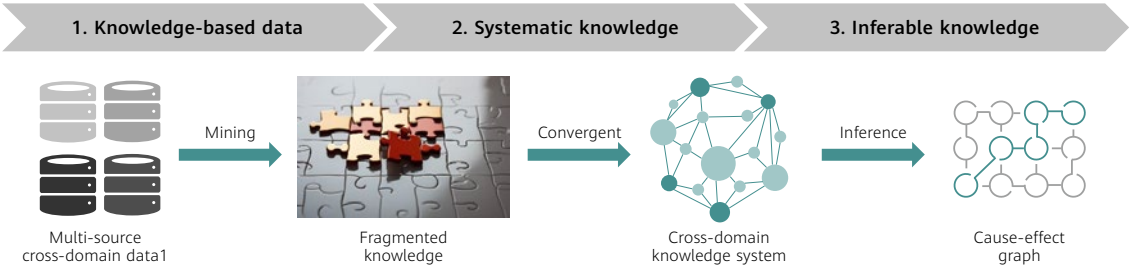
Previous knowledge systems were designed from a human perspective, focusing on learning materials like Word documents, slides, case libraries, and FAQs. Seasoned engineers compile this knowledge through extensive practical experience, and pass it down through generations. Unstructured knowledge makes up over 50% of existing knowledge (Accenture 2024 Insight). Only 5% of unstructured data is effectively managed by enterprises. In the landscape of large models, we need to convert the unstructured knowledge and experience into a format that large models can use, for example, token (for LLMs) and pitch (for videos and images). Besides, professional teams are often lacked during the need identification, import, review, display, operations, and consumption management of traditional knowledge and experience. To equip large models with human-like intelligence, their training data must be up-to-date and highly accurate.

In addition to the mechanism of knowledge operations, a Gen AI knowledge management platform is essential for quick knowledge mining, convergence, and inference. By 2030, Accenture and PwC predict that 55% of enterprises will have deployed knowledge management systems.

1. Knowledge mining: Quickly convert daily expert knowledge into standard data through the portal website and platform, and sort out fragmented knowledge.

2. Knowledge convergence: Associate multi-dimensional tags based on different roles, support keyword-based, tag-based, and association search, and integrate them into the enterprise production flow.

3. Knowledge inference: Perform inference on the knowledge required by various roles, push the knowledge based on scenarios based on the copilot, and automatically recall and generate the knowledge based on the role feedback.



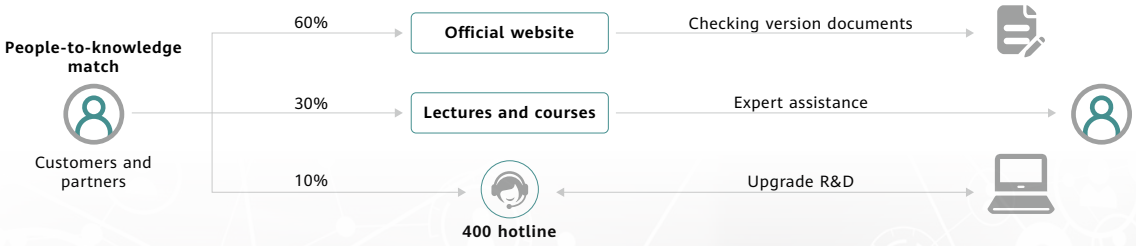
Enablement and AI+: From People-to-Knowledge Match to Knowledge-to-People Match

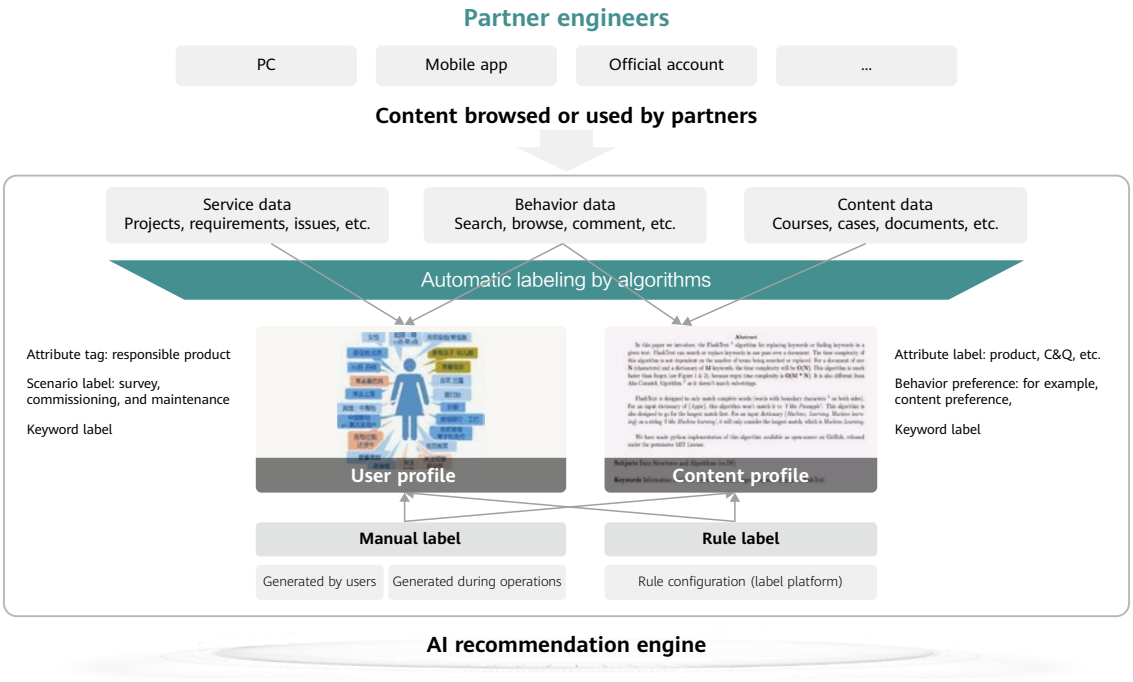
The traditional enablement system is designed based on the fragmented time of human. From the perspective of knowledge acquisition of industry users and partners, we can learn the following information:

- 60% of knowledge is obtained through online searches. General knowledge is typically acquired through industry cases, product manuals, and vendor websites. However, this process is time-consuming and often yields unsatisfactory results.

- 30% of knowledge comes from expert lectures and online and offline courses. Advanced knowledge, such as strategies, new platforms, and new technologies, is acquired through expert lectures and communications. The effectiveness of this enablement depends on the experts' level of knowledge. Furthermore, enablement in this form is not frequent enough, with an average of less than 10 enablement sessions per person per year.

- 10% of knowledge involve private technologies in the industry. Common practitioners do not have the enablement mechanism and can only seek help from vendors' R&D personnel.





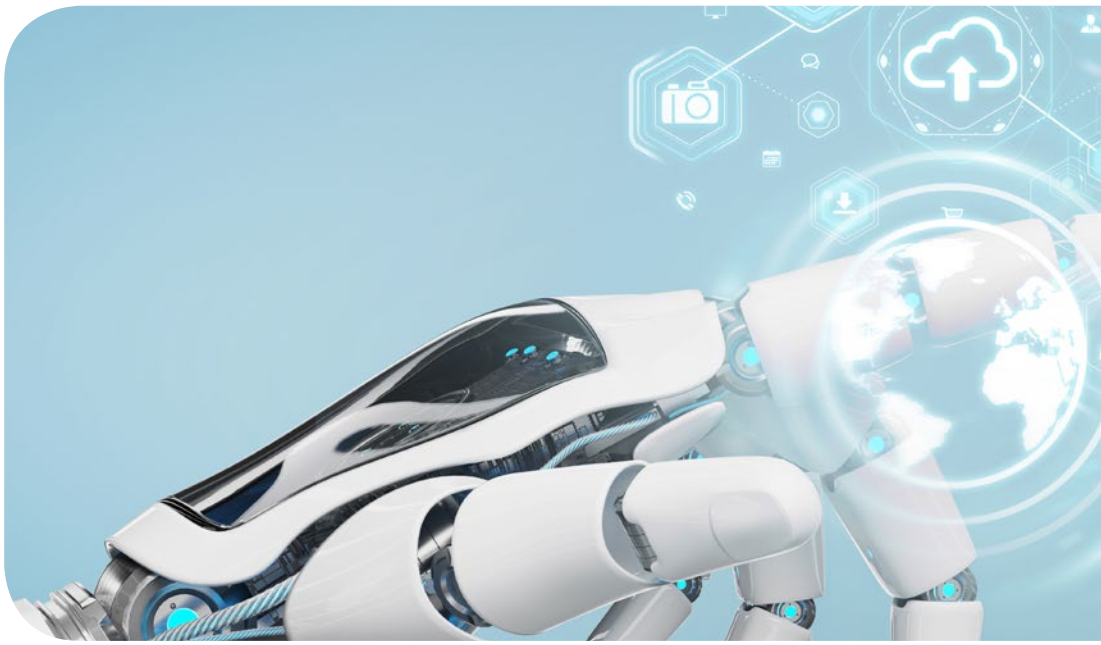
As Gen AI technology gradually penetrates into various industries, knowledge management is becoming an essential element for utilizing large models. From the production of enterprise knowledge to its storage and application, the process will become more standardized and intelligent. Various knowledge application assistants will be integrated into the production flow, providing real-time push notifications to employees with different roles at each stage, thereby significantly enhancing their work efficiency. At the same time, the knowledge of the enterprise can be continuously and rapidly converted into tokens, providing precise knowledge materials for feeding into large models, making the models smarter and more intelligent. With the aid of knowledge management systems and knowledge assistants, employee enablement in enterprises will shift from the traditional "people-to-knowledge match" to "knowledge-to-people match", and from previous fragmented enablement to lifelong enablement based on real-time push notifications aligned with production workflows. The platform identifies business, behavioral, and content data for each role based on employee/partner access records, creating user and content profiles for each role. The system will integrate with employees' daily ERP systems, including OA, OSS, and BSS, to proactively push relevant knowledge to those tho need it most at the right time.

By 2030, Accenture predicts that 80% of enterprise knowledge acquisition will move online through automated and interactive methods. Offline learning will focus on hands-on training, experience, debates, and other practical courses. This integration will help industries and companies continuously promote TECH4ALL and adopt intelligent technologies.

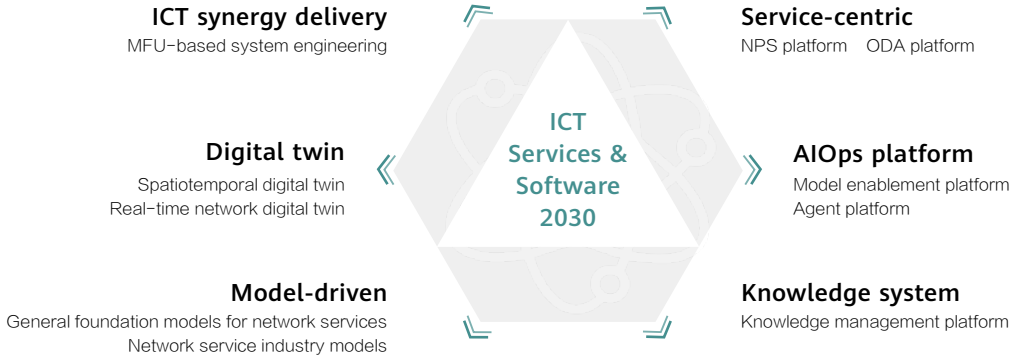
The next-generation enablement platform/ community product will feature online, open, and orchestration as its three key characteristics, aiming to facilitate rapid acquisition, categorization, sharing, and distillation of knowledge experiences across global domains.

- Online real-time learning will reduce the time it takes for humans to acquire knowledge by 90%.
- Cross-regional knowledge sharing and collaboration efficiency will increase by 200%.
- The training period for mid-level engineers will be reduced from 3 years to 6 months.

03



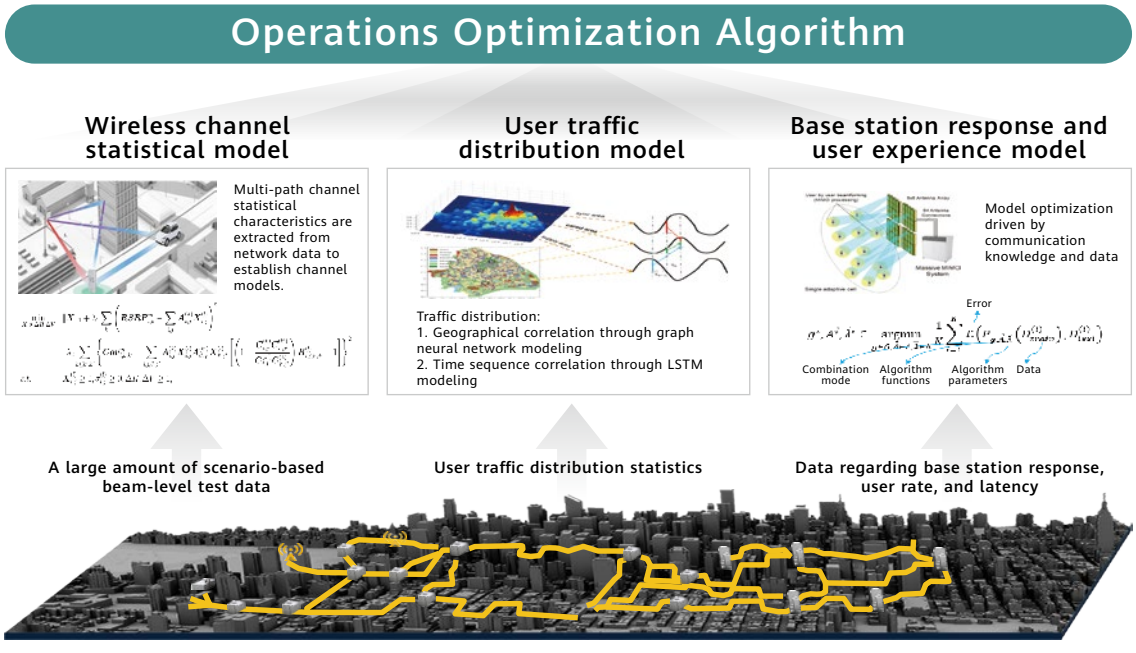
Vision and Core Technologies of ICT Services & Software 2030



Digital Twin

1. Spatiotemporal digital twin: The network optimization of mobile communication systems involves a category of technical issues that are difficult to be modeled in a unified manner. Parameters are mutually dependent or contradictory, making it difficult to establish global optimization models. TAZs identify the distribution differences of service types based on users, services, and networks. Differentiated rate assurance objectives are configured across clusters to maximum the potential of each cluster, improve carriers' revenues, and provide real-time data for network service models. The collection time required is reduced from hours to minutes. The following describes related models:

- Wireless channel statistical model: A large amount of scenario-based beam-level test data is collected across networks. Then, multi-path channel statistical characteristics are extracted from the collected network data to establish channel models.
- User traffic distribution model: User traffic distribution statistics are collected to establish geographical correlation through graph neural network modeling, and time sequence correlation through LSTM modeling.
- User experience model: Data regarding base station response, user rate, and latency is collected to enable model optimization driven by communication knowledge and data.

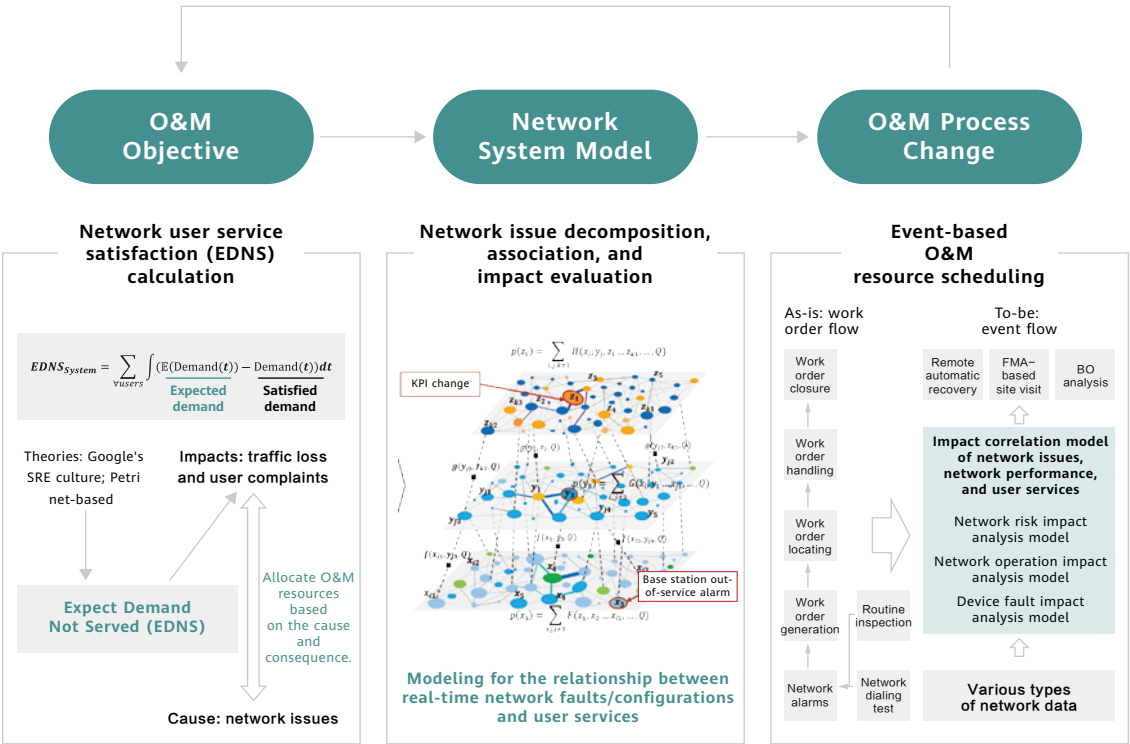


2. Real-time network awareness twin: It measures the impact of network issues on services, and associates the physical world with the digital world for different services. It can also review KQIs and KPIs, as well as visualize and manage the direct impact of each operation and change on services.

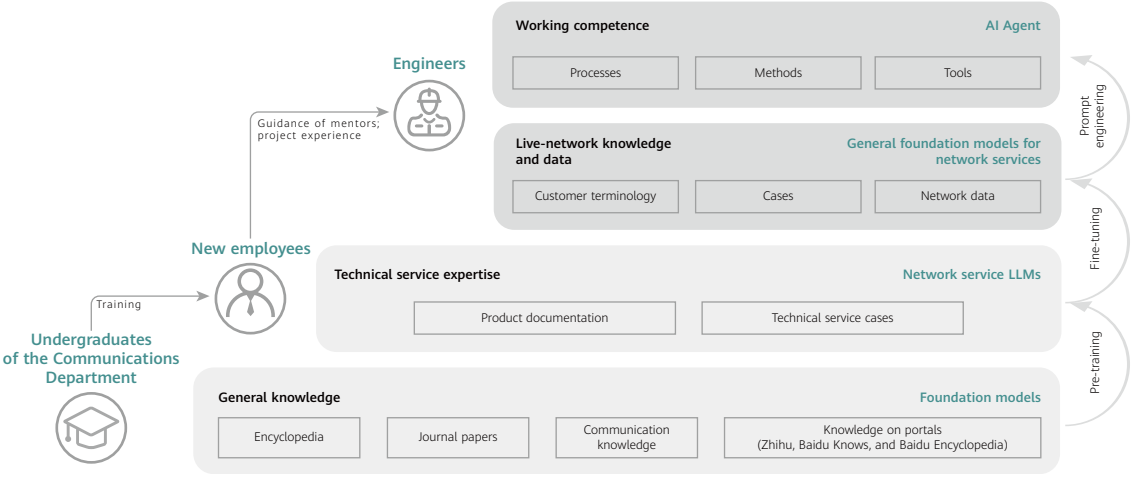
- Algorithm selection based on service objectives: Based on service requirements and the SRE reliability theory, the EDNS calculation logic is established to analyze the logic of service loss reduction.

- Network system model: The impacts on networks, resources, and services are analyzed through modeling. Models and algorithms are designed to analyze the association between each network fault, operation, and service.

- Accurate fault handling based on event flows: The proportion of false alarms is reduced, and precise service-centric O&M is implemented.

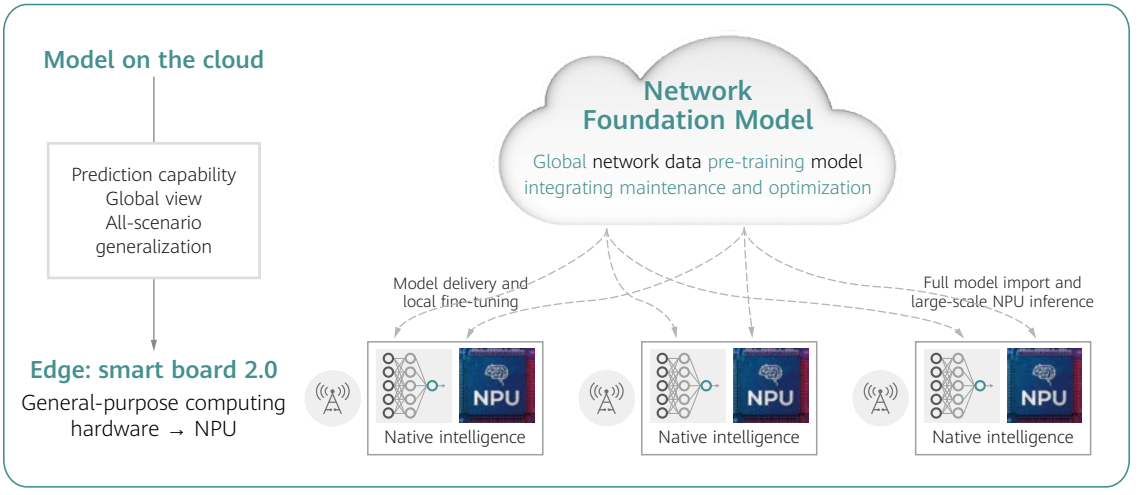


Model-driven



1. Network service LLMs: Global mainstream foundation models, such as Llama 3, Mistral, ERNIE Bot, and Zhipu, are evaluated, adapted, and pre-trained in advance. Common industry knowledge, including information on the C114 website and Wanfang Data, as well as product guides, case libraries, and exam libraries, is aggregated. By doing this, network service LLMs are equipped with basic ICT service knowledge and industry knowledge mastered by high school students. This lays a solid foundation for industry upgrade, and enables models to process and train unstructured knowledge.

2. General foundation models for network services: A general industry model that understands ICT protocols, signaling languages, planning, construction, maintenance, optimization, operations, and OSS/BSS is needed to solve increasingly complex service issues. The model should be able to understand the CoT of complex tasks and achieve an accuracy of over 99%. It needs to provide a knowledge system that consists of structured data and is equivalent to the industry knowledge level of undergraduates. The number of training parameters required for the general industry model is much less than that for LLMs. Small models with 7–10 billion parameters can be fine-tuned and RAG can be quickly deployed to support the production of carriers and industry customers.



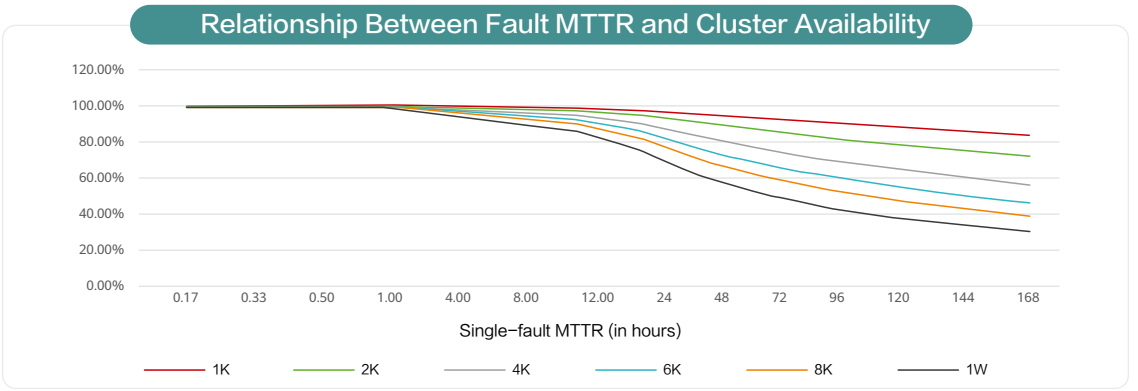
ICT Synergy Delivery

A larger cluster scale (n) indicates a greater impact of single-fault indicators (such as MTBF and MTTR) on cluster availability. According to a report of Meta Llama 3's training clusters with hundreds of thousands of cards in 2024, 466 interruptions occurred during the 54-day training, with an average of 8 interruptions per day. 41% of the interruptions were caused by software exceptions, cable issues, and network faults. OpenAI's Stargate Project will need 1 million cards, which will pose higher requirements on cluster performance (MFU). Each day of interruption will cause an economic loss of tens of millions of dollars. To solve this issue, we need to continuously improve the following two indicators regarding cluster integration and O&M:

1. The cluster linearity is related to network link stability (latency, performance, and configuration consistency) and NPU subhealth. It is strongly dependent on deployment models, configuration optimization, as well as routine subhealth governance and maintenance.
2. The cluster availability is measured by four indicators: single-fault MTTR (h), single-node exception rate (f), number of nodes occupied by jobs (n), and job duration (x).

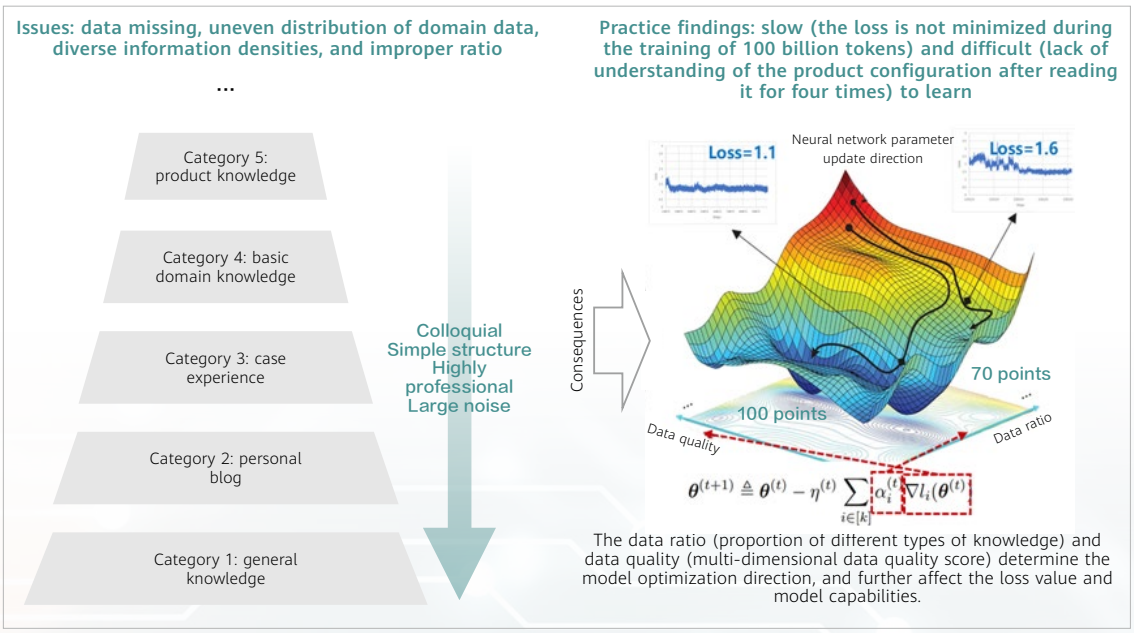
To achieve high availability of training and inference as well as the optimal availability and MFU of computing clusters, three integration capabilities need to be built:

1. Load scheduling at the global, region, and node levels
2. Determined and competitive cluster availability: Faults and subhealth can be detected in advance for proactive maintenance. More than 4000 clusters can run stably for over 30 days. Resumable training after breakpoints is supported, and single-card exceptions have no impact on jobs.
3. Cost-effective linear network, AI computing, and storage performance optimization; hardware pipeline for detecting, isolating, and rectifying subhealth and faults.



Data Engineering

Data is the foundation for the training and inference of large models. In existing practices of large models, 85% of enterprises encounter difficulties in learning. The core reason is that the data engineering required by large models needs to build core advantages of domain-specific models in terms of data scope, data quality, and training efficiency. To achieve this, the following nine core data engineering technologies are required:



The data ratio (proportion of different types of knowledge) and data quality (multi-dimensional data quality score) determine the model optimization direction, and further affect the loss value and model capabilities.

1. Parsing of domain-specific complex process data: Extracts and outputs complex text information of diverse types (signaling protocol interaction, alarm cause-and-effect diagram, and process), with an extraction and parsing accuracy of over 80%.

2. Multi-modal complex information tokenizer: Extracts texts based on the layout analysis of visual models, as well as modules like text matching and table parsing; achieves a PDF content extraction accuracy of over 95%.

3. Efficient domain data synthesis: Achieves self-growth of high-quality domain data — from unlabeled to labeled and from no CoT to CoT — and expands data in typical domain scenarios by 10 times.

4. Automatic data quality evaluation: Automatically evaluates the integrity, accuracy, consistency, timeliness, and quality of pre-trained data, as well as improves efficiency.

5. Domain data augmentation: Augments samples to expand the training dataset, enhances the accuracy and completeness of content semantics as well as the text diversity, improves the model training efficiency, and increases the data diversity by 10 times.

6. Knowledge locating and sourcing: Establishes the association among model capabilities, parameters, and data for overall improvement, locates capability weaknesses, and augments weakness-related data to improve the efficiency of locating data bad cases by 10 times.

7. Optimal data ratio: Breakthroughs have been made in model scaling technologies that obtain the optimal ratio of general data to domain-specific data as well as the optimal ratio for data of different domains in incremental training scenarios. The loss is minimized, and the model training efficiency is improved by 50%.

8. Data curriculum learning: The learning sequence affects the final model effect. The scaling law is used to find the optimal learning sequence and improve the domain knowledge answer accuracy by more than 30%.

9. Data annealing training: Builds the optimal domain-specific data subset, greatly improves model capabilities through multi-stage training and optimal data annealing at the end of training, and achieves a five-fold increase in the annealing rate.

Service-centric

NPS digital analysis platform: To implement service-centric experience management, the major challenge lies in how to identify issues, determine the impact scope, and develop high-quality simulators through digital methods.

1. Promoters and detractors are identified based on the Kano theory. With promoters and detractors as the sampling frame, simple linear regression analysis is performed to identify the reward and penalty driving forces of different factors.

2. Based on the driving forces of depreciation and recommendation, the overall impact on the NPS is deduced. That is, the NPS increase achieved by improving each experience indicator by 10%.

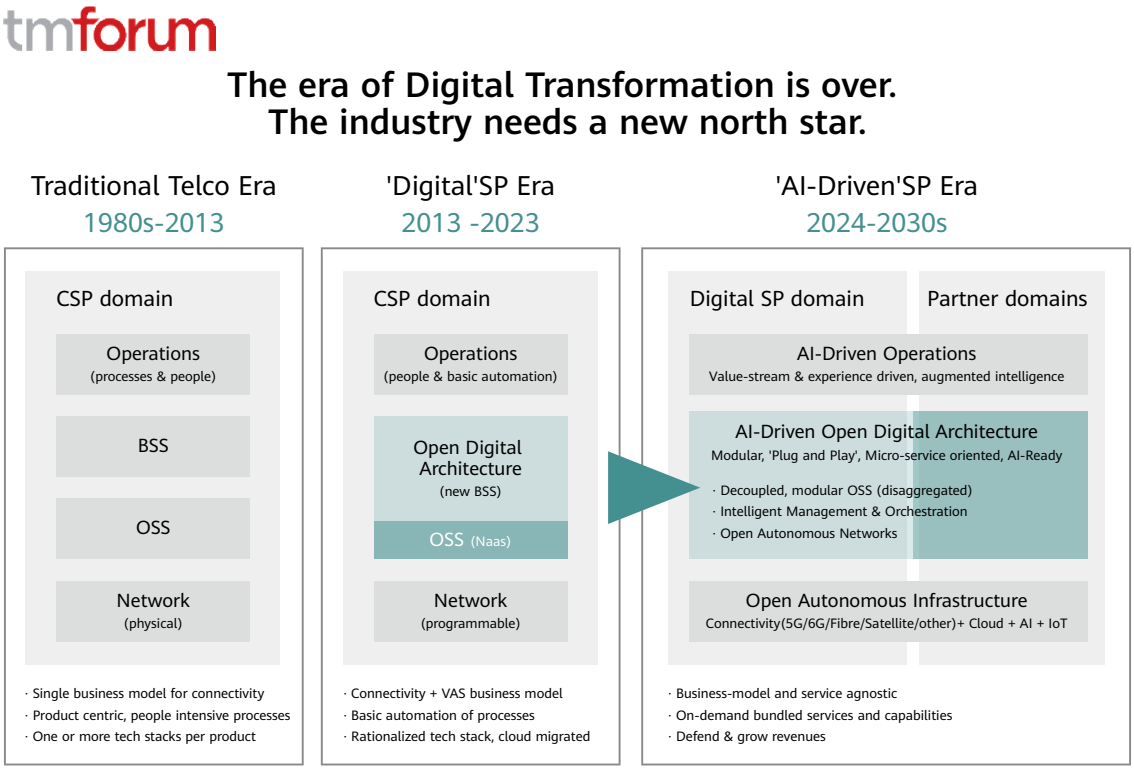
3. Simulators are required to help evaluate the NPS changes caused by the improvement or deterioration of related indicators, in order to develop targeted policies.

The common practice in the target industry is to specify the driving forces of reward and penalty for each indicator through modeling and analysis. Through cross-analysis with the satisfaction matrix, the reward factors that need to be guaranteed and the penalty factors that need to be preferentially improved are identified. The improvement priorities and measures of specific indicators are analyzed based on the driving forces of satisfaction, pain point occurrence rate, and reasons for customer recommendation and non-recommendation.

ODA O&M platform: Although CSPs have built IT systems that can integrate components from multiple vendors, these components are generally provided by communications and software vendors. As CSPs now compete on a larger stage, the ability to integrate components of vendors from industries other than communication becomes essential. When a new way (for example, Agents) of interacting with customers is available, it is unrealistic to wait for its version dedicated to the communication industry. In addition, as open source projects like open network automation platform (ONAP), Open Source MANO (OSM), and Open Baton emerge, any multi-vendor definition must include open source software. Many CSPs also begin to embrace open source in their future-oriented IT systems. TM Forum's next-generation OSS architecture ODA is also designed in attempts to solve such technical issues.

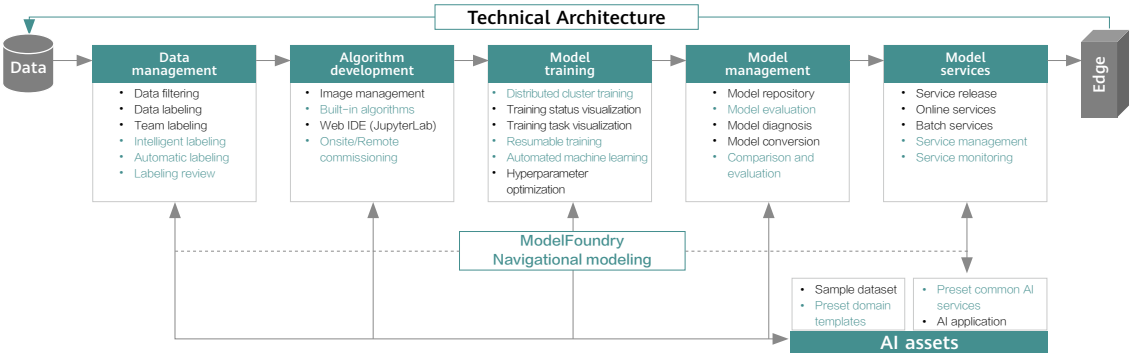
- Cloud components: Newly developed applications are managed in the common repository as cloud components that comply with ODA standard interfaces and specifications.
- DevSecOps: automatic software version verification and release platform.
- The ODA component repository consists of six subdomains: Party, Core Commerce, and Production that carry core processes, as well as Decoupling & Integration, Engagement, and Intelligence that support service processes. The component repository presents a panorama of BSS & OSS components, including 28 components whose functions have been preliminarily defined (that is, numbered components). Most of them are distributed in the Production and Party subdomains.

These new features make it possible for carriers to manage the digital ecosystem on a large scale across borders. For example, globalized automakers can reach an agreement with multiple CSPs on autonomy and IoV agreements. ODA also considers future-oriented key requirements and concepts that will not be considered in independent projects, such as AI, unified BSS/OSS architecture, and unified data-centric approaches.



Source: TMF Strategic Review 2023,12

AIOps Platforms



To prepare for numerous agents in the future, we need to develop model enablement platforms and agent platforms oriented to ICT services and software. With these platforms, customers can quickly develop and operate their own industry-specific models and agents based on the framework. In addition, model training assets can be quickly replicated across scenarios.

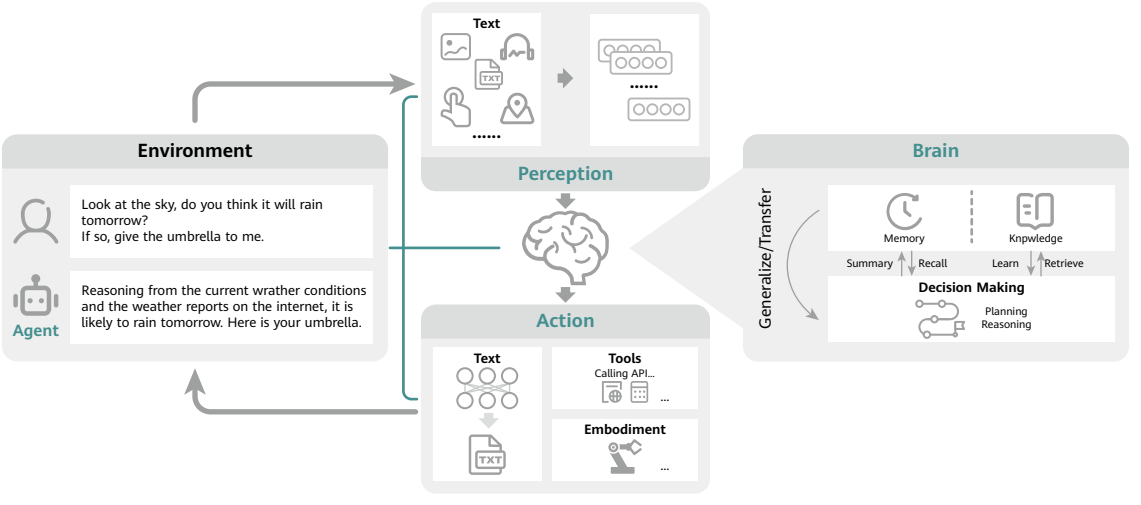
A model enablement platform provides five major services: model selection, knowledge management, model training/fine-tuning, model evaluation, as well as model compression and inference. A comprehensive LLMOps system and open-source tool chain have been established outside China. In China, Baidu and Zhipu are building their own tools based on open-source systems, and developing model enablement service packages and AIOps tool chains based on industries like banking and Internet.

An AI agent platform provides three major capabilities: awareness, thinking, and execution.

1. Awareness: Includes user task acquisition, environment status awareness, and feedback detection. These functions enable agents to obtain required information regarding digital twins, smart boards, and IoT.

2. Thinking: Includes planning, inference, knowledge learning, and memory storage. These functions enable agents to make analysis and decisions by referring to the CoT of people.

3. Execution: Answers texts, uses programming tools (such as APIs) and entity tools, and invokes tools like traditional small models in real time.



04

ICT Services & Software 2030 Initiative

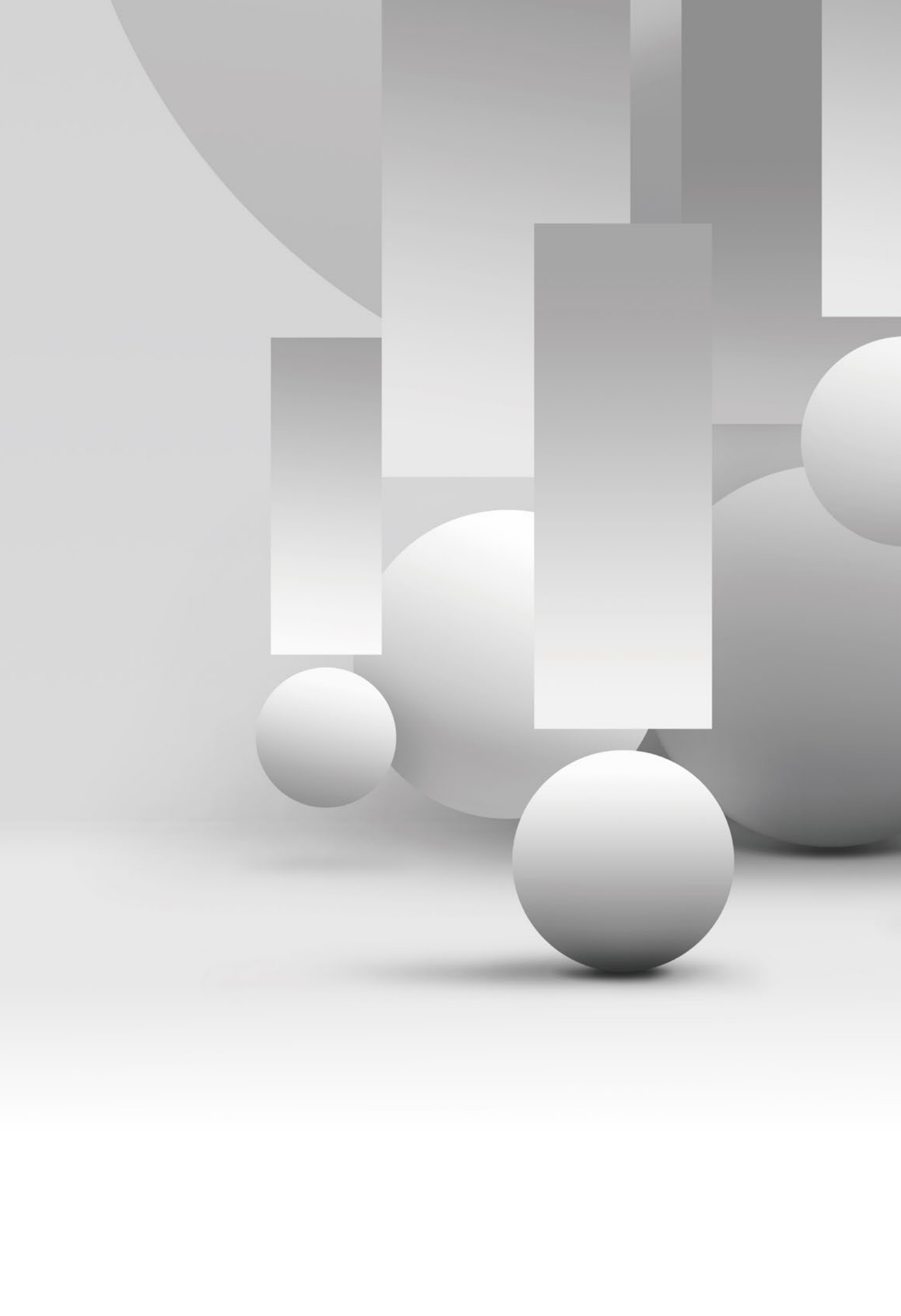
The key to intelligence is the collaboration between traditional business talent and new ICT talent. To enable machines to think like humans and drive intelligent transformation, business experts in each domain must actively learn and gain an in-depth understanding of AI.

By 2030, AI is expected to promote equality, openness, and security. It will enable every individual and organization to benefit from technologies and quickly upgrade infrastructure, businesses, and personnel.

Let's work together to usher in a new era of intelligence.

Glossary (Acronyms and Abbreviations)

Acronym	Full name
5G	5th generation mobile communication
AGI	artificial general intelligence
AHT	average handle time
AIOps	artificial intelligence for IT operations
COT	chain of thought
EDNS	Expected Demand Not Served
ERP	enterprise resource planning
FCR	first call resolution
GenAI	generative AI
IMT	International Mobile Telecommunications
IoT	Internet of Things
KQI	key quality indicator
LLM	large language model
MFU	model FLOPs utilization
MOP	method of procedure
MOS	mean opinion score
OA	office automation
RAG	retrieval-augmented generation
SLA	Service Level Agreement
SRE	system reliability engineer
TAZ	traffic autonomous zone
TECH4ALL	TECH4ALL initiative
TTM	TTM



HUAWEI

Version 2024

Intelligent Automotive Solution 2030

Building a Fully Connected, Intelligent World

Foreword

ICT enables an intelligent automotive industry and helps carmakers build better vehicles

The beginning of the 2020s has marked a rapid shift towards more intelligent electric vehicles within the automotive industry. A new era for the automotive industry is just on the horizon, and we will soon see these profound changes affect our daily lives.

There is an industry-wide consensus that vehicles will be more electric and intelligent.

Carmakers are embracing this trend by actively adjusting their business strategies and ramping up R&D investment. Many have made transformation a core part of their future development strategy and have already begun to take concrete steps in this area.

Technology and user experience are driving rapid growth in the new energy vehicle (NEV) market.

In 2023, China's passenger vehicle market continued to recover thanks to the introduction of next-generation intelligent NEVs. In China, sales of NEVs reached 7.94 million vehicles. This increase of 35% massively exceeded the 5% average growth rate seen across the automotive market. As a result, the market share of NEVs increased from 27.6% to 35.5%. This can be attributed to two key drivers – technology and user experience – that NEV companies have been able to leverage through heavy investment into R&D and closer analyses of user requirements.

Data and software are turning traditional vehicles into intelligent and software-defined vehicles.

Data and software support faster iteration of vehicle functions, helping vehicles deliver experiences beyond consumer expectations. New, ever-evolving functions and services are also providing stable revenue streams for carmakers, pushing the industry to move away from product-centered operations towards user-centered operations.

What it means to "build better vehicles" is changing dramatically for carmakers.

Users are increasingly focused on intelligent and electric features, rather than the traditional mechanical aspects of a vehicle. To make great intelligent electric vehicles, carmakers need to use digital platforms to achieve faster vehicle development and improve efficiency at lower costs. They also need to enable fast software iteration, ensure vehicle safety and trustworthiness, and address other challenges that consumers might face. These are what it means to "build a better vehicle" in the era of intelligent electric vehicles.

In the future, the market for new intelligent connected vehicle components will be worth trillions of dollars. Huawei hopes to bring its decades of ICT expertise to the automotive industry **as a provider of new components for intelligent connected vehicles.** As vehicles become more electric and intelligent, Huawei wants to **help carmakers build better vehicles.**

01

Macro trends: Cross-sector collaboration for shared success

The automotive industry is changing rapidly, and so are its products and industry landscape. As ICT is integrated into the automotive industry at an increasing speed, cross-industry collaboration becomes increasingly important. Huawei is committed to researching basic ICT technologies and bringing its ICT expertise to the automotive industry through partnerships with carmakers.

1.1 Faster industry upgrade brings a bright future for electric and intelligent vehicles

1.1.1 Favorable policies create new opportunities for electric and intelligent vehicles

Carbon neutrality has become a globally recognized mission. Many countries are racing to become carbon neutral. The transportation industry plays a key role in this process as it presents huge opportunities to conserve energy and cut emissions. This in turn makes the NEV industry very promising.

The EU has tightened carbon emissions standards and increased penalties, significantly driving up compliance costs for traditional fossil fuel vehicles. As part of its broader efforts to stimulate the NEV market, the EU also now offers purchasing incentives and tax benefits for those who buy electric vehicles. Similarly, the US has released a 2030 plan for electric vehicles, and is currently accelerating the deployment of charging infrastructure.

In China, low-carbon vehicles are playing a key role in the government's carbon peak and carbon neutrality ambitions. The transportation industry – the automotive industry in particular – is setting out a roadmap for reaching its carbon peak and carbon neutrality goals. The Chinese government has also tightened its dual-credit policy, which assesses carmakers according to their efforts to cut fuel consumption and produce NEVs. This policy has yielded positive results and stimulated significant investment into NEVs. China's big push for the electrification of public transportation is also driving NEV sales.

Many governments are fostering positive policies and regulatory environments for their intelligent automotive industry through independent research and policy guidelines. China, for example, has introduced many policies on intelligent connected vehicles in recent years, including specifications for vehicle quality and safety, functional safety, cyber security, data security, and road tests, which has facilitated the construction of demonstration zones for new products. Looking ahead, standards and regulation for intelligent vehicles will continue to develop in line with new technological advances. This kind of regulatory environment is critical to commercializing mature technologies and promoting sustainable growth in the intelligent automotive industry.

China's New Infrastructure Plan has laid out the requirements for enhancing the top-level designs for information, convergence, and innovation infrastructure, while improving underlying infrastructure like 5G, big data centers, artificial intelligence (AI), charging infrastructure, and integrated vehicle-road-cloud for NEVs. China is also promoting a new development model – the dual circulation model – which aims to create a powerful domestic market while promoting consumption and creating more space for investment. This new model will allow Chinese carmakers to compete globally while

also increasing internal circulation by stimulating domestic demand.

1.1.2 ICT accelerates upgrades in the intelligent automotive industry

The new vehicle lifecycle sees core functions continuously upgraded. This means vehicles need more sophisticated electrical/electronic (E/E) architecture, system on chip (SoC) computing power, software and data use, and cyber security. This is changing the automotive industry at a fundamental level, as it embraces more advanced ICT technologies and solutions.

Moore's law has long been the golden rule for the semiconductor industry, profoundly influencing the development of PCs, digitalization, Internet, and more for over 50 years. The next 10 years will continue to see Moore's law governing the development of computing power required for intelligent vehicles. Huawei predicts that that a vehicle will require more than 5,000 trillion operations per second (TOPS) of computing power by 2030 to enable the further advancement of telematics applications like intelligent driving, intelligent cockpits, and extended reality (e.g., augmented reality [AR] and virtual reality [VR]).

5G (including 5G-A) promises high bandwidth, low latency, and ultra reliability, making it possible to meet the essential connectivity requirements of intelligent vehicles. By 2030, intelligent digital platforms, powered by emerging technologies like 5G, cloud, big data, Internet of things (IoT), and optical technologies, will connect the physical and digital worlds for vehicles. This will greatly drive innovation and upgrade in the automotive industry.

1.1.3 Changes in supply: Vehicle sales will surpass those of fossil fuel vehicles by 2030

Huawei predicts that NEVs will account for 82% of China's total new vehicle sales in 2030. Electric vehicles will become much cheaper than fossil-fuel-powered vehicles. As charging and battery-swapping become more efficient, electric vehicles will be able to drive for one kilometer after just one second of charging. In addition, as more people switch to electric vehicles and China's New Infrastructure initiative is implemented, the number of charging and battery-swapping stations will continue to increase. This will make it easier for electric-vehicle drivers to travel longer distances without worrying about running out of power.

Carmakers both in and outside China are accelerating the deployment of NEVs. By 2030, Volvo, Bentley, Jaguar, BYD, and Geely will be exclusively manufacturing NEVs. Volkswagen and BMW have committed to making at least 50% of their new vehicles NEVs by 2030.

1.1.4 Changes in demand: Stimulating the intelligent electric vehicle market

User demand for intelligent electric vehicles is on the rise. As electric vehicles will soon cost much less and become more convenient, China is set to benefit greatly from its large consumer market. This gives the nation a base from which it can further develop intelligent electric vehicles. The Chinese market has long been less saturated than more developed markets, and Chinese consumers are also proving to be more receptive to new developments like electric vehicles and intelligent driving.

Due to China's constantly changing demographics, income structure, and consumer purchasing behavior, its consumption structure is also changing at a rapid pace. China will soon become a middle- or high-income economy, and consumption is expected to continue to grow as per capita GDP and household disposable income increases. Consumer distribution is also changing, which means demand is becoming more diversified. China's Post-2000 generation, a Gen Z analogue of true Internet natives, are big fans of new technology and individual expression. They represent a huge engine for growth in domestic consumption. A sliver economy – all economic activities linked to China's older age groups – is also emerging. The two- and three- child policy is also reshaping consumer demand.

Such changes in consumption structure will also affect car consumption, both directly and indirectly pushing China's vehicle market from relying on traditional models to digital models, from commodity-oriented to experience-driven, and from valuing commonality to valuing individuality.

1.2 Changes in product attributes: Reshaping the automotive value chain

1.2.1 Key vehicle differentiators: From powertrain and chassis systems to intelligence

As vehicles' power systems become electric, their powertrain and chassis systems will gradually become more standardized. This makes intelligent cockpits, intelligent driving, and other intelligent functions the key differentiators of vehicles. The intelligence level of vehicle cockpits and driving systems will become key factors in users' purchasing decisions. Over-the-air (OTA) updates will be used to deliver superior user experiences and further increase users' uptake of intelligent functions.

Such shifts also present an opportunity for carmakers to expand their hold on the vehicle market. As the laws, regulations, and policies on intelligent vehicles continue to improve and intelligent driving technologies become mature, autonomous driving will enter large-scale commercial use via robotaxis and closed or semi-closed low-speed driving scenarios by 2030 before gradually being implemented in passenger vehicles. In addition, human-machine interaction technologies will continue to advance and the intelligent cockpit application ecosystem will continue to improve, making vehicles an intelligent mobile "third space" outside of home and workplace.

1.2.2 A wider industry: From automotive products to all-scenario mobility services

5G (including 5G-A), IoT, AI, edge computing, and low-carbon technologies are still rapidly developing, converging, and iterating. They are accelerating the automotive industry's CASE transition. How carmakers will commercialize intelligent vehicles in specific scenarios is becoming increasingly clear.

As intelligent driving technologies continue to improve in different market segments, scenario-specific autonomous driving applications will be become more widely adopted. New forms of autonomous vehicles will emerge for specific scenarios, and the connection between transportation tools across different scenarios will become more seamless. Autonomous mobility services will appear in every link of people's travel. This will fundamentally change how people travel, how people interact with transportation tools, and how transportation tools interact with each other, greatly improving "mobility-as-a-service".

People's basic mobility needs have gradually changed from owning different transportation tools for different scenarios to using integrated mobility solutions for complex mobility scenarios. Many third-party application developers are mobilizing industry resources to develop new service applications for different scenarios, with the purpose of seamlessly connecting different transportation tools and providing end-to-end intelligent mobility services for users. These mobility solutions and services are providing new revenue streams for the automotive industry.

1.2.3 New profit models: From hardware to software and services

As key differentiators for vehicles change and the automotive industry's reach expands, individual intelligent vehicles will become platforms for continuous value creation. This will reshape the automotive industry's standard business model and value distribution pattern.

Carmakers have long profited from one-off deals – multiplying the unit price by the total number of vehicles or hardware units they sell. Software-defined vehicles will turn software and services into new revenue streams for carmakers, and their profits will be determined by software fees and car parc. Moving forward, data and software will support ongoing, OTA iteration of vehicle functions by allowing carmakers to remotely repair and upgrade products, and improve user experience. This will give users more flexible and operable service models, driving a shift in the industry from product-centered operations to user-centered operations. These revenue streams will also be more stable for carmakers.

The automotive industry as a whole will focus on the new operation and charging model created by intelligent driving as it greatly expands profits for carmakers. In addition, software-defined vehicles will reshape the value chain, and creating more opportunities to unlock value. This will attract more third-party developers and innovators to invest in the intelligent automotive industry, which will help improve the intelligent connected vehicle ecosystem and build a positive cycle of value creation.

1.3 Cross-sector collaboration will define the new industry landscape

1.3.1 Carmakers and tech companies work together to maximize their strengths

Intelligent vehicles are the product of multiple industries, built on the integration of core digital technologies (e.g., ICT, software, big data, and AI) and traditional mechanical technologies. Emerging carmakers are the frontrunners in the first phase of CASE journey. However, other carmakers are also joining the trend and beginning to improve their core competencies in software, electronics, and big data.

Auto underbody solutions are slowly standardized, becoming a shared platform on which other industries can grow. Tech companies, from

consumer electronics manufacturers to Internet companies, are taking advantage of this trend and expanding into the automotive industry either on their own or through alliances. These companies have large amounts of capital, strong experience in ICT, significant technological innovation capabilities, and huge brand recognition. Their entry into the industry is driving rapid development in intelligent connected vehicles and pushing the automotive industry towards CASE 2.0.

The automotive industry has been around for over 100 years, and carmakers have emerged as leaders in manufacturing, quality control, safety, and reliability capabilities. Tech companies, on the other hand, have amassed extensive experience and advantages in intelligent technology applications, such as AI algorithms and big data. Software-defined vehicles will significantly change how companies capture value, serve their customers, and build their talent mix. To meet increasing user requirements, all companies along the value chain should become more agile and adapt to this new environment.

Software and hardware will decouple and general platformization and standardization will continue.

This means the only way forward will be to foster a more open supply chain system and adopt more flexible vehicle models. Carmakers and tech companies will need to maximize their respective strengths, while also relying on cross-industry partnerships to find new and innovative ways to achieve business and social value.

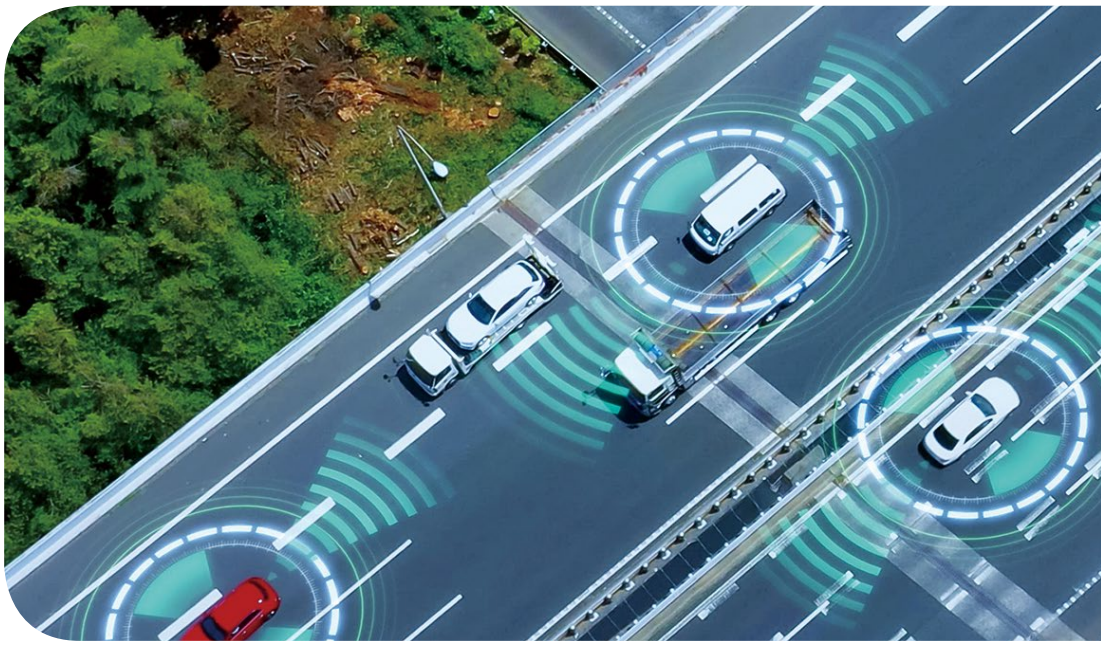
1.3.2 ICT is key to better travel experiences in the growing mobility sector

As the automotive industry's reach expands, demand for transportation and mobility services in new market segments will continue to increase. This will drive exponential growth in the forms and quantity of vehicles, as well as their related infrastructure. More and more traditional vendors are announcing transition plans to become "mobility solution providers", intending to tap into this huge new market. Different players will have different roles to play.

Mobility solution providers provide end-to-end solutions to meet user requirements across different mobility scenarios and control user traffic. Operators in closed scenarios understand operation requirements, customize the form factors of vehicles, and deploy infrastructure in closed scenarios. Carmakers use their existing manufacturing platforms and supply chains to manufacture these vehicles. Tech companies provide solutions like intelligent software and hardware, intelligent driving, and cockpit connectivity and control. Third-party ecosystem developers provide massive numbers of apps and deliver seamless travel experiences to users.

As the forms and quantity of vehicles continue to grow alongside their related infrastructure, we need to connect the vehicles and infrastructure in specific scenarios. Scheduling and connection across different scenarios, data sharing between different vehicles, and scenario-based smart service applications will need to be hosted on a cloud-based "brain". ICT will play a key role in connecting these disparate points and enabling scenario-based digital service sharing.

02



New scenarios: Bringing intelligence to every vehicle

As digital technologies are widely adopted and carbon neutrality has become a globally recognized mission, it is becoming an obvious trend that vehicles will become more electric and intelligent. Bringing intelligence to every vehicle will empower intelligent driving, intelligent spaces, intelligent services, and intelligent operations. This will allow for safer and more efficient transportation, greener and more convenient travel, more fun and intelligent lifestyles, and more efficient and lower-carbon operations.

2.1 Intelligent driving: Safer, smoother, and more efficient travel experiences

Intelligent driving can be categorized into six levels by automation, ranging from 0 to 5. Level 0 (L0) refers to traditional human driving with no automation. L1 offers AI-assisted driving with low automation, meaning that continuous driver assistance is required; L2 offers a moderate level of driving automation; L3 provides conditional driving automation; L4 delivers a high level of driving automation; and L5 represents full driving automation, which means that the vehicle can

be entirely controlled by the AI system with no human operation needed. Intelligent driving can be applied to almost any kind of road or area that meets the necessary level of automation. L2 intelligent driving is already available in passenger cars, providing consumers with a safer and smarter driving experience. L3 and higher levels of intelligent driving are still undergoing testing. L4 and L5 intelligent driving will first be seen on highways, campuses, and other closed roads, and will gradually expand to public roads such as streets in urban areas. We believe that advanced intelligent driving will revolutionize mobility and our entire society.

In 2030, robotaxi services provided by self-driving fleets are expected to cut labor costs and provide 24/7 mobility services that are more flexible and affordable.

Intelligent driving technologies will be integrated into existing modes of transport to provide safe, efficient, and affordable mobility service solutions that meet different travel needs and deliver the best possible experience. Mobility resources will be centrally managed and data will be shared in real time, making it possible to build an end-to-end, point-to-point, and door-to-door mobility network. This will in turn help maximize the use of all available mobility resources.

When a user plans a trip, a cloud-based brain weighs up all the possibilities and, based on real-time awareness of the traffic situation, offers the optimal route and mode of transport. Diverse mobility resources will allow users to enjoy efficient, green, and safe travel while maintaining a dynamic balance in urban transportation capacity, contributing to sustainable urban development.

2.2 Intelligent spaces: From a flexible mobile space to an intelligent living space that integrates the virtual and physical worlds

Vehicles are no longer just a tool for transport. Their relationships with people and with their surroundings are changing dramatically.

Advanced intelligent driving technology will free commuters to enjoy work, study, entertainment, and much more within their vehicles. When vehicles serve as mobile offices – or even mobile living rooms – it will change how people think about their daily commute.

Powered by human-machine interaction, in-vehicle optical technologies, and immersive AR/VR technologies, intelligent cockpits will become multi-functional units. People will find themselves spending more time in their vehicles even when they do not want to go anywhere. For example, it will not be uncommon to see people sitting in parked vehicles watching movies.

The way we interact with our vehicles is about to experience three major changes. First, a cockpit will no longer be a combination of a steering wheel, an instrument panel, and a screen; it will integrate the virtual and physical worlds. Features such as voice control, facial recognition, and gesture interaction will make interactions simpler, more natural, and more efficient. What's more, it will not be long until brain-computer interfaces start seeing commercial application. Second, our vehicles will no longer passively await our instructions; they will intelligently anticipate our needs. Technologies such as AI, biometric recognition, emotion perception, and vital sign monitoring will allow vehicles to better understand drivers' behavior, habits, and thinking, and become close partners that truly understand drivers' needs. Third, in-vehicle optical technologies will offer a richer spatial optical experience, and AR/VR technologies will further transcend the barriers of time and space. Such an immersive experience will drive broader and richer vehicle applications in both mobile and static scenarios.

An intelligent vehicle will become a truly intelligent space that integrates the physical and virtual worlds.

(1) Driving: The combination of in-vehicle sensors and wearable devices can accurately monitor drivers' health indicators, recognize fatigue, and send timely reminders to ensure safe driving.
(2) Entertainment: Passengers will be able to have true-to-life experiences of concerts and sports events without having to be there in person. A cinema will no longer be the best place to watch movies. AR technology will make gaming more immersive. Vehicles can become a personal entertainment space, a private cinema, an outdoor cinema, and a preferred place to play games with friends.
(3) Mobile office: Seats can be adjusted and rotated and windows can be used as large projection screens. A conferencing stream on a smartphone can be easily transferred to the vehicle, and the shielding function of the vehicle's sound zone ensures the privacy of the conference. Vehicles will become mobile offices, allowing professionals to get work done on the way to the airport, a restaurant, or their homes.
(4) Social networking: Drivers will not miss the beautiful scenery outside the window. Cameras mounted on the outside of the vehicle can be used to record, edit, and share beautiful moments. Getting stuck in a traffic jam no longer has to be boring. You can watch movies, play games, and make friends with nearby drivers using the head unit. AR/VR brings your friends within easy reach. With separate sound zones created within your vehicle, you will even be able to keep conversations private from other people in the vehicle.

2.3 Intelligent services: More intelligent scenario-based services

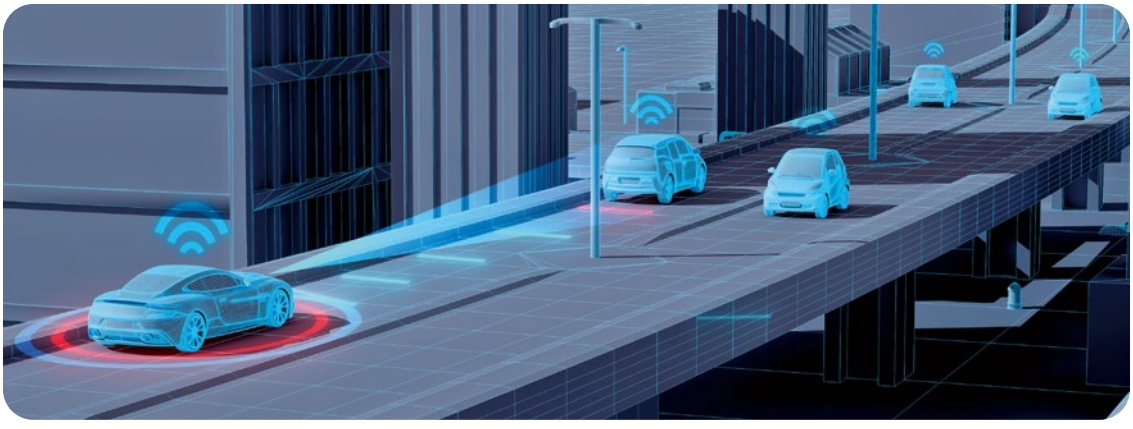
Digitalization is reshaping the world, and as a result, consumption patterns will change dramatically over the next decade. Services in all kinds of industries will be available online and become more customized, personalized, responsive, and scenario-based. As digital technologies are deeply integrated with vehicles, services will be intelligently and rapidly pushed to users based on the scenario they are in. This will be achieved in three ways.

First, as vehicles become more intelligent, interactions between users and vehicles will inform the services to be provided. Intelligent algorithms can identify, analyze, and understand users' interactions, predict their behavior based on their basic information and historical preferences, and provide the right services. Intelligent vehicles will continuously improve their understanding of users, in order to deliver better services.

Second, intelligent vehicles will make it possible to efficiently and accurately identify user needs in real time. By identifying and analyzing vehicle data, location information, and surrounding environments, intelligent systems can determine what scenarios users are in, proactively predict user needs, and provide the right services.

Third, brand-new, interconnected operating systems (OSs) can create more service scenarios, giving rise to an application ecosystem that is based on new modes of interaction. As the intelligent world approaches and the digital economy develops, a richer digital ecosystem is emerging to support all scenarios. In the connected world, more services will be provided by intelligent vehicles. The scenario-driven functions and services offered by connected vehicles will become increasingly intelligent, efficient, and convenient.

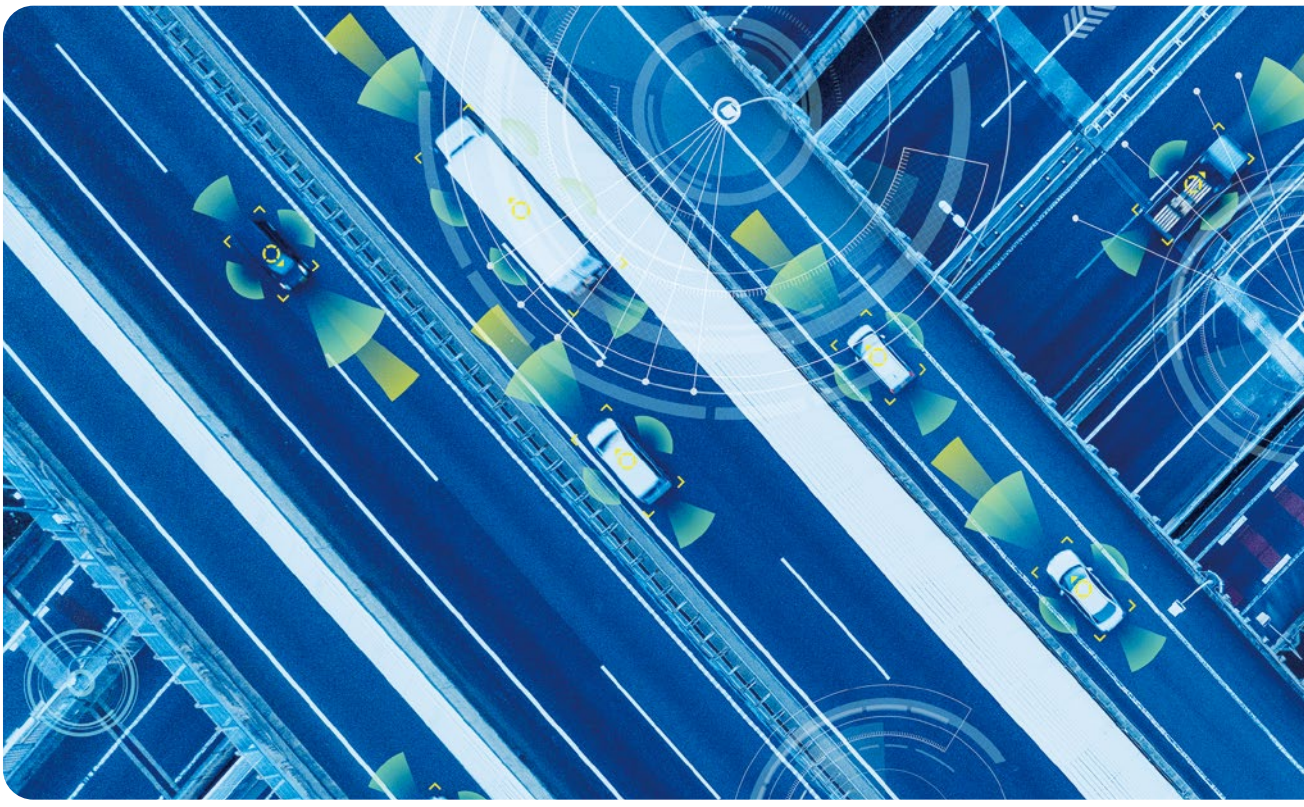
We can even imagine a situation in which a group of people would want a pizza while driving across town. Mobility-as-a-service providers will provide a shared vehicle that perfectly matches the passengers' preferences, select a high-rated pizzeria located along the planned route, and order the pizza in advance. The restaurant will then prepare the food, which will be collected by a drone. When the vehicle arrives at the designated handover location, its sunroof will automatically open and the drone will lower the food inside. This is a level of service that can only be achieved when every part of the process is seamlessly connected.



2.4 Intelligent operations: Autonomous driving is expected to be applied in commercial vehicles first, boosting the productivity of intelligent operations

Commercial vehicles are important tools for transport and the functioning of modern society. Their evolution into intelligent autonomous vehicles can help achieve the goal of carbon neutrality and boost work and operation efficiency while contributing to a more mature intelligent vehicle ecosystem. By 2030, commercial autonomous vehicles will be used on trunk lines and public roads in addition to closed areas and dedicated roads. This will make intelligent operations a reality and greatly increase productivity.

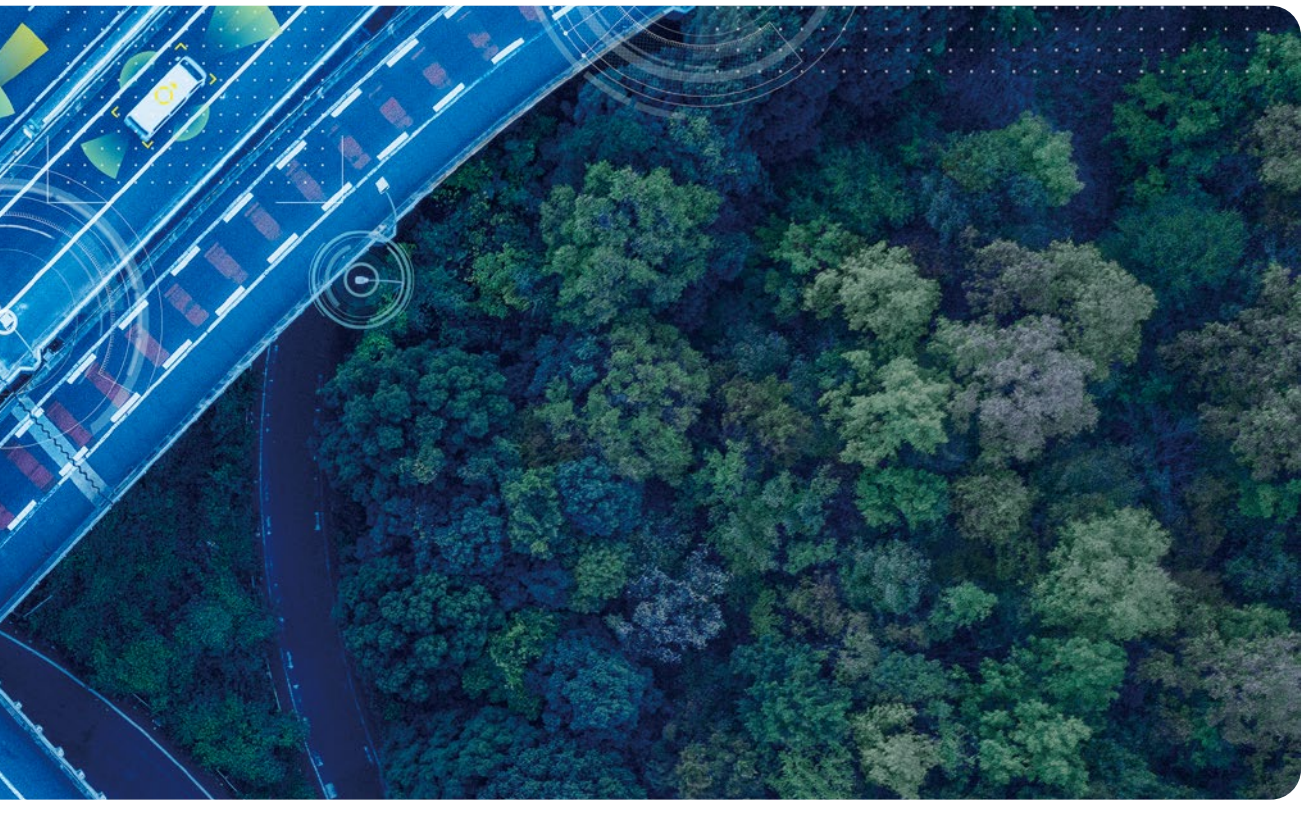
When autonomous vehicles are operated in a closed area, it is possible to enumerate all the scenarios in which a vehicle might find itself and foresee potential emergencies. For this reason, commercial autonomous vehicles will find their first large-scale commercial applications to be in closed areas like ports, mines, farms, campuses, airports, and closed scenic spots. In these areas, intelligent commercial vehicle technologies will not only be applied to transport vehicles; they will also be integrated with operations management systems to build unmanned manufacturing systems where autonomous driving applications have been integrated into the core production systems that support transportation and distribution. This means intelligent vehicle technologies will be commercially used on a large scale.

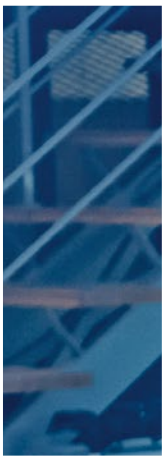


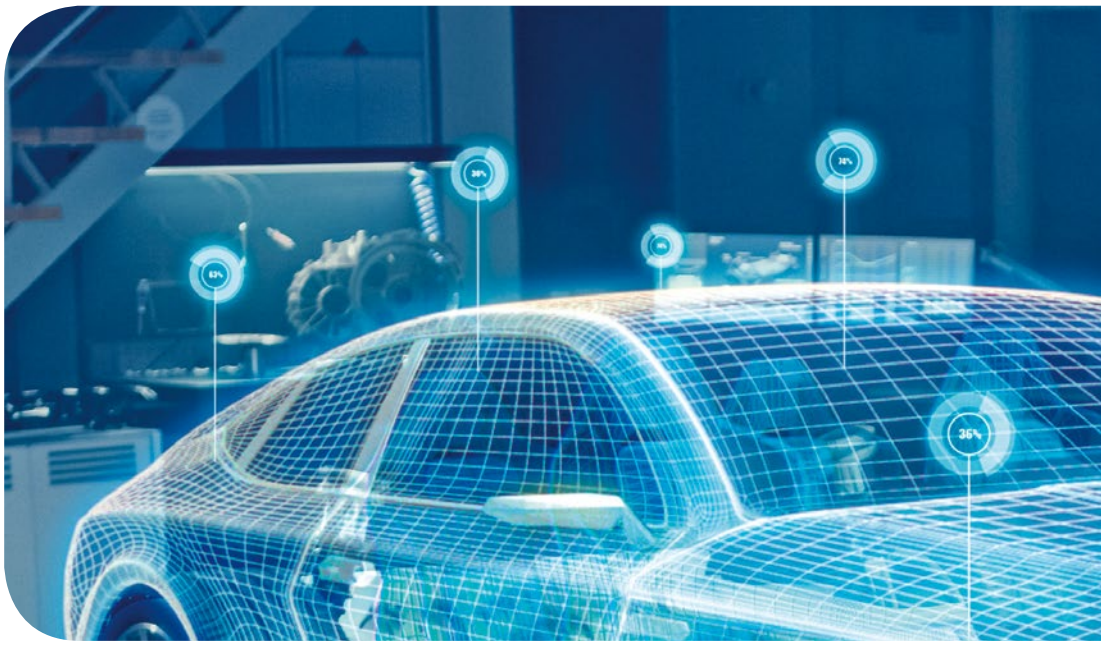
By 2030, vehicle-road-cloud collaboration solutions will make it possible for multiple autonomous vehicles to collaborate in closed areas, which means autonomous driving has the potential to be commercialized in vertical industries. Service capabilities, like comprehensive environment perception, global resource scheduling, dynamic service mapping, multi-vehicle cooperative driving, lane-level route planning, coordinated signal control, and service simulation testing, will further streamline service processes, make the collaboration between multiple autonomous vehicles a reality, and increase scenario-based operation and transportation efficiency. All of these will help cut costs and boost productivity.

Cloud scheduling will be critical to the service management and scheduling of autonomous vehicles. When intelligent commercial vehicles are used in closed areas, operation managers will use the vehicle cloud control management system to schedule and monitor those vehicles, and support service and safety management using end-to-end large models. In a port, for example, the operation control platform of an intelligent horizontal transport system will be connected to the terminal operating system (TOS), which means the scheduling of autonomous container trucks will be fully integrated into the automatic port scheduling system. This level of integration will help fully automate port operations.

As road infrastructure is upgraded over the next decade, on trunk lines, commercial vehicles will gradually shift from assisted driving systems to fully autonomous driving. As electric vehicles are widely used in urban short-distance transportation, and roadside network infrastructure becomes more intelligent, the penetration rate of intelligent commercial vehicles is expected to rise sharply on more complex public roads, including urban roads. Building on the basic capabilities of autonomous vehicles and their potential commercial application, carmakers can work with ecosystem partners to build more viable intelligent driving applications to overcome the challenging scenarios faced by commercial vehicles.







03

Technology projection: New components will drive sustained innovation in intelligent vehicles

3.1 Evolving to a central computing and communications architecture for software-defined vehicles

Today, most vehicles still use a decentralized E/E architecture. Under this architecture, each separate function has an independent controller, so a vehicle has almost 100 controllers and over three kilometers of wiring. This makes vehicles costly and heavy, and difficult to automate vehicle assembly. In addition, electronic control units (ECUs) are often developed by different vendors, meaning they have inconsistent standards which carmakers struggle to develop new functions or perform OTA updates. In the future, intelligent connected vehicles will be even more complex. The volume of data collected by vehicle-mounted sensors will dramatically increase. This will raise the bar for real-time data transfer and data processing. These trends mean that vehicles' E/E architecture must evolve.

With the rapid development of digital and intelligent technologies, vehicle functions are becoming more integrated and centralized. There is a widespread understanding in the automotive industry that a decentralized architecture is no longer viable, and that it must evolve first into cross-domain architecture, and eventually into a single central computing platform. Vehicle functions will become applications loaded onto a central processor so that they can share a single set of sensors and actuators. Components will gradually become more standardized. This will help reduce the cost and complexity of new function development in the long run. Meanwhile, domain controllers will evolve through the addition of new software features. By 2030, the

E/E architecture of vehicles will be a computing and communications architecture that consists of a central computing platform, zonal control, and high-bandwidth in-vehicle communications.

3.1.1 High-performance central computing platforms power software-defined vehicles

Central computing plus zonal control, either in a hub-and-spokes or ring model, offers architecture stability and functional scalability. Within this architecture, new external components can easily be added from the gateway of the nearest zone, and with pluggable hardware, computing capacity can be upgraded as necessary, enabling simple iteration and upgrades of application software on the central computing platform.

The mobility scenarios are complex and subject to frequent change. New functions such as intelligent cockpits, intelligent driving, and vehicle control are constantly being developed. A high-performance central computing platform is required to support them. This platform can perform several thousand TOPS. It must be based on a powerful SoC with a vehicle-specific operating system, middleware, toolchain, and a centralized architecture. Such platforms will offer a stable architecture for software-defined vehicles, while still allowing room for smooth evolution. The chassis, powertrain, cockpit, and intelligent driving system will each place different demands on the central computing platform in terms of security, latency, dynamic response, and the supporting software ecosystem. A high-performance in-vehicle central computing platform will use hardware virtualization and a central functional safety framework, as well as AI algorithms to deliver the necessary levels of security and ensure that hardware resources are available as required for each domain. The technologies required include:

- **High capacity processors:** SoC will deliver the thousands of TOPS of computing power required by the vehicle, including the chassis, powertrain, cockpit, and intelligent driving system. Key technologies needed to build SoCs include computing in memory and trustworthiness and functional safety islands.

- **High-speed concurrent processing for guaranteed low latency:** High bandwidth is only part of low latency; an even more crucial factor is the capacity to process data in real time. High-speed concurrent processing enables central systems to simultaneously receive data from multiple sources. It prevents data surges, and will enable the vehicle to handle the data generated by an ever-growing number of apps and the ever-increasing demands on the system.

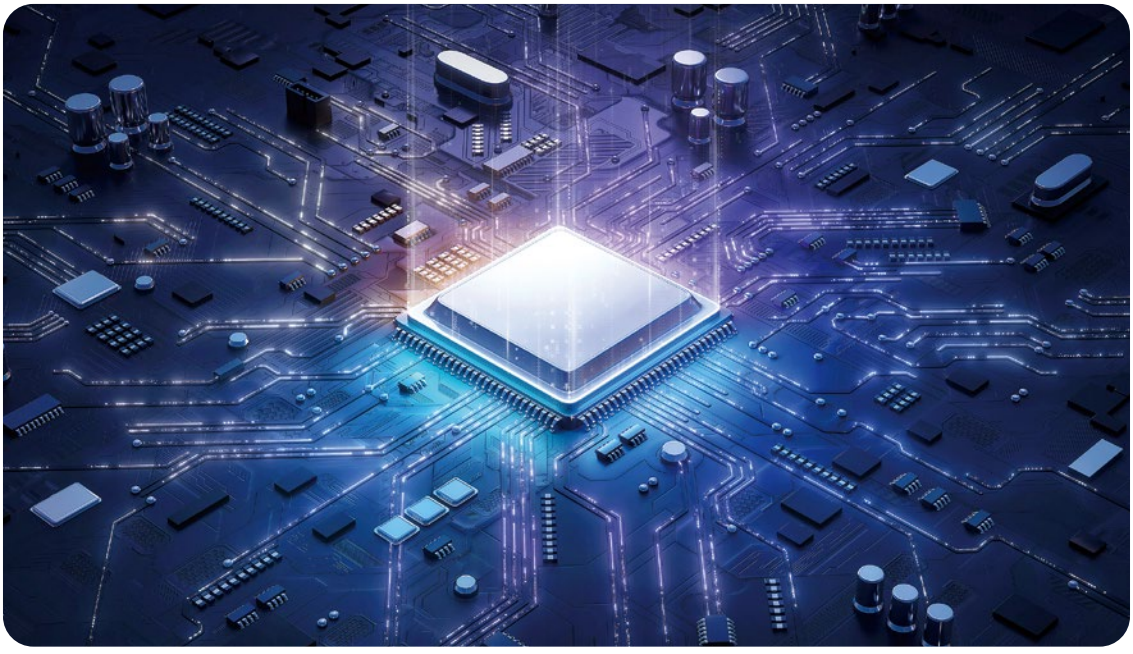
- **Hypervisor secure partitioning of hardware:** The Hypervisor allows one physical server to function as multiple virtual servers and delivers customized functional safety for the different domains of a vehicle. AI engines monitor and forecast the workloads for different virtual partitions and dynamically schedule hardware resources, thereby achieving secure partitioning and load balancing.

- **Inter-app freedom from interference (FFI):** The Hypervisor partitioning function delivers a secure silo of resources for the applications, communications mechanisms, OS, and hardware accelerators. Within the processors, a dedicated safety island provides a safety system that reaches the standards of 3-Level Safety Monitoring. The safety island's intelligent fail safe and fail operational functions enable coordination of the safety responses with other vehicle functions.

Building on a powerful central computing platform, the software-defined vehicle sector will concentrate efforts on agile development and real-time release of new functions to deliver the diverse experiences that users will demand of mobility.

3.1.2 In-vehicle, high-bandwidth, and multi-protocol networks for software-defined vehicles

The centralization of vehicle functions will drive substantial changes in the access approach and method used for communications. Vehicles will be divided into a number of zones, each with its own gateway. Zones will be defined by function, physical position, criticality, and safety. Sensors and



actuators will be connected to the nearest access points to transfer data to the backbone network and then to the central computing platform. This approach will reduce the total amount of wiring required, and support the development of new functions. Sensors will no longer be limited to a single function, and actuators will not be bound to a directly connected controller.

By 2030, multiple access protocols will be in concurrent use. Local Interconnect Network (LIN), Single Edge Nibble Transmission (SENT), and Peripheral Sensor Interface (PSI5), though slow, will still be used because of their cost advantages. But ultra-high-definition (UHD) cameras, 4D imaging radars, and high-resolution lidars will require much more bandwidth. According to Huawei, in-vehicle network transmission speed per link will exceed 100 Gbps by 2030. Vehicle Ethernet will become standard, and optical technologies will be widely deployed in vehicles because of their high bandwidth, light weight, insensitivity to electromagnetic interference, and low cost.

Conventional communications technologies are predominantly signal-oriented, using protocols such as Controller Area Network (CAN) and LIN. This approach deeply integrates communications with vehicle component deployment and routing, creating a problem: A change of the transmit/receive nodes will lead to changes of all nodes along the route.

Ethernet communications are service-oriented and can effectively address this problem. When the transmit/receive nodes are changed, no other nodes on the route will be changed. This will:

Decouple communications from vehicle component deployment and routing, making it easier to scale up.
- Make interfaces standardized and contractual.
- Achieve interconnectivity of in-vehicle services.

Once these technologies are realized, a point-to-point backbone network for software-defined vehicles can be created. The technologies that would be required for this include:

- **High bandwidth copper communications:** Signal attenuation is significant in copper cable over even short distances. Enhanced coding and algorithms will be required for intelligent power distribution and high-speed, high-bandwidth Ethernet transmission (10 Gbps to 25 Gbps). This will provide the high bandwidth required by in-vehicle applications for backbone networks.

- **In-vehicle fiber communications:** For bandwidths over 25 Gbps, copper is no longer an option, because of cost, engineering, and electromagnetic compatibility (EMC) challenges. This makes fiber an excellent solution. Fiber is cost-effective, light-weight, and is not affected by EMC issues. If solutions can be found for vehicle-related problems around temperature, vibration, and service life, optical fiber communications will be widely used in in-vehicle applications and support the evolution to higher-bandwidth communications.

- **Deterministic latency:** Real-time communications protocol stacks as well as time-sensitive network (TSN) protocol suites at the transport layer will need to ensure end-to-end deterministic latency at the microsecond level for vehicles. Transmission policies can be designed for specific service scenarios to meet the needs of different communications functions.

3.1.3 New wireless communication technologies for high-quality in-vehicle connectivity

By 2030, in-vehicle wireless communications will remove all barriers to connection within the vehicle. Any component will be able to connect using sliced wireless capabilities, so that new vehicle applications can call on them as needed. A new air interface will be required to deliver extremely low latencies of less than 20 microseconds for unidirectional transmission, five nines reliability, synchronization accuracy within one microsecond, up to hundreds of connections

and concurrent service provisioning, plus end-to-end cyber security. This is the level of quality required for vehicle connectivity. Service-specific resource management mechanisms will support in-vehicle wireless connections. Wireless slicing will make many things possible, like lossless audio streaming, UHD video apps, and ultra-low-latency interactive games. By taking the collaboration of multiple information domains to the next level, the interior of a vehicle will become an infotainment center offering immersive sound, video, images, and even light applications and sensations.

In-vehicle wireless communications technologies will transform the in-vehicle networking and enable simple upgrades of various vehicle modules. The use of wireless in place of wired connections will address design, production, assembly, and maintenance challenges created by vehicle wiring, and put an end to the highly-coupled architecture that wired connections create. In its place will be a platform + modular communications architecture, which can be replicated across many different models of vehicle. The flexibility of wireless communications allows for a range of different architectures, providing standardized wireless access interfaces. When vehicle-mounted devices become modular, standardized, and plug-and-play, the costs of vehicle development will fall, and smooth and ongoing evolution of the foundational platform will be supported.

With its extensive experience in wireless communications, Huawei will develop the next-generation wireless short-range communication solution to further improve the in-vehicle communications architecture and create greater value for customers.

3.1.4 Decoupled, service-oriented architectures for software-defined vehicles

Vehicles are now digital, intelligent products. User values, preferences, and needs no longer require vehicles to be a tool for transport; now, as with the phone, users want personalized experiences. Smart technologies, personalized features, and user experience are now the key factors guiding consumers' vehicle choices. At the same time, hardware and the associated technologies that go into vehicles are becoming less easily distinguishable, and carmakers are looking to

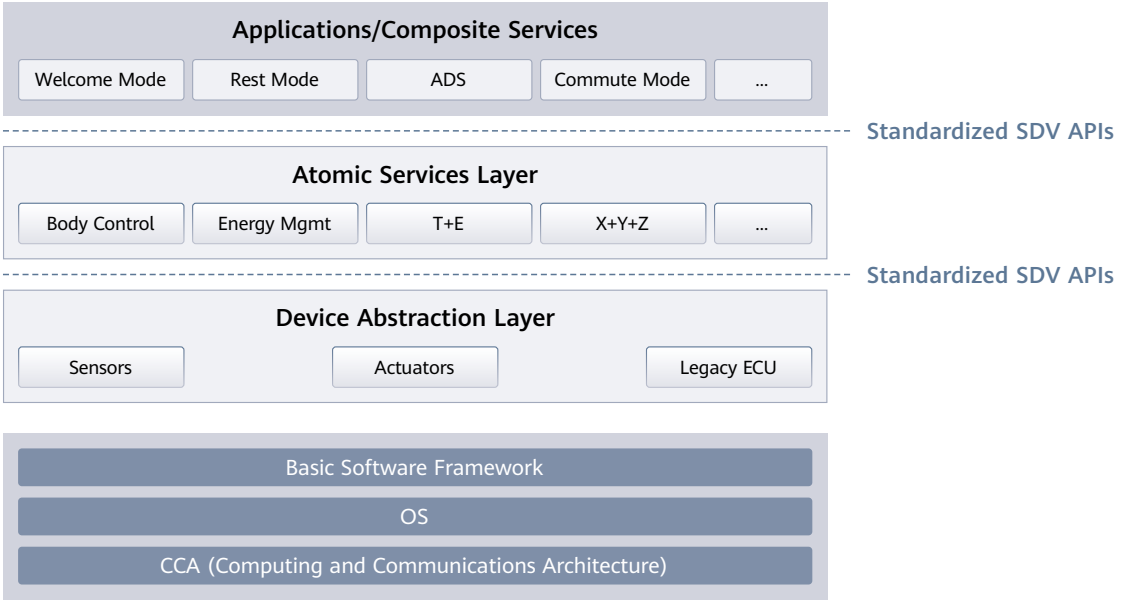
software and algorithms to create competitive advantages and deliver more value. All industry players are now pursuing software-defined vehicles.

Being software-defined means that software is a key feature in a vehicle's concept, development, testing, sales, and after-sales services. It also means that the entire process will be constantly refined, with refreshed experiences and continuous value creation. A key feature of these vehicles is the decoupling of the software from the hardware. In terms of physical delivery, that means that the hardware and software are delivered separately. In essence, vehicle functionality can be expanded, software can be replicable and upgraded, and hardware can be swapped out or scaled up.

OTA updates can keep software at peak performance; plug-and-play components can be freely added to expand functionality. Flexible, scalable software-defined vehicle platforms will be used to help intelligent vehicles meet the challenges of complex use cases and growing demands on functionality. A service-oriented architecture (SOA), with decoupled layers of software, has also been recognized as the best option for general software architecture. To realize this architecture, the system will need to add a device abstraction layer and a layer of atomic services.

Atomic services provide basic service capabilities, enabling upper-level applications to be replicable and portable across different vehicle models. The device abstraction layer normalizes the capabilities of underlying hardware, so that software can be decoupled from the hardware, enabling plug-and-play replacement and upgrades of hardware modules.

Services will be decoupled from system design to create basic service units. Each separate hardware component will be abstracted into a standardized service component, and each service component will provide one atomized function. These can be called on recursively and combined to produce complex functions, thus reusing the software as much as possible. Ecosystem partners can develop vehicle applications using platform components and standardized APIs that will then be managed by the vehicle platform. The platform will carry out app authentication, granting of access privileges, API call, security checks, and emergency management. Users will be able to choose vehicle apps the same way they do for their mobile phones. This will give users access to new vehicle experiences at very low cost: same vehicle, but a new journey every day! Developers, in turn, will benefit from consumers downloading their apps.



3.2 Intelligent driving: Making autonomous driving a commercial reality faster

Huawei uses sensor fusion technology to deliver superior safety in intelligent driving. Different types of sensors, including lidars, mmWave radars, ultrasonic radars, and cameras, are used to support the fusion and reconstruction of multi-dimensional environmental information. By continuously improving the sensing network, Huawei has eliminated the need for high-definition (HD) maps in vehicles. Intelligent driving, powered by autonomous sensing and navigation maps, is being widely adopted, and it enables vehicles to drive on any road with ease. The rapid development of AI is speeding up the deployment of end-to-end intelligent driving networks, which will soon be widely used in commercial settings. The continual and rapid iteration of these networks will ultimately enable a human-like driving experience even on roads with complex conditions. Looking ahead, Huawei's Global Industry Vision (GIV) predicts that by 2030, a staggering 90% of passenger vehicles in China will be equipped with L2 intelligent driving systems and 30% will be capable of L3 intelligent driving.

3.2.1 Continuous algorithm upgrades and digital materials for better user experience and autonomous driving

There are still many technical challenges that need to be overcome if we want to make autonomous driving a commercial reality. Given the complexity of corner cases in real-road conditions and how difficult it is to collect long-tail data, perception algorithms, planning and control algorithms, and simulation and training algorithms will be crucial for autonomous driving experience.

If we look at sensor fusion algorithms, there are many technologies that determine a vehicle's ability to perceive and understand its surroundings. These technologies include a vision-based perception framework, lidar point cloud generation and enhancement, lane-level traffic light processing in complex light environments, flashing and fuzzy light source processing, color processing, light source luminance differences, overlapping object recognition, and vehicle interaction prediction. Huawei's GIV predicts that by 2030, the algorithms will achieve a general object detection accuracy of over 99% which further enhances the vehicle's perception, and this goes hand in hand with their human-like understanding capabilities that enable them to understand complex traffic flows.

As perception and prediction both involve an element of uncertainty, the industry needs to further develop core planning and control algorithms that involve multi-object and multi-stage game-theory decision making, motion planning, human-like decision making and planning models focused on risky and complex interactions, and the identification and automatic labeling of key scenarios based on massive amounts of data. According to Huawei's GIV, these algorithms are expected to have an end-to-end response time of 400 milliseconds or less by 2030, which is twice as fast as that of humans.

Intelligent driving algorithms will evolve towards an end-to-end large-scale model which will enable the creation of a lossless information transmission channel that connects the sensing system to the planning and control system. By utilizing NN-based collaborative sensing, fusion, prediction, decision-making, and planning, intelligent vehicles will be able to drive just as well as humans, and ten times more safely. Huawei's GIV predicts that by 2030, this will increase the miles per intervention (MPI) to 621 miles (equivalent to 1000 km), reflecting the improved safety, comfort, and efficiency of intelligent driving.

Data is essential for rapid algorithm iteration, and virtual simulation systems offer a cost-effective and

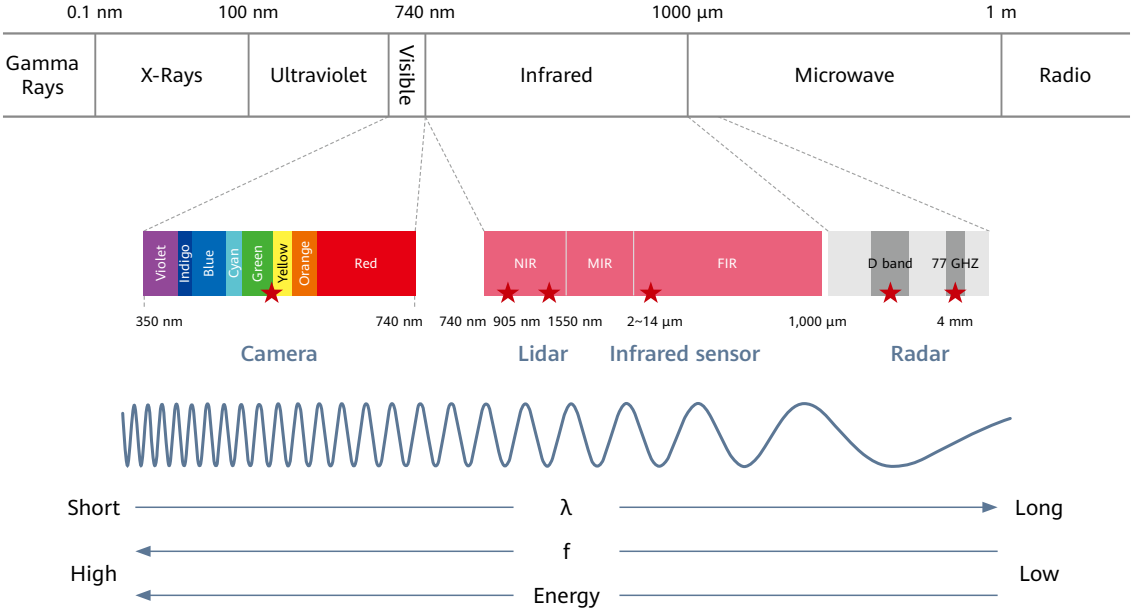
efficient way to obtain it. A comprehensive, high-fidelity simulation system can provide a constant supply of digital materials to facilitate the iteration of intelligent driving algorithms. Huawei's GIV predicts that by 2030, such a system will be capable of simulating a daily driving distance of 6.21 billion miles (equivalent to 10 billion km), and this requires a model for the various interactions between different traffic participants in many different large-scale simulation scenarios. With this simulation system, the end-to-end large-model algorithms can be continually improved, and this will ultimately improve autonomous driving capabilities.

3.2.2 Developing full-spectrum perception capabilities to make everything sense

As the automotive industry becomes more intelligent, perception systems become increasingly important. One day, they will be a cornerstone of intelligent driving. Ideally, the sensors that enable these systems will reach full coverage and cover all objects, all scenarios, and all weather conditions.

- **All objects:** People, vehicles, obstacles, road facilities, structures, etc.
- **Full coverage:** 360-degree coverage without dead zones.
- **All scenarios:** Highways, urban areas, traffic jams, accident scenes, and construction zones.
- **All weather conditions:** Day, night, rain, snow, fog, strong light, low light, and other harsh environments.

However, the industry still has a long way to go before it can make ideal sensors a reality. To make this happen, perception capabilities need to be built based on all sections of the spectrum.



1. Radars: The shift from the 77 GHz band to D band (110 GHz to 170 GHz) will significantly improve resolution.

Radar sensors can perform consistently in all weather conditions because they work on super-long wavelengths. These systems excel in measuring velocity, so they can create unique value in dynamic and static separation and simultaneous localization and mapping (SLAM). However, currently, their poor resolution is limiting their use scenarios.

Radar resolution can be improved by utilizing either ultra-high bandwidth or large-scale antenna arrays.

According to existing international standards, the 76 GHz–81 GHz frequency bands are already allocated to automotive radars. This means high-range resolution can be achieved through the higher 4 to 5 GHz spectral band. Angular resolution is determined by antenna arrays, which means the more antennas allocated for transmission and reception, the higher the angular resolution. Current radar systems still use three transmitting antennas and four receiving antennas (a 3T4R antenna array). Huawei recently improved on this by launching a 12T24R antenna array radar. However, the antenna arrays used in wireless communications have already reached 128T128R configurations.

Automotive radar sensors need to be physically small, which means the antennas used in wireless communications systems are unsuitable for automotive applications. These size restrictions and the 77 GHz wavelength mean that antenna arrays for radar sensors would be 48T48R to 64T64R, at maximum. The shift towards higher frequency bands will continue. The D band (110 GHz to 170 GHz) provides ultra-high bandwidth and has generally not yet been allocated or used for other services. The 140 GHz band is still being researched, but has a relatively suitable atmospheric window, so its propagation is less attenuated. What's more, wavelengths in this band are reduced by half. That means imaging radars that use an ultra-large 128T128R antenna array can be used in a limited space while still delivering high resolutions.

2. Lidar systems are moving from the 905 nm wavelength with time of flight (ToF) to the 1550 nm wavelength with frequency-modulated continuous waves (FMCW). Lidars are being integrated with chips, and high-performance 4D lidars will be widely adopted.

Thanks to their relatively mature components, 905 nm lidars have already been widely adopted and are ready for mass production. From a technical perspective, as the industry moves from analog to digital and from discrete to integrated, the transmitting and receiving components are being arranged in arrays, and core modules will be integrated directly into chips.

These trends mean high-performance, low-cost, highly integrated, and highly reliable lidars may be the way forward:

- Transmitters will go from discrete edge emitting laser (EEL) components to EEL arrays, and finally to large vertical-cavity surface emitting laser (VCSEL) arrays.
- Receivers will shift from avalanche photodiodes (APDs) to single-photon avalanche diodes (SPADs), and finally to SPAD arrays. This will help improve receiver sensitivity and support long-range and high-precision depth measurements.
- Scanners will move from mechanical scanning to micro rotating mirrors and finally to being made up exclusively of solid-state components.

It is worth mentioning that the transition from APDs to SPADs is not just about performance improvement; more importantly, it marks the shift from analog to digital. SPAD is a pixel structure, so SPAD chips using complementary metal-oxide-semiconductor (CMOS) technology can evolve into photosensitive chips like the ones used in cameras. According to Moore's law, pixel size will continue to increase, possibly even to tens of millions of pixels, to support higher-resolution lidars.

Most light at 1550 nm, a near-infrared band, is blocked by human cornea before it reaches the human retina, which means it does not cause much damage to human eyes. Because of this, the 1550 nm band allows lidars to transmit at higher powers, which can greatly increase coverage. In terms of modulation, FMCW's use for radars can also be applied to lidars. FMCW lidars deliver better performance through high-performance 4D imaging (which can be used to measure velocity), strong anti-jamming capabilities, and higher sensitivity and dynamic range. In addition, FMCW lidars can be mass produced at lower costs, when combined with silicon photonic and optical phased array (OPA) technologies.

However, FMCW technology at the 1550 nm wavelength is far from ready for commercial use and will require concerted efforts from across the entire value chain to develop further. Further exploration of silicon photonic technology in line with Moore's law is one way that can help support FMCW. More complex and discrete optical functions can also be integrated into silicon-based chips, making lidar sensors more integrated, more affordable, and smaller.

3. Cameras will integrate visible light and infrared thermal imaging technologies to work in all weather conditions.

Cameras are a type of passive sensors and are most similar to the human eye. They can sense surrounding objects through catoptric imaging. As one of the three major types of sensors used in vehicles, cameras are the most critical for identifying elements in a static environment, like traffic lights and road signs.

However, cameras also have drawbacks. The performance and confidence of catoptric imaging suffers at night and in low-light conditions, and heavy rain and snow can impede a camera's line of sight, greatly reducing its visibility. Cameras cannot independently overcome these harsh weather conditions.

However, in the infrared spectrum (in the 2–14 μm band) right next to the visible spectrum, a thermal emission imaging system can be used. Sensors that work on this band have effective night vision, and can detect objects through rain, snow, sand storms, and fog. They even have certain perspectives to further meet the requirements of all weather conditions. Vehicles are now starting to come equipped with infrared thermal imagers that have night vision, but this application still needs a low-cost solution for mass production.

3.2.3 Full convergence: Accelerating innovation in sensor form factors for simplified deployment

As vehicles become more intelligent, the number of sensors per vehicle is increasing dramatically. Vehicles come equipped with many sensors, from 1V and 1R1V to 5R5V and 6R13V3L configurations, and more will be deployed in the future. Here, R refers to radar, V means vision (i.e., camera), and L means lidar.

However, a vehicle's body is limited in size, and the installation and deployment of sensors raise even high requirements for body design, such as strict requirements on fascia, thickness, installation intervals, and flatness. This makes the entire style and design process much more difficult, because designers must balance vehicle appearance and sensor performance.

To make sensors easier to deploy in vehicles, innovation in form factors is a must. Miniaturization of sensors will be the way forward. In addition, sensors will need to be designed to better fit vehicle styles. Integrated designs that consider both sensors and vehicle style can greatly reduce the constraints on the vehicle body. This requires joint efforts from multiple sectors including materials, processes, and engineering.

1. Distributed antennas and central computing

Radars generally come in an integrated design that encapsulates front-end antennas and back-end signal and perception processing units into one box, to support point cloud generation, object detection, and perception processing. As central computing becomes more common, radar signals can be segmented and extended using the techniques already utilized in Huawei's distributed base stations. Conventional monolithic radars can be used only to generate point clouds, while perception processing units can be integrated with domain controllers. Radars that use distributed antennas and central computing deliver better performance, consume less power, and occupy less space than current radars. Solid-state lidars can also be deployed in a similarly distributed manner.

2. Integration into the vehicle body

An alternative solution for separate deployment

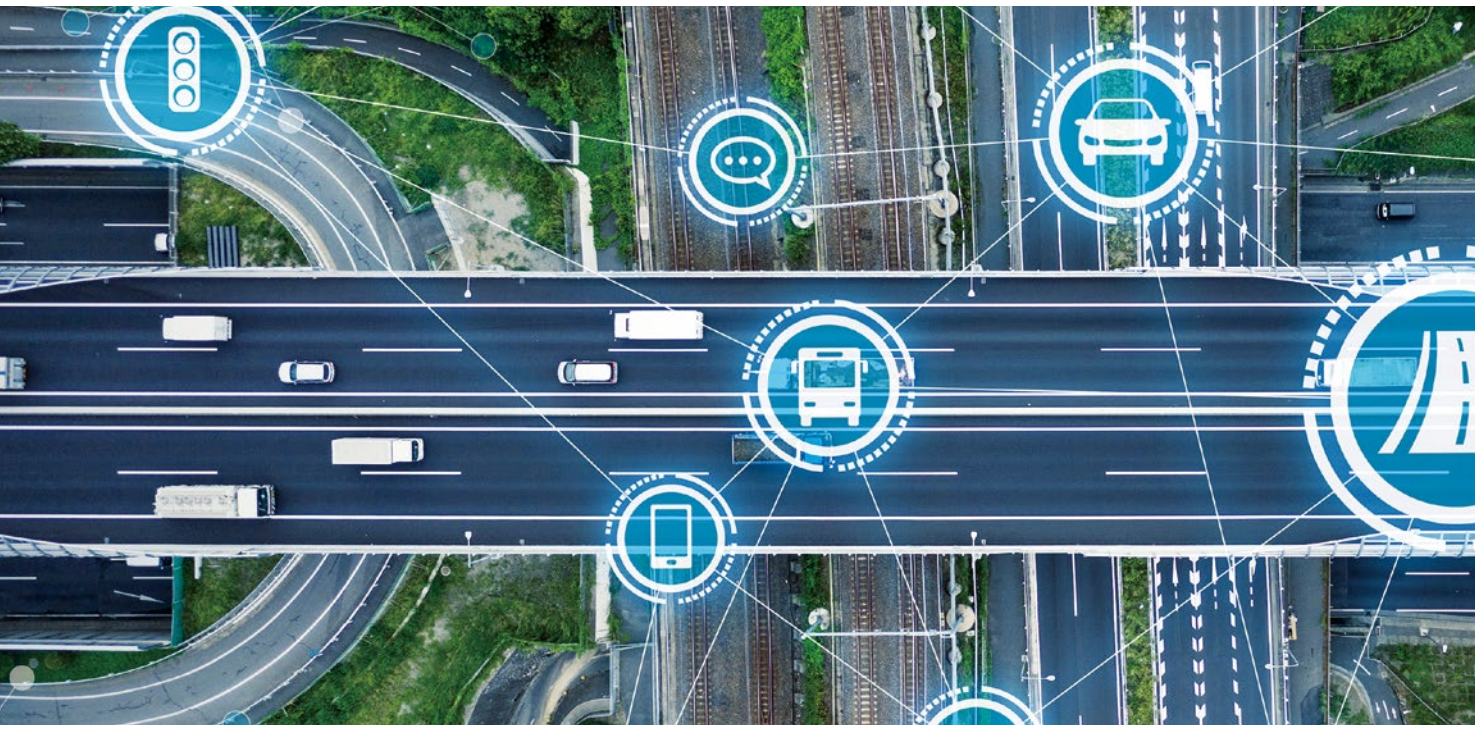
is integrating sensors with existing components into vehicle bodies. Shark fin antennas already support GPS, 4G/5G, and frequency modulation. Surround-view cameras are also already integrated into rearview mirrors. Similar solutions can be implemented in the future, such as installing lidars into front-view headlights, integrating distributed antennas into glass components and doors, and combining far-infrared sensors with existing cameras. Such combinations can make sensors more adaptable and easier to deploy, but they also raise the bar for sensors. More effort is needed to address heat dissipation, interference, EMC, and other related issues that may arise from such convergence.

Sensors can also be integrated with each other. Low-cost distributed inertial measurement units (IMUs) can be physically integrated with sensors, which can help change sensors' motion compensation from inter-frame compensation into intra-frame (signal-level) compensation and improve the accuracy of sensor attitude perception. This model can help further improve sensors' overall performance and safety in scenarios like vibration, dead reckoning, and slopes.

3. Surface-mount sensors will reshape sensor deployment.

Surface-mount sensors are the ultimate vision for future sensor deployments. Such sensors would need to be smaller and flatter, and eventually plug-and-play. Highly integrated chips would be necessary for this solution, as well as:

- **Microlens array technology:** An assembly of precision-manufactured microlens, just a few millimeters deep and high, used to project a tightly-focused beam. This can massively reduce the focal length between light sources and lenses, making flat designs possible. Though this technology is now mainly used in projectors, it also provides a new possible path for miniaturized lidars and surface-mount sensors.
- **Smart Skin (conformal antenna) technology:** An antenna array designed to conform to the shape of the carrier. This allows an antenna array to be directly attached to the surface of a carrier like a Band-Aid, so that the antenna array can be integrated with the platform structure. This makes sensors more adaptable,



and adds more flexibility in vehicle style design.

In the future, sensors will be like a layer of skin attached to the outer surface of intelligent vehicles. To make this vision a reality, all players along the value chain need to work together to advance in multiple fields, including materials, processes, and engineering.

3.2.4 Central computing platforms: Providing large computing power to support intelligent driving

Powerful computing platforms will provide the fundamental computing power needed for intelligent driving. More sensors, both in type and number (over 50), will be deployed in a single vehicle to enable intelligent driving in all scenarios in complex road conditions. This includes 100-million-pixel cameras, event cameras, 4D imaging radars, high-resolution lidars, ultrasonic radars, infrared detectors, and sound detectors. In addition, all of these sensors will be increasingly accurate. These sensors will generate massive amounts of data that needs to be analyzed and processed in real time.

Huawei's GIV predicts that by 2030, a vehicle will come equipped with a computing power of over 5,000 TOPS and 3 million DMIPS with less than 150 W of power thanks to improved chip processes. When these advancements are combined with improvements in memory technology, they will further boost energy efficiency ratio of new vehicles. These central computing platforms will provide the computing power needed to enable intelligent driving.

3.2.5 V2X cloud brain for multi-vehicle collaboration and intelligent driving

Connected infrastructure is continuing to improve, and intelligent driving is already being adopted in more scenarios. The challenge we face is no longer just making single vehicles intelligent. Now, we want to make vehicles that cooperate with each other. This will push commercial intelligent driving and intelligent transportation to the next level.

Ubiquitous (vehicle-to-everything) V2X connections need to be built to intertwine people, vehicles, and road infrastructure. A vehicle-road intelligent cooperative driving platform needs to be established on the cloud to streamline end-to-end application scenarios. Thankfully, multi-vehicle cooperative intelligent driving will quickly become a commercial reality, thanks to services like comprehensive environment perception, global resource scheduling, dynamic service mapping, multi-vehicle cooperative driving, lane-level route planning, coordinated signal control, and service simulation testing.

A cloud-based brain can integrate all the necessary information elements of people, vehicles, roads, and environments, improving vehicles' ability to perceive dynamic traffic environments. A cloud-based brain can also share driving strategies between different vehicles, allowing vehicles and traffic infrastructure (e.g., traffic lights and signs) to work in synergy. This allows us to optimize overall (not partial) driving strategies and further promote the development of intelligent transportation systems.

3.3 Intelligent cockpits: AI speeds up software and hardware upgrade

3.3.1 Open cockpit OSs for an ever-evolving application ecosystem with brand-new experiences

Compared with smartphones and other consumer devices, an intelligent cockpit has more peripherals and supports multiple users, multiple concurrent operations, and multiple modes of interaction. The design and development process of OSs for intelligent cockpits must consider these factors. In addition, an ever-evolving application ecosystem is vital for cockpits to bring brand-new experiences to users. This raises the bar for the consistency and stability of a cockpit OS application interfaces.

At present, the cockpit OS market is plagued by the fragmentation and customization of solutions across the industry. For example, when a carmaker is developing a function that needs to work with a camera or microphone, they have to create unique versions of the function for different vehicle models because each hardware platform is different. This fragile approach to software development prevents different hardware components from effectively connecting with each other or sharing software capabilities.

There is yet another challenge for carmakers in software development. The work is outsourced to multiple vendors, with each being responsible for different functions, and redundant copies of the same function may co-exist. This creates chaos in version management, leads to a lot of extra development work, and complicates software upgrade and maintenance.

In the future, there could be three primary OSs in the intelligent cockpit field with unified APIs to support a thriving and ever-evolving application ecosystem. Developers will be able to quickly develop applications to improve the cockpit ecosystem by utilizing the basic platform-based capabilities that APIs provide.

3.3.2 Platform-based interaction algorithms, quickly improving cockpit application development efficiency

Integrating the fundamental algorithms for human-machine interactions in cockpits into a single OS can help ecosystem partners improve development efficiency. In future intelligent cockpits, technologies that support primary interaction capabilities — such as voice, visual, and audio — will remain essential for human-machine interactions. Specifically:

Voice: When it comes to experience, voice will be the most important interaction capability of intelligent cockpits. As foundation models and computing power continue to improve, voice capabilities will advance to allow for more human-like interactions , intelligent services, and personalized experiences.

(1) Human-like voice interactions: With voice technology, cockpits can recognize emotions and engage in emotional interactions. The system can utilize visual and voiceprint recognition technologies to quickly detect users' emotional state and provide emotional support through human-like tone of voice, emotional responses, and interactive UI images. In this way, voice interaction truly transforms from human-machine interaction to human-human interaction.
(2) Intelligent services: Generalization is the key for voice technology to understand more. As foundation models become widely implemented, developing voice technology that understands more and provides more services requires the integration of a variety of ecosystem applications. Cockpits need extensive ecosystem applications and accessible interfaces that can handle vague or complex commands like "navigate to the nearest fast-charging station" or "play the most recently watched episode of Peppa Pig."

These capabilities are essential for cockpits to provide intelligent services.

(3) Personalized experience: In the future, in-vehicle voice technology will increasingly rely on multimodal fusion. Integration of visual and voiceprint recognition capabilities allows the voice assistant to accurately identify each occupant and their preferences, providing personalized services.

Visual: The second key way that vehicles interact with users will be through vision. Currently, in-vehicle visual recognition technology is mainly used for the driver monitoring system, cockpit monitoring system, and detection and recognition of human-machine interactions (e.g., through gestures). In the future, more advanced visual recognition technologies will emerge to deliver functions such as detecting living beings in the vehicle, monitoring users' health, and enabling secure payment, entertainment, and integrated audio-video services.

Users will be able to interact with their vehicles in new ways, with vehicles providing more precise and convenient services. For instance, eye tracking technology and AR-HUD will work together to identify objects within eyesight in real time, and then project information about the objects, such as their details and relevant ads. When vision-based gesture recognition technology is used in conjunction with mmWave radars, the accuracy of gesture recognition will be improved and users will be able to smoothly interact with their devices. When there is a lot of background noise, integrated visual-audio technology will read the user's lips and translate lip movements into commands, supporting speech recognition and vehicle control across all scenarios.

Audio: Cockpits of the future will possess more advanced audio capabilities in terms of smart sound effect, smart sound field, and active noise cancelation.

(1) Smart sound effect: Software algorithms automatically separate sound elements and add angles and trajectories to create a 3D surround sound effect, even for 2D audio sources.

(2) Smart sound field: Intelligent processing of audio signals allows occupants to listen to personalized content thanks to independent sound zones while experiencing immersive multi-channel surround sound.

(3) Active noise cancelation: This will continue to be a major area of focus for in-vehicle audio technology over the next ten years. More advanced hardware and algorithms will be needed to thoroughly cancel out the noise generated by the vehicle engine, road traffic, and wind, so as to continuously improve the ride comfort.

In the next 10 years, breakthroughs in components, algorithms, and architecture, along with improvements in high-performance computing chips and digital signal processors (DSPs), will take in-vehicle audio to new levels, transforming vehicles into mobile entertainment hubs.

3.3.3 Multi-device collaboration through distributed technology for seamless intelligent experiences

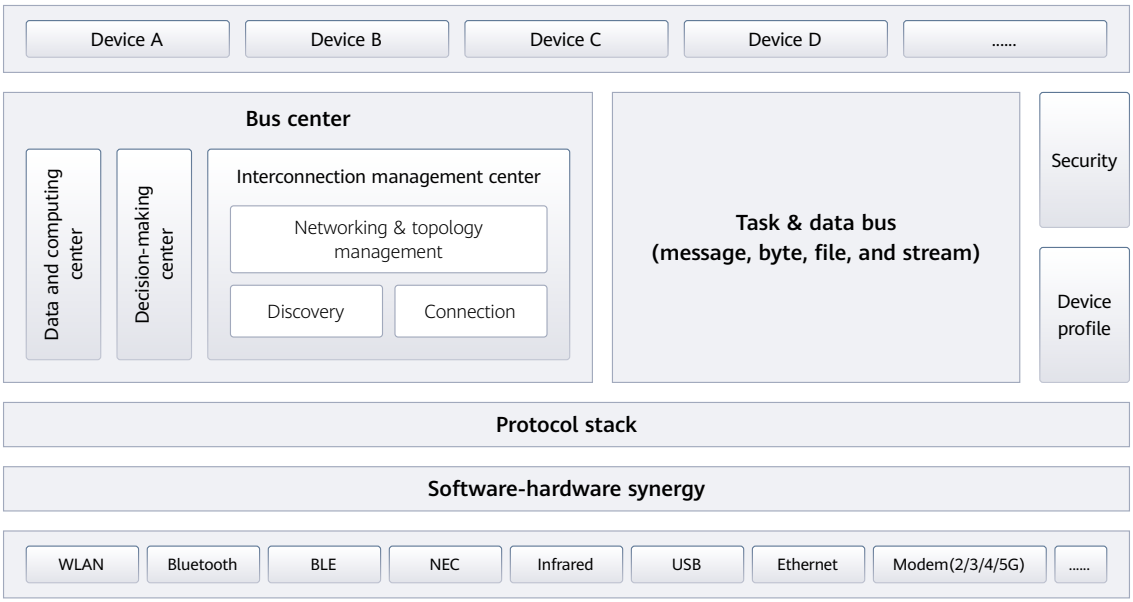
Intelligent vehicles are integrated systems, and how they interact with users will involve the broader surroundings. Connection and interaction of devices depends on both general-purpose cloud technology and a distributed software bus. Huawei's HarmonyOS for Automotive provides a distributed software bus to create a seamless experience across nearby devices, making it easier and more comfortable for users to interact with their devices.

Imperceptible sensing and zero-wait transmission are the deciding factors for realizing seamless experience across nearby devices. If we are to meet these preconditions, we must first devote efforts to answering the questions of how different devices should discover and connect with each other, how connected devices should come together to form a network, and how transmission between devices that use different protocols can be realized. The key technologies here include automated device

discovery, connection, networking (e.g., multi-hop automated networking and multi-protocol hybrid networking), and transmission (e.g., diverse protocols and algorithms, and intelligent perception and decision making).

A distributed software bus (DSoftBus) connects different types of devices by using a "protocol shelf" and a software-hardware collaboration layer, despite the different protocols used by the devices. The hub module of the bus analyzes commands to discover and connect devices. The task bus and data bus provide other functions, such as transferring files and messages between devices.

The following figure shows the HarmonyOS DSoftBus architecture.



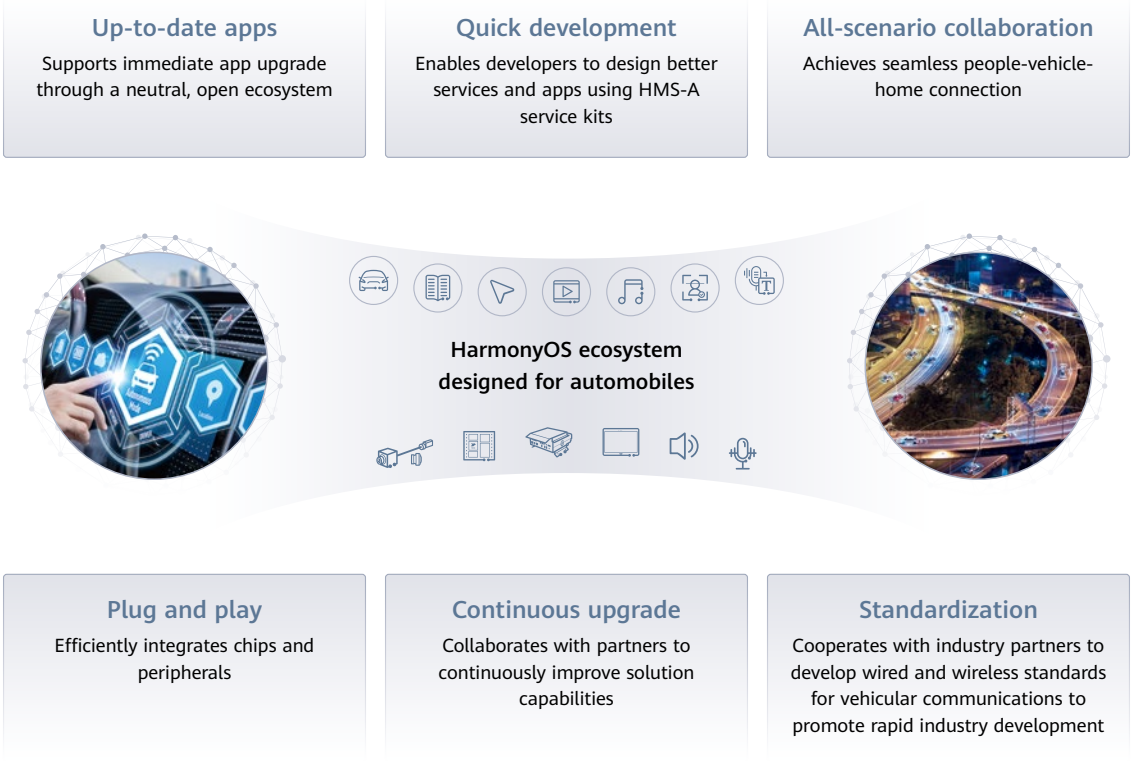
Several preconditions have to be met before an intelligent vehicle and IoT devices can work synchronously to offer an interactive experience. In terms of design logic, it is important that interactive experience of the intelligent cockpit be consistent with that of mobile phones or other such devices. Regarding operating logic, intelligent cockpit applications must provide the unified functions of smartphone applications, and may be designed based on the hardware capabilities of cockpit peripherals. In addition, users want a seamless experience when they switch between cockpit and smartphone applications, especially when it comes to their calendar, navigation, music, video, and conferencing applications.

3.3.4 Standardized interfaces keep user experiences fresh throughout hardware lifecycles

Consumer devices like smartphones usually last two to three years, and their software and hardware integration packages are small. Vehicles have a longer lifespan, including a sales window of 5 to 10 years and a useful life of 10 to 15 years. Carmakers tend to research and develop multiple vehicle models at the same time, so they have no choice but to simultaneously maintain a whole range of software and hardware versions.

As novel applications appear, hardware performance needs to keep up. For example, chips and other key components like cameras and displays will have to stay up-to-date throughout a vehicle's lifecycle. Cockpit hardware upgrading will lead to new business models in the post-sales stage.

Open and Specialized Hardware and Software Streamline New Functions in Vehicles



Inside a vehicle, chips or chip modules must be able to meet the computing needs of software and hardware for the next three to five years. The chip module itself needs to be designed for backward and forward compatibility, so as to ensure easy upgrade (e.g., through pluggable components) and strike a balance between hardware lifetime and computing demand. Hardware plug-and-play is essential for this type of chip module.

When a key peripheral is replaced by a new one, it is necessary to install a new driver to support the new peripheral. Carmakers should aim to make this process as simple as installing a new driver on a Windows operating system. Also, it should allow OTA updates for certain parts of the cockpit OS. To make this possible, unified standards should be established for the interfaces of cockpit hardware, in order to eliminate issues caused by customized interfaces of head units, cameras, displays, HUDs, intelligent seats, intelligent steering wheels, in-vehicle robots, intelligent windows, holographic projection hardware, etc. To standardize cockpit hardware, the automotive industry needs to double down on standardizing communications interfaces, including those for short-distance wireless communications, wired communications, video, and audio. Standardization is key to driving down component costs and cultivating a hardware ecosystem.

3.3.5 Up-to-date OS and diverse ecosystems enable a sophisticated and engaging cockpit experience

Intelligent cockpits are driving the need for diverse user experiences, requiring collaboration from various ecosystem partners. However, adapting applications for each OS proves costly due to the relatively small number of vehicles compared to smartphones. This is why mainstream car models often have a limited selection of applications, leading to usability difficulties and low adoption rates. To address these challenges, cockpit vendors require an integrated OS platform and ecosystem that is supported by multiple carmakers and

models, empowering application developers to create value and attract more users. Additionally, cockpit vendors should provide ecosystem partners with a streamlined development platform that requires minimal input while delivering substantial output, ultimately offering users more sophisticated experiences.

A unified OS is essential for cockpits, as it can foster a cohesive user experience while enabling seamless cross-model upgrades that guarantee access to cutting-edge system capabilities. In addition, open APIs are integral for empowering application developers to create applications that can be deployed across multiple devices with minimal additional effort. By minimizing the complexities associated with diverse hardware configurations, transmission protocols, OSs, and functional modules, a unified OS enables exceptional experiences to be swiftly replicated across various vehicles.

Today, HarmonyOS and its ecosystem applications have been deployed for 23 vehicle models, providing over 300,000 users with diverse experiences through nearly 150 apps across 12 categories. Huawei works closely with content providers to enhance user experience and explore new business opportunities in navigation, music, video, gaming, themes, and charging. Notably, the company partners with game content providers to design "steering wheel racing games," audio-visual content providers to create "cockpit concerts," and theme content providers to develop "customizable themes." All of these help to deliver unique experiences to users. These innovative applications can be developed for a single vehicle and upgraded on many, receiving iterative updates every quarter on the HarmonyOS platform. This approach fosters an ongoing cycle of innovation and improvement, providing users with a fresh and up-to-date experience.

3.4 In-vehicle optical applications: Lighting up a new vision for drivers and passengers

3.4.1 Viewing: Panoramic, immersive holograms for eye-opening experiences

Humanity's desire for superior visual experiences knows no limits. As vehicles become more intelligent, their front windshields, side windows, and panoramic sunroofs are quickly becoming displays that can show information through lifelike holograms. As intelligent laser and pixel technologies continue to evolve, they are expanding the roles of headlights, from being mere sources of lighting to projecting 3D information in all directions around the vehicle.

In-vehicle optical applications are designed to create superior visual experiences as they support information display, interaction, and entertainment. In terms of navigation, the windshield can be used to enhance driving safety by displaying essential information and safety warnings like road directions and obstacles. The windshield and even rear side windows can be used for entertainment, serving as holographic screens that offer the kind of immersive 2K/4K viewing experiences that you get at the cinema. Curved panoramic sunroofs can project customized light patterns, to mimic everything from meteor showers and constellations to deep-sea coral reefs. Going forward, intelligent vehicles will also be equipped with headlights capable of wide-gamut, high-pixel projection that will allow users to project and watch movies outdoors.

Vehicles are quickly becoming the third living space for people after their home and workplace. This is why user demand for visual experiences on the go has been increasing. Users expect immersive experiences that deliver images and video with higher resolution and broader view. Users are also looking for new eye-friendly technologies that

can help with carsickness. This means interactive functions not only have to create truly immersive experiences, but also help users avoid getting carsick while making long video calls or watching movies. In addition, rear-seat passengers expect optical display technologies to deliver a wide array of entertainment functions without fatiguing their eyes.

Looking to the future, spatial optical technologies – in tandem with human-focused experiences – will reproduce the real world with extremely high-resolution images that are sharper than even the human eye can process. This will require:

- **Wide-view, immersive technology:** Spatial optical technologies like freeform mirrors, diffractive optical waveguides, and polarization beam splitting can be used to project images up to 100 inches in size from a 10-inch display. With directed light field technology, users can watch 3D movies on in-vehicle displays without 3D glasses. Also, directed sound field technology offers amazing acoustics that would previously have been available only to best seats in the cinema.

- **True-color UHD displays:** Displays will soon come fitted with optical engines for 2K, 4K, and 8K video, and diffuser film displays based on a micro-nano structure. These UHD displays will significantly enhance pixels and brightness, making the text and images they display so crisp that the pixels will not be visible to the naked eye. With RGB lasers, UHD displays will support DCI-P3 color space or even BT.2020 color space for 8K video, perfectly showcasing the true colors of objects being displayed.

- **Visual health:** Virtual imaging systems can display images at a 3-meter distance from the viewer's eyes, thus eliminating the risk of myopia. Passive cool tint technologies also make it possible for displays to emit zero radiation and reduce the amount of blue light that reaches the eye.

- **Human-focused experiences:** Carsickness is caused by conflicting information the human body receives from eyes and ears when a vehicle is in motion. Staring at an on-board screen in a moving vehicle often exacerbates this problem. When engineering technologies that focus on dynamic human factors are incorporated into on-board screens, this problem can be minimized or avoided completely. Eye fatigue occurs when the eyes' ciliary muscles contract too tightly. The technologies that enable eye tracking and diffuser film displays that automatically adjust the distance between your eyes and the images being projected can all help relax ciliary muscles.

3.4.2 Connecting: New interaction methods ensure better driving safety and stronger emotional bonds

In-vehicle optical applications provide new ways for vehicles to interact with their users and the world around them. These applications are also crucial for driving safety. Inside vehicles, augmented-reality head-up displays (AR-HUDs) are an intuitive tool for vehicle-driver interaction. They can directly display information on the windshield, enabling the driver to more easily view information, instead of having to look down at various instruments. The AR-HUDs can display real-time information, such as AR navigation directions, alerts for obstacles, vision augmentation in rain, fog, and darkness, as well as information about nearby gas stations and other services.

Intelligent lighting systems also provide a new way for vehicles to interact with the outside world beyond just their horn and signal lights. When a vehicle is in motion, intelligent lighting can project interactive information onto the road, such as vehicle width, alerts for rain and fog, and night vision, to help drivers make more informed decisions and enhance safety. In addition, intelligent lights can project useful information for pedestrians, such as turn and right-of-way signals. Intelligent lights are also capable of projecting emotions, showing customized information such as patterns, emojis, texts, and weather data, and can even enable other forms of interaction through light shows and concerts.

In the future, a variety of in-vehicle optical applications will create even more ways for intelligent vehicles to interact with people. This can happen through:

- **HUDs:** Currently, AR-HUDs use megapixel-level optical modulation engines and spatial optical technology, but future adoption of dual-focal plane technologies will pave the way for even more advanced multi-layer AR-HUDs that will be able to project dashboard information two to three meters in front of the driver, and navigation and other useful information seven to ten meters in front of the driver. Future naked-eye 3D technology will also further improve the interactivity and experience of HUDs.

- **Lights:** With megapixel-level modules and optical lens, automotive lights will be transformed into projectors, displaying information such as vehicle-to-vehicle distance alerts and animated greetings. Vehicles that use precision laser lighting and sensing technologies will be able to interact with the environment through methods like dynamic ground projection (dubbed "dynamic light carpets"), to illuminate the surrounding area outside the vehicle and provide centimeter-level precision lighting. This will undoubtedly make driving safer and more fun. In the future, current/voltage modulation will also make it possible to display information in beams, and visible light communications technology will be able to support vehicle-to-vehicle communications.

- **Windows as displays:** Ultraviolet light projectors and fluorescent film glass will turn vehicle windows into colorful, full-size displays where notifications of the driver's intent, ads, and other types of information can be shown to pedestrians.

3.5 Intelligent vehicle cloud services: Providing ultimate experiences and attentive services throughout the lifecycle of intelligent vehicles

As vehicles become increasingly intelligent and connected, data has transformed from a mere utility into a vital asset and key differentiator in the automotive industry. In response to this shift, carmakers must adapt to meet the growing demand for diverse intelligent applications, foundational models, efficient intelligent vehicle services spanning the entire product lifecycle, and algorithm training for advanced driving systems and cockpits. This has paved the way for data-driven intelligent applications and comprehensive lifecycle services to emerge as the lifeblood of the automotive sector, guiding the development of intelligent connected vehicle cloud platforms.

The intelligent connected vehicle cloud platform is a comprehensive service platform that seamlessly integrates cloud computing, big data, Internet of Things (IoT), and AI technologies. It provides efficient, secure, and intelligent data and application services for intelligent connected vehicles, delivering an unparalleled experience of interconnectivity between the vehicle, cloud, and mobile device. Furthermore, by offering efficient and attentive services throughout the entire lifecycle of the vehicle, it caters to diverse user needs and preferences.

As the automotive sector undergoes rapid transformation, traditional carmakers' Telematics Service Platform (TSP) is being completely transformed, evolving into a sophisticated, intelligent, and all-encompassing cloud service platform tailored to next-generation connected vehicles. Early TSP platforms offered basic functionalities like vehicle access management, location tracking, and emergency response systems. But these platforms were hampered by

technological limitations, restricting their scope and capabilities. And with the widespread adoption of cutting-edge technologies such as cloud computing, big data, and AI, coupled with the integration of intelligent components into vehicles, carmakers are compelled to proactively explore innovative ways to craft more intuitive and personalized service experiences. As a result, traditional TSP platforms are being replaced with modern intelligent connected vehicle cloud platforms, which not only elevate user satisfaction but also significantly enhance carmakers' offerings in terms of intelligent services. This has helped to inject fresh dynamism into the entire automotive industry.

3.5.1 Proactive services for vehicles throughout their entire lifecycle with innovative big data and intelligent applications

The advent of intelligent connected vehicles has ushered in a new era of unprecedented challenges and opportunities for the automotive industry. While traditional component failures can be readily identified by using specific fault codes together with local diagnostic equipment, and then resolved through replacement or repair, the complexities inherent in intelligent vehicle components present a daunting task.

With a high degree of hardware concentration and a substantial software proportion, these advanced components exhibit significantly increased functional complexity. This makes pinpointing and diagnosing intelligent component faults an arduous process, which is made worse by the increasing number of experience-related problems that arise from the growing variety and complexity of automotive applications. This creates significant challenges for after-sales service of intelligent vehicles.

Nevertheless, the deep integration of ICT technology and the automotive industry has yielded substantial benefits through provisioning efficient services for intelligent connected vehicles. By establishing a comprehensive vehicle data application service system in the cloud, carmakers can consolidate data from various vehicle components to enable functionalities such as real-time vehicle status monitoring, alarm management, fault warning, remote diagnosis, and predictive maintenance. This not only enhances service quality and operational efficiency but also gives their services a competitive edge,

ultimately contributing to increased market share and improved brand reputation.

To fully integrate intelligent components with cloud-based technologies like big data and AI, several key considerations must be taken into account within the cloud environment. These include collecting more maintainable data from intelligent components and collecting and aggregating signals from automotive components more efficiently. Furthermore, incorporating service logic, industry expertise, and industry practices is essential for driving innovation in intelligent applications. In particular, the following specific areas require attention:

(1) Efficient collection and aggregation of vehicle data and signals

By leveraging open and standardized V2X technology, it is possible to access a massive number of vehicles and provide concurrent services to millions of users through the cloud. A data pipeline should be built to support efficient collection and aggregation of vehicle data, signals, and other relevant information from various automotive components, such as the advanced driving system, cockpit, vehicle control module, T-Box, powertrain, and body.

(2) Vehicle data analysis and processing

It is necessary to leverage vehicle data to develop various services, including vehicle status inquiry, data dashboard, fault management, alarm management, problem management, signal analysis, log management, data analysis, and more. These features will facilitate user-centric services, risk mitigation, fault location, remote diagnosis, predictive maintenance, and other critical operations.

(3) Efficient remote fault diagnosis and warning model

A cross-domain diagnostic and analytical model should be constructed for fault scenarios, utilizing vehicle signals and data in conjunction with industry knowledge and experience. In this way, a fault diagnosis model can be built, facilitating rapid identification and resolution of complex problems and faults across domains, ultimately benefiting both carmakers and consumers.

(4) Specialized large model

Proactively exploring the potential application of large models within the automotive sector is a worthwhile endeavor. Currently, Huawei's intelligent vehicle cloud service stands at the forefront of intelligent diagnostics, underpinning the Yunque model — an L2 large model tailored specifically for the automotive industry. Built on top of a foundational large model with 10 billion parameters, the Yunque model integrates cumulative fault diagnosis data and expertise alongside an industry-specific knowledge base to facilitate automated fault diagnosis. Through question-and-answer interactions, the Yunque model can identify faults or problem descriptions, perform intelligent triage, and formulate diagnostic solutions. At the end of the automated diagnosis process, it generates conclusive findings and comprehensive diagnosis reports. This not only enhances the efficiency of remote vehicle problem diagnosis but also paves the way for more sophisticated services and applications in the future, ultimately elevating user experience.

3.5.2 Accelerating application and business model innovation for an ultimate device-cloud connection experience

In today's interconnected world, the device-vehicle-cloud synergy has given rise to an array of innovative experiences, such as digital keyless entry, smartphone-based car finding, scheduled charging, one-touch vehicle preparation, and remote parking. And as the automotive industry continues to evolve, we can expect to see even more sophisticated applications to emerge. Such applications will leverage the power of intelligent vehicles to create new possibilities for on-board video conferencing, social networking, real-time environmental awareness, live streaming of scenic views, smart route planning, smart

home automation, e-commerce and logistics management, seamless multimedia entertainment, and personal health monitoring.

The intelligent vehicle cloud plays a pivotal role in enabling carmakers and developers to create superior device-cloud interconnectivity experiences from the following three aspects:

(1) Ensuring a seamless service experience of V2X applications

Underlying IoT communication technologies and optimized algorithms enhance Quality of Service (QoS), reduce end-to-end latency, and ensure the reliability and efficiency of operations, enabling seamless interactions among mobile devices, IoT terminals, vehicles, and the cloud.

(2) Optimizing service experiences with data-driven closed-loop capability

Carmakers can analyze user feedback on service experience using cloud service data. By delving into the service logic segment by segment, they can identify areas for improvement and implement targeted optimizations to enhance overall user satisfaction.

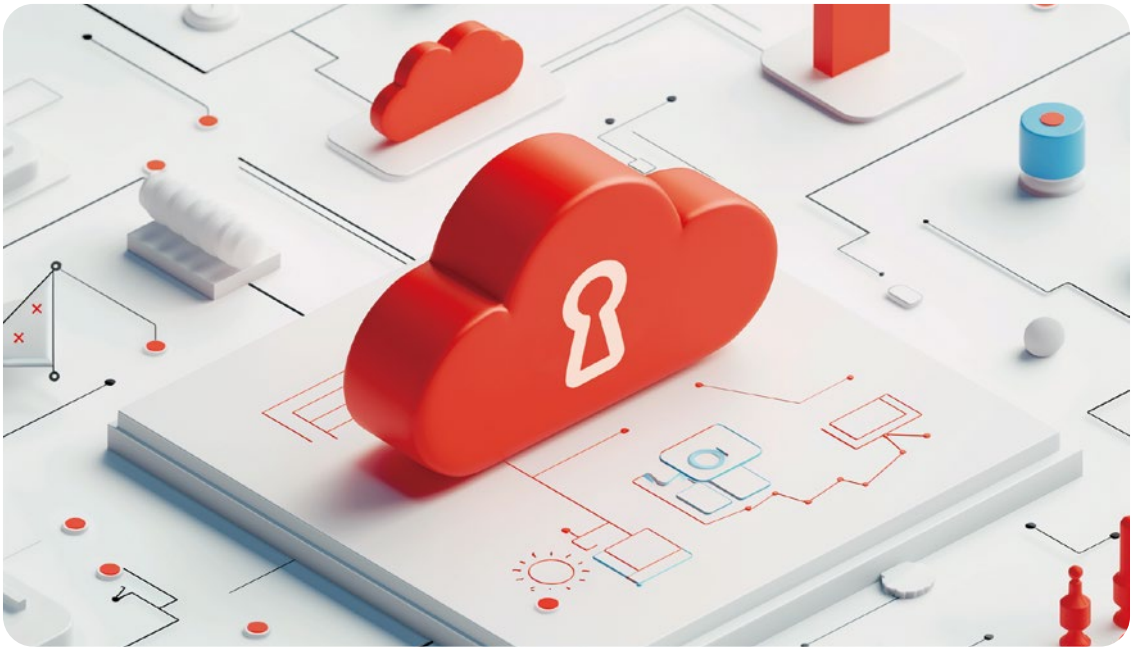
(3) Enhancing service and business model innovation

The intelligent vehicle cloud's open data service capabilities enable the development of intelligent applications, allowing carmakers to build an innovation ecosystem that revolves around users' mobility and lifestyle needs and unlock data value faster than ever through data-driven services.

Through the intelligent vehicle cloud, carmakers and developers can offer users extensive, tailored services and intelligent applications, design novel service scenarios and business models, undergo service-oriented transformations, and ultimately boost service revenue.

3.5.3 Creating a secure defense system for vehicle-cloud collaboration to protect data and user privacy

As V2X applications are widely adopted, protecting vehicle data and user privacy becomes of vital importance. Considering this, the intelligent connected vehicle cloud platform must establish a robust security defense system that encompasses data, cyber security, and consumer privacy. Key aspects of this endeavor include:



(1) Developing cutting-edge encryption technologies and authentication mechanisms

This entails leveraging sophisticated encryption algorithms to ensure end-to-end secure communications, as well as introducing the two-way or multi-factor authentication mechanism to ensure secure interactions between vehicles and the cloud.

(2) Building a multifaceted security defense system across domains

By integrating automotive components with the cloud, a comprehensive vehicle security defense system can be built to guarantee the security of both software and hardware, as well as in-vehicle communication of intelligent connected vehicles through mechanisms like cross-component authentication, process-level software security protection, and vehicle-specific authorization.

(3) Establishing a cloud-based 24/7 security operations center to facilitate proactive defense and emergency response

As AI and machine learning continue to advance at an unprecedented pace, the intelligent connected vehicle cloud platform must be capable of detecting and responding to security threats with greater intelligence. Through the harnessing of real-time data analysis, tool-assisted inspections, and other methods, anomalies can be identified in time, triggering prompt defensive responses. Additionally, developing a robust security vulnerability management and emergency response system will be a major focus for the advancement of V2X security technologies.

(4) Enhancing privacy protection technology

Ensuring the confidentiality and security of personal information is a top priority for carmakers. To address this, intelligent connected vehicle cloud platforms must leverage advanced technologies such as data anonymization, intelligent data masking, differential privacy, and other robust measures to safeguard user privacy throughout the entire data lifecycle — from collection and aggregation to processing and application. Furthermore, stringent data access controls and auditing mechanisms should be deployed to prevent unauthorized access or exploitation of sensitive data.

3.5.4 Creating a cloud-based simulation platform for efficient intelligent driving algorithm training and iteration

To solve the long-tail problem with intelligent driving, cloud service providers need to continuously enrich corner case datasets and simulation scenario libraries for iteration of intelligent driving algorithms. Throughout this process, they need petabytes of data and a huge amount of computing power (more than 1,000 GPUs) for algorithm training, and must simulate driving astronomical distances (as far as 10 billion miles) to validate an algorithm. In addition to large storage capacity and computing power, the iteration of algorithms also requires reliable, secure, and scalable infrastructure services. The

conventional model for data center construction puts the costs and O&M responsibility on intelligent driving developers, so we expect that cloud computing technologies will be widely applied in intelligent driving to address these challenges.

Cloud service providers will need to be capable of providing one-stop intelligent driving development platform on the cloud that provides a complete and automated development toolchain to help address complex engineering problems of intelligent driving, like data collection, data replay, automatic labeling, identification of corner cases, incremental dataset generation, model management, training task management, model delivery, simulation scenario library building, simulation test, and algorithm adaptation. Carmakers and developers will then be able to quickly build up their intelligent driving development and testing capabilities and allow for faster algorithm development and iteration. Specifically, cloud service providers need to:

(1) Provide scalable, secure, and compliant infrastructure for intelligent driving algorithm development

Hyperscale data storage and computing centers, built based on cloud platforms, can provide the massive uploading capacity, compliant storage services, and massive computing resources needed to handle the huge volumes of data that will soon be generated by intelligent vehicles. In this way, carmakers developing algorithms for intelligent driving will be able to access affordable, scalable, reliable, and secure infrastructure.

(2) Address engineering incoherence and support the DevOps of intelligent driving algorithms

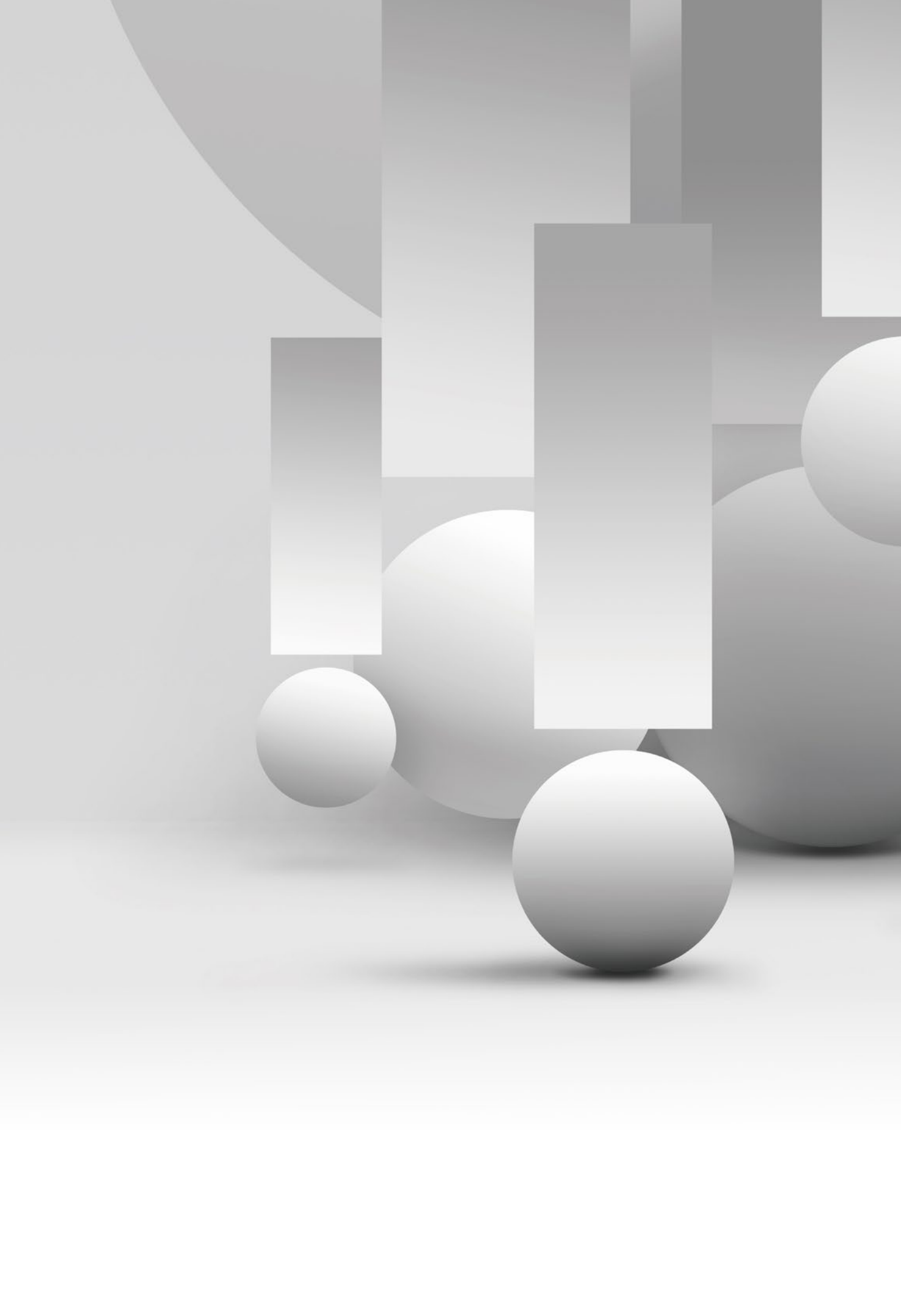
A comprehensive development toolchain, preset algorithms, datasets, and scenario libraries, along with simulation and validation services, must be provided to ensure a closed-loop intelligent driving development process, from data collection and processing, and identification of scenarios (especially corner cases), to algorithm management, training, simulation, and validation. These capabilities will help enhance the efficiency of intelligent driving algorithm development and iteration.

(3) Enhance scenario construction for improved cloud-based parallel simulation and accelerated algorithm verification and iteration

To expedite cloud-based parallel simulation and accelerate algorithm verification and iteration, scenario construction capabilities must be enhanced. This involves developing quick test and verification capabilities on the cloud to meet the stringent demands of testing 10 billion miles of driving. Specifically, the cloud should possess the ability to rapidly build scenario libraries and generalize scenarios, as well as construct scenarios tailored to functional safety, safety of the intended functionality, and V2X applications. At the same time, vast cloud resources and container technologies should be employed for more efficient large-scale parallel simulation. Huawei's GIV predicts that by 2030, cloud-based daily simulations will cover tens of millions of miles of driving, greatly enhancing the efficiency of intelligent driving test and verification.

Notes on the update:

Huawei collaborates with industry experts, customers, and partners to explore the intelligent world. The progress towards an intelligent world has accelerated significantly, with new technologies and scenarios emerging constantly, and industry-related parameters changing exponentially. As a result, Huawei has updated the *Intelligent Automotive Solution 2030* report released in 2021, providing insights into the scenarios and trends towards 2030, and adjusting the relevant forecast data.



HUAWEI

Version 2024

Digitalization Trends in the Electric Power Industry 2030

Building a Fully Connected, Intelligent World

Foreword 1

Carbon peak and neutrality are the national strategic goals. To achieve them, a new power system that uses new energy as the primary energy source is absolutely crucial.

China's electric power industry accounts for about 40% of the country's total carbon emissions. As such, it will be elemental to achieving carbon peak and neutrality, facing heavy tasks and significant responsibilities. In addition to ensuring a secure and stable power supply, the power system must accelerate its emission reduction. Thus, it is urgent to develop renewable energy, improve energy efficiency, and digitalize the power system.

A new power system that integrates electricity and computing will be the driving force behind the industry's transformation. Prioritizing new energy sources will profoundly change the modes, characteristics, and mechanisms of traditional electric power systems. Power generation, grid, load, and storage will need to be converged, forming a system that features a large power grid + active power distribution network + micro power grid. In addition, a powerful software platform based on digital data will need to deliver the necessary computing power. The new software-defined power system will integrate information, computing, and sensing technologies with control theories, AI, and the Internet. Ultimately, we will shift toward digital, information-based, and intelligent transformation. In this way, a visible, perceptible, and controllable transparent power system will be established.

Under these circumstances, power consumption of terminals is greatly improved. In addition to direct use of electricity, it can be converted into other forms of energy, implementing wide-area electrification. Therefore, boundaries of traditional power grids and industries are broken. Various industries interact with the electric power field to form a new energy ecosystem that features flexibility, openness, interaction, cost-effectiveness, and sharing. The resulting power system can be intelligent, secure, reliable, green, and efficient.

Digitalization Trends in the Electric Power Industry 2030 lays out a blueprint for building a digital twin of the power system through the in-depth integration between digital and electric power technologies. The book analyzes six core service scenarios: digital green power plant, intelligent power grid inspection, multi-source self-healing distribution network, coordination and complementarity of multiple energy sources, cross-region power dispatching, and green and low-carbon enablement. To support these scenarios, the authors draw on four key technological areas: digital edge, ubiquitous communication network, computing power & storage, and algorithm & application. The book also describes the necessary features for digital power technologies, which include green network, security and reliability, ubiquitous sensing, real-time network connection, endogenous intelligence, and service openness. Ultimately, this book recommends that an open, efficient, and intelligent electric power digital engine be built based on cloud-edge synergy. This type of architecture will support and drive electric power system transformation,

accelerate new energy consumption, and promote the achievement of carbon peak and neutrality goals.

Digitalization Trends in the Electric Power Industry 2030 draws on extensive expertise in relevant technologies, best practices and lessons learnt, and an in-depth understanding of the power industry to develop a detailed plan for electric power digitalization 2030. It also quantifies the objectives of digital development, making them more attainable. It is practical and forward-looking, meant to serve as a reference guide for the development of the electric power industry and cross-industry collaboration.

Academician of Chinese Academy of Engineering
Professor of the South China University of Technology
Chairperson of the Expert Committee of China Southern Power Grid

Foreword 2

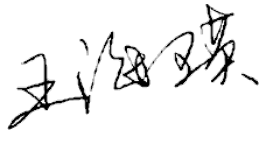
In 1875, the world's first thermal power plant was built in Paris. Since then, the electric power industry has been developing for nearly 150 years, becoming the pillar of the energy sector and society.

In the past 20 years, the global energy industry has undergone dramatic changes. The world is striving to reduce the impact of human activities on climate change by cutting greenhouse gas emissions. At the same time, the petrochemical industry is slowing down due to limited natural resource reserves. Significant advances have been made in new materials, engineering, and power generation technologies, significantly improving the efficiency of renewable energy power generation while also reducing the costs. As a result, new energy is slated to become the principle source of power in the near future. In this, technology will play a key role. Smart grids and power system technologies facilitate more flexible, robust, and intelligent power grid operations and management. Breakthroughs in battery technologies and industrial manufacturing also drive the popularization of electric vehicles. At the same time, digital technologies — such as digital twins, cloud-edge integrated IoT, AI, and high-bandwidth wireless communication — are enabling IoT sensing, data collection, edge computing, and intelligent analysis, driving innovation in the power industry.

That said, the power industry faces many challenges. How can the industry move away from its dependence on fossil fuels through renewable energy? How can it respond to more diverse power consumption demands while ensuring security and reliability? How can electricity costs be reduced? To address these challenges, in the next 10 years, the global power industry will witness a new round of rapid development and transformation, driven by the advance of new energy and digital technologies.

Beyond the big questions, there are also practical challenges arising in the power industry supply chain. In the past, processes followed a singular path — power generation, transmission, transformation, distribution, consumption, and finally dispatching. Today, the industry is shifting towards the collaboration and convergence of power generation, grid, load, and storage. The past few years have seen the rapid development of industry-specific digital technologies, represented by cloud, big data, IoT, intelligence, and mobility. We are also seeing the emergency of cutting edge tech like the metaverse, Web 3.0, edge intelligence, 6G, 10GE wireless communication, quantum computing, and quantum communication. Based on these technologies and the preceding challenges, the work group for this white paper took specific industry scenarios as the starting point for the transformation of the power industry. The white paper breaks down and analyzes each challenge from multiple dimensions. It also looks at the future industry requirements and the best practices that can inform how we address them. Ultimately, the white paper helps readers clearly understand the root causes of these industry challenges, accurately identify the key technologies that can help solve them, and successfully develop effective solutions.

Ernst & Young (China) Advisory Limited has a long-standing history of providing consulting services for electric power customers both in China and around the world, contributing to the industry's development and transformation over the past decades. The application of digital technologies is now a key contributor to the development of the global electric power industry. As digital technologies develop rapidly, there are more and more technological segments. For power companies, selecting the right technology at the right time is becoming increasingly important as well as difficult. We hope that the Digitalization Trends in the Electric Power Industry 2030 white paper can help power customers seize the development opportunities of the industry in the next decade to scale new heights.



EY Greater China Managing Partner of Consulting

Foreword 3

Green development is key for the future of our planet. Looking to achieve carbon peak and neutrality, countries and regions around the world, including China, the European Union (EU), and North America, have all released strategic initiatives to transform their energy structure through low-carbon energy, electrification, networking, and intelligence. In this fight toward a greener future, energy is the main battlefield, electric power is the main force, and digitalization is the key.

Major electric power markets around the world have proposed the goal of building "SmartGrid" and "IntelliGrid." Leading electric power enterprises in China also have the vision of building digital power grids. These new grids will support the reliable, flexible, and economical transmission of power and information flows. They will assure secure and stable network communication and system control along with comprehensive data integration and digital monitoring. The grids will also enable free power trading and distributed power access, facilitating the interaction between power grids and users.

The ultimate goals of digitally transforming electric power are to ensure efficient and stable operations, improve energy efficiency, and eventually achieving carbon peak and neutrality goals. By connecting the physical world with the digital space, device information and production processes in the power system are converted and expressed in a digital way. Essentially a digital mirror of the power system is built in the virtual world. What used to be virtual-physical mapping will evolve toward the in-depth interaction between the physical and the digital, resulting in a digital twin of the entire power system. In turn, the power digital twin will enable security and intelligence. Right now, clean energy, such as wind and solar power, can be intermittent and random, resulting in discrete distribution and uncontrollable fluctuations. Based on large-scale interconnection and dispatching, the digital twin will also facilitate efficient development and utilization of clean energy.

To create the power digital twin, power electronics and digital technologies need to be deeply integrated. For this, we need to build a more open, efficient, and intelligent digital platform for low-carbon development using digital technologies. We must also drive standardization of the electric power industry, including communications and control, and promote the interconnection of power system devices, thus embedding intelligence into the industry. As for the entire process from power generation, transmission, and transformation, to distribution and use, we should continuously create ICT value for connection, computing, and intelligence and build a modern device asset life cycle management system, so as to reduce power asset operation risks, prolong service life, and improve device security and operational efficiency. Then, we need to provide an all-round energy consumption service system for customers to recommend the optimal power consumption solution, maximizing energy utilization and reducing energy waste. Also, a new dispatching control support system is required to improve flexibility and stability of the power system, enabling source-grid-load-storage interaction and overall efficient energy utilization. In addition, a new electric power trading system is necessary to make electricity

become commodities. Green power trading is an important means to achieve carbon peak and neutrality and promote clean energy consumption.

Ongoing innovation in digital technologies will drive sustainable development. Our research shows that by 2030, over 95% of terminals deployed by industry-leading digital electric power enterprises will be connected to the network. The cloud penetration rate will exceed 75%, the edge intelligence adoption rate will exceed 75%, the reliability of electric power communication reaches 99.99999%. Electric power digitalization will feature green network, security and reliability, ubiquitous sensing, real-time network connection, endogenous intelligence, and service openness. Unlike the rigid systems of the past, power supply and demand will become more flexible.

The future of electric power digitalization is full of possibilities and challenges. Traditional electric power enterprises (such as power generation companies and power grid operators), new businesses (such as electric vehicle vendors), technology enterprises, campus operators, and platform service providers must work together toward source-grid-load-storage transformation. Let's explore and innovate together to build electric power digitalization 2030.

Executive Director of the Board
Chairman of ICT Infrastructure Managing Board, Huawei

Introduction

Today, green and low-carbon development is now a trend. Many countries regard the development of green and low-carbon industries as an important measure to promote economic restructuring. The deployment of green and low-carbon infrastructures is also accelerating across the world. In September 2020, China vowed to strive to achieve the dual-carbon goals of carbon peaking and carbon neutrality by 2030 and 2060 respectively. North America is pushing for a US$555 billion clean energy program, to increase investments in key domains such as infrastructure and clean energy and subsidize electric vehicle (EV) buyers and households with solar roofs. The EU plans to invest more 350 billion euros each year from 2021 to 2030 to promote EVs and public transportation to achieve emission reduction targets. Germany has vowed to abandon fossil fuels by 2035 and planned to accelerate the construction of renewable energy infrastructure such as wind and solar energy to achieve 100% renewable energy supply.

To fulfill the green development goals, the energy industry — especially the electric power sector— will play a vital role. By continuously improving the electrification rate of terminals, we can effectively reduce the dependence on traditional fossil fuels and improve the penetration of high-quality electric power. New energy represented by photovoltaic (PV) power and wind power will play an important role.

With the continuous increase of installed new energy capacity, a high proportion of renewable energy will lead to a high proportion of electric and electronic devices on the power system. On the grid side, energy resource allocation will be highly elastic. On the load side, electrified energy consumption is highly interactive due to the bidirectional and multi-source feature. On the transaction side, local low-carbon energy supply will make the cost of energy consumption lower. On the operation side, various energy systems will be highly integrated based on digital twins.

In 2030, the continuous development and in-depth application of electric power digital technologies will become a key success factor for the upgrade of power systems, helping them better adapt to and cope with the trend of 'four highs and one low'. In the future, driven by the new digital engine of electric power digitalization, power systems will become greener, and more secure, efficient and friendly, better serving industries and households.

Chapter 1: Background and Objectives of Electric Power Digitalization

(1) Background of electric power digitalization

The electric power industry is undergoing profound transformation. In the future, the power system will experience two important changes, which specify the basic direction for the development and application of digital power technologies.

1. Green energy structure

From the perspective of power supply, according to IRENA statistics, 3.865 billion kW of renewable energy was newly installed worldwide in 2023, accounting for 43% of the global power generation installation, with a year-on-year increase of 14.0%. It is estimated that by 2030, the world's energy system will undergo significant changes, and renewable energy will account for nearly 50% of the global power structure.

In 2023, China's secure and stable energy supply capability has been steadily enhanced, and the pace of green and low-carbon energy transformation has been accelerated. Renewable energy installed capacity has consecutively exceeded 1.3 billion and 1.4 billion within the year, reaching 1.45 billion kW, which accounts for over 50% of the country's total installed power generation capacity (achieving the goal 6 years ahead of schedule). Renewable energy generated 3 trillion kWh of electricity, approximately one-third of the total electricity consumption, becoming a new force in ensuring power supply. It is estimated that by 2030, China's installed clean energy capacity will reach 2.57 billion kW, accounting for more than 70% of the total. Clean energy is expected to generate 5.8 trillion kWh, accounting for 52.5% of the total energy yield.

During the construction of new energy, there is a mismatch between the abundance of wind and solar resources and the level of economic

development in certain regions. Take China as an example. Most of the key new energy bases are distributed in areas with low population density and power load, such as western and northern China. The demand for energy mainly comes from densely populated areas in the coastal eastern and southern regions. Therefore, remote power consumption poses new requirements on the long-distance transmission and consumption capabilities of the power grid, auxiliary construction of energy storage devices, and flexible operations.

In addition, most EU countries are small in land area and can use existing resources nearby. A large number of distributed PV and distributed wind power devices emerge at the user side. These power supplies in the load system bring new challenges to the stable operations of the power distribution network.

2. Flexible power grid regulation

As renewable energy continues to rapidly develop, the fluctuating and unstable power generation will also increase from the perspective of the power grid. This presents new and more complex demands for power control, voltage regulation, and communication methods. The traditional power grid is mainly facing two major challenges: horizontal balance stability and vertical operation control. In recent years, the power distribution network has seen a significant increase in the number of distributed PV modules, with over 98% of them being connected to 400 V. This has transformed the distribution network from a passive to an active network with numerous distributed power sources. However, synchronizing the PV with local load requirements in terms of time has become challenging. This leads to the "duck curve" phenomenon in the net load curve, which increases uncertainty in power generation and consumption balance and real-time dispatching. As a result, load-based power generation is changed to source-grid-load-storage interaction, and the power flow between the transmission network and the traditional distribution network frequently turns and fluctuates greatly, which severely affects the secure operation of the power system. Therefore, the key challenge in solving the new power system lies in the distribution network.

3. Interactive power supply

From the perspective of power consumption,

in recent years, China's electricity demand has shown a clear "winter-summer" peak characteristic, with cooling in summer and heating in winter accounting for an increasing proportion of the load. According to statistics, in 2023, more than 10 provincial-level power grids in China had a summer "cooling load" (additional electricity consumption due to rising temperatures and increased use of air conditioning) that accounts for over 40% of the highest electricity demand. Factors such as the increasing frequency of cooling loads have led to a larger and more intermittent peak-valley difference in electricity demand, significantly increasing the difficulty of balancing supply and demand in the distribution network system. This further highlights the problem of insufficient flexible regulation resources in the power system, leading to increased pressure on peak regulation and voltage control. In addition, the number of electric vehicles increases rapidly. In 2023, the number of electric vehicles in China exceeded 20 million. It is estimated that the number will exceed 100 million by 2030. The global number of electric vehicles will be nearly 10 times the current number. The uncertain load generated by charging these vehicles may increase the regulation pressure on the power system, posing new challenges to the stable operation of the distribution network. As society's understanding of sustainable energy development deepens, the penetration rate of electric vehicles is gradually increasing. Consumers are increasingly demanding more reliable and transparent local energy consumption. This will lead to the emergence of more and more distribution network-level distributed energy systems in the power system. The power supply mode will change from "being centered on large power plants" to "being centered on prosumers".

With the large-scale application of distributed energy resources such as PV and wind power, the traditional distribution system will be upgraded from a one-way power supply mode to a source-load integrated mode. The power supply mode will also shift from the past "load-based power generation, production plan-led, unidirectional step-by-step power transmission" to "source-load interaction, more consumption of renewable energy, and bidirectional flexible allocation based on supply and demand changes." Consumers will have greater autonomy and a wider range of choices when it comes to energy consumption. This shift from passive electricity usage to active consumption will lead to new demands for accurate electricity demand prediction, flexible allocation of electricity resources, resilient and balanced distribution networks, and unified IoT and interaction among a large number of devices.

(2) Key objectives of electric power digitalization

The ultimate goal of digital transformation and development of the electric power system is to absorb more green and low-carbon electric power represented by wind and light, promote efficient interaction between the source network and load storage, ensure efficient and stable operation of the electric power system, and improve energy efficiency, promote the realization of carbon peak and carbon neutrality.

As the "last mile" of the power system, the power distribution network plays a major role in the construction of the new power system. The development and innovation of power distribution networks will be the core of the construction of new power systems. More than 50% of China's total power grid investment is allocated to power distribution, while in Europe, power distribution investment accounts for over 70%. Power distribution networks need to have stronger capacity to accommodate the needs of new entities and formats such as large-scale distributed new energy and electric vehicle charging facilities. This will help increase the proportion of green power consumption.

We believe that there are five goals for the digital transformation of electric power:

1. The shift from the traditional load-based power generation to source-grid-load-storage interaction

Large-scale new energy facilities, distributed energy systems, and energy storage devices of different scales are widely used. Take China as

an example. It is estimated that by 2030, there will be over 600 million users, 100 million electric vehicles, 50 million charging stations, 10 million transformer districts, and more than 300 GW of installed new energy storage capacity.

The unpredictable availability of natural resources for power generation and the real-time demand for power consumption by users no longer perfectly match, resulting in a potential mismatch between the supply and demand of electricity, known as the "scissors difference". To address this issue, the power supply mode needs to shift from the traditional load-based power generation to a source-grid-load-storage interaction mode.

Supply and demand can be balanced through source-grid-load interaction. AI is a key measure to achieve autonomous management of transformer districts and balance power supply among different cities. Multiple large models, such as meteorology, prediction, power flow calculation, and solver, need to be integrated to build a four-level AI+ scheduling system. The system should start by balancing supply and demand within transformer districts, using customized policies for different transformer districts. Then, it should gradually balance power supply among different regions, starting from the city level and moving up to the provincial and grid levels. To ensure safe and economical electricity usage within transformer districts, an intelligent hub for microgrids can be established. This hub would provide unified scheduling and management of energy flow, information flow, fund flow, and control flow. Additionally, devices within transformer districts can be monitored using Power Harmony+, HPLC, and AI inference, achieving grid-based intelligent management.

2. 100% grid connection and consumption of new energy

In 2023, China had a total installed capacity of around 300 million kW for distributed energy. The penetration rate of distributed energy is rapidly increasing and has even surpassed 50% in certain regions. Moving forward, the expansion of distributed clean energy access is expected to experience rapid growth.

Compared to traditional energy sources, new energy generation has characteristics such as randomness, volatility, intermittency, and uncontrollability. This highlights the active feature of the distribution network, which brings about issues such as bidirectional flow, voltage limit violations, and isolated operation. Additionally, new energy sources have weaker tolerance to extreme weather conditions, leading to uncertain factors in electricity production and output. This causes fluctuations in voltage, frequency, and greatly impacts the reliability of grid power supply. Low-voltage distribution networks have become the primary connection path for distributed power sources.

With the construction of large-scale new energy bases, there have been frequent curtailments of wind and PV power, as well as disconnection of new energy from power grids. New energy power generation and grid connection have become a key challenge to achieving low-carbon, green power supply.

Therefore, the key to the power industry's active implementation of carbon peaking and "carbon neutrality" lies in the application of a series of digital power technologies such as sensing, prediction, control, and dispatching. They will enable the 100% grid connection and consumption of new energy from both the source and load ends, and offset the fluctuations caused by new energy grid connection.

3. High asset security, efficient operations, and the shift from planned to condition-triggered repair

For instance, China's two power grid companies had total assets of CNY5.75 trillion by the end of 2022, marking a 7.4% increase from the previous year. In 2023, the total investment of major electric power enterprises in China reached CNY1.5502 trillion, a YoY increase of 24.7%. The investment in power grid construction reached CNY527.7 billion, a YoY increase of 5.4%. This indicates that China's power grid assets are still expanding, and there is a rise in investment to facilitate the advancement and upgrade of power systems. It is expected that by 2030, China's power grid assets will continue to expand to

meet and support the development requirements of new power systems.

With the development of power grids, new energy power plants, distributed power supply, and electronic devices will gradually replace traditional outdated devices. However, the power industry is a traditional heavy-asset industry, so existing power assets still play an important role during the transformation. On the power generation side, traditional energy is still the main power source. Some large-capacity, high-efficiency, and low-consumption thermal power plants will remain the main power generation units in the foreseeable future. Thermal power units will also be responsible for peak-load modulation and frequency modulation and balancing the fluctuation of new energy output. On the power transmission and distribution side, the power transmission network consisting of a large number of AC and DC UHV power grid infrastructure is still the fundamental guarantee for cross-regional power transmission. In terms of incremental assets, pumped storage stations, compressed air, and electrochemical energy storage will be the key to implementing integrated coordination and interaction of power source-grid-load-storage. In the future, a large number of energy storage devices will be constructed and put into operations.

Distribution networks will also play a vital role in the construction of new power systems. According to the guidance of the National Development and Reform Commission and National Energy Administration, power distribution networks will be flexible, intelligent, and digital by 2030, effectively promoting the convergence of distributed intelligent grids and large grids.

Therefore, with security as the core, the primary objective of electric power digitalization is to improve asset utilization efficiency at the lowest cost, reduce the operation risks of power assets, prolong the service life, improve security and operation efficiency, and ensure the smooth power decarbonization transformation and secure and reliable power supply.

Digital methods are used to support the transformation of device O&M management, from passive emergency repair to proactive O&M. Devices are managed throughout the lifecycle, and device status is detected in a timely manner. The maintenance mode is changed from planned repair to condition-triggered repair.

4. 100% visibility and traceability of green power transactions

From 2021 to 2023, the transaction volume of green power in China was 8.7 billion kWh, 18.1 billion kWh, and 69.7 billion kWh, respectively, with an annual growth rate of 283%. This indicates that green power trading plays an increasingly important role in promoting clean and low-carbon transformation and encouraging green power consumption.

Currently, green power market-oriented transactions have been formed. However, there are multiple transaction entities and complex authentication processes, resulting in potential risks such as high costs, difficult traceability, and easy tampering. Therefore, blockchain technologies are required to support green power transactions. The blockchain technology features traceability, anti-tampering, openness, and transparency. It generates a unique proof of green power consumption that complies with transaction and review specifications, making the entire green power transaction chain traceable, reliable, and verifiable. It is estimated that all green power transactions will be visible and traceable by 2030.

The application of digital power technologies allows power companies to voluntarily pay premiums for green power and stimulates the enthusiasm of various market players to proactively participate in green electricity transactions. As a result, green electricity trade will become a key driver for achieving dual-carbon goals.

5. Innovation and development of new business forms such as virtual power plants

Diversified power supply modes, changing power consumption modes of various users, and increasingly diversified power consumption

requirements generate some new application scenarios and services, such as virtual power plants, orderly charging of electric vehicles, and response to user requirements.

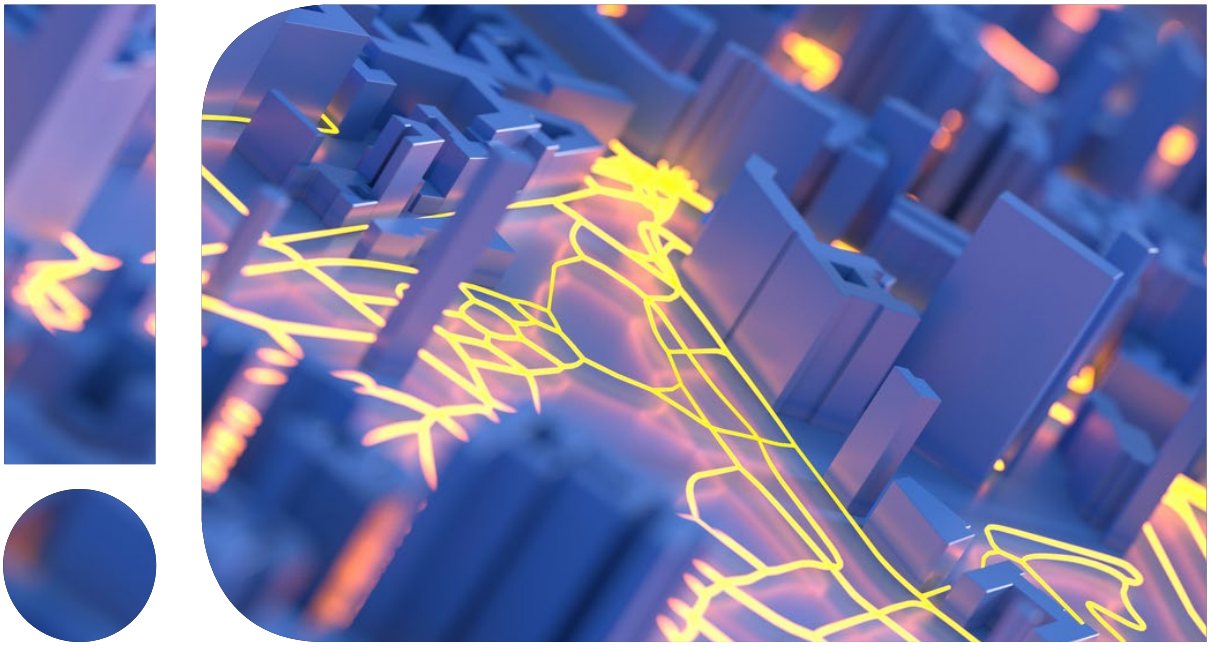
Virtual power plants are rapidly developing in China with the support and response of government policies. Take Shenzhen virtual power plant as an example. By July 2024, over 50 operators have joined the Shenzhen virtual power plant management platform, with a resource capacity exceeding 2.75 million kW. The platform has a maximum regulation capacity of over 600,000 kW, which is expected to reach 1 million kW by 2025. This supports the stable regulation capacity of around 5% of the annual maximum load. Shanghai, Tianjin, and Zhejiang all propose that the peak load response capability on the demand side should not be lower than 5% in 2025.

It is estimated that in 2030, China's virtual power plants are expected to regulate 5% of the country's total electricity consumption, creating a market space worth CNY200 billion.

Through intelligent optimization policies, virtual power plants integrate and coordinate different types of distributed resources, such as distributed power generation (such as wind and PV power), distributed energy storage facilities, and controllable loads. They can optimize operation control and market transactions. Virtual power plants can integrate resources like wind and PV power, which have a good output time period but low self-consumption ratio, with energy storage or controllable loads. This integration enables local consumption and improves the utilization efficiency of renewable energy.

The use of digitalization technologies in the electric power industry is driving the development of new scenarios and business models, increasing the intelligence of power systems, reducing energy costs, and promoting efficient energy utilization and low carbon emissions.

In general, the goal of digitalizing the power industry is to address the issues related to energy security, green development, and efficiency, and to create an effective means of building a new power system.



Chapter 2: Scenarios of Electric Power Digitalization

(1) Power digitalization blueprint

The entire industry is going decentralized, and terminal devices are being electrified. This trend drives the interaction among source, grid, load, and storage in the future power grid and breaks the traditional value chains. Instead of being load- based and plan-driven, power generation, as well as the power supply-demand relationship, will be more flexible and random.

By developing and applying next-generation digital enablement technologies, such as digital edge and device (edge/device collection and control), ubiquitous network communication network (terrestrial and satellite communications), computing power and storage (cloud platform, cloud-edge-device synergy, space computing, and blockchain), and algorithm and application (AI, graph computation, and advanced analysis), we can further connect the physical world to the digital one. Digitalized presentation of device information, production process and other information of the power system builds a digital image of the system in the virtual space. A leap forward in the digital capabilities, such as digital monitoring, intelligent analysis, and digital and intelligent autonomy, accelerates the in-depth interaction between the physical and digital worlds, building a digital twin of the entire electric power system.

Specifically, the electric power digital twin can be divided into three forms:

Digital surveillance

The purpose of monitoring is to comprehensively and accurately monitor the running process and status of power equipment assets in the

digital space through ubiquitous sensing, high-speed communication, and platform storage, and dynamically monitor and diagnose device assets throughout the lifecycle based on multi-dimensional data. In this way, we can use bits to perceive watts in various scenarios. The demand for capacity and flexibility in power distribution networks is expected to significantly increase, leading to a comprehensive push towards digital transformation in these networks. It is projected that by 2030, the digitalization rate of equipment will exceed 90%. The establishment of perception networks and mechanism models is the basis for efficient digital monitoring of power systems. In addition, data interworking and ubiquitous IoT also require data encryption technologies to ensure information security.

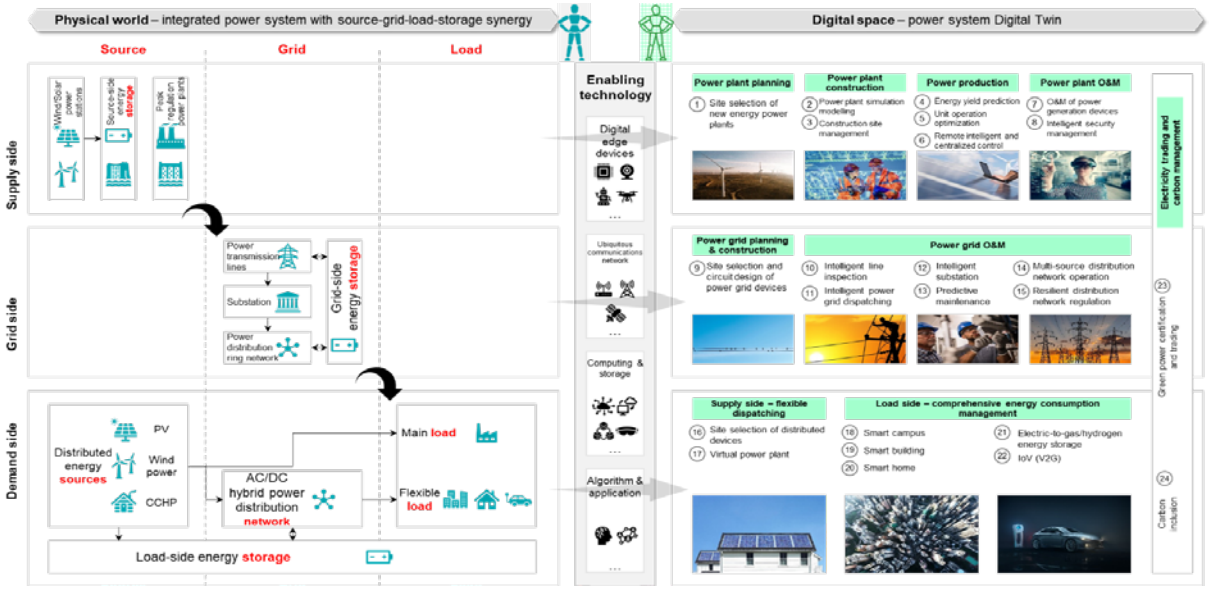
Intelligent analysis

The purpose of intelligent analysis is to analyze, predict, and simulate future operation changes of generator sets, power transmission and distribution networks, and power loads based on the determined operation mode and mechanism rules, provide decision-making support for operation optimization and system control based on the existing system, and implement 'bit manages watts' in various power scenarios. It is expected that by 2030, over 90% of digital power businesses will be moved to the cloud. Computing power and algorithms are core technologies for improving the accuracy of intelligent analysis of power systems. By constructing complex data models covering multiple domains and disciplines and simulating digital space, physical entities can be optimized and an effective closed loop can be formed.

Digital and intelligent autonomy

Based on cross-system and cross-module massive data interaction, adaptive and self-evolutionary complex algorithm models, and intelligent achievements shared by digital space, proactively identify the bottleneck of the current physical world running mode, issue decision-making instructions or propose predictive reconstruction plans to promote in-depth interaction between the physical world and the digital space through decision-making autonomy and achievement feedback of the digital space, so that 'bits can add value to watts'. It is projected that by 2030, over 80% of digital power solutions will be powered by AI. A massive amount of cross-system data needs to be exchanged and shared. Therefore, in addition to AI technologies such as advanced analysis, technologies such as blockchain and privacy computing are also the key.

Figure 1: 2030-oriented digital twin of the power system



(2) Typical scenarios of electric power digitalization

Selection of typical scenarios

Judged by the value and technology maturity, electric power digital scenarios based on the power digital twin are classified into three types:

1. **Current hot scenarios:** The market scale of these scenarios is growing rapidly and mature digital solutions are available. Power enterprises can start their digital capability building from these scenarios.

2. **Future focus scenarios**: Power supply reliability and dual-carbon target achievement are two value points in these scenarios. However, we need breakthroughs in key energy and digital technologies to support the large-scale application.

3. **Other scenarios**: Other scenarios are either red ocean markets or those with vague business values. They are not discussed in this White Paper.

Current hot and future focus scenarios are further analyzed, and six core business scenarios of electric power digitalization oriented to 2030 are formed:

Scenario 1: digital green power plant — 3 4 5 6 7 8

Scenario 2: intelligent power grid inspection — 10 12 13

Scenario 3: multi-source self-healing distribution network — 14 15

Scenario 4: coordination and complementarity of multiple energy sources — 18 19 21

Scenario 5: cross-domain power dispatching — 11 17

Scenario 6: green and low-carbon enablement — 23 24

Figure 2: Power digital scenario evaluation model

Analysis of typical scenarios

Transforming to the 2030-oriented electric power industry, we have three focuses: security, efficiency, and environmental friendliness. To build the new power system, we are now faced with many challenges: secure power supply, lifetime extension of key devices, plant network operation efficiency, new energy consumption and transaction, etc.

The core values and technologies of electric power digitalization in the six business scenarios of the future power grid are detailed as follows:

Figure 3: Key development and transformation challenges of the electric power industry

2030-Oriented Key Scenarios	Key Industry Transformation Challenges		
	Security	Efficiency	Environmental Friendliness
Digital green power plant	▸ Occasional security problems occur in the factory, resulting in damages to personal and financial safety	▸ Independent plant operation, resulting in inconsistent business management	▸ Inadequate response to weather changes, resulting in power loss
Intelligent power grid inspection	▸ Insufficient awareness of security risks and passive response to faults; dangerous manual inspection	▸ Ineffective manual inspection	▸ Delayed response to fault location, affecting energy consumption
Multi-source resilient distribution network	▸ Poor ability to withstand extreme situations and slow power supply recovery	▸ Slow fault locating and delayed response	▸ New energy power access, disturbing the distribution network operation
Multi-energy synergy and complementarity	▸ Energy storage, the core, is prone to fire and explosion	▸ Lack of effective linkage among multiple energy forms, resulting in inefficient energy cascade use	▸ Inadequate source-charge-storage interaction and inadequate carbon emission reduction/offset
Cross-domain power dispatching	▸ Frequent power rationing and production curtailment	▸ Limited load-side response to power demands	▸ Insufficient new energy power consumption capabilities of large power grids
Green and low carbon enablement	▸ Information security issues and data distortion	▸ Time-consuming review and certification process	▸ In adequate mechanism, affecting the participation enthusiasm

Scenario 1: digital green power plant

The power system is being reshaped, influenced and driven by environmental sustainability, energy security, and other factors. The key of the power system is shifting to new and distributed energy.

Large-scale wind farms and PV plants will be the key to improving the proportion of green electricity in the entire power system. Large-scale new energy stations are geographically remote and sparsely distributed, and their yield is subject to weather and environment. Besides, the on-grid energy also depends on the consumption capacity of the large power grids. To improve effective energy yield, prolong the service life of devices, and adjust generated energy in response to the capacity of power grids, intelligent management is required in terms of device inspection, power plant O&M, and remote control.

In the management of future new energy power plants, digital power technologies will be applied to scenarios such as digital twins throughout the lifecycle of power plants, and remote centralized control based on cross-domain IoT, improving the intelligent management level of power plants.

E2E digital twin

The full-lifecycle digital twin of new energy power plants will cover three phases: planning and construction, planned production, and O&M. In the planning and construction phase, we can effectively promote project implementation through onsite digital twins. In the planned production phase, we can optimize production policies through digital twins. In the O&M phase, we can improve device status in time through digital twins for production devices. In the entire lifecycle, the digital twin of the production environment ensures asset and personal safety.

In the management of new energy power plants based on digital twins, edge side data collection (including traditional production information monitoring and management systems and diversified sensor devices) has a certain foundation. How to effectively use accumulated massive data and fully explore the value of data assets is the key for new energy power plants to improve efficiency. Among them, spatial computing and machine learning will play an important role.

Key technology application 1: Spatial computing and 3D modeling facilitate scenario simulation and improve efficiency

1.**Full-cycle BIM support**: In the planning and construction phase, based on device parameters, onsite images, and surrounding environment data, we can restore the construction site through spatial computing and 3D modeling simulation, dynamically monitor and manage the entire project process based on the BIM model, as well as warn and analyze deviations from the planning, project construction risks, and security risks in order to ensure the project progress and quality.

At the same time, the BIM model not only guides construction, but also provides visualized management support for site production and operation, continuous upgrade and reconstruction, and device change and retirement through electronic handover, effectively solving the problem of cross-domain data silos and collaboration problems. It is worth noting that electronic handover not only improves the management efficiency of power plants, but also plays an important role in scenarios such as power grid construction and operations.

2.3D dynamic security management: Displays the power plant panorama in 3D, monitors employees' locations and heights through electronic fences, and automatically triggers security warnings. In addition, based on the real-time monitoring results of key areas and danger sources, the system accurately identifies security risks, generates alarms in a timely manner, and automatically plans the optimal evacuation path when an emergency occurs to minimize the probability of security accidents and ensure personnel and asset security.

3.Immersive skill training and remote inspection: Use XR terminals to simulate device fault scenarios and provide high-quality immersive inspection and maintenance training for employees, effectively improving employees' professional capabilities and device maintenance efficiency. In addition, with the help of smart wearables, convenient remote expert inspection can be implemented, and the combination with device monitoring can further improve accuracy of device inspection.

Key technology application 2: Machine learning supports decision making and optimizes operations for power plants

1.**More accurate prediction of generated electricity**: Different from the stable output of traditional energy, the features of new energy greatly increase the difficulty in formulating the new energy power generation plan. The production plan based on the historical data may deviate greatly from the actual situation.

Machine learning provides an effective solution for new energy power plant operators. Based on massive data such as historical weather conditions and historical output levels of new energy devices, learning and modeling are performed. Based on multi-dimensional variables such as weather forecast and actual device running parameters collected at the edge side, short- and long-term predictions on the future power output and energy yield of new energy devices can be developed.

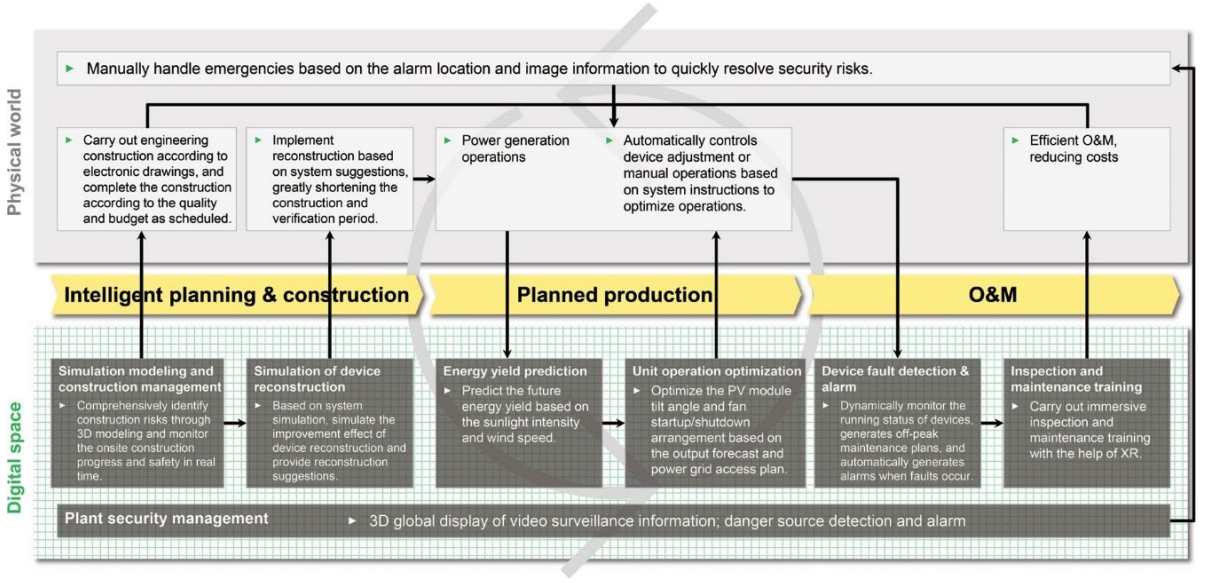
On one hand, this solution can provide decision-making support for the formulation or adjustment of new energy power generation plans. On the other hand, it can optimize the operation strategy of new energy equipment based on the prediction results. In addition, the energy storage charging and discharging policies can be flexibly adjusted based on the energy yield prediction result and power market price changes to maximize economic returns.

2.More efficient self-operation control: Based on the dynamic monitoring results of edge devices on fan wakes and dust accumulation on PV modules, climate change, short-term energy yield prediction, and algorithm models and digital space simulation results obtained through machine learning, we can generate automatic control instructions for a single device based on the PV module tilt angle, fan blade speed and angle, and fan startup, shutdown, and output status to formulate optimal operation policies for new energy stations.

3.More timely device defect warning: Monitors the running parameters of power generation devices, evaluates current and upcoming device defects based on the learning model of historical device defects and inspection and maintenance records, and generates warnings in time. Properly arranges off-peak inspection and maintenance to reduce unplanned shutdowns.

The digital twin of power plants based on spatial computing and machine learning can help new energy power providers implement virtual-physical interaction and closed-loop management throughout the lifecycle of power plants. Based on the current reflection and future prediction of digital space, the digital twin provides decision-making support and guidance for the physical world to take corresponding measures. Finally, the operation efficiency of the power plant is improved.

Figure 4: Full-lifecycle digital twin operation mode of new energy power plants



Remote intelligent and centralized control

Large-scale electric power enterprises face new challenges. On the one hand, new energy power plants are scattered in remote areas. The inspection costs of stations and devices are high, and the management of different power plants is relatively independent. As a result, the unification and collaboration are insufficient. On the other hand, for cross-domain power plant investors, they also lack effective operation and management methods for self-built new energy power plants. By building and applying the cloud-edge synergy technical architecture, we can build a new energy power plant operations platform that supports remote, intelligent and centralized control. The operations platform can manage units across regions, effectively reduce the operating expense (OPEX) of the new energy power plant, and improve the operating efficiency.

Application of key technologies: cloud-edge synergy, bringing the value of three elements into full play

1.Ubiquitous IoT and data convergence: Ubiquitous IoT is the first step to achieve cloud-edge synergy. Currently, many device vendors have embedded various sensors in power devices or components. However, different vendors use different technical roadmaps, resulting in inconsistent standards and data silos. Therefore, an enterprise-level IoT cloud platform needs to be built to unify the data and communication standards collected by different devices, remove data barriers between devices in different power plants, and implement comprehensive data access, openness, sharing, and coordinated management. In addition, the processing and analysis of a massive amount of cross-domain data also depends on the support of more reliable low-latency network communications technologies.

2.Edge computing power improvement: Edge computing power improvement is the second step to achieve cloud-edge synergy. Storing and computing massive real-time data collected by various sensors on the cloud will deplete cloud resources and affect the timeliness of data processing. Once the network is faulty, the running of the entire station will be affected. By deploying smart edge terminals, computing resources on the cloud are flexibly allocated to the edge to support distributed computing. This not only improves the timeliness and response speed of local data processing, but also effectively avoids security risks caused by data

transmission, achieving instant interaction and stability and security.

3.**Moving core algorithms forward**: Moving core algorithms forward is the third step to implement cloud-edge synergy. Based on edge data processing, algorithm models are required to quickly and accurately identify device faults or control and adjust device running in a timely manner. Modeling and machine learning are performed based on global device data resources aggregated on the cloud to form a global cognitive algorithm model. The algorithm model is deployed on smart edge terminals to function as the 'brain', achieving logical centralization and physical distribution. This way, precise analysis and efficient processing of edge data can be realized.

The cloud-edge synergy mode enables data convergence between power plants, supports model building, and improves edge data analysis and response capabilities through the deployment of smart edge terminals. In this way, new energy enterprises and power plant operators can implement cross-domain remote control and coordinated management through mobile terminals, shifting from 'partial improvement' to 'global optimization'.

Best practice: real-time monitoring and control of nearly 1,100 wind turbines and over 150,000 PV panels

Based on IoT and cloud-edge synergy solutions, platform service supplier A in China provides services supporting features of edge computing, warning and prediction, domain synergy, and open source architecture. It helps customers monitor wind power and PV stations in a centralized manner and implement unattended operation through end devices. It helps customer automatically identify and predict the abnormal status of devices and maintain devices in stations based on their health status, reducing the OPEX by 20% while improving the yield by 10%.

Component-level sensing: Smart microphones are installed on the door frame of fan towers to monitor the swing sound and send warnings of blade defects. All-round perception of key components of the drive chain, pitch yaw control, and other components makes device operation visible in real time. Fault-triggered inspection and repair replace component replacement.

Cloud-edge synergy: The latest edge computing technologies are used to process a large amount of data at the edge layer by means of direct device connection and preprocessing of station and device data, improving efficiency, reducing cloud load, and achieving high-caliber data accuracy. Through data standardization and integration, the cloud dynamically monitors and controls various operating indicators, and visually aligns the performance of each power plant to facilitate indicator assignment.

Machine learning: Based on deep learning about environmental factors' impact on energy yield, 5000-core parallel computing is used to predict the energy yield. The average wind power prediction accuracy reaches 90%, which is 7% higher than the industry average, and the optical power forecast accuracy reaches 93%, which is 1% higher than the industry average. The deep learning results can provide reference and decision-making support for new projects.

Summary of the application of digital technologies in digital green power plants

Judging by the current application of key enablement technologies, most enterprises still construct and operate different stations separately, resulting in many data silos and non-centralized control. In addition, although the plants are mostly covered by the communications networks, the WAN communications capabilities are insufficient. The key directions of research and breakthroughs in the future are as follows:

- **Improving the level of device connection**: Enhance the data standardization of various sensor devices to implement data interworking.
- **Improving WAN communications capabilities**: Adopt low-latency and high-reliability communications technologies to ensure the collection, processing, and analysis of massive real-time data.

- **Building a centralized control operation platform**: Build an operation platform based on the cloud- edge synergy architecture to improve edge computing power and intelligence, and achieve fast and accurate local response.
- **Improving AI coverage**: Strengthen AI training to improve the maturity and accuracy of key models such as energy yield prediction and device fault diagnosis.
- **Improving the security management efficiency of stations and devices**: Strengthen the application of 3D modeling in security management and immersive inspection and maintenance training to improve security assurance and employees' troubleshooting capabilities.

Scenario 2: intelligent power grid inspection

Power grid lines are to the electric power system what the skeleton is to a human being, and substations are to the system what joints are to the skeleton. Similarly, a healthy power grid is the prerequisite for efficient operation of the electric power system.

The traditional operation mode of power grid devices will be changed in the future. The access of large-scale yield of new and distributed energy will affect the health and service life of devices. The traditional power grid inspection mode, which mainly includes regular spot checks and emphasizes transmission lines while ignoring distribution lines, can no longer meet the new requirements of the upgraded power system. A faster inspection method that is more comprehensive, frequent, and accurate is urgently needed.

Electric power digitalization technologies can help implement all-dimensional and around-the-clock automatic inspections. By digitizing the power grid inspection, these technologies ensure the safe and efficient operation of transmission hubs, extend the service life of power grid devices, and ensure the power supply security of the entire system.

Intelligent line inspection

Transmission lines are long and the coverage of distribution lines is extensive, posing significant challenges to line inspection. At present, in many areas (especially in remote and mountainous areas where public network signals are weak), transmission and distribution line inspections still rely on manual operations with low efficiency. Unmanned aerial vehicles (UAVs) are deployed in some areas to perform on-site operations, but they are, too, controlled manually. Employees have to determine exceptions based on collected images. In addition, UAVs have dead spots and cannot be used in no-fly zones. They can hardly be used in extreme weather conditions and other line faults that are prone to occur.

To improve inspection efficiency and risk check accuracy, and ensure secure line operation and reliable power supply, we need to deploy smarter and richer inspection methods and faster and more reliable communications networks.

Key technology application 1: space-air-terrestrial integration + edge intelligence, making inspection intelligent

1.**Space and ground side**: Various edge-end collection devices, such as UAVs, radar + PTZ dome cameras, and non-electrical sensors, complement each other to comprehensively sense and monitor tower foundation intrusions and power grid operating abnormalities.

2.**Air side**: The remote sensing capability of LEO satellites achieves full coverage and all-weather high-precision monitoring in real time, complements UAV inspection, which is restricted by region and climate. Satellite remote sensing and telemetry technology will play an irreplaceable role when power grid operation is in urgent need of disaster recovery due to the huge hidden danger caused by extreme weather.

Table 5: Application of various edge-side collection devices in intelligent line inspection

Core Collection Device	Objective
UAV	Determine the distance between the line and the obstacle in real time through infrared thermal imaging/3D modeling.
Radar + PTZ dome camera	Radar: Utilize millimeter wave technology with high precision and strong anti-jamming capability to realize all-weather, high-resolution, and multi-target recognition for dynamic intruding objects. PTZ dome camera: Proactively take snapshots and monitor and track the area coordinates detected by the radar.
Non-electrical quantity integrated sensor	Line running status: temperature, movement, etc. Route safety risks: icing, bird's nest, tree barrier, etc.

3.**Edge intelligence**: To meet the requirements of power supply reliability, power grid lines have high requirements on timeliness and accuracy of monitoring data processing and analysis. The edge side needs to be able to quickly respond to any abnormalities that occur at any time and make decisions accordingly. Therefore, cloud-based training and edge-side execution will be deployed as the standard mode of intelligent line inspection in the future — relying on the powerful computing power of the cloud to achieve learning and modeling of massive unstructured image data and structured running monitoring data and enabling edge-side devices through remote deployment. On the edge side, machine learning-based standardized diagnosis replaces differentiated judgment based on personal experience. It accurately identifies exceptions and automatically generates alarms, improving the power grid inspection efficiency and ensuring tower foundation security and power supply reliability.

Key technology application 2: developing power private networks for reliable communications

Power grid lines have different data communications characteristics due to different geographical locations, voltage levels, and transmission distances. For example, for UHV backbone grids or remote areas without signal coverage, the MS-OTN-based next-generation optical communications technology or OPGW can be used to ensure ultra-long-distance communications transmission over 1000 km. For medium- and long-distance lines, microwave technologies with high reliability, strong anti-interference capability, and high availability in harsh climates can be used to reduce high-cost optical fiber investments and achieve rapid deployment. For low-voltage lines, cameras can be connected in wireless chain mode to implement lightweight deployment, because the communications distance is short. The construction of dedicated power networks based on actual local conditions helps eliminate signal blind spots and improve data transmission efficiency.

In addition, considering the impact of information interruption caused by the network quality of a single network node on the operation of the entire power system, the multi-path transmission protection solution should be implemented, and the capability of edge-side data buffering and resumable data transmission should be deployed to ensure the data continuity and consistency during path switchover.

Key technology application 3: achieving harmonized communications and sensing for next-generation communications

By 2030, with the maturity of 5.5G/F5.5G/6G/F6G technologies, access networks will integrate communications, sensing, and computing capabilities. Key technologies such as sub-millisecond latency, centimeter-level positioning, millimeter-level imaging, and fiber-based precise sensing are adopted to implement harmonized communications and sensing and open a new channel for real-time interaction between the physical world and the digital space.

In the future, the application of high-bandwidth, new edge-side collection devices with high real-time performance, model training based on image recognition and anomaly awareness, and highly reliable communications network that meets power requirements can further solve the problem of power grid line inspection efficiency and achieve intelligent remote inspection.

Intelligent substation

Substations are the transportation hub of electric power transmission and are vital to the power grid system. At present, the substation management mainly uses cameras at fixed surveillance locations. There are dead spots in patrol inspection, and high-risk manual operations are required to supplement the substation management. At the same time, the preventive maintenance method of planned maintenance is still adopted for the substation devices, which leads to the shortening of the service life and frequent replacement of the devices.

Smart substations in the future need to rely on more powerful and flexible sensing devices and more advanced fault prediction models to achieve automatic operation and management, improve the positive effect of device maintenance, prolong the service life of devices, and eliminate unplanned downtime.

Key technology application 1: cloud-device synergy to build a digital team

In the future, smart substations will be unattended. Inspection tasks will be completed by various robots inside and outside substations, such as UAVs, wheeled robots, rail-mounted robots, and cameras at important points. Collection devices as such will implement comprehensive inspection of substations without dead spots and improve the coverage and efficiency of inspections.

In addition, to meet the requirements for high real-time troubleshooting performance of substation facilities, we need to further improve the computing power of devices on the basis of cloud-edge synergy. Replying on edge-device synergy solutions, we need to deploy big data and AI technologies in field operations. With the support of these technologies, intelligent robots can complete tasks such as data collection, filtering, storage, analysis, and mining, diagnose security risks and device exceptions in substations, send corresponding warnings, identify fault areas, locations, and reasons, and complete automatic inspection and maintenance task assignment and emergency handling.

To avoid the impact of communications interruption caused by network interference on substation inspection and management, the local data buffering capability and offline computing capability of devices need to be enhanced to ensure that devices can perform intelligent inspection of substations when they go offline, further ensuring power supply security.

Key technology application 2: advanced intelligence for predictive maintenance and life extension of substation devices

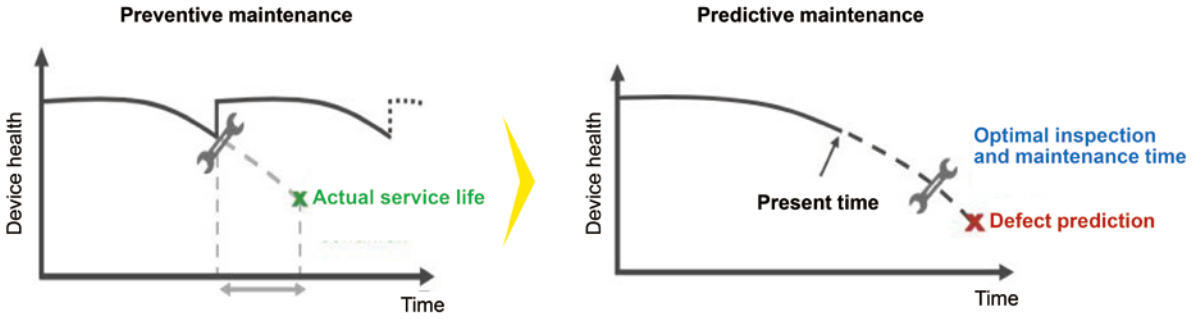
The key to prolonging the service life of substation devices is to accurately determine the time when the device fault occurs and perform targeted maintenance before it occurs. But at present, the maintenance of substation devices is based on planned maintenance, so the mathematical prediction model based on historical inspection and maintenance records cannot reflect the change and trend of device running status.

In the future, more advanced AI technologies need to be explored, and related technologies such as graph computation, advanced analysis, and unsupervised learning can be used to perform deep learning on various related factors, including historical device defect records, test records, and operating status and mechanism models, so as to build more complex device defect diagnosis and prediction models. In addition to comprehensive evaluation and health analysis of the current running status of the devices, based on multi-dimensional influence factor simulation, more accurate prediction of possible future device failure risks and their causes is made, the optimal time point for manual intervention is determined, and corresponding inspection and maintenance solutions and material requirements are provided.

This facilitates the preparation and training of electric power operators in advance.

Robots and AI technologies can be integrated to monitor the substation security and device running status in an all-round manner, and implement fault diagnosis and alarm sending. On top of that, advanced AI technologies support predictive maintenance, build an intelligent 'brain', and implement the fast sensing, decision-making, and response of substation device faults.

Figure 6: From preventive to predictive maintenance



Summary of digital technology application in intelligent power grid inspection

Judging by the current application of key enablement technologies, the usage of intelligent devices varies in the power grid inspection in different areas. Currently, the identification of and response to faults are passive, and plan-driven preventive maintenance inspection is adopted for device inspection and maintenance in general. The key directions of research and breakthroughs in the future are as follows:

- **Improving the intelligent device inspection rate**: Adopt intelligent devices for the inspection of all power grid lines and substations.
- **Improving the communications efficiency**: Achieve full coverage of power private networks, reduce the transmission latency to milliseconds, and eliminate signal coverage holes. In addition, explore and promote the application of next-generation technologies, and further promote harmonized communications and sensing.
- **Improving the edge autonomy capabilities**: Build a cloud-edge-device synergy architecture, move the application of computing power and algorithms forward, improve the timeliness of exception diagnosis and response, and enhance local cache and resumable data transfer capabilities.
- **Improving the coverage of AI**: Strengthen AI training to improve the accuracy of anomaly identification of different power grid devices and reduce the rate of false alarm reporting and missed detection rate. Explore advanced AI applications (for example, graph computation), and improve the proportion of predictive maintenance of substation devices.

Scenario 3: multi-source self-healing distribution network

Urban distribution network is a bridge between the power grid and end users. The safe and stable operation of the distribution network is the key to ensuring the daily operation of enterprises and the living and working of residents. Along with changes in the energy structure and supply mode, the focus of power supply has shifted from the backbone network to the power distribution network, requiring a flexible and stable future urban distribution network that can respond to emergencies. Applying digital and intelligent technologies can build a distribution network that can satisfy rigid demands through flexible adjustment, and improve the power grid's capability of consuming new energy at the load side.

Multi-source distribution network operations

With the access of a high proportion of distributed power supplies and diverse loads, the urban distribution network will become active, multi-directional, and involve high proportions of renewable energy sources and power electronic equipment.

Intermittent outputs of distributed new energy devices, sudden voltage changes and power flow changes caused by reverse power supply, and harmonic pollution caused by power electronic devices to power grid operations pose new challenges and requirements on the operation and management of the urban distribution network. Therefore, enhancing the stability of the power grid in normal scenarios is one of the core objectives of the operation and management of the future active distribution network. Deeply integrating power electronic technologies and ICTs will support flexible power access, ensuring that the power grid runs properly and orderly in response to power source and load fluctuations and random disturbances, reducing the fault rate to the largest extent.

Key technology application 1: standard access for plug-and-play

Distributed power supplies are converted through power electronic devices such as inverters first and then input into the power distribution network. However, designed based on actual application scenarios, distributed power supply devices feature varying topology architectures, electrical ports, and communications protocols. This increases the difficulty in distribution network management.

Therefore, power electronic device technologies need to be upgraded to develop unified power conversion devices that support high power density, multiple electrical interfaces, and independent parallel connection of multiple modules. This will help expand the access capacity of the power supply system and improve the standardization, maintainability, and interchangeability of various switch-mode power supplies. Unified information models and IoT protocols will eliminate device differences locally, significantly shortening the access commissioning time and providing a hardware foundation for ICT-based global monitoring, coordinated scheduling, and collaborative control of source, grid, load, and storage in new distribution network systems that are active and multi-dimensional.

Key technology application 2: from terminal intelligence to edge intelligence

Some problems arise during the upgrade of the new power distribution network. For example, the interfaces of primary and secondary equipment do not match, and the devices from different vendors are incompatible. These problems hinder devices' function expansion and the improvement of the running level and efficiency of power

distribution devices.

Fusion of primary and secondary equipment is an effective way to solve these problems. This means integrating primary equipment with some intelligent units of secondary equipment to make the equipment more intelligent. For example, the switch device can be integrated with the measurement and monitoring function to proactively detect and monitor power quality parameters such as the phase voltage, phase current, zero-sequence voltage, zero-sequence current, and harmonic pollution on both sides. Optimization algorithms and feature libraries are used to analyze and locate pollution sources after distributed power supplies are connected to the power grid, supporting precise management of power quality.

However, involving a wide range of areas, equipment upgrade is costly and difficult to implement. Edge technologies may provide more economical and practical solutions. Devices adopt unified information models and IoT protocols to ensure real-time data monitoring and effective data aggregation. In addition, the edge computing capability is improved, and device-end data is processed in a centralized manner to ensure better computing performance while reducing the requirements for intelligent terminals.

Edge intelligence also plays an important role in orderly and flexible charging of electric vehicles. When an electric vehicle accesses a power distribution network through a charging pile, the charging pile collects vehicle-end data such as the charging power and remaining battery capacity in real time, and uploads this data to the edge gateway. Within the maximum available capacity allowed by the power distribution plan, the edge gateway flexibly adjusts and controls the charging power and charging start/stop time of each single charging pile based on the actual access status of different charging points to develop the optimal charging policy in the transformer district.

Key technology application 3: from cable communications to fiber communications

Traditional edge-side device interconnection depends on communications cables. However, as there are many distributed power supplies that are located across a wide area, a large investment is required, and cabling is very difficult, raising high requirements on local communications. High-speed power line communications (HPLC) integrates communications cables and power cables. This way, devices can be connected as long as power cables are available. This effectively solves the problem of difficult interconnection between far-end devices, implementing efficient interconnection and high-frequency communications with zero wiring. Compared with the traditional low-speed narrowband power carrier, HPLC uses the 2 MHz–12 MHz frequency bands to achieve a transmission rate greater than 1 Mbit/s and a network latency less than 30 ms.

In the future, as more power supplies are connected to the power grid, the amount of data needing sensing and monitoring will increase exponentially, and the power distribution network will have higher requirements on the communications bandwidth and latency. Fibers support an access speed ranging from the Gbit/s level to the Tbit/s level and a network latency less than 1 ms. After the communications network is upgraded from the WAN to the transformer district that is deployed closer to the edge, the operational efficiency of new multi-source distribution networks will be significantly increased.

Converging power electronic technologies and ICTs enhances edge intelligence and communications upgrade, so as to achieve wide access of distributed power supplies and stable running of power distribution networks, increase new energy consumption, and ensure reliable and secure power supply.

Self-healing distribution network control

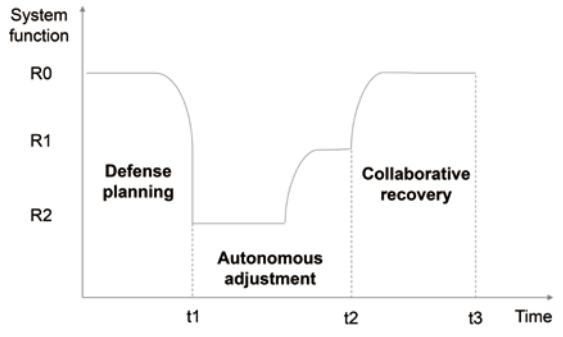
In addition to ensuring power grid stability in normal scenarios, the urban distribution network needs to improve the power grid's response to emergencies and recovery from extreme events.

Extreme events feature a small probability, large impacts, and uncertainties, easily causing large-scale power outages. Therefore, enhancing the self-healing capability of power distribution networks is critical for preventing the impact

degree and scope of faults caused by extreme events from increasing and for reducing the impacts on power users.

Self-healing refers to power distribution networks' capability to quickly adapt and respond to and recover from faults when extreme events occur. The construction and implementation of self-healing power distribution networks include three phases. The first phase is before an extreme event occurs, when the power distribution network needs to be upgraded through line reinforcement, distribution network architecture enhancement, and distributed power supply access, so as to improve the resistance of the power distribution network to extreme events. The second phase is after an extreme event occurs but before it ends, when the faulty devices or lines should be quickly isolated to ensure secure and stable running of unattacked modules, accurately locate faults, and promptly rectify them, improving the adaptability of the power distribution network to extreme events. The third phase is after an extreme event occurs, when flexible power supplies such as distributed power supplies, energy storage devices, and electric vehicles are used to support load recovery based on power supply priorities through demand management, improving the recovery capability of the power distribution network.

Figure 7: Power supply curve of the self-healing distribution network



Key technology application 1: Comprehensive sensing and detection provide a data foundation for self-healing distribution network construction.

A device and line status awareness system is the basis for improving the self-healing capability of the distribution network. This system uses sensor terminals to comprehensively detect and monitor the internal running status, external situation changes, and users' energy consumption status of the distribution network system, including electrical, status, and environment parameters, in real time. This data provides decision-making support for device risk identification, fault locating and repair, and post-disaster recovery.

Best practice: AIoT helps identify risks and locate faults more quickly and more accurately.

An IoT solution provider in China uses high-precision electronic current transformers to detect line currents and ground electric fields in real time, comprehensively supporting distribution network fault identification and diagnosis. When an anomaly is detected, the terminal automatically triggers high-precision sampling of transient recording to collect the anomaly data and then sends this data back to the primary station. The cloud platform deployed at the primary station quickly classifies the recording files received through machine learning, accurately locates the faulty section and identifies the fault cause based on the line topology, generates alarms in a timely manner, and notifies maintenance personnel of the fault locating result. This significantly shortens the fault locating and response time and quickly restores power supply in the faulty area.

This solution provider has deployed nearly 5000 sets of stable and reliable devices, achieving a 1% line current measurement precision, a nearly 90% grounding fault detection accuracy, and a 100% short-circuit fault locating accuracy.

Key technology application 2: Optical fiber communications networks implement fast and accurate load control.

Distribution network communications technologies are not as advanced as those for power transmission lines. Inadequate optical cables and poor communications conditions restrict the power supply security of power distribution networks. Wireless communications technologies represented by 5G have helped improve the secure operation of distribution networks. 5G drives the development of next-generation optical communications technologies. OLT is integrated with OTN, improving the information communications capability of the

distribution network. With no protocol conversion required between the access network and the transport network, data is transmitted at an ultra-low delay. When extreme events occur, non-critical and unnecessary loads are quickly cut off to ensure power supply reliability of the power distribution network and defend against various faults or disturbances. Sensing and digital technologies will drive the preceding three capabilities of self-healing power distribution networks to improve.

Key technology application 3: Machine learning helps with better response and faster recovery.

1. **Extreme event prediction**: Machine learning helps model the occurrence probability and frequency based on different event types and construct the association between different events and the component failure rate of distribution network devices. Based on the calculation result of event and fault prediction models, power companies can quickly take measures when extreme events occur and better allocate maintenance engineers.

2. **Isolated island division policy**: The self-healing distribution network implements isolated island management when an extreme event occurs. This means dividing the power outage area of the target distribution network into several isolated islands based on the type, capacity, location, and importance of the local power supplies and loads connected to the distribution network. Each of these islands covers one or more power supplies.

The corresponding algorithm model is continuously optimized through machine learning to develop the optimal island division policy for different extreme events or faults. This ensures that the power supply of as many loads as possible is restored in the shortest time and helps quickly switch back to the grid-connected mode after faults are rectified, minimizing the loss caused by power outages on the distribution network.

The status sensing and communications infrastructures for distribution network devices are constantly upgraded so that the distribution network can defend against and respond to faults more proactively and more timely. In addition, AI and big data further improve the effectiveness of fault response and shorten the fault duration to quickly restore power supply.

Summary of digital technology application in the multi-source self-healing distribution network

Judging by the current application of key enablement technologies, the sensing and communications infrastructures of the current urban distribution network faces many difficulties in promptly and effectively responding to changes and harmonic pollution caused by grid connection of distributed new energy sources and storage devices.

The key directions of research and breakthroughs in the future are as follows:

- **Improving the edge adoption rate**: Enhance the edge computing capability to process the running data of massive access devices in a centralized and efficient manner, reducing the cost of reconstructing power distribution devices. It is predicted that by 2030, ECU edge intelligence will achieve 100% coverage.

- **Improving the real-time performance of communications**: Upgrade the power distribution network with optical fibers to achieve the high- speed transmission of massive data with a low latency (Tbit/s-level rate and ms-level latency) and improve the system operating efficiency.

- **Predicting local communications coverage**: Enhance the communications capabilities of the 400 V low-voltage distribution network, achieve 100% coverage of the next-generation HPLC, support the transparency of the distribution network, and promote the development of new businesses such as virtual power plants and orderly charging of electric vehicles.

- **Improving AI coverage**: Enhance machine learning training to accurately locate faults and predict extreme events, supporting response to faults and accelerating recovery from disasters. It is predicted that by 2030, AI will achieve 85% coverage in the distribution network.

Scenario 4: coordination and complementarity of multiple energy sources

As terminal electrification accelerates and energy storage technologies develop, electric power will be the core of the future power supply system. In a certain region, constructing micro-grids or micro-energy grids to further integrate electric power digitalization with energy technologies can ensure flexible conversion and synergy among various energy sources, such as electric power, heat, and gas. This helps improve the reliability of regional power supply and the integrated energy utilization efficiency and service quality, finally achieving net zero emissions.

Micro-grids or micro-energy grids feature the integration of power source, grid, load, and storage. They can be operated independently or connected to upper-level power grids through switches for power exchange. Smart campuses and smart buildings are two typical application scenarios of micro-grids or micro-energy grids.

Smart campus

Now, the focus of urbanization changes from high speed to high quality, making comprehensively promoting digital transformation of cities an important task of city construction. Smart campuses are major venues for industry convergence, production, and life activities in cities, so they lie in the heart of smart city construction.

Next-generation digital technologies like cloud, big data, IoT, AI, and mobile communications are being applied in more and more scenarios and will be integrated with various power electronic devices in the future power grid. This will help streamline the energy consumption management process of monitoring, analysis, prediction, and optimization and the power management process of inspection, warning, and handling for smart campuses. (For details about key technical support for power quality monitoring and fault locating on the distribution network, see scenario 3.) Besides, multi-energy collaborative dispatching and tiered energy utilization are implemented to improve comprehensive energy utilization, finally building smart zero-carbon campuses.

Key technology application 1: Energy routers integrated with ICTs facilitate flexible device access and precise carbon emission measurement.

1. Data collection: There are many types of energy supply and consumption systems in a campus, each of which includes many devices. For example, the power supply system includes distributed PV devices, wind power devices, and electrochemical energy storage facilities; the heating system includes combined cooling, heating and power (CCHP) supply devices, heat pumps, and heat storage facilities; the gas supply system includes gas supply stations and hydrogen energy storage facilities; and the power load system includes factories, buildings, electric vehicles, and street lamps.

   Energy routers capitalize on power electronic conversion and control technologies to provide diverse electrical interfaces for various devices, and adopt standardized protocols to let different terminals integrate and interconnect. This way, energy routers become core devices and energy hubs in micro-grids. When combined with digital technologies such as 5G, energy routers will be capable of communications and intelligent decision-making support in addition to their basic metering and control functions. They can collect and transmit information such as device running status and energy usage in real time to implement unified data collection

and classified metering. They can also actively or independently manage the energy flow direction and power as instructed by users or dispatching centers.

2. Data application: The energy consumption data that energy routers collect and summarize in real time can be used to accurately capture and follow carbon footprints. Measurement factor rules can be preset to accurately measure and monitor a campus' total carbon emissions in real time, ensuring that the data is reliable and trustworthy. This lays a foundation for IOC-based visualized carbon management, carbon quota management, carbon asset trading.

Key technology application 2: Intelligent algorithms and big data help achieve automatic energy optimization and multi-energy collaborative dispatching.

1. Energy saving: Massive energy consumption data of various loads in a campus is collected in real time and classified for measurement. Energy consumption models of different loads are trained, which, together with big data analysis and edge computing, helps intelligently control and optimize energy consumption behavior such as air conditioning and lighting in the campus, achieving energy saving.

2. Flexible supply: AI algorithms are used to predict the output of various distributed power supplies and the load of various power consumption systems in a campus. This data is then considered together with various other factors such as meteorological prediction, power price changes, and response requirements of power users to calculate a global intelligent dispatching solution and deliver dispatching control instructions for the output time and power of various energy supply systems in the campus. Besides, the dispatching policy is constantly optimized and adjusted based on the prediction of the impact of short-term climate changes on power outputs and loads.

   When the power supply exceeds the demand, the excess power can be reused many times level by level through energy transformation technologies and various energy storage devices. In addition to directly storing the excess power through electrochemical energy storage, the power can be converted into hydrogen energy through new technologies such as Power2Gas. This has two benefits: One is storing energy using hydrogen storage facilities, and the other is reducing gas procurement costs through hybrid gas supply. When the power supply is less than the demand, the power stored can be released, or redundant heat energy and hydrogen energy can be transformed into power. This way, multiple energy sources, such as electricity, heat, and gas, complement each other, increasing campuses' self-sufficiency capability of energy supply.

With energy routers, telemetry, teleindication, and remote management are performed on massive power-consuming terminal devices in a campus. Based on energy consumption analysis and prediction, energy supply devices and loads are remotely adjusted to implement optimized dispatching and energy sharing based on a global policy in the WAN, improving comprehensive energy utilization.

Smart building

Buildings are one of the core infrastructures of a city, and their intelligentization is also an important part of smart city construction. Buildings empowered by digital technologies such as Digital Twin and AI become sensitive and smart. Together with high-quality construction and operations of energy storage devices, smart buildings will create a more comfortable, secure, energy-efficient, and environmental-friendly working and living space.

Key technology application 1: energy consumption management based Digital Twin

Device awareness and BIM help accurately measure and clearly display the real-time energy consumption and changes of each area. Energy consumption behavior is analyzed to facilitate operation decision-making. Based on this energy consumption data and some other data such as the people flow, environmental detection, air quality, and sunlight intensity, the system automatically generates and delivers control

instructions like device startup and shutdown as well as temperature and light brightness adjustment using advanced algorithms, and implements flexible control and adjustment by device and area through smart gateways.

Best practice 1: The digital building platform helps partners reduce power consumption and costs.

A high-tech manufacturing enterprise independently developed a digital building platform that integrates software and hardware, cloud and edge, and strong and weak currents. Owning multiple innovative patented technologies, this enterprise has become a leader in smart store operations and management.

The enterprise's digital building platform solution includes the following key features:

1. The solution supervises and controls energy consumption by socket. Specifically, it monitors power and temperature in real time, and generates real-time alarms for exceptions. It remotely controls each socket in real time or as scheduled to ensure security.

2. The solution draws on intelligent algorithms to automatically control the running status and pre-settings of the air conditioning and fresh air systems as well as the illumination and startup and shutdown time of lighting devices. It also supports one-click remote management by area or device.

3. The solution supervises key indicators such as environment and device running parameters, and diagnoses, evaluates, and predicts the environment and key device health based on algorithmic models and multi-dimensional data analysis. For predicted defects, the system generates O&M alarms four days to two weeks in advance to avoid abnormal downtime.

The enterprise has cooperated with many large shopping malls, supermarkets, and franchised stores around the world. One of its customers expects to save 11,000 kWh energy annually after deploying the digital building platform, reducing the annual energy consumption per unit area from 343.7 kWh to 281.1 kWh per square meter.

Key technology application 2: Diverse smart energy storage methods facilitate multi-energy, bidirectional, and flexible allocation.

With the emergence and increase of passive houses, heating, ventilation, and air conditioning (HVAC), which accounts for 50% of a building's energy consumption, will be replaced by a new heating mode that mainly uses natural heat sources and is supplemented by electrical heat pumps. This will significantly reduce the energy consumption. For buildings deployed with distributed PV devices, there is a high probability for them to change from controllable loads to adjustable power sources. These buildings can be used as large-scale energy storage facilities to participate in the operation of the power distribution network, achieving peak load shaving.

Power consumption and reverse supply policies for smart buildings can be developed based on the observation and prediction of power price changes and users' response requirements. Smart gateways can be used to adjust the charge and discharge switches and directions of energy storage devices behind meters. This way, power can be purchased and stored during the valley period and released for usage or sold during the peak period to reduce energy costs and increase the cost-effectiveness of smart buildings.

Besides, the waste heat recovery and storage facilities for passive ultra-low energy consumption buildings can implement the reverse supply of heat energy when the heat source is adequate. They even support the mutual transformation between heat energy and electric power to supply and store energy for smart buildings in a way with multi-energy synergy and optimal benefits, upgrading from electric power prosumers to multi-energy prosumers.

Best practice 2: a multi-energy complementary building with near-zero energy consumption

A world-leading industrial digitalization solution provider from Europe built a passive house technology center covering a total area of 13,800 square meters in a northern coastal city of China. This passive house building can keep the room temperature above 20°C in winter without using

the active heating and air conditioning systems. It creates a comfortable indoor environment with nearly-zero energy consumption.

According to statistics, the building can save primary energy of nearly 1.3 million kWh each year, reducing carbon emissions by 664 tons, equivalent to the carbon sink of 53,000 trees.

This company is developing the smart building solution business in many regions around the world. In some regions where electric heating systems are widely used, it has begun to explore how to use buildings as batteries and connect them to the power grid as part of the distributed energy system. The excess heat energy of buildings within a period of time is recycled and transformed into electric energy for storage. During peak hours, this stored electric energy is sold as a supplement, implementing peak load shaving.

Summary of digital technology application in multi-energy coordination and complementarity

Judging by the current application of key enablement technologies, devices in the incremental campus network are well connected. However, there are still a large number of dumb devices and heterogeneous networks in existing campuses, which hinders global management and precise control. Distributed power supplies and load terminal devices are still connected through unidirectional switches, which cannot support the synergy of power source-grid-load-storage to improve the energy efficiency. In addition, lacking the computing power support of data centers, campuses and buildings cannot fully respond to the changes of power source, grid, and load in real time, or efficiently make dispatching-related decisions.

The key directions of research and breakthroughs in the future are as follows:

- Accelerating the upgrade to bidirectional energy control: Deploy more energy routers based on flexible substation technologies to meet the on-demand control and adjustment of the electricity flow direction and power, help distributed power supplies better connect to the grid, and provide power electronic technical support for various micro-grids to participate in the operation of the power grid through reverse power supply.
- Building edge capabilities: Improve the computing power at the edge to support data storage, processing, and application, reduce bandwidth occupation and latency caused by the backhaul of massive device running and load data, and improve the real-time computing performance. Strengthen the AI training based on the energy dispatching model where various energy sources are complementary and matched with storage to optimize operations. Simplify the edge-side application of models trained on the cloud through cloud-edge synergy to improve edge intelligence in scenarios lacking the computing capability support from data centers.
- Accelerating the upgrade of campus networks: Increasing access devices on the campus network require an upgraded campus broadband communications network to ensure a high bandwidth and a low latency for data collection and transmission on the edge side.

Scenario 5: cross-region power dispatching

Since 2021, energy shortages have occurred in many countries and regions around the world due to various factors, such as frequent extreme weather conditions, rising fuel prices, and continuous spread of the pandemic. As a result, these countries and regions have to take non-seasonal power rationing and other measures to ensure power supply, causing a serious impact on urban production and people's lives.

Cross-region power dispatching is a solution. Through reasonable dispatching, the power in low-load areas or areas with power surplus can be transmitted to power-rationing areas to relieve the power shortage. There are two ways to implement this. One is using the intelligent power grid dispatching system to connect and coordinate the power grids in different regions. The other is using the virtual power plant platform to redistribute power.

Intelligent power grid dispatching

Different countries and regions have different concerns about power grid dispatching. For example, in China, large wind and solar power bases are constructed and put into operations, and power supply and demand mismatch in many areas. Given this situation, UHV backbone power grids will become the most important way to consume the power generated by those large wind and solar power bases and transmit the large amount of power generated using new energy sources over a long distance. In 2022, the Chinese government clearly proposed to invest more in planning and building a new energy supply and consumption system based on large wind and solar power bases, the clean, efficient, advanced, and energy-saving coal power plants around them, and stable, secure, and reliable UHV power transmission and transformation lines. Investment in UHV lines and power transmission and dispatching around UHV power grids will become a trending topic in the industry.

Computing capabilities are the core drivers for efficient, intelligent dispatching of power grids in an environment with massive data, regardless of the power system form and the focus of power grid dispatching.

Key technology application 1: Strong computing capabilities support massive data processing.

Power generation devices that use new energy sources feature large-scale, intermittent, and fluctuating outputs. After they are connected to the power grid, the amount of data to be processed during power dispatching increases exponentially, urgently needing large and ultra-large data centers to provide powerful computing support for stable storage, high- performance computing, and precise analysis of massive power grid operation data. The convergence of electricity+computing provides space for applying technologies such as deep learning, model training, and graphics processing. This helps unveil the power grid operation rules in the new energy era behind massive data, promoting the construction of digital grids.

Key technology application 2: Optical networks support secure, reliable, and real-time communications.

A highly secure and reliable communications network with a large bandwidth and a low latency is required for data centers to invoke massive power grid operation data and for power transformation hubs to receive real-time power dispatching data. Next-generation optical communications networks have strong anti-interference capabilities, can recover from faults in 50 ms, provide a large single-fiber capacity, and support long-distance data transfer. These characteristics make them the best communications method for future digital grids,

ensuring quick response and execution of power grid dispatching.

The strong computing and efficient communications capabilities will support the power grid to implement more accurate power dispatching globally.

Virtual power plant

A virtual power plant does not change the structure of physical networks. Instead, it aggregates scattered and independent power supply devices, energy storage devices, and controllable loads through the software platform. Through flexible dispatching management and efficient interaction with power grids, the virtual power plant can supply power to the power system as a "positive power plant" and consume surplus power in the system as a "negative power plant" to realize multi-spatial power balance, improve power grid security and new energy consumption capabilities.

Three core technologies are required to efficiently operate a virtual power plant: metering, communications, and dispatching.

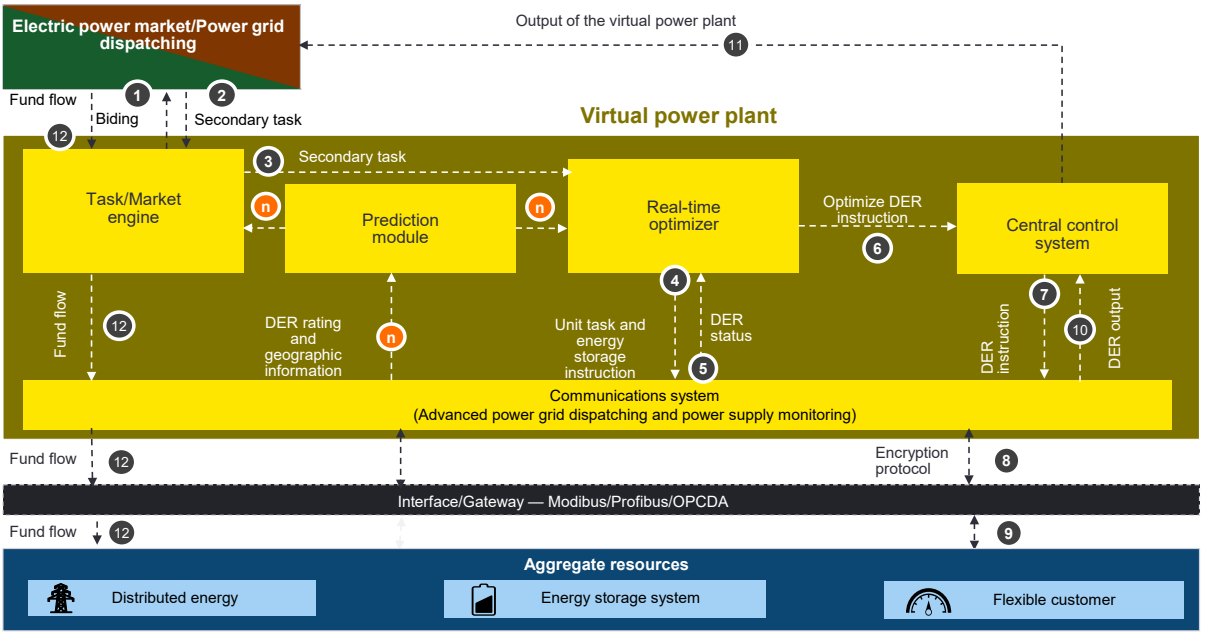
Key technology application 1: AI and big data enable optimal dispatching. Intelligent dispatching decision-making is the core capability of a virtual power plant.

Proper and effective dispatching arrangements for aggregated power supply and energy storage devices can ensure balanced operation of the power grid, improve energy utilization, and improve the economic benefits and participation enthusiasm of each party.

Dispatching of the virtual power plant is decided based on two elements:

1.**Supply-based dispatching**: There are two ways for virtual power plants to participate in electric power market trade. First, determine the adjustable power that is directly included in the bidding scope in the electric power market based on the power generation capacity of adjustable power supplies. Second, provide auxiliary power services such as peak shaving, valley filling, and frequency modulation based on the demand response requirements of the power grid to suppress power grid fluctuation and ensure power grid balance. In this mode, a key factor to support dispatching decision-making is to accurately judge the output change of the adjustable power supply or the requirement change of the controllable load.

Figure 8: Operation mode of the virtual power plant



AI technologies will greatly improve the accuracy of prediction models. The mutual impact relationship can be established between climate, power generation performance, and power consumption requirements through sparse modeling, ensemble learning, or other machine learning methods. The supply and demand curve can be accurately predicted based on future weather changes.

As the multi-source distribution networks are complex, the running status of distributed power supplies or energy storage devices may change or adjust at any time. Therefore, virtual power plant operators also need to monitor the status of the aggregated adjustable power supplies in real time. When the device status or output status changes, the prediction model parameters should be updated and adjusted in a timely manner, and the output policy should be dynamically optimized based on the real-time prediction result to support flexible dispatching.

2.**Price-based dispatching**: Since the power supplies aggregated in virtual power plants have different characteristics, their output curves are different. Based on the output prediction of the adjustable power supply and the price prediction of the electric power market, virtual power plant operators have different dispatching arrangements for different units through data modeling to improve the economic benefits of related parties and themselves.

Take the German electric power market as an example. For biomass, combined heat and power (CHP), and other units with stable output, the output change of the units can be consistent with the spot price trend in the electric power market. Therefore, such units can be used to generate electricity only during peak hours (when the price is higher). For new energy devices such as PV and wind power that are dependent on weather conditions, dispatching needs to be arranged based on the prices updated every 15 minutes in the electric power market.

AI and big data technologies help virtual power plant operators accurately predict and dynamically optimize adjustable power. In addition, parties of the virtual power plant can get more benefits by matching the adjustable power output and the price trend in the electric power market.

Key technology application 2: Unified terminals and standard protocols facilitate remote dispatching and control.

Comprehensive sensing, precise metering, and efficient communications of aggregated power supplies and energy storage devices provide data basis for intelligent dispatching decision-making of virtual power plant operators. Leading virtual power plant operators usually install unified terminals at each power supply or energy storage device before reaching an access agreement, and provide standard communications protocols such as Modbus, Profibus, and OPCDA. In this way, data of scattered and different types of devices can be efficiently obtained and dispatching instructions can be delivered in a timely manner.

Best practice 1: virtual power plant operations on the power generation side

The core objective of virtual power plants that mainly aggregate resources at the generation side is to improve the grid connection and consumption of new energy.

A virtual power plant in Europe mainly uses adjustable power resources such as distributed new energy devices, biomass, CHP, and small hydropower. The plant installs remote control devices at the power supply end to evaluate the energy yield free of charge. In this way, the power supply and other running parameters can be integrated into the central control system of the virtual power plant platform, and data about generation capacity is obtained.

Virtual power plant operators remotely control power supplies based on the running parameters, power grid status, and power prices in the electric power market. Distributed new energy devices generate power intermittently and the power will be directly included in the bidding scope in the electric power market.

Power generated by adjustable power supplies can be included in the bidding scope of the electric power market and sold in the power balancing market to gain the electrical capacity charge and frequency modulation service fee.

The power resources aggregated by the virtual power plant operator are equivalent to four 600,000-kW thermal power units, accounting for 10% of the local power balancing market.

Best practice 2: virtual power plant operations on the load side

The core objective of virtual power plants that mainly aggregate resources at the load side is to respond to demands and reduce energy consumption.

A virtual power plant focuses on controllable loads. Based on the behind-the-meter energy storage devices at the user side, the virtual power plant builds an ecosystem that integrates "vehicles, piles, PV, storage, load, and intelligence". The virtual power plant operator invites load-side users who meet the requirements to join the platform. When the power supply is insufficient to meet the increasing power consumption demand, the virtual power plant mobilizes the power in the energy storage devices to support the operation of the power grid. Users who respond to the load can obtain direct economic incentives from the virtual power plant.

In less than two months after the project was started, 2500 users have been connected to the grid, and the grid-tied solar power capacity reaches 16.5 MW, providing power equivalent to that of a small power plant. In the future, the virtual power plant operator will promote this operation mode in a wider area.

Best practice 3: operations of a virtual power plant that integrates source, grid, load, and storage

The objective of a virtual power plant that integrates source, grid, load, and storage is to increase the overall energy utilization efficiency.

In the first phase of China's first virtual power plant pilot project, the intelligent management and control platform is used to connect and control adjustable resources across three provinces in real time, covering 11 types of objects, such as regenerative electric heating, adjustable smart campus, smart building, smart home, energy storage, electric vehicle charging station, and distributed PV. The total capacity is about 160,000 kW. Through computing and storage of device data and interactive information, the plant integrates multiple functions such as energy operation management, trading, and services, enabling real-time interaction with the power system.

From the perspective of operation effect, the virtual power plant supports nearly 10% of the load required by air conditioners for the local power grid in summer through demand response. In winter, it is estimated that 720 million kWh of clean energy can be generated but 637,000 tons of carbon dioxide less will be produced, efficiently utilizing clean energy and flexibly adjusting power.

Summary of digital technology application in cross-region power dispatching

Judging by the current application of key enablement technologies, the construction of data centers and the upgrade of communications networks in the electric power industry are still at an early stage. Besides, virtual power plant operations are also at the exploration and pilot stages.

The key directions of research and breakthroughs in the future are as follows:

- **mproving the computing capability and real-time communications performance to support power grid dispatching**: Improve the computing capability and computing efficiency of data centers. Continuously improve the power communications network.
- **Improving the application of AI in virtual power plant operations**: Improve the prediction accuracy of output of adjustable power supply devices, and reasonably develop dispatching plans.

Scenario 6: green and low-carbon enablement

To achieve dual-carbon targets is extremely challenging in the electric power industry. In addition to replacing traditional energy sources with renewable ones and building a circular economy, digital technologies and platforms play an increasingly important role by drawing on their advantages in terms of coverage, real-time performance, credit enhancement, positive incentive, etc.

Carbon trading

The core objective of carbon trading is to transform the environment into a paid factor (cost) of production with the help of market forces. Valuable assets such as carbon emission rights and green electricity will be regarded as commodities in market trading to stimulate new energy power generators and power users and gradually increase the proportion of green power consumption.

In China, for example, the current carbon trading system includes three commodity forms: carbon credit, China Certified Emission Reduction (CCER), and green electricity. The carbon credit specifies the upper limit of carbon emissions for cap-limited enterprises. Enterprises not limited can apply to regulatory authorities, and after being approved, use the CCER to offset the excess emissions of cap-limited enterprises. Green electricity can reduce the demand for power generated by traditional energy sources, thereby directly reducing carbon emissions. The three types of commodities complement each other and play a positive role in the low-carbon production and operation of the entire power system. Green electricity, which is directly related to the electric power industry, is an independent trade commodity designed for new energy such as PV and wind power under the mid- and long-term electric power trade framework. In addition to the acquisition by the power grid, a new channel is created for supply and demand parties to directly trade new green electricity. To further improve the consumption level of new energy electricity, the Chinese government has given many priorities to green electricity in terms of policies, such as preferential organization, arrangement, execution, and settlement. In the future, more entities will be attracted to participate in the market-oriented trade.

Key technology application: Blockchain enables authentication of green electricity and accelerates green electricity consumption.

In the early stage of green electricity trade, there are many challenges during green electricity production, trade, settlement, and certification, such as repeated metering, check, and verification, and even data falsification.

The blockchain technology features decentralization, anti-tampering, openness, and transparency to ensure data and transaction reliability. Specifically, the distributed ledger function of the blockchain technology helps store data in the entire process. The consensus mechanism supports mutual verification and proves data authenticity. The smart contract function is used to complete trade automatically and settlement efficiently. The electronic signature is used to issue green certificates, supporting accurate tracing and full-lifecycle tracing of each trade of green electricity and improving the authority of green electricity consumption certification.

The blockchain technology is of crucial significance in building an efficient, fair, and open green electricity trade market.

Carbon inclusion

Carbon inclusion refers to the quantification of green and low-carbon behavior of the public and small and micro enterprises, and the

establishment of a positive guidance mechanism combining policy incentives, business incentives, and certified carbon emission reduction trades.

Currently, each carbon inclusion platform independently develops the methods for energy saving and carbon emission reduction in the electric power industry. There is no unified industry standard. As a result, some green and low-carbon behavior is not included in the carbon inclusion statistical scope. For example, a user installs distributed PV devices and uses new energy to save energy and reduce carbon emissions. However, from the perspective of the power system, the user only saves electricity fees. A unified standard is need to promote the positive guidance of carbon inclusion.

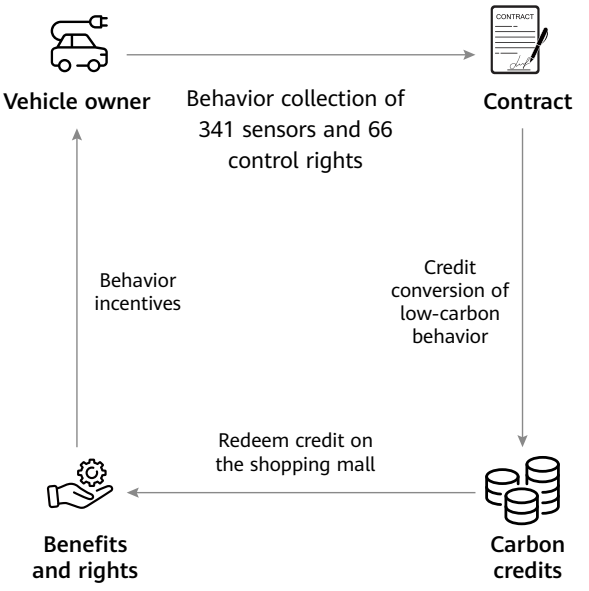
If a methodology is formed, a challenge we will face in promoting the carbon inclusion mechanism is how to take unified and authoritative certification of eligible public behavior.

Key technology application: Blockchain standardizes carbon reduction behavior certification in the electric power field.

The carbon inclusion ecosystem has many participants and complex processes and covers a wide area. The blockchain technology matches the ecosystem. To enhance data authenticity, distributed ledgers, consensus verification, smart contract, and other blockchain-related technologies are used to record energy saving and carbon emission reduction behavior that meets the methodology of the electric power industry, as well as reduced quantity of carbon emissions, and considerations for reduced carbon emissions and flow directions. At the same time, IoT technologies are used to automatically sense, collect, and upload data to the chain, further reducing manual intervention and ensuring data authenticity. In addition, privacy protection technologies can effectively protect personal information while making data open and transparent.

Best practice: carbon credit application system for new energy vehicle enterprises

A new energy vehicle enterprise in China uses an open ecosystem platform built in the vehicle to connect 341 sensors and 66 control rights of the vehicle and record related information in the decentralized blockchain in real time. The smart contract function automatically converts vehicle driving data and emission data into carbon credits. In the carbon credit application system, vehicle owners can redeem credits to enjoy benefits and rights brought by carbon emission reduction, which provides positive guidance and incentives for low-carbon behavior.



Summary of digital technology application in green and low-carbon enablement

Judging by the current application of key enablement technologies, the blockchain technology is widely used in carbon inclusion scenarios, but most leading players are government agencies or platform service providers. The green electricity trade market is still in the pilot phase, and the blockchain technology is in the technical research, patent application, and standard formulation phase.

The key directions of research and breakthroughs in the future are as follows:

- **Improving the coverage of blockchain**: Establish a green electricity trade mechanism,

and develop regional power wholesale markets to implement on-chain transaction. In scenarios that directly involve end users, such as smart campuses and load-side virtual power plants, explore the application of the regional blockchain in fields such as power retail and carbon behavior certification.

Summary

Based on the application of technologies enabling electric power digitalization in six core business scenarios of the future power grid, the value of core technologies is described below:

1. **Device**: The proportion of connected devices decides the data collection efficiency and device controllability at the edge side. In the future, we need to unify interfaces and standardize protocols to support ubiquitous perception of various power devices and connect to all key devices.

2. **Edge**: Cloud-edge-device synergy can improve data processing and analysis. In the future, we need to improve the coverage of the cloud-edge-device synergy architecture, improve the edge and device data adoption rate, and balance the timely business response and accurate data analysis based on business needs.

3. **Pipe**: A private communications network for electric power can improve data transmission efficiency and support reliable power supply. In the future, we need to achieve high bandwidth (Gbps- or Tbps-level) and low latency (ms- or μs-level) to ensure the efficient transmission and processing of numerous data. The construction of the power IoT will improve the access and security of the vast amount of sensor data on the device side. In the future, the transmission bandwidth will be more than 20 Mbps to ensure the interconnection of numerous devices.

4. **Cloud**: The widespread AI technologies have strongly supported power source, grid, load, and storage in the digital twin era. In the future, we need to further improve AI-based computing capability and the application of AI technologies in different scenarios such as yield prediction, predictive maintenance, power grid monitoring, energy consumption analysis, and flexible dispatching.

5. **Data**: Blockchain ensures the mutual trust during the exchange of energy and data flows. In the future, we need to strengthen the application of the blockchain technology in power trade and energy metering.

Enablement Technology		Digital Green Power Plant		Digital Inspection of Power Grids		Multi-Source Self-Healing Distribution Network		Multi-Energy Synergy and Complementarity		Cross-Region Power Dispatching		Green and Low Carbon Enablement	
		E2E Digital Twin	Remote Intelligent and Centralized Control	Intelligent Line Inspection	Intelligent Substation	Multi-Source Distribution Network Operation	Self-Healing Distribution Network Regulation	Smart Campus	Intelligent Building	Intelligent Power Grid Dispatching	Virtual Power Plant	Carbon Trade	Carbon Inclusion
Digital Edge & Device	Edge & device collection	★	★	★	★	★	★	★	★	△	★	△	△
	Control terminal	△	★	△	△	★	★	★	★	△	★	–	–
Ubiquitous Communication Network	Terrestrial communication	★	★	★	★	★	★	★	★	★	★	△	△
	Satellite communication	△	△	★	△	△	△	△	–	–	–	–	–
Computing Power and Storage	Cloud resource platform	★	★	★	★	★	★	★	★	★	★	–	–
	Cloud-edge-device collaboration	△	★	★	★	★	★	★	△	△	△	–	–
	Spatial computing	★	△	△	△	△	△	△	★	–	–	–	–
	Blockchain	–	–	–	–	–	–	△	△	–	△	★	★
	AI	★	★	★	★	△	★	★	★	△	★	–	–
Algorithm and Application	Graph computation and advanced analysis	△	△	△	★	△	△	△	△	△	△	–	–

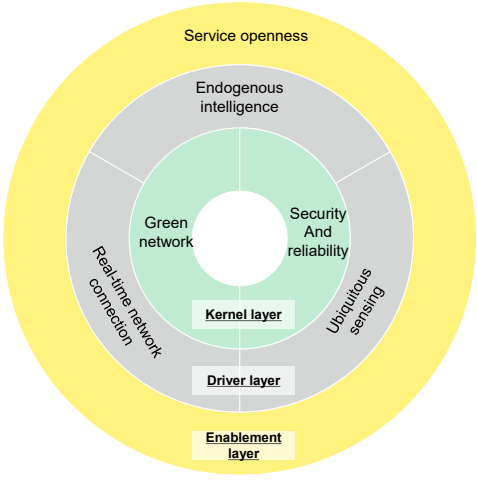
★: Core technology; –: Assistive technology; -: N/A

Chapter 3: Technical Features of Electric Power Digitalization

(1) Key technical features

In view of the requirements for digital technologies, such as IoT sensing, network communication, cloud computing, big data, AI, and blockchain, from the in-depth development of the future power system in various service scenarios, we believe that electric power digitalization technologies will have six key features at three layers.

Figure 10 Key technical features of electric power digitalization in 2030



Kernel layer

Building a green and secure network environment is the core objective and basic principle of power system development in the digital twin era.

Feature 1: green network

Developing and applying electric power digitalization technologies not only help the power system improve new energy consumption capability, but also reduce the energy consumption increase caused by high penetration rate of power electronic equipment and high data processing efficiency. The all-optical network can build an underlying communication network that fully supports service and energy efficiency

requirements at the physical layer, making the entire electric power system greener and more low-carbon.

All-optical networks provide green transport capacity assurance for computing power.

An all-optical network (AON) refers to that E2E information transmission and exchange between communication nodes are implemented by photons without the intervention of electronic signals. This reduces the negative impact of electronic devices or optical-to-electrical conversion on the transmission rate. It effectively meets the requirements of various services such as power generation, grid, load, and storage for high bandwidth and low latency. The AON greatly simplifies the deployment of communication sites and equipment rooms. It can reduce the space required for deploying traditional communication devices by 70% to 80% and power consumption by more than 40%. This contributes to low-carbon development of electric power.

The construction of an all-optical network includes optical lines and nodes.

(1) All-optical transmission lines: Optical communication networks have become the best transmission solution for future ultra-broadband communication technologies due to their huge available spectrum (10 THz), ultra-large capacity (100 Tbit/s), and ultra-high rate (1 Tbit/s). From industry development trends, the optical transmission technology has two breakthroughs: transmission rate and single-fiber capacity.

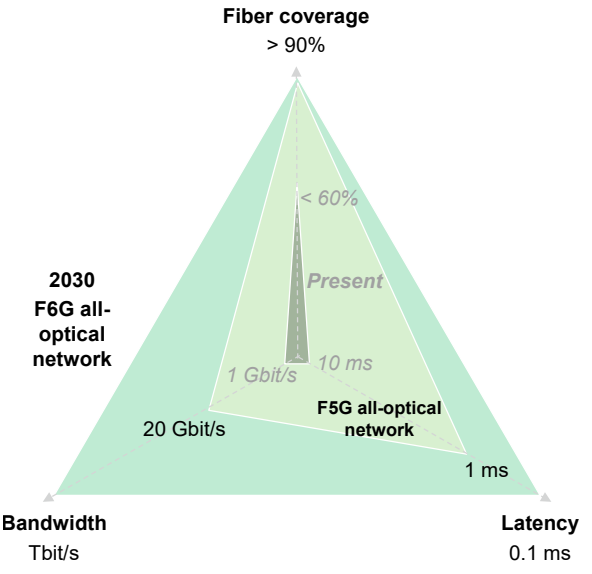
G.654.E fibers have an ultra-low attenuation coefficient and ultra-large effective area. They can significantly extend the transmission distance without relay, greatly reducing the construction requirements of relay sites. G.654.E fibers are suitable for carrying UHV systems that require 400G or ultra-400G transmission performance and ultra-long haul transmission. The power communication and transmission capabilities are greatly improved. At the same time, the optical spectrum will be expanded from C band to C+L band, which will soon achieve an ultra-large capacity of 32T per fiber. The backbone fiber network will enter the 80 x 400G era, building powerful infrastructure capabilities for digital transformation of electric power and digital economy development. In addition, the optical transmission network (OTN) that uses the oDSP algorithm can further improve the communication performance. With the same transmission distance, the capacity increases by 40%. With the same transmission capacity, the distance increases by 20%.

(2) All-optical transmission nodes: The optical cross-connect (OXC) technology is the core solution for implementing a complete all-optical network. An optical backplane integrates hundreds of optical fibers to implement connection-free and zero fiber patch cords. This greatly reduces the space occupied by devices and power consumption, improves system reliability, and provides more flexible configuration capabilities. Only the control system is required to control wavelengths on the optical backplane. In this way, new services can be quickly provisioned.

Key measurement indicators and references for green networks

Bandwidth and latency are core indicators for measuring network performance. Therefore, improving the coverage of optical networks and upgrading high-quality and deterministic all-optical networks are critical to achieve green development.

Figure 11 Key measurement indicators and target references for green networks



Feature 2: security and reliability

In the digital twin era, the power system connects to many devices with heterogeneous protocols, which generate hundreds of millions of data records per second. The data is transmitted at a high frequency through complex communication networks and invoked by various software systems or application services. With the transformation of power digitalization, power communication networks are facing unprecedented security challenges. The future power system requires not only secure and reliable communication networks, but also more digital and trustworthy ones. In terms of cyber security, build a three-layer defense system consisting of the transport layer, network layer, and data layer. As for digital trustworthiness, build a data security framework consisting of the root of trust, distributed trust, and data security & privacy. The two have their own focuses and there is also collaboration between them.

Three-layer defense ensures intrinsic security.

Now firewalls are protecting the running security of network systems. In the future, the traditional external architecture with centralized protection will be evolved to a new architecture to ensuring intrinsic network security.

(1) Anti-interruption at the transport layer, ensuring service continuity: As the key infrastructure of the power system, the power communication network also needs to build three lines of defense for network security to enhance the support for power digitalization. To be specific, device-level redundancy protection enables services to be quickly switched to the standby component when a component in the system is faulty, ensuring device running. Link-level redundancy protection enables services to be quickly switched to the protection link based on the negotiation mechanism when an optical fiber is damaged, implementing quick service recovery. With network-level redundancy protection, when the active network breaks down in a large scale, services can be quickly switched to the standby network based on the independent dual planes, improving the capability of coping with emergencies.

In the redundancy design of power communication networks, the hierarchical redundancy protection mechanism can be adopted according to the actual situation. For the backbone power transmission network, three lines of defense can be configured to ensure reliable communication and secure dispatching. For the power distribution network, link-level or network-level redundancy protection can be flexibly configured in addition to device-level redundancy protection.

(2) Anti-attack at the network layer, preventing network breakdown: First, communication protocols and network devices are reconstructed, and trusted identifiers and password credentials are embedded in IPv6 packet headers, which can help network devices verify the authenticity and validity of requests, preventing forgery and spoofing. Fine-grained access verification and source tracing capabilities are built, improving architecture resilience. Second, build a security service architecture featuring cloud-network-security integration and synergy to automatically respond to network attack threat events and handle them in seconds, implementing global defense. Third, use AI technologies such as graph computing and federated learning to analyze associated events that generate threats, complete threat identification model self-evolution, and continuously improve the threat event detection rate. This implements dynamic detection and intelligent analysis of threat events.

In addition, the number of security policies increases exponentially due to the increase of user scale and complexity. As a result, the traditional manual planning and management mode cannot adapt to the increase. In the future, we need to further study AI-based self-learning and modeling technologies with traffic and service features, feature model-based risk prediction and security policy orchestration technologies, as well as security policy conflict detection and automatic optimization technologies.

(3) Anti-ransomware at the data layer, preventing data loss: On the basis of filtering and identifying ransomware through firewalls and sandboxes at the network layer, build the last line of defense at the data layer. In the production

area, ransomware is detected and intercepted in a timely manner with AI models, and local storage security snapshots are used to implement service recovery in seconds. In addition, local backup and isolated storage in the backup area and isolation area further prevent data loss and ensure data security.

Technology convergence promotes native trustworthiness.

The combination of blockchain and privacy computing technologies can effectively resolve industry problems such as key leakage, privacy data disclosure, and algorithm protocol vulnerabilities, thereby vigorously promoting the development of data elements as assets. It is an important technical roadmap to ensure security compliance throughout the data lifecycle.

(1) Root of trust: Credible data sources are the basis for security and trustworthiness. Trusted Execution Environment (TEE) at the component (chip and operating system) level is a widely recognized and used solution. Moving forward, chip-level trustworthiness computing technologies will be introduced to NE devices on power communication networks. This will help build a secure and trustworthy running environment for software and hardware within the underlying NEs, thus enabling level-by-level verification of chips, operating systems, and applications to ensure data authenticity.

(2) Distributed trust: To meet the complex security and trust requirements of the cloud, pipe, edge, and device, the blockchain technology will be introduced to build a trustworthy service system for basic digital resources (including connectivity and computing) for future networks. Distributed accounting, consensus mechanisms, and decentralized key allocation will help ensure the authenticity of resource ownership and mapping relationships and prevent anonymous tampering and illegal hijacking.

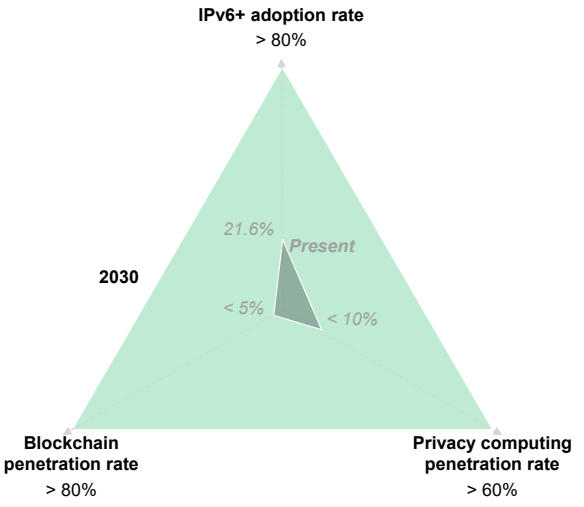
(3) Data security and privacy: User data is accessed during user access and at service awareness points. To ensure user information security, the capability of encrypting user IDs and communication data needs to be enhanced. Technologies such as pseudonymization and encrypted computing are used to implement transparent user information. That is, data is calculated and analyzed on the premise that data privacy is protected from being disclosed, promoting highly reliable sharing and exchange and achieving data availability but invisibility.

Key measurement indicators and references for security and reliability

The IPv6+ based network intrinsic security, AI-based management security, and network redundancy protection design improve the security and reliability of electric power communication networks from different levels of network architectures.

Data trustworthiness through technologies such as root of trust, blockchain-based distributed trust, and privacy computing ensures security and confidentiality of data processing and use, improves the penetration rate and adoption rate of related technologies, and implements wider and more secure data collaboration.

Figure 12 Key measurement indicators and target references for security and reliability



Driver layer

Ubiquitous sensing, real-time network connection, and endogenous intelligence demonstrate the core service process of electric power digitalization from collection, transmission, to processing and analysis of massive electric power data. Together, they form a technical foundation that will drive

the intelligent and digital transformation of the future power grid.

Feature 3: ubiquitous sensing

With the deepening of electric power digitalization, two networks will be gradually generated for the future power grid: a physical network that connects various electric power devices and an information network that links massive production, operation, and consumption data. The in-depth integration and interaction of the two networks can greatly promote the operation efficiency of the electric power system.

Building an efficient sensing network to improve the quantity and quality of collected data, and realize information interaction and intelligent processing is the basis for supporting the two digital electric power networks.

Intelligent terminals help build a multi-dimensional sensing network that intelligently connects things and data.

Intelligent terminals are the foundation of ubiquitous IoT construction for electric power. By 2026, it is expected that the penetration rate of intelligent terminals with a unified operating system will exceed 16%, and by 2030, it will surpass 60%. This will require the standardization of different operating systems and communication protocols.

Intelligent terminals enable all things connected. With the continuous convergence and collaborative upgrade of power source, grid, load, and storage, the power system becomes more and more complex. Massive data is generated in every aspect of the power system, such as the power supply side, backbone network nodes, transformer district side, load side, and energy storage side. Ubiquitous sensing is to collect key information such as energy flow changes, running status changes of various devices, and external environment changes that affect the running of the power system through various means. All power grids are covered with various sensing devices, indicating parameters of electricity, status, environment, space, and behavior. Various perception devices, such as video cameras, infrared spectrometers, lidars, low-orbit satellites, and communication sensing devices, perform comprehensive inspection on large-scale infrastructures such as power plants, power transmission and distribution networks, and substations. Primary and secondary convergence devices or energy gateways are fully connected to distributed power supplies, energy storage devices, energy consumption devices, and measurement meters. All-round and all-weather holographic sensing of the power system lays the physical foundation of ubiquitous sensing.

Intelligent terminals enable data connection of everything. On the basis of IoT awareness on the device side, the edge capability needs to be further enhanced. The protocol converter is used to understand differentiated communication protocols of devices and systems with various specifications and interfaces. Algorithms are used to implement conversion between different protocols. Thus, unified access standards and unified communication languages are implemented to convert, translate, and collect massive heterogeneous data, break the siloed system and application architecture, transform the information from single data to data sets. In this way, various terminals in the power system are transformed from physical connection to chemical convergence, providing data support for global service processing and dispatch decision-making. In the future, with the mature application of the distributed soft bus technology, more proactive heterogeneous networking can be implemented by automatically discovering new surrounding devices, breaking through the restriction.

As for more diversified data types and larger data volumes, traditional narrowband communication cannot meet the requirements of all things connected and ubiquitous sensing. The broadband capability of the edge and sensing terminal communication module needs to be further improved to ensure that "data vehicles" of different types and capacities are accessible. Currently, China's high-tech manufacturing enterprises have started to seize the "HPLC+HRF" dual-mode communication market. The broadband power line carrier and high-speed RF communication technologies complement

each other to expand the scenario versatility of communication modules. Also, dual channels simultaneously transmit and receive data to expand the communication bandwidth and provide more robust communication guarantee for ubiquitous sensing.

Intelligent terminals enable intelligent connection of everything. The power system has high requirements on quick response and timely handling of exceptions. Traditional cloud-based collection, summary, analysis, training, and result feedback cannot ensure real-time services due to massive data. Therefore, cloud models need to be deployed on edge intelligent terminals, which calculate the data collected by sensing devices and deliver decision-making and control instructions. This enables edge power plants and stations to be autonomous locally, achieving intelligent control.

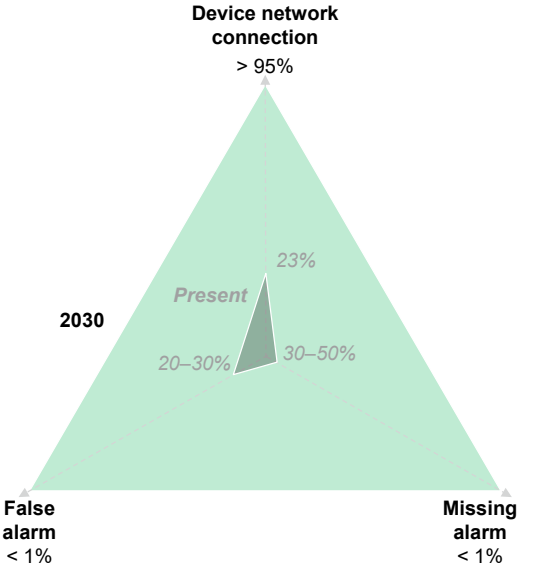
With the continuous development of IoT, a large amount of unstructured data will be collected by terminals. It is estimated that more than 80% of the data will be related to images in the future. The intelligent vision technology provides a method for computer systems to automatically identify, measure, locate, and detect the internal and external environments of factories and stations. Continuous algorithm optimization and interaction and collaboration between the cloud and edge, can help quickly and accurately determine locally collected images. Edge power stations and sensing terminals are enabled to operate independently, achieving intelligent identification.

Finally, service data is locally collected and processed, while devices are locally controlled, forming ubiquitous sensing closure. This feature greatly improves the quantity and quality of data to be collected and ensures efficient service response and processing.

Key measurement indicators and references for ubiquitous sensing

The development goal of ubiquitous sensing is to increase the number of connected terminals, ensure the quantity and quality of collected data, use intelligent vision to better identify image information, and improve the identification rate and accuracy of power system running exceptions, device exceptions, and environment exceptions.

Figure 13 Key measurement indicators and target references for ubiquitous sensing



Feature 4: real-time network connection

Electric power communication is an important part of the electric power system. We rely on the power communication network to send and receive information, such as power quality supervision data, switch control instructions of electrical devices, and reasonable power allocation and scheduling instructions.

Building reliable transmission channels is the foundation for automatic and intelligent control and dispatching of the power system, and also an important technical means to ensure secure and economic power grid dispatching. There are many types of devices with different output features in the future power system, so it is difficult to accurately predict the power flow and flow direction. To cope with various disturbances and emergencies in a timely manner, real-time communication will be increasingly demanding.

5G + Wi-Fi/GWL, reducing E2E latency

The mobile communication technology and the wireless local area network (WLAN) communication technology complement each other. Flexible networking can be performed in

each node of power generation, grid, load, and storage, achieving E2E low-latency and real-time communication.

5G uRLLC: As a synonym for mobile communication technologies, 5G is critical in wide areas and in scenarios that require mobile and high-speed data transmission, such as inspection using drones and intelligent robots.

Based on network slicing technologies, 5G supports three typical application scenarios: enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (uRLLC), and massive machine-type communications (mMTC). uRLLC enables a bidirectional transmission latency between a communication base station and a terminal to be less than 0.5 ms. It provides services and support for latency-sensitive scenarios, effectively meets operation requirements of a power system, helps the terminal device respond to and handle detected security risks and system running exceptions in a timely manner, and ensures secure and reliable power supply.

Since 2018, 3GPP has released three consecutive versions of mobile communications standards: Release 15, Release 16, and Release 17. The introduction and iteration of sub-technologies, such as flexible frame structures, slot/mini-slot-based dispatching mechanism, PDCCH listening period configuration, and PUCCH dispatching-free grant mechanism, continuously enhance the performance of uRLLC in ultra-low latency and ultra-high reliability.

In addition to optimizing the low latency technology, Release 17 also has two new features. One is the non-terrestrial communication network (NTN) technology, which can enable direct communication with satellites or between any two terminals, providing a more flexible solution for emergency communication in the case of extreme events. Second, the millimeter-wave band is increased from 52.6 GHz to 71 GHz, so the service capability of uRLLC is further enhanced through bandwidth extension.

In the future, with the deep application of millimeter-wave spectrum in mobile communication technologies, the communication spectrum and sensing spectrum will overlap. In the 6G/F6G era, communication sensing convergence will be implemented, and the communication latency will be improved to the sub-millisecond level, further supporting the digital construction of electric power.

Wi-Fi: Wi-Fi is synonymous with the WLAN communication technology. Compared with the xG technology, Wi-Fi is more applicable to intra-station communication or scenarios with high device density in a certain area, such as smart campuses.

With the IEEE 802.11 protocol standard, Wi-Fi has developed to the sixth generation, featuring high rate, low latency, and multiple connections. Wi-Fi 6/6E supports 2.4 GHz and 5 GHz frequency bands, and the theoretical rate can reach 9.6 Gbit/s. In terms of latency, the OFDMA and MU-MIMO technologies used by Wi-Fi 6/6E support simultaneous access of multiple devices, greatly increasing the number of concurrent connections and reducing the latency by about 30%. In addition, Wi-Fi 6/6E introduces a target wakeup time (TWT) dispatching mechanism, which enables flexible on-demand wake-up of Wi-Fi through negotiation with terminals, reducing power consumption by about 30%.

In the future, with continuous breakthroughs in communication speed and coverage, Wi-Fi technologies will evolve towards higher frequency bands, higher rates, and lower latency. According to the released technical features, the next-generation Wi-Fi technology will further expand the bandwidth to 320 MHz, increase the modulation mode to 4096 QAM, and achieve 30 Gbit/s ultra-high-speed communication through wider channels and higher traffic density. In addition, it will introduce the Multi-RU and MLO technologies to enhance network anti-interference capability through multi-spectrum resource allocation and dynamic switching between multiple Wi-Fi networks, thereby reducing the communication latency.

GWL: With the rapid development and application of Wi-Fi, it also faces security risks caused by forgery of management frames and authentication frames or information leakage. To meet the security requirements of the electric

power industry, Chinese vendors take the lead in developing the grid wireless LAN (GWL) security access solution based on the WAPI security protocol and state cryptography system. It can be applied in power grids, power plants, and integrated energy services. GWL integrates technologies related to datacom, optical network, and security chips, including software and hardware devices and security management platforms such as wired access networks, WLANs, and mobile terminals. Compared with Wi-Fi, GWL also encrypts the entire process from user access authentication to data transmission (except for public information specified in protocols). In addition to implementing large-bandwidth, full-coverage, and low-latency communication, security performance of wireless communication is significantly improved.

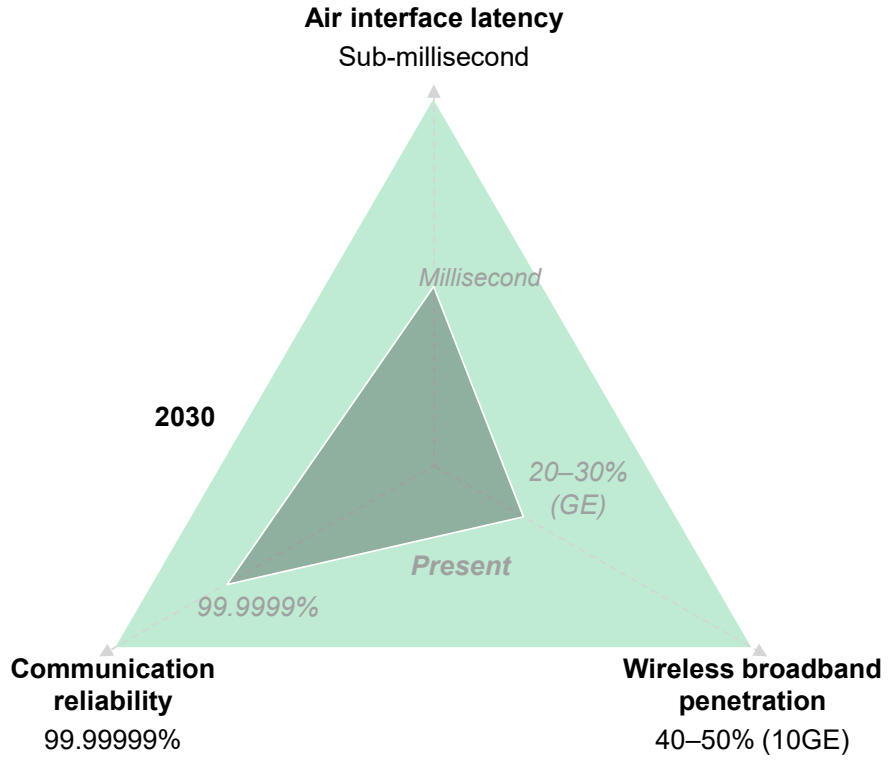
HPLC: It is a communication technology that uses power lines as communication media for data transmission. A next-generation HPLC technology can implement high-speed and stable data transmission on a power line, and has advantages of no re-cabling, low costs, and wide coverage. With the support of 5G technologies, the next-generation HPLC is expected to increase the bandwidth from 2 Mbit/s to more than 10 Mbit/s by 2025, improve the communication reliability from 99% to 99.9%, reduce the latency by 80%, and increase the number of concurrent channels from 1 to 4. By 2030, it is expected that this technology will be fully implemented.

The next-generation HPLC technology is used to implement interconnection between user-side devices and intelligent terminals at the power edge, providing support for intelligent power systems. It is estimated that by 2030, the next-generation HPLC will fully support electric power digitalization services, achieving network availability when electricity is available.

Key measurement indicators and references for real-time network connection

The objective of real-time network connection is to continuously improve the communication latency and network reliability, realizing real-time data transmission and service response in each service phase of the digital twin power system.

Figure 14 Key measurement indicators and target references for real-time network connection



Feature 5: endogenous intelligence

In future power systems, gigawatt-level thermal power units will be gradually replaced by megawatt-level or even smaller-capacity new energy units. In addition, the emergence of distributed energy systems will greatly increase the number of power supply, energy storage, and controllable load devices.

Given ubiquitous sensing and real-time network connection, the data to be processed by the entire power system will also soar exponentially.

In addition to providing powerful computing power support, a computing power network that can implement on-demand allocation and flexible dispatching of storage and computing resources between the cloud, edge, and device based on service and latency requirements is the assurance for precise prediction, effective control and collaboration.

Intelligent computing on one network for electric power helps build the powerful computing foundation.

In the future, the computing center of each regional electric power company will no longer be independent. New network technologies can be used to connect geographically dispersed computing center nodes and dynamically detect the computing resource status of each node in real time. In this way, computing power can be coordinated and allocated globally, computing tasks can be scheduled, and data results can be transmitted and shared, accelerating distributed parallel computing and building a single network for electric power and computing. This solution can effectively solve the problem of resource scarcity when a single AI computing center goes online with full workload.

To handle the complex dispatching in integrated source-grid-load-storage scenarios, more powerful solving capabilities are required to help electric power operators accurately plan the optimal dispatching and countermeasures based on the exponential growth of massive service data and variables. The development and use of traditional solvers face high barriers and rely on expert experience. Therefore, it is difficult to dynamically adjust parameters. The combination of AI and operations research enables AI to replace human brains, implementing the upgrade from expert modeling to intelligent modeling and from manual parameter adjustment to AI adaptive dynamic optimization. This lowers the threshold for solver use, making new breakthroughs in modeling efficiency, solving efficiency, solving scale, and solving speed.

With the increasing demand for computing power in the AI era, semiconductor chips based on Moore's Law will also face bottlenecks in the future. In view of this, photonic chips, as a next-generation chip technology, can carry and implement quantum computing, improving computing power by hundreds of times. At present, the optical quantum chip technology is still in research. We are looking forward to its future.

In addition, the large-scale construction of computing power centers provides support for endogenous intelligence, but also brings large power consumption. According to statistics, the power demand of global data centers accounts for about 1% of the global total power consumption, and the power usage effectiveness (PUE) reaches 1.65. Technology breakthroughs are expected to improve computing efficiency by three to five times, but the period is long. In addition to improving computing efficiency, AI helps reduce energy consumption of computing centers. Sensors collect various data such as temperature, electricity, pump speed, power consumption rate, and preset value. AI is used to analyze the data and automatically adjust the running control threshold based on the model calculation result. In this case, energy consumption for cooling is effectively controlled, reducing the PUE.

This year has seen significant breakthroughs in AI technology, particularly with the emergence of large models. This significantly accelerates construction of AI computing power foundations, bringing it on par with general computing power. Most applications can now be reshaped using AI capabilities, and the development of massive edge inference is rapidly advancing. With the growth of large models, it is expected that inference computing power will exceed training

computing power by more than 10 times in the future. It is estimated that the proportion of AI computing power in the total computing power will be adjusted to 90% by 2030.

Tiny machine learning (TinyML) makes edges smarter.

According to Gartner's prediction, by 2025, 75% of data will be generated at the edge outside data centers. Digital transformation is moving from the cloud to networks and terminals. The electric power industry is no exception. For the electric power industry, various services have high requirements on real-time performance. To respond to various source-grid-load-storage requirements in a timely manner, the power system has increasingly strong requirements on cloud computing power closer to users and edge intelligence.

The effectiveness of edge intelligence depends on algorithm model performance, which is determined by training effect. Also, the effect relies on injection and calculation of a large amount of data. However, on the edge side, there are usually not so many data samples that can be trained, and sample features have certain limitations. As a result, a model trained completely by using edge data cannot effectively satisfy precision requirements of service decision-making. Therefore, how to solve the "last mile" problem is the key to edge intelligence. TinyML, an emerging field dedicated to designing, training, and optimizing algorithms and models for edge scenarios, is attracting increasing attention.

Both TinyML and traditional machine learning perform initial training on large models through data migration to the cloud. The biggest difference between them lies in the deployment and optimization of models after training. To be embedded into edge devices and adapt to limited computing resources, TinyML must perform in-depth compression on large models, including model distillation and model quantization. Finally, TinyML completes edge deployment after encoding and compilation.

As AI enters the core business areas, the demand for practicality increases, and data needs to be processed nearby in a real-time closed loop. The previous cloud computing framework is unable to meet the demand, leading to an increase in edge intelligence services. It is expected that by 2030, the adoption rate of edge intelligence will exceed 75%. The fast-paced growth of AI has brought about sustainable development potential for edge intelligence through software and hardware decoupling. In the electric power industry, the cost of edge intelligent devices is lower than that of secondary deployment and reconstruction, which will inevitably speed up the deployment of edge intelligence.

Model distillation: After the initial training of a large model is completed on the cloud, TinyML modifies it and builds a model with smaller memory usage and more compact form. Model distillation includes knowledge distillation and model pruning. The core of knowledge distillation is to migrate prediction results of one or multiple large models to a lightweight single model. Generalization capability of large models helps train small models. A principle of model pruning is to remove redundant parameters whose activation value of neurons from the convolution layer to the fully connected layer approaches 0, so that simplified neurons have the same model expression without affecting output prediction. After the pruning is complete, you need to retrain the model and fine-tune the output result.

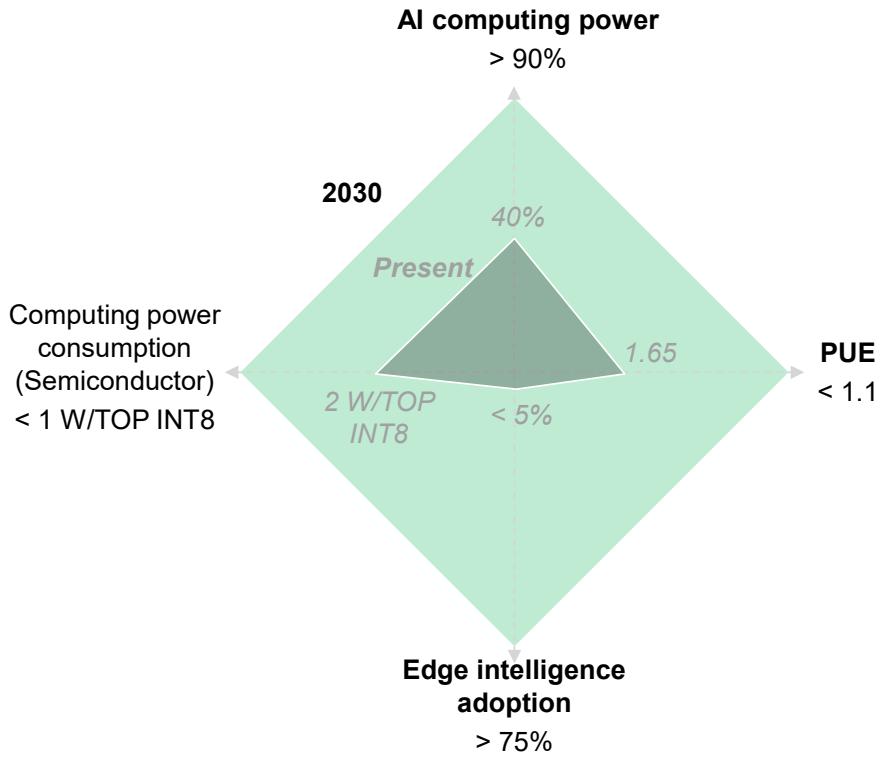
Model quantization: After model distillation, small models need to be quantized to approximately represent 32-bit or even 64-bit floating-point data trained on the cloud in a data type that is compatible with the edge device format and has fewer bits. In this way, the model size and memory consumption are reduced, and the inference speed is accelerated.

Within the acceptable precision loss range, the size of the small model after distillation and quantization can be reduced by about 30 times, and the inference speed can be increased by 4 times. Through the collaboration mode of cloud training and edge inference, TinyML significantly improves the reliability of edge intelligence and supports the implementation of endogenous intelligence.

Key measurement indicators and references for endogenous intelligence

It can strengthen the application foundation of AI technologies to build a green computing network and enhance the support for services by enabling intelligence at the front end.

Figure 15 Key measurement indicators and target references for endogenous intelligence



Enablement layer

Service openness can catalyze capability sharing and ecosystem co-construction, maximizing the value of technology foundation.

Feature 6: service openness

With the continuous upgrade of digital infrastructure construction, digital transformation has shifted from a "minority matter" to a shared goal of the whole society. Many enterprises have huge demands for digital transformation. However, due to challenges such as insufficient capabilities, limited development costs, and poor scalability, the transformation slows down.

Suppliers who master core technologies can access various links and carriers in the integration industry through service platforms and sharing. This can help the industry save infrastructure investment, lower the application development threshold, and reduce operation and management costs with diversified resources, tools, or application services available on the platform. Technology inclusiveness enables access enterprises to enjoy efficient and standard electric power services or quickly customize personalized service requirements. In addition, data accumulated through application services can become new learning materials for the open platform to iterate and upgrade its capabilities, providing access parties with better services.'

The open cloud-network architecture achieves data sharing, capability sharing, ecosystem co-construction, and industry prosperity.

During digital transformation, cloud is the core, data is the key, and network is the foundation.

Through continuous upgrade, openness, and sharing of cloud and network capabilities, digital transformation will shift from single-vendor and single-system capability building to industry-wide and ecosystem-wide capability improvement. Leading ICT capabilities are integrated to enable the electric power industry. Thus, IaaS, PaaS, SaaS, and NaaS services are provided for participants in the upstream and downstream of the industry chain. This improves digital capabilities and supports digital transformation of the electric power industry.

(1) Cloud openness

Through openness and integration, services of enterprises can move to the cloud in a diversified manner. Open tools help enterprises achieve cost-effective, efficient, and agile development.

Non-intrusive integration: With accelerated penetration of digital technologies, different electric power enterprises have their own digital footprints. Thorough innovation not only wastes historical investments, but also decreases the efficiency of enterprise digital transformation, which is inconsistent with the original intention of service openness. Therefore, the openness of cloud services needs to complete application adaptation and data integration between the cloud platform and existing systems through APIs, data interfaces, and messages, support cross-cloud connections, and provide multiple cloud migration paths. This can help enterprises quickly improve efficiency based on existing application systems and gradually upgrade, reconstruct, and migrate applications in the future.

Cloud native technologies: The application of cloud native technologies enables enterprises to maximize the use of cloud capabilities for agile development and replicable and scalable features, greatly improve development efficiency, and reduce development and O&M costs. Cloud native technologies will become the best path for enterprises to achieve digital transformation and an important driving force for service openness.

Cloud native technologies include containers, container orchestration and management, and DevOps. The container technology represented by Docker provides enterprises with an environment dedicated to developing, testing, and deploying new applications. Through loose coupling and execution environment isolation, applications can be iterated and updated frequently without affecting the use of other application services. The container orchestration technology represented by Kubernetes can effectively perform operations on containerized application services, such as load balancing monitoring, scheduling management, fault isolation, and automatic recovery. DevOps development and O&M collaboration is also an important embodiment of cloud native technologies, which include low-code/zero- code development, continuous integration, continuous delivery, and continuous deployment. Low-code/Zero-code development tools provide various preset components and orchestration capabilities such as UI orchestration, process orchestration, logic processing, and model building for enterprise service personnel. Quick modeling is implemented through graphical drag-and-drop, helping enterprises quickly customize O&M management based on digital twins. Continuous integration (CI) uses the CI server to automatically compile and test new code and output results each time to determine whether old and new code is correctly integrated. Compared with traditional phase- based integration, it is easier to locate errors and improve code merging efficiency. Based on CI, continuous delivery (CD) ensures the availability of new code in production through automatic test, simulation, and feedback in the production-like environment. After confirmation, related personnel manually deploy the code in production. Continuous deployment improves continuous delivery to a new level. The entire process from new code submission to new function deployment and rollout does not require manual operation. Instead, the process is fully automated.

Inclusive AI: AI is the core driving force for the electric power industry to enter into a comprehensive digital twin era. However, according to the current industry penetration, AI applications are distributed in a scattered manner with small volume due to challenges such as high costs and difficult training. In the future, AI will become a common capability for thousands of industries, which is an important goal of digital service openness.

Inclusive AI means to build industry-level ultra-large pre-training models based on multiple algorithm capabilities, such as natural language processing (NLP) models, intelligent vision (CV) models, multi-modal models, and scientific computing models. Centralized pre-training helps continuously accumulate common industry knowledge, form standard common models, and generalize the models to more scenarios to reduce human intervention and consumption. AI accelerates transformation from traditional workshop-based development to modern industrial development. So, when a universal and easy-to-use AI development pipeline is formed, more service personnel can quickly customize deployment and continuously iteratively provide feedback on pre-trained models, making AI application training faster with better performance, efficiency, and quality.

To sum up, the open access of cloud platforms, cloud native technologies, and inclusive AI, facilitate the collaboration and shared success between platforms and operators.

The maturity of cloud-based and service-oriented technologies has greatly improved, resulting in a significant increase in the efficiency of cloud technology application development and a qualitative leap in cloud security. The increasing number of services and data running on cloud platforms has significantly improved overall IT efficiency and flexibility. The development of networks has made cloud-network synergy smoother, while the rapid expansion of cloud-based development and edge applications has accelerated the increase in cloud penetration rate.

(2) Data openness

Although the cloud platform solves the problem of data flow within enterprises or alliances, it causes new problems. In the multi-cloud and edge cloud era, distributed data storage results in new data silos between clouds. Storage resources cannot be efficiently shared, resulting in low resource utilization. Also, data cannot be associated and seamlessly transferred across clouds, and data migration, replication, backup, and integration are difficult.

In the future, by building a unified underlying data storage resource pool, diversified storage hardware devices, such as all-flash, converged, high-density, and Blu-ray devices, can be integrated into an open software architecture. By connecting and converging storage resources and multiple clouds, the data silos can be eliminated, and data interaction in each phase of power generation, grid, load, and storage can be streamlined. This promotes efficient data use. In addition, the unified storage resource pool can match different resource service levels based on the requirements of different upper-layer electric power service scenarios, and allocate and provision storage resources on demand. This greatly improves storage resource utilization and ensures consistent user experience.

(3) Network openness

With the in-depth application of AI, the "autonomous driving" capability is derived from the network. According to the definition of autonomous driving network levels, the current network automation level is in the L2.5 phase, which can implement autonomy for specific environments and network units with the support of AI models.

With the combination and application of AI and simulation technologies such as high-performance approximate network measurement, the real network can be comprehensively analyzed and reasoned based on multi-dimensional information, and the unknown network status can be accurately determined and effectively predicted. This enables the network to continuously evolve to more advanced cognitive intelligence, and implements L5 autonomous driving for networks throughout the lifecycle. L5 autonomous driving network can provide cognitive intelligence in two dimensions. One is the time dimension. Future performance deterioration can be accurately predicted based on historical network performance changes and alarm information. The other is the function dimension. Based on the situation awareness and understanding learning of multiple network environments, upcoming changes of network functions can be determined (such as channel changes and security situation changes). Based on cognition of change trends in the two dimensions, intelligent decision-making is implemented through models, and dynamic

network configuration and reconstruction are performed based on the decision-making results, enabling the network to adapt to decision-making autonomy.

Looking into the future, network platform service providers can open APIs to carriers to provide highly reliable and automatic network performance services and assurance for access users.

Finally, based on the construction, integration, and capability output of the open cloud network software and hardware architecture, the overall digital capability of the electric power industry is improved, achieving data sharing, capability sharing, ecosystem co-construction, and industry prosperity.

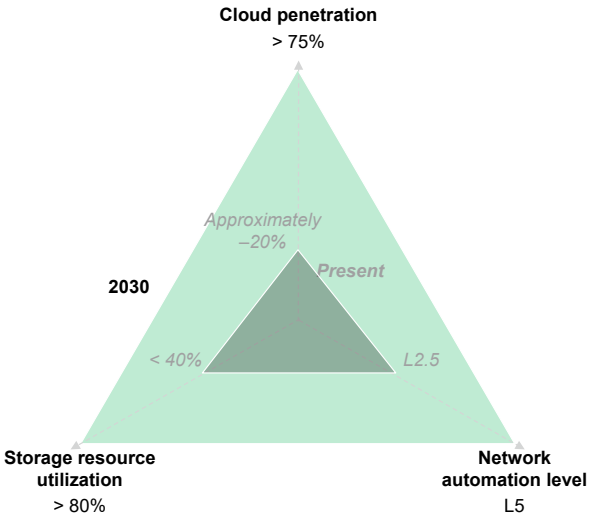
Figure 16 Definition of autonomous driving network levels at the Telecommunications Management Forum (TMF)

Level	L0: manual inspection	L1: auxiliary O&M	L2: partial autonomy	L3: conditional autonomy	L4: high autonomy	L5: full autonomy
Service	N/A	Single use case	Single use case	Multiple uses cases available	Multiple uses cases available	Any
Execution	Manual	Manual/ Automatic	Automatic	Automatic	Automatic	Automatic
Awareness	Manual	Manual	Manual/ Automatic	Automatic	Automatic	Automatic
Analysis/ Decision-making	Manual	Manual	Manual	Manual/ Automatic	Automatic	Automatic
Intent/ Experience	Manual	Manual	Manual	Manual	Manual/ Automatic	Automatic

Key measurement indicators and references for service openness

The construction of the data storage resource pool can implement cloud-based data integration, and the construction of the network automation level can ensure the continuous improvement of the service level, thereby improving the cloud penetration rate of electric power enterprises and achieving service openness.

Figure 17 Key measurement indicators and target references for service openness



(2) Target technical architecture

A comprehensive architecture is required to support the implementation of the six key technical features of electric power digitalization.

Currently, large energy and electric power enterprises are planning their own technical architectures in light of their business development and infrastructure informatization status. Although enterprises are becoming more and more digital, in terms of "building one integrated network", there are still key challenges, such as heterogeneous edge devices, insufficient cloud-edge synergy, and long application development period. Separate construction and independent running of systems restrict resource and data sharing. In this process, data silos may be formed, leading to the Matthew effect of electric power digitalization. This will hinder the common development of the entire industry, integrated source-grid-load-storage collaboration across time and space, and the achievement of the electric power digital twin blueprint and objectives.

In the future, the electric power industry needs to comprehensively integrate and upgrade the current technical system to build a more open, efficient, and intelligent technical architecture, supporting the digital transformation of the electric power industry, and fulfill the ambitious vision of E2E electric power digital twins.

The target technical architecture contains scenario-based applications and cloud-pipe-edge-pipe-device layers. The architecture implements northbound enablement (five aspects) for service scenario digitalization, supporting the rapid utilization of industry capabilities and improving service digitalization efficiency. In the southbound, the architecture provides application, data, O&M, and intelligent synergy, enabling efficient and real-time interaction of cloud services. This improves efficiency, reduces costs and supports software-defined services. A three-layered OS is built that unifies standards and supports digitalization of services, while embracing an open and adaptable digital future. As the saying goes, "If you want to get rich, build roads first." Without communication, there is no way to achieve informationization, digitalization, or intelligence. Only with an efficient communication system can we make digitalization and intelligence accessible. Communication construction should be guided with a target communication network architecture in mind, looking ahead to tomorrow to determine what needs to be done today. It is important to not only focus on current issues and challenges, but also anticipate the needs of the next five to ten years.

The target technical architecture is a complete and systematic architecture that features software and hardware decoupling, openness, and evolution, including:

(1) Northbound enablement:

Development enablement: API orchestration and development are implemented through the development center, application delivery in DevOps mode is implemented through the tool chain, and secure and standard application development is implemented through the software pipeline.

Application enablement: Supports quick deployment and management of applications.

Data enablement: Integrates data and provides powerful data development, data governance, data operations, and service modeling capabilities to provide data services for applications.

AI enablement: AI platforms and assets are used to implement out-of-the-box usability and quick incubation of AI.

Integration enablement: Tools quickly implement data, service, and message communication and convergence between systems.

(2) Southbound synergy:

Application synergy: Applications are packaged and delivered through cloud-edge synergy platforms to implement development on the cloud and deployment at the edge/device sides.

Data synergy: Industry gateways are used for data access and aggregation to implement cloud-

side analysis and edge/device-side control.

AI synergy: Models are trained on the cloud to implement AI synergy and edge/device-side inference.

O&M synergy: Device and configuration information is aggregated to support cloud-based O&M and edge/device-side management.

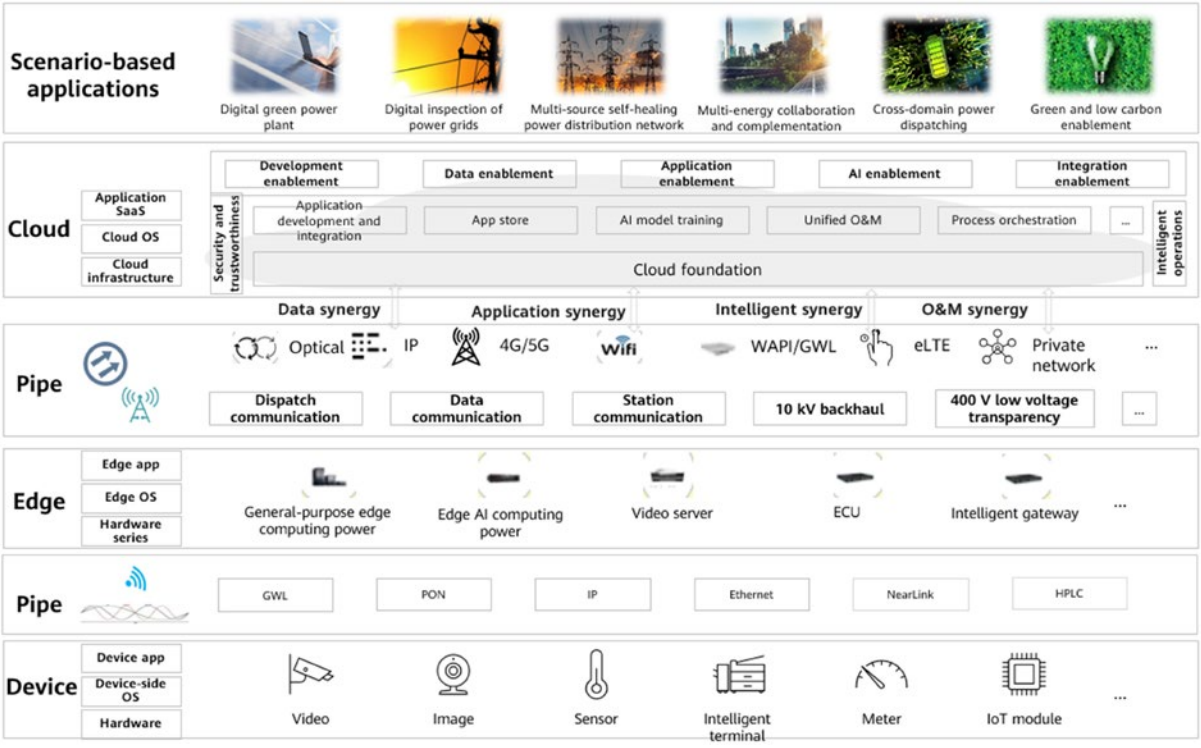
(3) Three-layer OS

Cloud OS: Based on the cloud platform, we can build a cloud OS, application development platform, IoT management platform, and AI training platform to provide unified digital capabilities.

Edge OS: Shields differences and supports plug-and-play, replacement instead of repair, software and hardware decoupling, platform-based hardware, and app-based software.

Device-side OS: Meets the access requirements of different hardware devices, shields differences, and enables device plug-and-play, device-device interconnection, and convenient O&M.

Figure 18 Target technical architecture of electric power digital transformation



(4) Target communication network

Currently, the focus of the power communication target network is distribution network communication. This is due to the growing number of distributed PV access and electric vehicles, which has led to a greater need for communication within the distribution network. Currently, the coverage rate of communication in the distribution network is less than 10%.

It is expected that by 2030, the coverage rate will exceed 70%. This will lead to a more comprehensive and diverse selection of communication technologies for the distribution network. The target distribution communication network needs to mainly solve two challenges: 10 kV backhaul and low-voltage 400 V transparent communication. 10 kV backhaul can be implemented through optical networks and

wireless networks. We must ensure that they are systematic, operable, and manageable and support smooth upgrade and evolution. Wireless networks or power line communication can be used to enable communication from 400 V grids to users. The demand has shifted from being observable and measurable to being adjustable, controllable, and traceable, to meet the high-speed, reliable, and cost-effective data collection needs of users. This supports a clear topology for low-voltage power grids and encourages the development of new user-side services, such as virtual power plants, orderly charging of electric vehicles, and market-based power trading.

To ensure proper usage, the target network should serve as the guide, regardless of whether it is a wireless public or private network, optical fiber backhaul, or carrier communication. It is important to consider specific scenarios, hierarchical structures, security and reliability requirements, cost-effectiveness, openness and evolution of communication technologies, and full-lifecycle collaboration across various domains. To make the right decisions today, we need to take a forward-thinking approach and consider the future from the perspective of the day after tomorrow.

(3) Digital innovation system

The electric power industry is asset-heavy that is experiencing an unprecedented growth trend in renewable energy. This has led to the digitization of the entire power grid, from the main network to the distribution network, with equal emphasis on both. To effectively utilize digital applications for a large number of assets and users, and to take advantage of rapidly advancing digital and intelligent technologies, a systematic approach to innovation is required. This primarily involves three types of innovation: architecture innovation, mode innovation, and ecosystem innovation.

Architecture innovation

The electric power industry is characterized by numerous points, long lines, and extensive coverage. To cope with external changes such as uncertainty, diverse service needs, autonomy, and complex ecological environments, the Spark architecture that enables deep cloud-edge-device synergy is required. It enables the shift from single-point digitalization to architecture-supported and systematic digitalization. This way, the architecture can significantly reduce scale, replication, and time costs, which is more practical for electric power enterprises with heavy-asset operations and new power system challenges.

Mode innovation

How can we keep core digital capabilities in the hands of operators to ensure that related digital work is continuous and sustainable? Electric power enterprises are different from digital native enterprises such as Internet enterprises. Their production materials are naturally in the digital world. Therefore, we need to stand in the present and bring historical achievements to the future to help power companies go digital as they grow rapidly. This is like changing the tire of a speeding car, which required mode innovation.

To achieve mode innovation, it is essential to establish systematic and targeted digital talent and organizations. Operators must master the core capabilities of digitalization to ensure the sustainability of related work.

First, we should empower electric power experts with digitalization expertise. We need to focus on key industry scenarios and bring together operators, industry companies, and the ecosystem to cultivate a first group of talent through training & practice and build systematic digital capabilities around them. Second, we need to pool the wisdom of the grassroots. We must lower the threshold for digitalization, and incorporate and stimulate the creativity of all employees. Everyone can be a mini-CEO.

Ecosystem innovation

In terms of ecosystem innovation, James Moor proposed the Hawaii mode and the Costa Rica mode. The latter means an open ecosystem model, which features rapid localization of external ecosystems and rapid evolution of local ecosystems, leading to interaction and co-evolution.

In the digital age, no one can dominate the market completely. The traditional approach of "introducing, digesting, absorbing, innovating, and replacing" may work in the informatization age, but it is not feasible in the digital age. We need to shift from the traditional "main business companies + industry companies" model to the "main business companies + industry companies + ecosystem" model. Main business companies are responsible for the success or failure of digital transformation, and industry companies should focus on addressing key challenges related to their main business to create value.

Take the power distribution IoT project in a province as an example. 29 ecosystem partners were involved, including 12 cloud partners, 9 edge partners, and 8 device partners. Throughout the project construction, ecosystem partners leveraged their strengths and collaborated extensively in production, education, research, and application. This enabled operators to quickly benefit from industry and cross-industry capabilities. Any issues that arose during the day could be resolved at night and implemented the following day.

The Costa Rica mode can not only give full play to their respective strengths, but also enable both industry and cross-industry capabilities to be used by operators and evolve together.

Our Proposals

Great success can be achieved through mass efforts. To adapt to the development trend, which features the green energy structure and interactive power supply mode, in the view of new power system features, all participants in the electric power industry need to put efforts together. We aim at asset security and efficiency improvement, new energy grid-tied consumption, source-grid-load-storage coordination and interaction, green power market-based transaction, and low-cost and efficient energy use. We should focus on following core service scenarios: digital green power plant, intelligent power grid inspection, multi-source self-healing distribution network, coordination and complementarity of multiple energy sources, cross-region power dispatching, and green and low-carbon enablement. We can seek technical support from the digital technology application fields of digital edge, ubiquitous communication network, computing power and storage, and algorithm and application to implement green network, security and reliability, ubiquitous sensing, real-time network connection, endogenous intelligence, and service openness. In this way, we can build a set of cloud-edge collaboration technical architectures, an open, efficient, and intelligent new engine for electric power digitalization, thereby fully supporting and driving electric power system upgrade and transformation, accelerating new energy consumption, and finally promoting the achievement of the carbon peak & neutrality goals.

However, the digital transformation of the electric power industry faces many challenges. It requires both key technological breakthroughs in fields such as energy and informatization, and policy support.

Integration is the key word for technological development. For the future power system with frequent interaction between source, grid, load, and storage and highlighted features, electric power digitalization cannot be achieved without the in-depth integration of IT, communication technology, and electric power technology. ICT infrastructure vendors, software vendors, power electronic equipment vendors, and energy technology institutions need to quickly break the inherent industry barriers, accelerate cross-

industry technical integration and interaction by establishing industry alliances or promoting strategic cooperation.

At the same time, emerging businesses (such as virtual power plants) based on the digital twin also require faster standardization in industries such as communications and controls.

Improvement is the key word for policy assurance. With the emergence of new transaction forms such as green power transaction, cross-border dispatching of virtual power plants, and point-to-point transaction of power distribution networks, it is urgent for policy agencies and regulators to figure out how to establish or further improve a more flexible matching transaction mechanism based on the existing framework to build a power transaction system in which multiple forms coexist.

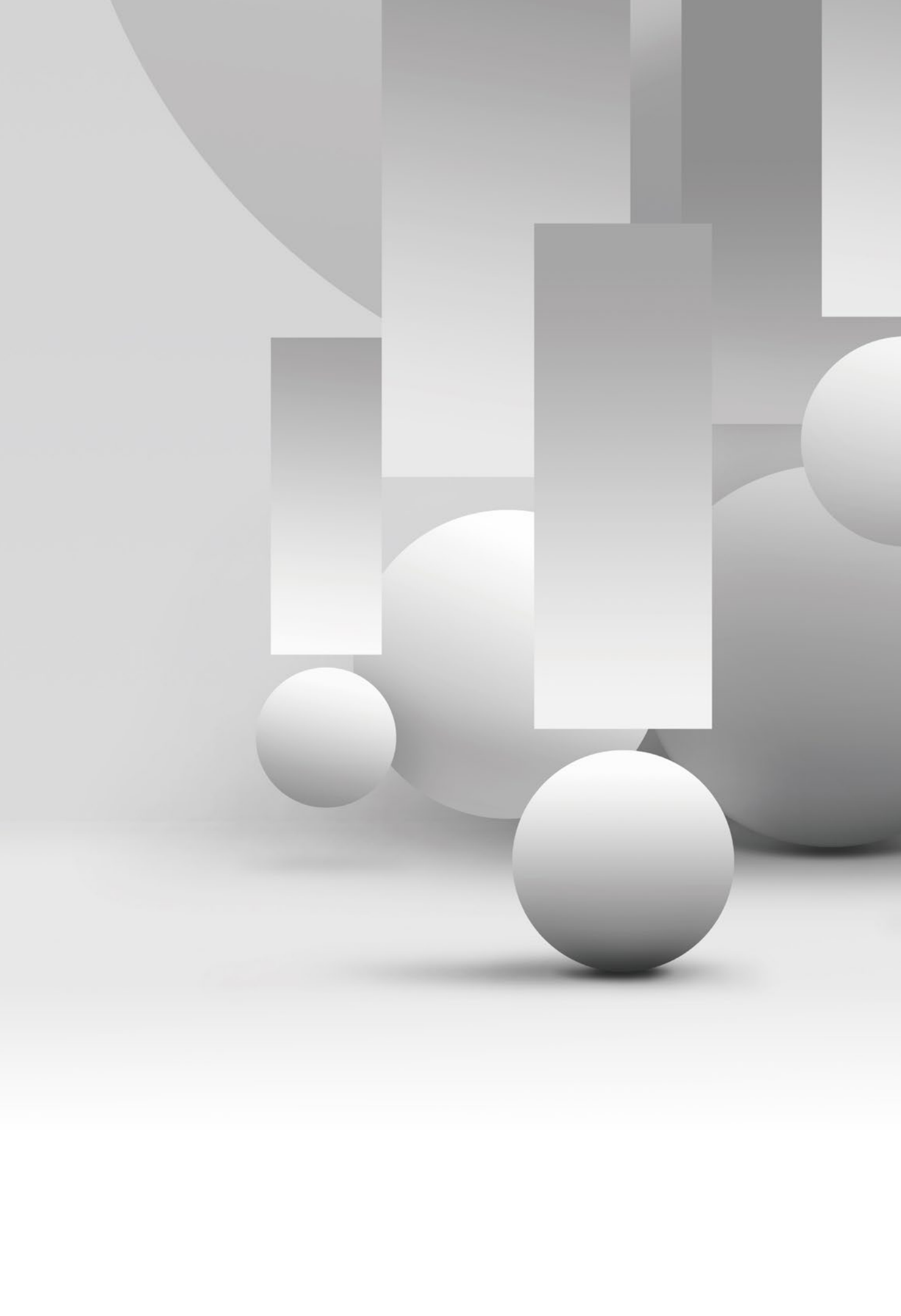
The digital transformation opens more possibilities for future power systems. With the continuous development of electric power digitalization, the vision of the electric power digital twin will be clearer and more accessible.

The future has come. This digital revolution enabling green and efficient power systems that involves traditional electric power enterprises such as power generators and power grid carriers, new businesses such as electric vehicles, and cross-industry participants such as technology giants, campus carriers, and platform service providers, can only be achieved with the joint efforts of the entire industry and society. Let's work together to drive the electric power digitalization by 2030.

Note: Huawei is constantly engaging with industry experts, customers, and partners to delve deeper into the intelligent world. We have observed a significant acceleration in the development of the intelligent world, with new technologies and scenarios emerging rapidly, and industry-related parameters changing exponentially. In light of this, Huawei has systematically updated the Digitalization Trends in the Electric Power Industry 2030 report, which was released in 2022. With a view to the power scenarios and trends that will shape the year 2030, we have made adjustments to the forecast data.

References

1. White Paper on the Digital Technology Support System for New Power Systems, State Grid, 2022
2. Digital Grid Practice White Paper, China Southern Power Grid, 2021
3. Green Path — Insights into the New Power System Industry, Schneider, 2022
4. Digital Twin Grid White Paper — Digital Transformation of Electric Power Enterprises, Academy of Institute 4.0, 2022
5. Li Ruisheng, Development and Prospect of Plug-and-Play of Random Power Supplies, DU Magazine, 2017
6. He Weiguo et al. Construction and Development Path of the Urban Resilient Distribution Network, Power System Technology, 2022
7. Xu Yin et al. A Review on Distribution System Restoration for Resilience Enhancement, Electrical Engineering Magazine, 2019
8. Xu Yong, AI Is Undergoing Significant Changes from Computing Centers to Computing Networks, People's Posts and Tele-communications News, 2022
9. Typical Scenarios and Key Capabilities of 6G IMT-2030 (6G) Promotion Group, 2022
10. Three-Year Action Plan for New Data Center Development (2021–2023), Ministry of Industry and Information Technology, 2021
11. Industrial Internet Platform Application Data Map (2021), China Industrial Control Systems Cyber Emergency Response Team, 2021



HUAWEI

Version 2024

Data Center 2030

Building a Fully Connected, Intelligent World

Data Center 2030

Exploring the Future of Data Centers
to Lead the Intelligent Era

Foreword

David Wang

The emergence of an intelligent world

When large AI models reach a certain size, there's a sudden change in performance where the capabilities of systems go far beyond the confines of their training data – a phenomenon described by researchers as "emergence". These emergent properties have taken artificial intelligence to a new level, from perceiving and comprehending to creation itself. This evolution in AI is behind the popularity of ChatGPT, and it has spurred the emergence of hundreds of industry-specific foundation models.

Today, a vast array of models and modalities are being applied across different scenarios and industries, addressing specific issues that organizations face and speeding up the intelligent transformation process. AI's moment of emergence is here, setting the stage for a magnificent new age of intelligence.

In the intelligent world to come, demand for computing power will be unprecedented, and data centers will become the world's most critical infrastructure. According to Huawei's *Intelligent World 2030* report, the volume of data generated globally will exceed one yottabyte (i.e., a quadrillion gigabytes) by 2030. The global computing power will reach 3.3 ZFLOPS (FP32), and the demand for AI computing power will increase sharply. By 2030, the computing power will reach 864 ZFLOPS (FP16). Moving forward, every 10 years we're set to see a hundred-fold increase in computing power.

Modern data centers are the conduit for new information and communications technologies, like AI and cloud computing. In effect, data centers have become the computing backbone of new digital infrastructure, playing a role of unprecedented strategic importance – the engines of digital economy.

The future of computing power supply will be bound by considerable resource constraints. In terms of computing power, demand is already surging beyond the projections of Moore's law, and individual chips will struggle to keep up. At the same time, pressure to reduce carbon emissions is growing as the world struggles to meet its sustainable development goals. Future data centers will need far more optimal computing architectures to generate even greater computing power while consuming less energy.

If we look back on the history of the ICT industry, every major development has been fueled by resource constraints. Over the past three decades, finding a way to deliver ultra-large bandwidth under considerable cost restraints has supercharged the connectivity industry, leading to the development of technologies like 5G and F5G. In the next three decades, providing strong computing power with limited resources is the next challenge. These constraints will drive the computing industry full speed ahead, paving the way for AI and cloud computing to reshape the world around us.

The conflict between computing demand and resource constraints will give rise to technological, product, and solution innovations at system and architecture levels. Resolving this conflict will be the through line of efforts to build data centers of the future.

Choosing the right direction and the right way forward is about making informed decisions. Looking at the world as it stands right now, it's clear that the ICT industry has enormous development opportunities ahead. Everything is going digital and intelligent. So what will the world look like in 2030?

In September 2021, Huawei published the *Intelligent World 2030* report, alongside a series of reports on different focus domains. *Data Center 2030* is the latest report in this series. It's a collection of thoughts from hundreds of academics, customers, partners, and research institutions, as well as industry experts both inside and outside of Huawei, on one simple but infinitely relevant question: What will data centers look like in the decade to come?

This report opens with the most pressing challenge ahead – the spiking demand for computing power and related resource constraints. It outlines five major scenarios that will affect data center development over the next 10 years, followed by targets for improvements in efficiency, including the efficiency of data, operations, computing, energy, and transmission.

The report is the first in the industry to propose the technical features of future data centers. It systematically details possible challenges for all relevant tech domains, including cloud services, computing, storage, networks, and energy, as well as how we should innovate to address these

challenges. In this report, we also provide a reference architecture for future data centers. It's our hope that this report can help inform the future construction and development of data centers around the world, and help lay the foundations for a booming digital economy.

Today, we're in the process of connecting everything. Tomorrow, everything will be intelligent – and intelligently connected. A better, intelligent world is approaching, and it's the forerunners who will make it happen. In the Intelligent Era, the acclaimed researcher and writer Wu Jun wrote, during each technological revolution, people, businesses, and even countries only have two options. They can either choose to ride the tide and become the top 2%. Or they can wait, hesitate, and be left behind. The next 10 years will overflow with fundamental breakthroughs and world-changing marvels that are set to reshape every major industry.

People tend to overestimate more immediate, short-term changes, but underestimate those that are coming in the next ten years. It's because these are harder to see. *Data Center 2030* seeks to demystify what's coming. Making bold assumptions and accurate predictions can often be dialectically opposed. But it's in exploring the overlap that we can create a better future. Looking forward, we will still have a many challenges ahead of us. But if we work together – and innovate together – we can bring about a better, more intelligent world for all.

David Wang
Executive Director & Chairman
of the ICT Infrastructure Managing Board
Huawei

Industry Trends

01

As an important part of the foundation of next-generation information and communication technologies (ICT) such as artificial intelligence (AI) and cloud computing, data centers have become the computing backbone of new digital infrastructures. Data centers have therefore taken on a role of unprecedented strategic significance and have been deemed engines of the digital economy. Looking ahead to 2030, we have identified the following data center development trends.

The demand for computing power will increase by a factor of 100 over the next decade, and the distribution of computing power will become more polarized

According to Huawei's *Intelligent World 2030* report, the world will usher in the era of yottabytes in 2030. The global computing power will reach 3.3 ZFLOPS (FP32), and the demand for AI computing power will increase sharply. By 2030, the computing power will reach 864 ZFLOPS (FP16). The global data center industry is currently entering a new cycle of rapid development. We predict that there will be more than 1000 hyperscale data centers worldwide within the next three years and that their number will continue to grow rapidly. At the same time, due to the popularization of applications such as autonomous driving, smart manufacturing, and the metaverse, the number of edge data centers will also grow rapidly. According to third-party predictions, more than 10 million edge computing nodes will be deployed in enterprises by 2030.

The scale and efficiency of computing power will become core indicators of a country or business' competitiveness

In the agricultural economy, the main factors determining competitiveness are the size of the labor force, large-scale water conservancy facilities, and high production efficiency due to continued mechanization. Similarly, competitiveness in the digital economy is defined by the scale and efficiency of computing power. We're in a new phase of global and intelligent industry transformation, and AI foundation models featuring hundreds of models and thousands of modalities have become the focus of development. It is predicted that the computing power demand of the Generative Pre-trained Transformer 5 (GPT-5) training cluster will be 200 to 400 times higher than that of GPT-3. Almost all scientific fields and major industries are moving towards multi-dimensional, high-precision, large-scale data analysis. For example, in scenarios such as depth migration in oil exploration, the computing power demand per unit area of the exploration zone will increase by more than a factor of 10. Industry-specific intelligent transformation scenarios powered by technologies such as AI and blockchain will also generate a lot of demand for computing power. Efficient computing power is needed for everything from the sensing, recording, and processing of each swipe of a digital racket, to customer profiling and credit assessment for each micro-transaction in inclusive finance. In the future, many industries will allocate an increasingly large proportion of their investment budget to computing. Take the banking industry as an example. It is predicted that China's banking industry will invest more than CNY400 billion in technology in 2024. More than half of that will go to AI and cloud computing as they are key areas of investment.

AI will revolutionize practically every scenario in data centers

Huawei predicts that the global AI computing power will exceed 105 ZFLOPS (FP16) by 2030. AI computing power will become the most critical factor in driving data center development. The development of general-purpose foundation models over the next five to ten years is likely to bring AI to a point where it can understand texts, music, painting, speeches, images, and videos better than humans can, and deeply integrate with the Internet and smart devices to profoundly change the consumption patterns and behavior of our whole society. The effect of the significant "diffusion time lag" between AI technologies and productivity is gradually weakening. The capabilities of general-purpose foundation models will be embedded in productivity and production tools, industry foundation models, and scenario-based AI applications. Innovation in AI technologies will have an unprecedented impact on business value. With the multi-modal generalization of general-purpose foundation models, the demand for training computing power will continue to increase sharply beyond the levels predicted by Moore's Law. Data centers need to be continuously innovated and quickly iterated in aspects such as computing power scale, architecture, algorithm optimization, and cross-network collaboration. In the future, the development of AI will accelerate the construction of super data centers for platform-based enterprises and computing networks in different countries.

Data centers are shifting from consuming a lot of power to prioritizing green development

Data centers consume more than 80% of the ICT industry's total power consumption. To ensure the sustainable development of the data center industry, it is crucial that we improve the power usage effectiveness (PUE) of data centers to reduce the carbon footprint. Many countries and international organizations have released related data center policies. For example, the U.S. government has established the Data Center Optimization Initiative (DCOI), which requires a PUE of less than 1.4 for new data centers and less than 1.5 for existing data centers. Data center operators and industry associations in Europe signed the *Climate Neutral Data Centre Pact* and pledged to make data centers carbon neutral by 2030. China issued the Implementation *Plan of Computing Power Hubs for National Integrated Big Data Center Collaborative Innovation Systems* to promote the construction of national integrated big data centers, and launched the "Eastern Data, Western Computing" project. These efforts

aim to promote the green and sustainable development of data centers, accelerate the R&D and application of energy-saving and low-carbon technologies, and achieve a PUE of less than 1.3 for new large-scale data centers by 2025. In the future, as more policies are enacted and technology continues to develop, more advanced energy-saving technologies will be used in data centers, further reducing the PUE. It is estimated that the PUE will enter the 1.0x era by 2030. As the proportion of power that comes from renewable sources increases, the data center microgrid featuring collaborative "source-network-load-storage" can further reduce carbon emissions and work towards achieving the zero-carbon goal. In addition to reducing their own carbon emissions, data centers can also facilitate intelligent transformation in other industries and support carbon reduction across society. The Global Enabling Sustainability Initiative (GeSI) predicts that due to their impact on other industries, ICT technologies will help reduce global carbon emissions by 20% by 2030. This reduction is ten times the emissions of data centers themselves.

Promote data centers featuring multi-flow synergy beyond the boundaries of physical data centers

On one hand, most major data center operators and leading digital enterprises face the same challenges in predicting large-scale, medium- and long-term demand and accelerating technology iteration. By 2030, there will be cloud data centers running millions of servers and industry data centers running hundreds of thousands of servers. More gigantic, ultra-intensive tasks like ChatGPT will emerge. Moreover, due to uncertainty in land and energy acquisition, traditional data center planning based on monolithic facilities and predictions of demand over the next 10 years will become obsolete. In the future, phased, modularized, clustered, and service-oriented data centers that are logically unified and physically distributed will become the new norm.

On the other hand, the requirements for high-performance computing are increasing. Batch computing tasks such as film and image rendering, scientific computing tasks such as gene sequencing and wind turbine

simulations, and parallel computing tasks such as AI training often consume a large amount of computing power resources and computing time. Most such tasks are cost-sensitive, time-insensitive, and variable in computing scale. To better address such requirements, prices can be leveraged as a key factor to encourage users to perform their computing tasks in a time period with lower power prices. Other means, such as resumable training and renewable rendering, can be used to pause or even change the paralleling scale during the execution of computing tasks, in order to more effectively process the tasks between peak and off-peak power loads. Multiple flows – the energy flow, data flow, and service flow – can be precisely associated and coordinated to build a green data center with more efficient computing.

System-level innovation will become a mainstream of data center development

The brain of an ant typically consumes only 0.2 milliwatts of energy, but it is capable of doing many complex things, such as making nests, looking for food, and raising aphids. By contrast, the computing system in an autonomous car consumes dozens or even hundreds of watts. There is still a huge gap in energy efficiency between the technical and biological worlds. In view of the conflicts between the 100-fold increase in computing power demand over a decade and the energy consumption constraints, future data centers need to overcome the Von Neumann bottlenecks by seeking a new, highly adaptive and efficient computing model based on a new architecture and new components. In the field of information computing, more than a dozen computing models have been developed and are being widely used. For example, the butterfly computing model based on Fast Fourier Transform (FFT) algorithms is widely used in wireless and optical communication, and the finite-state machine computing model based on logic state transition is commonly used for routers. In the field of intelligent computing, the industry is exploring new computing models that are more efficient, such as mathematical logic computing, geometric manifold computing, and game computing, in addition

to statistical computing. In certain scenarios, these new computing models can improve computing energy efficiency by a factor of 100. Next-generation data centers will also call for a brand-new system featuring multi-technology collaboration between computing, storage, network, and security, shattering the constraints of the power consumption wall, I/O wall, and storage and computing wall that traditional computing devices face. This new system represents a shift from single devices to clusters and from single nodes to networked operations, and leverages system-level innovation and software-hardware synergy to make data centers much more efficient.

Continuous innovation targeted at computing power demand and resource constraints

By 2030, we expect demand for computing power to increase exponentially, by a factor of close to 100, in line with the accelerating development of the digital economy. At the same time, Moore's Law is nearing its limit on single chips, and mandatory requirements on carbon reduction have been introduced globally to promote sustainability. These will become the main factors defining the future development of data centers. Innovation targeted at computing power demand and resource constraints will likely be the key theme of future data center development. The best digital enterprises and digital countries will systematically innovate from a range of different aspects – on a micro, medium, and macro basis, and at different layers – within single data centers, within data center clusters, and between data centers – to build "one computer" at the enterprise or national level. This approach maximizes the difference between computing power supply and resource constraints through overall efficiency improvement and accelerates the move towards an intelligent world.

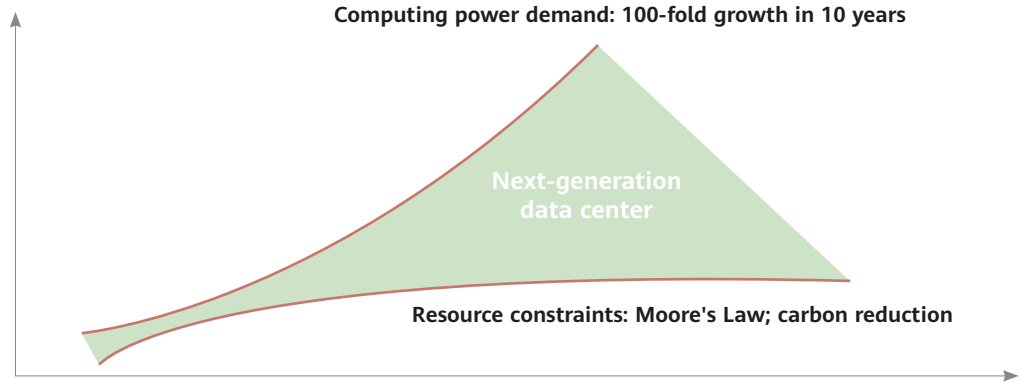


Figure 1-1 Challenges of computing power demand and resource constraints

Future Scenarios and Innovation Directions

02

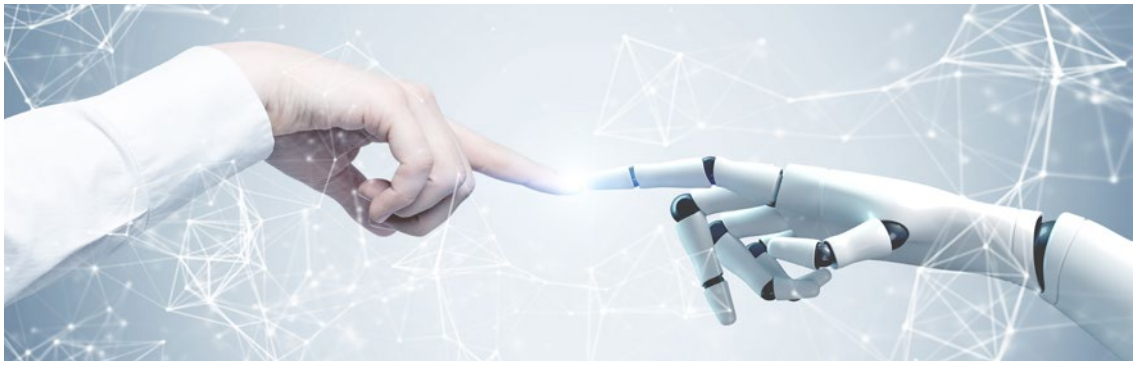
Data centers are almost everywhere in our digital lives. They support breakthroughs and innovation in scientific research and intelligent, efficient production for an intelligent, efficient life. So they need more computing power to process more data. The computing power requirements are expected to grow so fast they will outpace even Moore's Law. At the same time, the growth of computing power is subject to resources. To cope with this contradiction, continuous innovation to improve efficiency will become the core direction of future data centers.

AI for All: Creating new productivity

The history of science is a history of exploration into the laws of the universe. It has been a continuous process of discovering the laws of everything within the boundaries of science and creating new tools of production. This has driven our society to evolve from agricultural to industrial civilization, and now we are entering a new digital phase. We are becoming a digital civilization. In the future, AI will emerge as a new source of productivity. Within the boundaries defined by human beings, AI will analyze and create faster and more efficiently, and today's digital civilization must evolve into a phase of artificial intelligence.

Humans are good at analysis, but AI may be better. Analytical AI has been widely applied to the analysis of data sets or image sets. It is used to find patterns, and this ability to recognize patterns has been applied to fields such as fraud prevention and object detection.

Humans are good at creating, but AI may create faster. As generative AI is developing rapidly, AI has begun to create meaningful and beautiful things, such as poems and drawings, and with incredible efficiency too. Generative AI had made so much progress in image generation largely thanks to the application of the diffusion model, which is a deep learning technology that generates realistic images from noisy images. The progress in natural language processing (NLP)



has been driven by ChatGPT, a text generation deep learning model trained based on available Internet data. ChatGPT is a type of AI used for Q&A, article summaries, machine translation, classification, code generation, and chat. The progress in code generation is represented by two code generation systems, AlphaCode and Copilot. In February 2022, based on their latest research, DeepMind launched AlphaCode. AlphaCode is an independent programming system that has defeated more than 47% of the human competitors in a programming competition held by Codeforces. This shows that AI code generation has reached a competitive level, a new first.

AI technologies are infiltrating thousands of industries, and they are doing it faster and faster. For example, for meteorology, an AI foundation model can generate a 7-day weather prediction within 10 seconds. Compared with the traditional HPC numerical prediction models, AI foundation models generate predictions more than 10,000 times faster. In the securities sector, AI foundation models have helped a financial enterprise increase the accuracy of its intelligent financial warning system to up to 90%, up 11% from the traditional machine learning model. AI foundation models are not only applied to consumer applications like intelligent chatbots, short essay composition, and image generation, but also to business scenarios such as office work, programming, marketing, design, and search. In the future, these models will likely also see more widespread application to enterprise scenarios such as financial risk control, intelligent customer service, AI-assisted diagnosis, and medical consulting. These applications will improve productivity for nearly every industry.

Humans are shifting from understanding the world through analytical AI to creating a world with generative AI. In 2030, AIs with cognitive capabilities will be as ubiquitous as the land, plants, air, and sunlight that we are familiar with. Self-driving cars, robots that can cook, self-managing communications networks, and self-optimizing software platforms will become part of people's daily life, and support the human civilization to evolve continuously.

The fourth paradigm: Exploring the unknown with data-intensive computing

Thousands of years ago, science mostly relied on inductive methods. Experiments and observation of natural phenomena were used to learn about the world. Science was empirical. In more recent centuries, theoretical research was born and mathematical models started being used for analysis. In the past decades, computing emerged, and computers started being used for simulations and analysis of complex problems. In the early 21st century, new information technologies are leading to the birth of a new paradigm, a fourth paradigm, one based on data-intensive scientific research. This paradigm is about unifying theory, experiments, and computing simulations. Data is collected by instruments or generated by simulations, and processed by software. Information and knowledge are stored by computers. Scientists analyze data and documents with data management and statistical methods.

Data-intensive scientific research is generating massive data that needs to be analyzed and processed. For example, simulating the neural network of the human brain to explore how hundreds of millions of neurons connect and work will deliver up to 100 TB of data throughput per second. Self-driving vehicles generate dozens of TB of data every day for training image recognition algorithms. Reconstructing synaptic networks in the brain with electron microscopes requires over 1 PB of image data per cubic millimeter. Astronomical experts need to analyze dozens of PB of data to discover new celestial bodies. Petabytes of data enables us to analyze data without models and assumptions. After data is thrown into huge computer clusters, as long as there is interrelated data, statistical analysis algorithms can discover new patterns, extract knowledge, and even identify rules that cannot be discovered by using the scientific methods of the past.

Scientific data has become a key product of scientific research and an important strategic resource. As data is exploding in terms of

Category	Time	Research Method	Model
The first paradigm: empirical science	Before the 18th century	Mainly inductive methods based on blind observation and experiments	Experimental model
Second paradigm: theoretical science	Before the 19th century	Mainly deductive methods, not limited to experience and empirical evidence	Mathematical model
Third paradigm: computer science	Mid-20th century	Computer simulations and other forms of modeling problems in various scientific disciplines	Computer simulation model
Fourth paradigm: data-intensive science	Early 21st century	Data management and statistical tools are used to analyze data.	Big data mining model

both the demand and the volume, how to store, manage, and share scientific data has become a hot topic for scientists worldwide and an important application scenario for next-generation data centers. When there is more than 1 PB of data, traditional storage subsystems cannot meet the read and write requirements of massive data processing. I/O bandwidth bottlenecks become more and more prominent. Processing data by simply dividing it into blocks cannot meet the requirements of data-intensive computing and is contrary to the whole point of big data analysis. At present, the biggest problem faced by specific scientific research is not a lack of data, but rather a lack of knowledge of how to deal with that much data.

Currently, supercomputers, computing clusters, super distributed databases, and Internet-based cloud computing are unable to completely resolve this contradiction. Computer science is looking forward to a brand new revolution.

Spatial Internet: Supporting virtual-physical interaction

Virtual-physical integration is the next big trend for the next generation of the Internet. A multi-dimensional space offering highly immersive, highly interactive experience will enable people better to interact with information and economic activities much more efficiently.

There are two ways virtual-physical integration is developing.

The first is from physical to virtual. The virtual world imitates the physical world. The digital experience is enhanced by an immersive digital experience. This is mainly a process of digitalizing real experience. In the era of mobile Internet, the virtual world was mainly made of text, images, video and other 2D forms. In the metaverse of the future, the physical world will be digitally reconstructed in the virtual world to enhance multi-dimensional interaction.

The second is from virtual to physical. Here it is no longer about imitation of the physical world but rather about creativity based on the virtual world, which not only can give birth to a value system independent from the physical world, but can also influence the physical world and make digital experiences real. For example, an augmented reality (AR) game can help brands attract more consumers by cooperating with the brands to issue coupons that can only be collected at specific locations. In this way the digital experience can drive spending in the physical world.

Technologically speaking, a multi-dimensional interactive experience with virtual-physical integration depends on the multi-dimensional spatial computing capability of the computer. It depends on graphics and image processing and low-latency networks. In addition, it will require the assistance of powerful AI cognitive capabilities and ubiquitous, unobstructed data connections. Computing and network capabilities directly determine the depth and breadth of the virtual-physical integration.

Industrial digital twins: Promoting intelligent upgrade

Digital twins oriented to a range of industries are important application scenarios of data centers. According to third-party predictions, the compound annual growth rate (CAGR) of the global digital twin market space will reach 40.1% and is expected to reach US$131.09 billion by 2030. Digital twins involve integrated applications from next-generation information technologies, such as modeling, perception, simulation, rendering, big data, and AI. Digital twin is one of the key fields for digital economy development.

As various industries are becoming more intelligent, the requirements for digital twin applications in cities, manufacturing, transportation, water conservancy, and energy have been rapidly increasing, driving the computing power requirements of data centers on both the device and cloud sides. Fast-growing WebGL-based digital twin applications are driving a need for more powerful terminals. Digital twin applications based on cloud rendering have been driving a boom in demand for cloud-based computing power, and this boom in demand for computing power is creating a need to upgrade the cloud computing industry. The compute supply, resource utilization, and rendering algorithms will all need to be improved.

Inclusive cloud native: Bridging the enterprise digital divide

Over the past decade, smartphones and mobile Internet have reshaped our lifestyles and business production models. Today, intelligence and electrification are reshaping the core competitiveness and ecosystem of the automobile industry. Reshaping and reconstruction are supported by data intelligence consisting of powerful computing power, algorithms, and data, as well as cloud-native IT systems featuring agile iteration, elastic scaling, and resilient self-healing. In the future, supported by new technologies such as AI foundation models, Internet of Everything (IoE), socialized data collaboration, and digital twins, industries that are more closely related to the physical world will quickly embrace the intelligent world based on cloud native.

Forward-thinking leaders in various industries and game-changing players of current division of labor are digging deeper into intelligence, promoting the integration of information and operations technologies with distinct cloud native characteristics. In this way, they can develop more refined and agile products, processes, and organizations as their new sources of competitiveness. As digital systems become increasingly complex, releases and changes become more frequent, computing power grows more concentrated, while digital systems become more distributed. Enterprises will need to rely more and more on platform capabilities. More and more of them will find themselves fully embracing cloud native technologies.

Inclusive cloud native technologies bring opportunities to traditional enterprises and even individuals. Cloud native technologies help modernize production and operations. They help bridge the digital divide and provide simple, economical, professional, and personalized paths toward intelligence. Each enterprise that embraces change, combining cloud computing power, data service APIs, IoT control processes (like those offered by Tuya), and commercial industry AI algorithms; can obtain intelligent capabilities that help them better align with industry leaders.

Multi-flow synergy: Improving energy efficiency

Around the world, taking action to fight climate change has become a major priority and, as such, developing green and low carbon technologies has become an important goal for data centers. Most countries and regions have released corresponding policies for individual data centers. Backed by extensive research and testing, China has deployed eight national hubs comprising a national network for computing power. The plan is to help bring large-scale data centers together and promote a new infrastructure for improved computing power and to allow data to flow better. This plan is intended to enable data generated in the densely populated eastern portions of China to be processed in the western portions, where there are more renewable resources available.

Aiming at greener, more sustainable development, data center-related enterprises have developed a large number of innovative technologies for efficient low-carbon infrastructure and operations, and they have adopted the technologies in existing or new data centers. For example, Apple deployed distributed power generation facilities using renewable energy such as solar energy, wind energy, and biogas within their data centers. It also signed long-term procurement agreements with renewable energy power plants for power supply of its own data centers. These measures enable Apple data centers to run on 100% renewable energy.

During the construction of an intelligent cloud green data center, Microsoft proposed that the energy flow, data flow, and service flow of the data center be effectively synergized throughout the site selection, construction, and operation to ensure greener, more efficient processes. Gui'an Huawei Cloud Data Center uses free cooling technologies, including direct ventilation and taking advantage of nearby lake water to dissipate the heat from some high-density servers. Some of the heat from the data center is also collected and recycled to heat the office area. The design takes full advantage of the natural conditions in Guizhou and incorporates many green and low-carbon ideas for sustainable development.

Synergizing energy flow, data flow, and service flow is the key to building energy-efficient data centers by 2030.

Software and hardware synergy: Improving computing efficiency

Computing power has evolved through three stages: single-core, multi-core, and networked. Due to various technological and commercial limitations, the computing power for a single-core silicon-based chip will max out at 3 nanometers. Due to economic reasons, just adding cores to improve the computing power also reaches a practical limit at 128 cores. These challenges will put a lot of pressure on the architecture to evolve from one based on single, multi-core devices to an architecture that relies large networks of devices all working together. In addition, as network technologies are not as advanced as we need and bandwidth is expensive, edge computing power will become a new core scenario for data centers. Ultimately, we need an architecture with ubiquitous cloud-edge computing power, with differentiated levels of computing power deployment.

Over the past half century, the integrated circuit industry has been developing rapidly, following Moore's Law. Computing power has been increasing in leaps and bounds. In an era where hardware is the main force driving rapid improvement of computing power, there is too much reliance on the underlying computing power and not enough importance is placed on the optimization of the architecture and code. New high-level languages keep emerging, which make program execution less and less efficient. This leaves room for optimizing computing performance through software and hardware synergy. Vendors of mainstream chips and devices have started to improve the computing performance by optimizing their software and hardware together. It is believed in the industry that for every order of magnitude that the hardware can improve performance, software and hardware synergy can double that figure. The heterogeneous computing services of Huawei Cloud use software to optimize the hardware passthrough capabilities, significantly reducing the performance loss caused by resource virtualization. David Patterson, a Turing Award winner, once said that in the computing field, we will see more innovations in architecture optimization and performance improvement in the next decade than we have seen in the past 50 years.

By 2030, improving computing efficiency through software and hardware synergy and optimization in central clusters and cloud-edge multi-level computing resource collaboration will be an important direction for data centers.

Lossless networks: Improving transmission efficiency

As data centers continue to develop, they will continue to demand more of the networks they connect to. Traditional networks are not flexible enough in terms of service configuration and resource management. As a result, computing resources within and between data centers are not fully utilized, wasting plenty of resources. In particular, AI foundation model training scenarios require a lot of data, and model parameters get really big. To improve training efficiency, hundreds of GPUs are needed to place a foundation model as a data parallel group. Multiple such data parallel groups are required to shorten the time needed to train a foundation model. When the number of GPUs grows to the thousands, performance depends not only on the GPUs or servers, but also on the network.

To build a high-performance network for more efficient transmission between compute and storage resources, we need just more bandwidth and less latency; even more important is lossless packet forwarding. No data packet loss can be tolerated at all. Relevant experiments reveal that computing performance decreases by 30% for every 1‰ of data loss.

To achieve lossless networks, synergy between networks and computing and storage service systems is more than important ever. Some industry vendors have implemented innovative solutions for network-storage synergy and network-compute synergy in their data center scenarios, solutions involving distributed storage, centralized storage, and high-performance computing. Leading telecoms have also proposed a computing power network solution linking data centers based on application- and compute-aware requirements. Technologies such as all-optical, end-to-end slicing, and elastic scheduling are used in scenarios like distributed storage and cross-node distributed computing. These solutions aim to provide zero packet loss services and build an efficient and lossless network between compute resources.

Socialized data collaboration: Improving data efficiency

Raw materials for production reflect different levels of productivity that human society has achieved throughout different stages of history. Data is a new raw material for production. Data is the foundation of digitalization, networking, and intelligence. It has been rapidly integrated into production, distribution, circulation, consumption, and how we manage our social services. It has been profoundly changing the way we produce, live, and engage in society. Currently, the explosive growth of data is not only driving rapid growth in the digital economy. It is also impacting traditional forms of production. New industries, business models, and patterns are emerging and will become key resources driving economic and social development.

Bringing the benefits of industry digitalization to every corner of society will break down corporate boundaries, and the ability to obtain and use data is becoming the key to more innovative services and improved user experience. Sales platforms can engage in precision marketing, sending messages to potential customers based on their browsing history. Manufacturing enterprises can analyze production line data to adjust production in a timely manner to improve production efficiency. Smart home companies can analyze customer living habits and create smart homes to improve living standards. Various applications show that data can create a lot of value after being effectively mined and integrated. One common view in the industry is that data will gradually become a fourth core competitiveness, alongside people, technologies, and processes. Data sharing and exchange across corporate boundaries are now quite popular. In the future, we can expect to see multi-domain data aggregation, AI integration, and better privacy protection, and we can expect to see data treated more like a commodity. Take the rural household loans, a form of inclusive finance, as an example. Risk analysis involves household details, a government credit check, data about who they know, what their agricultural land is like, and any agricultural capital they may have. The data is collected from their peers, form government sources, agricultural capital suppliers, satellite remote sensing, and the Internet. With such a diverse range of sources, data transactions will no longer be point-to-point. The transactions will have to run on an intermediary-based multi-layer data transaction system.

Digitizing society for governments and public utilities can make social governance more targeted and people-friendly. For example, under the centralized urban management of the Chinese government, a platform for interconnection needed to be built. This platform would bind governments and enterprises together, integrating all of government and social data sources, to make

full use of public data such as that which comes from telecoms, and water and power utilities. In addition, governments need to create and share data better. To this end, cameras and sensors of different departments should stay on 24/7, in all scenarios.

Data differs from traditional resources in several ways. First, data is plentiful and reusable. Second, data is highly mobile. It can flow much faster, be made much more widely accessible, and more deeply permeate society than traditional resources. Third, data use is non-exclusive. Within certain restrictions, and with the right permissions, data can be reused. In the future, social data will be either available and visible or available but invisible. Their combination will contribute to a regular cross-enterprise and cross-industry interconnection mechanism, as the support for diverse, collaborative, and common governance in the era of digital economy.

Digitalizing society can create new value from data as it flows, is shared, and is processed. However, the aggregation of massive data may create serious security issues. Once the infrastructure is threatened by security issues, there will be serious consequences. For example, the data center of one European cloud service provider caught fire in 2021. As a result, 3.6 million websites were paralyzed and some data was permanently lost, which was a huge loss for society. How to effectively utilize and protect data has become a major concern for the secure and stable operations of the digital economy. Data security technologies and management methods should be continuously updated to meet rapidly changing security requirements, and data centers and related basic networks, cloud platforms, data, and applications should be integrated to better ensure security. Only with these measures can infrastructure and data security be guaranteed.

Human-machine collaboration: Improving operation efficiency

Traditional data centers are operated and maintained by people, and the limits of human capabilities will become an O&M bottleneck for the data centers of the future. According to the latest research from China Academy of Information and Communications Technology in 2023, more than 60% of data center breakdowns was caused by manual operations. As data centers continue to grow and provide more and more services, eventually, a human-based O&M model will be no longer viable.

According to the Chinese association standard Evaluation Method for Intelligent Operation and Management of Data Center Infrastructure, data center operation can be divided into five levels, where level 1 is purely manual operations and level 5 is fully automated. By 2030, the O&M level of leading data centers is expected to reach L4, the highly automated level. At this level, predictive troubleshooting and analysis, emergency handling, and AI energy efficiency management are all almost entirely automated in running state.

Operations to run unattended, data centers need to be digitalized, networked, and intelligent throughout their lifecycles. The planning, construction, and O&M of tomorrow's data centers will all be supported by intelligence. By 2030, with the rapid development of remote monitoring, data analysis, human-machine interfaces, and robotics, simplified and efficient intelligent data centers with human-machine collaboration will become a new direction for industry development.

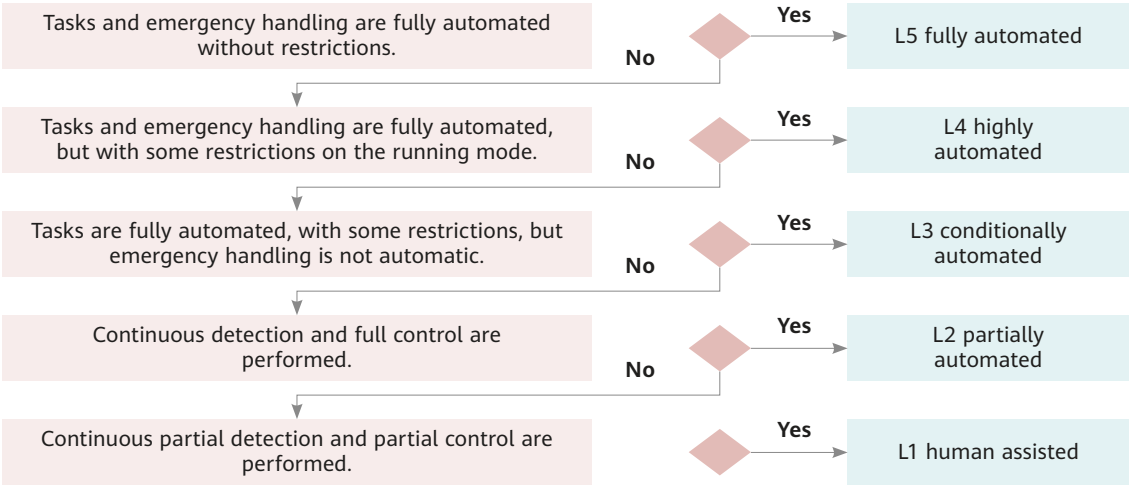


Figure 2-1 Five levels of data center automation

Our Vision for Future Data Centers and Their Key Technical Features

03

Vision

Society is accelerating towards an intelligent world. All industries are seeking ways to accelerate development. Data centers have emerged as the computing foundation of the new digital infrastructure and the engines that drive the development of the digital economy. Over the next decade, data centers will not only need to deliver 100 times more computing power to meet the requirements of fast-growing intelligent services, but also become 100 times more efficient in order to meet the long-term goal of green and sustainable development.

We believe that future data centers will be defined by six technical features: diversity and ubiquity, security and intelligence, zero carbon and energy conservation, flexible resources, SysMoore, and peer-to-peer interconnection.

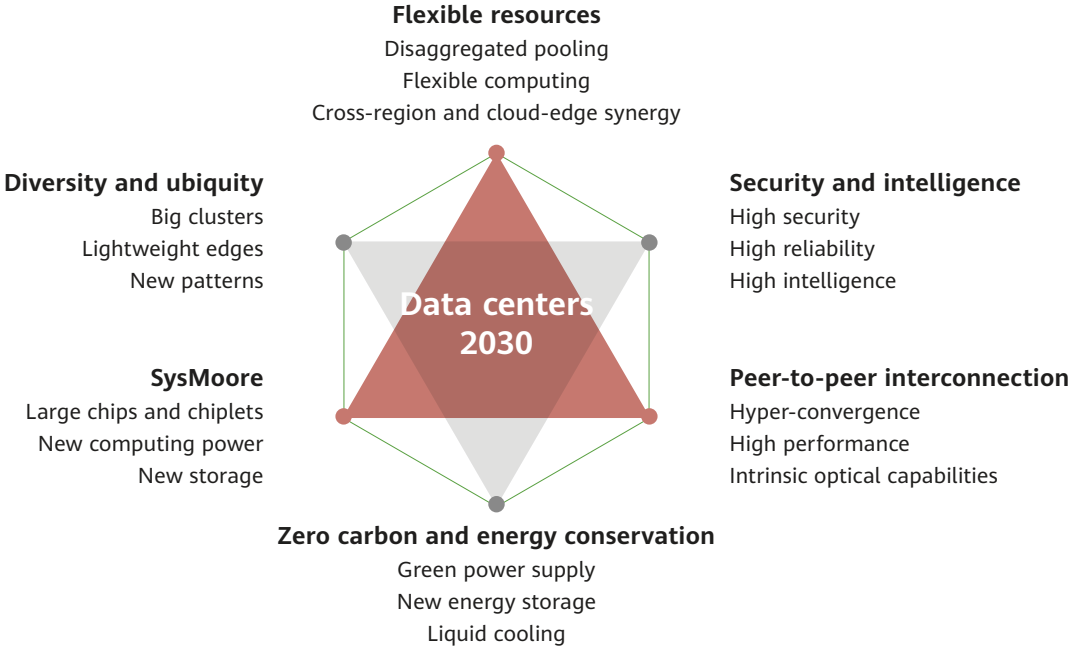


Figure 3-1 Key technical features of data centers 2030

Key technical features

Diversity and ubiquity

In the future, data centers will see polarized development. On one hand, the construction of hyperscale, intensive data centers will continue to grow. It is estimated that by 2030, the effective general-purpose computing power and AI computing power provided by a single cluster will reach 70 EFLOPS and 100 EFLOPS, respectively, with exabytes of storage capacity. On the other hand, lightweight edge computing nodes that can satisfy the low latency and stringent data security demands of a range of industries will be widely deployed. By 2030, over 80% of data will be collected and processed by lightweight edges, and over 80% of industrial production equipment will be connected to lightweight edges through IoT and digitization. Innovative data centers, such as space data centers and underwater data centers, will emerge to cater to new scenarios. Data centers in various forms can satisfy deployment requirements in different scenarios, and will maintain the development momentum of digital economy.

(1) Big clusters

A hyperscale, intensive data center, or a data hub, may contain 10,000 up to 100,000 servers. This requires highly efficient server deployment and streamlined O&M. However, in conventional data centers, servers are deployed one by one. Before a server can be rolled out online, it needs to be unpacked, installed in a cabinet, connected to power cables, network cables, optical modules, and optical fibers, and registered. From an O&M perspective, even if one person can maintain one thousand servers, an O&M team with nearly one hundred engineers would still be required for a hyperscale data center, considering work shifts. Therefore, conventional deployment and O&M methods will soon be unable to meet the requirements of hyperscale data centers in the near future. O&M is shifting from server-focused to cluster- and even data center-based O&M, and servers will be packaged, shipped, and

deployed in cabinets to significantly improve efficiency and lower human resource costs.

- **Pre-assembled delivery**
  Moving the server installation work from data centers to the factory manufacturing lines can improve efficiency and reduce costs throughout the entire process. Burn-in tests based on actual configurations can be carried out within the manufacturing facilities. Testing items that are usually unavailable at data centers, such as the temperature stress, can be easily added when done in manufacturing centers, to make the tests more complete and detect potential defects early. It is also more efficient to repair any faults that are identified in manufacturing centers. Moreover, shipping entire cabinets instead of individual servers is more cost-effective and can reduce the costs of packaging, storage, and shipping by almost 70%.

- **Integrated cabinet engineering**
  The global pooling of power modules provides centralized power supply for cabinets, and dynamically adjusts the power supply based on the load to ensure that cabinets work as efficiently as possible. Dynamic adjustment of power supply and energy storage makes it possible to cope with sudden surges in power demand during peak hours. For example, using built-in, liquid-cooled cabinet doors or liquid cooling technology can increase the heat dissipation capability of each cabinet to 60 kW.

- **Innovative cluster backplanes**
  Cabinets use cable backplanes, instead of optical modules or fibers, to connect servers with TOR switches. These are more reliable because cable backplanes are passive components that do not consume power.

These innovations, which include pre-assembled delivery, integrated cabinet engineering, and innovative cluster backplanes, can implement blind mating for servers and eliminate manual errors in cabling. Hyperscale data centers can automate O&M to increase scaling flexibility without increasing deployment and O&M complexity.

(2) Lightweight edges

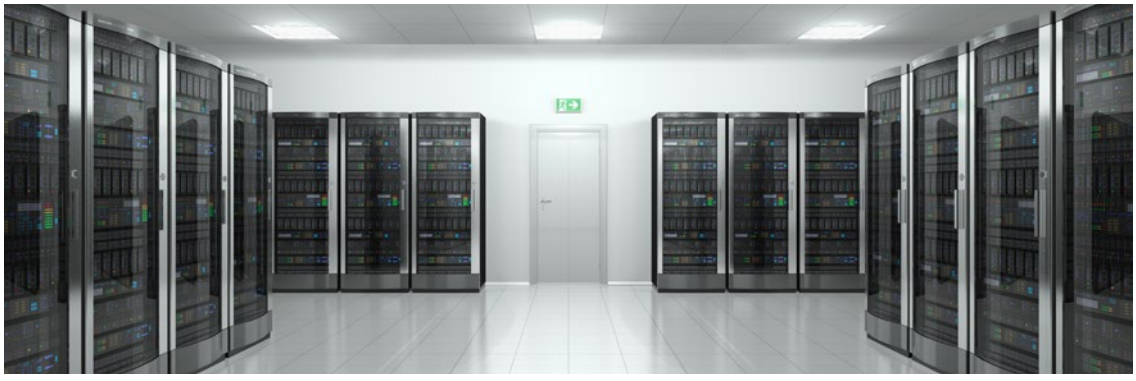
Cloud-based digitalization and intelligence are no longer benefits exclusive to the Internet industry, as they have penetrated into all sectors, and expanded their scope from non-real-time web transactions, social networking, search, and back-end IT support services to real-time interactive media, metaverse AR/VR, industrial manufacturing systems, robots, and even IoT. In this context, applications and data carried by hyperscale and intensive data centers cannot guarantee the low-latency access and processing needs of consumer smart terminals, industrial IoT terminals, and robots in any location. Extending the

cloud's elastic resources, application services, and intelligent inference capabilities from hyperscale data centers to lightweight edges that are closer to access terminals should be an urgent priority.

Lightweight edges refer to "lightweight edge clusters" and "lightweight edge services and applications". In terms of the former, cloud service vendors provide small-scale hardware computing clusters and distribute them in appropriate network locations. Then, through physical or logical private lines, some core capabilities of full-stack cloud services, such as elastic VMs/containers, storage, networks, middleware, databases, media processing, stream data processing, AI inference, and other latency-sensitive services and applications can be extended from regions to edge clusters. In terms of the latter, latency-sensitive services and applications such as middleware, databases, media processing, stream data processing, and AI inference are deployed in the form of lightweight containers or functions in hardware and OS environments. Such hardware and OS environments are provided by cloud service vendors, carriers, enterprises, households, individuals, and third parties, and are connected with central cloud data centers via the Internet and over HTTP/HTTPS protocols to penetrate firewalls. Lightweight edge services and applications are light and flexible, because they are not bound to edge computing hardware or the private lines that connect edges with data centers. Lightweight edge clusters which load full-stack cloud services from data centers, can provide richer cloud services and capabilities.

- **Lightweight edge clusters**
  Lightweight edge clusters can be classified into the following two types based on whether they have access to the Internet:

  Type 1: open public lightweight edges that have the ability to access the nearest Internet. They allow the offloading of public cloud resource pools, cloud services, and network access capabilities to city Internet data centers (IDCs), content delivery network (CDN) edge sites, 5G multi-access edge computing (MEC) devices, and other related locations. This



creates edge clouds that start small with just a few servers and can subsequently grow to thousands of servers, with high bandwidth, low latency, and high performance. The core technical features include: (1) **Low-latency access**: Local Internet service provider (ISP) ingress points are available to interconnect multiple carrier networks, providing urban areas with Internet access within 10 ms. (2) **Diversified edge computing power**: By offloading heterogeneous computing power, such as the ARM, GPU, and NPU, to the edges, scenarios such as video rendering, edge AI inference, and cloud mobile phones/games can greatly benefit from more efficient edge data processing, in addition to CDN-enabled acceleration of hot website and video content caches. (3) **Cloud-edge synergy**: Edge computing and central regions are interconnected through high-speed backbone networks or private lines. After the high-frequency, low-latency, and large-bandwidth hot data processing is achieved on the edge side, the less frequently accessed warm and cold data can be transmitted to the central cloud for processing and archiving, and this implements tiered processing. The central cloud's basic and advanced services are extended to the edge infrastructure to implement center-edge synergy, network-wide computing power scheduling, and unified network-wide management and control.

Type 2: lightweight edges that are exclusively used by specific enterprise cloud tenants and do not present the Internet egress externally. In addition to low latency assurance, these lightweight edges focus more on local compliance and multi-region branch deployment with cloud center-based unified management. By seamlessly integrating public cloud infrastructure and cloud services and deploying them to users' equipment rooms, these edges can provide standardized full-stack public cloud service capabilities on user premises. They can satisfy the diverse business and scenario needs of enterprise users and provide comprehensive and consistent cloud service experiences at locations closer to users' businesses through highly integrated hardware and adaptable cloud service software. The core technical features include: (1) **High integration**: There are various computing and storage servers that can be applied to multiple environments and scenarios and can reuse public cloud standards, and they provide standard elastic cloud

computing services, such as cloud hosts, cloud containers, and cloud storage. (2) **Elastic deployment**: Dedicated converged node models are independently designed for different edge scenarios, allowing 1 to 500 highly elastic nodes to be deployed at a single site. (3) **Customized hardware**: Customized lightweight hardware is provided for different equipment room environments (no equipment rooms, fields, standard equipment rooms, and micro modules) and scenarios such as secure and trusted computing, HPC, AI computing, and serverless. This lightweight edge type requires a collaborative effort between enterprise users and public cloud service providers to ensure the reliability of infrastructure and maintain normal power and network supplies in equipment rooms.

- **Lightweight edge services and applications**
  The cloud has extended its services and applications from large-scale central service areas to lightweight edges near end users. Since being decoupled from the hardware, the cloud is able to fully utilize thousands of heterogeneous edge nodes and millions of server resources globally, without relying on hardware, infrastructure, or deployment forms (bare metal, VM, and container), and this enables cloud-based applications to serve a wider range of industries. Real-time nearby access is implemented for new media applications, with an edge-cloud interaction latency of less than 1 ms. The lightweight advanced edge function capabilities allow media, robots, web 3.0, and other services with real-time interaction requirements to naturally run at the edges. This creates a novel real-time interactive operator computing mode that covers a variety of regions and services and that can be used within an edge, between edges, and between edges and clouds.

(3) New patterns

With the rapid development of digital economy (represented by big data, AI, and the metaverse), data centers must meet users' demands for low latency and deliver the ultimate experience, while overcoming challenges such as insufficient land resources and energy supplies. In addition to the two mainstream development trends of data centers, namely the big clusters and lightweight edges, the industry is also exploring new data center patterns to meet specific scenario requirements. These include carrier access network edge data centers, underwater data centers, and space data centers.

- **Carrier access network edge data centers**
  In recent years, the carrier access network edge data center has emerged as an innovative pattern that offers users desired services and computing functions on the edge nodes of access networks, which brings application services and content closer to them. Through network collaboration, it ensures reliable and optimal service experiences. According to Huawei's predictions, global carriers will deploy more than 10,000 MEC nodes by 2030. These nodes have the following advantages:

  Low latency: MEC is deployed closer to user access networks to provide efficient and high-performance heterogeneous computing power. This greatly reduces the latency of data distribution and provides users with faster and smoother service experiences.

  Data localization: MEC brings computing power directly to cities and counties. It allows local data processing at edge nodes and eliminates the need for data transmission across networks. This localized management approach ensures the security of enterprises' core data assets, while providing visibility, manageability, legality, and compliance to fully protect data security and privacy.

  High reliability: MEC provides optimal network connections at edge nodes and dynamically adjusts east-west connections and north-south paths based on service requirements and network statuses to prevent single points of failure (SPOFs). By supporting nearby access, MEC achieves optimal connections and provides users with more reliable and stable data transmission.

  Cloud-network integration: MEC provides a native and integrated cloud-network base and northbound interfaces that comply with the European Telecommunications Standards Institute (ETSI) and Third Generation Partnership Project's (3GPP) specifications; implements the servitization of multiple resources such as VMs, containers, bare metals,

networks, and storage; supports the self-integration of services; supports the automatic integration and provisioning of cloud, network, and service resources; minimizes edge data center management and maintenance costs; and implements the rapid development and deployment of edge services. Unified northbound interfaces are used to facilitate performance statistics and fault management of the device, network, edge, and cloud, and implement rapid fault localization and recovery for edge services.

MEC data centers have unique application scenarios and technical practices. For example, in content distribution, MEC can provide a smoother and clearer viewing experience for services such as Naked Eye 3D and Extended Reality (XR). In 5G convergence applications, MEC provides real-time and efficient data inference capabilities to provide enterprises with a more efficient and convenient digital industry experience. As of 2023, network operators in China have deployed more than 1200 MEC nodes, covering more than 90% of cities in China. With high-performance and highly integrated edge hardware, MEC data centers offer a new set of possibilities to explore in the future development of data centers.

- **Underwater data centers**

  Cooling systems are a typical method to dissipate heat in data centers, but the efficiency of conventional cooling systems has been called into question due to high power consumption, which usually accounts for one-third of the total energy consumption. One innovation to reduce energy consumption is to deploy data centers underwater, where cold seawater can be used as a natural coolant, and this helps the underwater data center to provide data storage, computing, and transmission services while achieving green, energy-saving, and efficient objectives.

  Underwater data centers have several key advantages over on-land data centers. As a natural resource, seawater can take away the heat generated by a data center, with

little to no impact on the environment due to its high specific heat capacity. Such resource-saving innovation can lower the energy consumed during heat dissipation and cooling. This is evidenced by China's first submarine data cabin. It runs a PUE of 1.076, which is much lower than that of conventional data centers. Underwater data centers are efficient and do not require evaporative heat dissipation, which means cooling towers or cold water systems are no longer necessary. As a result, zero water consumption can be achieved. In addition, since most facilities are located under the sea, an underwater data center requires on average just one-tenth of the land space that a conventional one needs.

Underwater data centers can deliver low latency. Many of the world's major Internet companies and high-tech enterprise clouds are located in developed coastal areas. For this reason, underwater data centers are superb options for latency-sensitive services because they ensure servers are closer to end users, which shortens the transmission distance and latency needed in fields like industrial Internet and telemedicine.

Another benefit is the low deployment and operating costs of underwater data centers. Land prices in developed coastal areas are high, so deploying data centers under the sea can greatly reduce land costs. When coupled with lower electricity fees due to lower energy consumption, it is easy to see that total operating expenses will also fall.

Since underwater data centers have such significant advantages, the industry is working to commercialize their use. In March 2023, the first cabin of the world's first commercial underwater data center was officially put into operation in Hainan, China. It is the world's largest underwater data cabin and is set to become a new paradigm for green data centers.

- **Space data centers**
  Unattended, self-driving data centers are becoming the trend for enterprises trying to maximize efficiency. Deploying data centers in space, the extreme edge of the network, may be an ideal option. The commercialization of space is gathering pace. There is already a foundation of

satellites that provide circuit, broadcast, and navigation services, making the next step in space data centers possible. The next decade is expected to see the launch of commercial space stations and thousands of satellites into low-Earth orbit. Many are hoping to send data centers into orbit to build the digital infrastructure that will fuel the space orbit economy.

There is great potential for space data centers, thanks to benefits in efficiency. Solar energy in space can provide a stable and continuous power supply to data centers, and this will reduce the energy pressure and carbon dioxide emissions on Earth. The low-temperature environment in space greatly affects the temperature control mode for data centers, and reduces energy consumption. Another advantage is the enhanced utilization and transmission rate of space data and the reduction in the volume of data transmitted between satellites and the ground. Despite the large number of satellites now located in low-Earth orbit, there is competition for resources that, when combined with an ever-growing volume of data, may cause delays in transmissions between satellites and the ground. By contrast, space data centers ensure data is collected and used directly in space, in close proximity to the computing and application ends. The combination of data centers and satellite communication networks provides enhanced edge computing that supercharges system efficiency and reduces service latency. Another key benefit is the high data security and low operating expenses. The main operating expenses of a data center come from maintenance and energy consumption, but the inherent environmental advantages of space greatly reduce operating expenses. In addition, for a space data center, there is a very low probability that the data will be tampered with or intercepted during transmission. Therefore, edge computing of satellite data is more secure.

The high cost is one of the main problems in constructing space data centers. The costs of everything from R&D, to the launch of a space data center, can add up to more than CNY1 billion on average, meaning that such projects must be undertaken by enterprises valued at over CNY10 billion. Other issues include resistance to space radiation and the exceptionally high reliability requirements for servers in these data centers. Innovations in dedicated computing chips and magnetic random access memory (MRAM) will be the catalysts for a new era in space data center operations.

As the cost of delivering payloads into Earth's orbit declines every year, there are expectations that by 2030, space data centers will have evolved enough to become a reality.

Security and intelligence

(1) High security

The digital economy relies on data, algorithms, and computing power. Ensuring their security and compliance is essential for digital economy development. Data centers provide the infrastructure that underpins the digital economy, so they must not only function as a platform for data, algorithms, and computing power, but also serve as a hub for data circulation and transactions. As such, data centers must have built-in security throughout their system design. That security must also be maintained throughout the data processing process from computing and storage to networking. Everything from the chips to the applications used in data centers must be able to defend against various security threats. Because of this, security investment in data centers is projected to account for 20% of the total data center investment by 2030.

The data infrastructure of data centers must be designed to protect all types and levels of data. Full-lifecycle data security, compliance, manageability, and controllability require multiple security policies that can be configured for data use, storage, and transmission based on data value and compliance requirements. In addition, data centers must be able to automatically adapt their security policies to new security levels and rules.

Data centers also need to provide zero-trust security solutions and use fine-grained access control models, such as role-based access control (RBAC) and attribute-based access control (ABAC), to strictly manage data in use.

Encryption should be utilized during data transmission and flushing to disk. The entire cryptography system must take into full consideration the risks posed by quantum computing attacks. For high-value data, confidential computing, federated learning, homomorphic encryption, and other protection solutions should be provided so that data can be made available without being exposed.

Better protection against malicious attacks will be crucial, especially for high-value data. High-value data and common data in fact should be stored separately. Security solutions for high-value data, such as write once read many (WORM) and key management and distribution mechanisms, will have to consider security design in both hardware and software to ensure integrity and confidentiality of high-value data.

As we look towards the year 2030, our security research will have to shift its focus beyond traditional equipment and network security, and put a greater emphasis on trusted computing, confidential computing, and the

development of a new security ecosystem for foundation models in the AI era.

- **Trusted computing in new computing paradigms**

  The Trusted Platform Module (TPM) emerged in the late 1990s as a sophisticated solution that uses cryptography to address software integrity protection and cryptographic key storage issues. However, it has become increasingly difficult for the TPM to effectively support software integrity measurement of cloud VMs and heterogeneous computing software. Mainstream open source communities therefore began to support software TPM (swTPM). However, the TPM services provided in most VMs are simulated using the libTPM software which does not use hardware root of trust (RoT). As a result, TPM service security is often not effective.

  Software integrity measurement for VMs and heterogeneous computing software will require extending the current trusted computing technology TPM standard to support multiple trusted computing instances starting from hardware. If this is achieved, the software integrity measurement service can be provided for multiple VMs as well as for trusted execution environments (TEEs) and XPUs in heterogeneous computing environments. This will allow trusted computing technologies and standards to eventually be built on the hardware RoT and support cloud and heterogeneous computing environments.

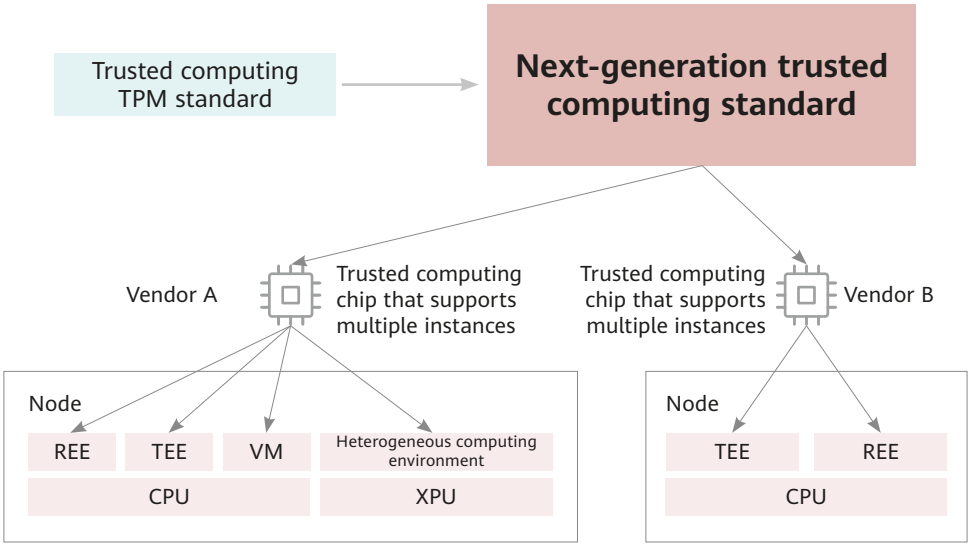


Figure 3-2 Trusted computing in the new computing paradigm

- **Heterogeneous confidential computing for the digital economy**

  Confidential computing is a new cloud computing technology in which code runs in a hardware-based TEE to ensure data confidentiality and integrity in the environment as well as data operation process confidentiality. Under the digital economy, confidential computing requirements will be driven by trust requirements between enterprises and users, internal enterprise security requirements, and organizational data-sharing requirements. First, user data must be made secure and private without relying on enterprise trusted computing environments. This data must be protected in all common risk scenarios, such as the scenarios caused by malicious cloud administrators. Second, enterprises will put in place effective measures to protect their own data security in untrusted environments. Typically, this means they will need to securely manage keys in edge computing devices deployed in public places. Third, organizations will need to conduct data cooperation without exposing their own data. A typical scenario related to this is multi-party computation (MPC) and modeling.

  This kind of confidential computing has gained increasing industry recognition and acceptance as governments around the world enact new data protection and privacy laws and regulations. An increasing number of enterprises now favor data security and privacy protection solutions that use both software and hardware protection technologies to build a secure and trusted computing environment for data sharing and exchange. However, confidential computing is mainly used in device applications (such as payment and facial recognition on mobile phones, tablets, and other devices) at the moment. These applications have high security requirements, but handle small volumes of data. The other main scenario where confidential computing is being used is in applications on the cloud that leverage confidential VMs or containers to implement block chain, key management, and other such functions. Confidential computing for general-purpose computing and AI computing involving large-scale data is currently in the trial and exploration stage. There is still a long way to go before it can be fully applied.

  By 2030, as big data applications and AI foundation models mature, it will become increasingly common for organizations to share data in order to explore its value and train more accurate foundation models. Confidential computing can ensure high computing performance, data security, and availability without risk of exposure during large-scale data sharing and computing. It is set to become the future mainstream technology for data security. To efficiently meet the demanding computing power

needs of foundation models and big data, CPU-centric confidential computing will gradually evolve to data-centric heterogeneous confidential computing while remaining compatible with existing foundation model software frameworks. This shift will allow a wide range of computing power devices (such as GPUs, DPUs, and NPUs) to collaborate and accelerate the safety computing power of confidential computing.

Specifically, the heterogeneous confidential computing architecture of 2030 is expected to have the following features:

(1) Safety computing power that is fully compatible with the common computing power ecosystem and can be flexibly configured. Users will be able to flexibly choose whether to utilize safety computing power for their computing tasks and data. This flexibility will prevent users from having to change their application and software ecosystems when they use confidential computing technologies, thereby preventing extra workloads.

(2) Safety computing power that is extended to various computing power devices instead of being limited to the TEE in the CPU. These devices include GPUs, DPUs, and NPUs. This will allow for more efficient use of heterogeneous hardware for computing acceleration and offloading, while remaining compatible with the existing software ecosystem. Additionally, this will make access control and communication encryption available to ensure the security and trustworthiness of heterogeneous computing.

(3) Safety computing power that is expanded from a single node to multiple nodes. This will allow for flexible scheduling and unified management of computing power resources that extend across entire data centers, effectively meeting the needs of different organizations to share and federate large amounts of data.

- **A new AI security ecosystem for foundation models**

  As we enter an era of foundation models, exemplified by ChatGPT, AI will play a critical role in more fields and fundamentally transform the way we live and work. As more and more AI applications are adopted in critical infrastructure, the business value of AI will increase. However, new security threats and attack methods will increase in kind. These attacks will target vulnerabilities specific to AI models, including data poisoning, model backdoors, adversarial examples, model extraction, and prompt injection in large language models. Additionally, there will be increased risk of conscious or unconscious abuse of AI technologies, such as using AI to commit fraud or to create misinformation and deepfakes. These could lead to massive data and privacy breaches.

  The security and trustworthiness of AI systems and applications have become a common concern for many countries, communities, industries, and users. Major countries, regions, and international standards organizations are also exploring new approaches to effectively regulate emerging AI systems, applications, and services. In 2021, the EU became the first to introduce significant legislation on AI with their draft EU AI Act. This Act sets out specific requirements for high-risk AI systems, including data governance, accountability, accuracy, robustness, and cybersecurity requirements. The Chinese government has released the *Provisions on the Administration of Deep Synthesis of Internet-based Information Service and also the Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment)* to address the urgent threat of AI abuse posed by the proliferation of foundation models and *AI-Generated Content (AIGC)*.

  Against this backdrop, it is imperative for the industry to propose innovative security technologies and develop security solutions to address AI security issues and threats.

(1) AI lifecycle security: Security must be built into the entire AI lifecycle, including security governance during the R&D and use of AI. The security of AI models must be improved through continuous model security assessments, and AI applications must also be continuously monitored throughout their use so that security issues can be promptly resolved.

(2) Securing AI using AI: Traditional security measures cannot identify or defend against emerging AI security threats, such as adversarial examples and prompt injection. To tackle this challenge, security attacks must be detected and countered by innovative AI security models that leverage advanced end-to-end learning and generalization capabilities based on deep neural networks.

(3) Transparency and traceability technologies for supervision over AI: There is a general international consensus that supervision over AI must be strengthened to prevent and minimize any of AI's negative impacts on society and ensure AI for the benefit of all. By utilizing innovative technologies that promote transparency, accountability, and traceability, all parties involved in the AI lifecycle can have their rights and responsibilities defined in a clear and trustworthy manner. This is the only way to ensure that AI is truly beneficial.

(2) High reliability

Highly reliable data centers are important to the development of the digital economy. As we approach 2030, data centers will transition from having high reliability at the device, node, and intra-city levels to achieving high reliability across multiple regions. They will also have to transition from data-level reliability to service-level reliability. Both system- and service-level availability will need to reach 99.999%.

To ensure the service-level reliability of data centers in different regions, new key technologies must be further researched, including data consistency assurance across multiple data centers, remote multi-active data centers, and AI-based high reliability.

- **Data consistency assurance across multiple data centers**

  Currently, technologies such as active-active and synchronous replication can ensure data consistency within a single data center cluster or between two data center clusters in the same city. However, balancing data consistency and long-distance latency remains a challenge.

  To maintain data consistency across multiple data centers over long distances in different regions, optical network transmission technologies and distributed database technologies will have to be further explored. Optical network

transmission delivers ultra-low latency, and distributed databases can be used across multiple data centers. Chronization and precise clock synchronization technologies will also be necessary. Finally, it will be important to take into account SLA policies such as latency and data consistency. By doing so, flexible and large-scale data consistency protection can be achieved for multiple data centers that do not share physical proximity.

- **Remote multi-active data centers**
  Implementing remote multi-active data centers will be a systematic project, requiring multi-active service sharing, precise scheduling, and traffic self-consistency from the network access layer to the data, storage, and compute resource layers.

  As cloud computing and low-latency, high-bandwidth network connection technologies continue to advance, resource pools across multiple data centers will be integrated into virtual data centers. This means that upper-layer services will not be aware of what regions they are operating in (which is considered "regionless"). This will lead to high data reliability and service continuity, regardless of the geographical locations of the data centers, laying a foundation for remote multi-active data centers.

- **AI-based high reliability**
  Service continuity is difficult to ensure using current data center failover and emergency management through preset operations, manual decision-making, and manual triggering.

In the future, data centers will use AI technologies to prevent and detect potential risks. These AI technologies will be integrated with internal environments (including data center IT health and power supply), external environments (including power supply networks and earthquake awareness systems), security posture, and other elements. AI-powered prevention algorithms for deep self-learning and big data analysis algorithms can be used to intelligently predict disaster correlation and enable automated prevention and response. In the event of a failure or disaster, data centers can automatically perform full-chain self-healing. They can also effectively predict and carry out both scheduled and emergency responses to

address potential risks before services are affected. However, a comprehensive and compliant disaster recovery (DR) operations management system will be required to visualize all of these elements, monitor the entire process, and perform intelligent decision-making, automatic failover, and visualized commands.

The three technologies can greatly improve service continuity assurance capabilities across multiple data centers, fully schedule data center resources, and improve resource utilization.

(3) High intelligence

Investment into data centers is rapidly increasing, resulting in larger data centers and a higher device density within data centers. The complexity of data centers is also on the rise, making traditional construction and operations methods less effective. AI and data will play a crucial role in the planning, construction, and operations stages in the data center lifecycle by improving the efficiency, power consumption, and intelligence of data centers.

- **Enablement with AI**

  The use of AI technologies in data center planning, construction, and operations can significantly enhance the efficiency of data centers while reducing costs. The integration of an AI-powered intelligent management system with data center power supply and cooling systems can significantly reduce power consumption and the likelihood of operational failures, while improving the operational efficiency. For example, applying AI to uninterruptible power supply (UPS) management can greatly improve the quality of a data center's power supply. A UPS system can monitor the main parameters related to the input power grid and output load quality in real time, and use AI algorithms to proactively learn and analyze historical data. AI can also be integrated with systems and components in data centers to improve the operational efficiency. Intelligent application O&M can make data center operations more efficient by improving the operations processes and standards. Intelligent network O&M is also one of the most important scenarios for intelligent data center O&M as it can continuously improve network visualization, manageability, and controllability. AI-powered network O&M technologies can implement network management, control, O&M automation, and optimization in data centers, helping them better recover from network failures, manage congestion, and achieve network self-optimization and self-evolution. The network itself then has execution, monitoring, analysis, and decision-making capabilities in any scenario, implementing closed-loop management and automation. Ultimately, this all provides users with better network services.

- **Digital twins in data centers**

  The digital twin technology uses historical data, real-time data, algorithms, and models to simulate, verify, predict, optimize, and control physical entities throughout their lifecycles. It greatly improves the automation and intelligence levels of data centers and offers competitive solutions to challenges related to secure operations, energy conservation, emission reduction, and more. In the data center design phase, the use of the digital twin technology mainly involves simulation evaluation and 3D visualization. This technology will eventually be able to automatically optimize data center design solutions. In the data center construction phase, the digital twin technology can be used to manage the construction progress, quality, and security, visualize the progress, and help coordinate human and material resources, making data center construction more intelligent. In the data center O&M phase, digital twin visualization uses the 3D technologies to perform data processing, modeling, and simulation and create digital mappings between twin objects, including campus buildings, equipment room layouts, infrastructure, cold aisles and cabinets, IT devices, and strong- and weak-current links. Visualizing information about IT facilities, power and environments, capacity, links, and alarms, and simulating, analyzing, predicting, and verifying data can provide a stronger basis for decision making, helping data centers improve and eventually downsize.

As larger and more centralized data centers continue to develop, traditional data centers will move away from their rigid structures and inefficient management and operations to digital, networked, and intelligent models. The use of AI and digital twin technologies will help maximize investment and operational efficiency throughout the entire data center lifecycle. The use of new technologies (such as intelligent O&M robots) will also help reduce O&M workloads by enabling independent diagnosis and automatic troubleshooting, and improving defenses. By 2030, industry-leading data centers are expected to reach L4 automated operations, and will be approaching truly unmanned status. The advancement of high-intelligence data centers will continue to help us move away from human labor towards technology-enabled operations that better support the digital economy.

Zero carbon and energy conservation

(1) Green power supply

Data centers consume high amounts of energy and account for a significant proportion of global carbon emissions, which leads to a high OPEX. As carbon neutrality gains momentum around the world, more data centers will accelerate their green and low-carbon transformation plans. Mirroring such positive trends and advances is the green power industry. In recent years, green power is undergoing a renaissance and expanding its global footprint with more cost-effective electricity, providing data centers with a viable option to achieve carbon neutrality. As more global green and low-carbon development policies are adopted, the proportion of clean energy such as wind and solar energy in the energy mix of data centers will increase. It is estimated that by 2030, large data centers will be exclusively run on green power.

- **Increasing green power utilization**

  (1) Wind power

  Wind power forms a major proportion of the renewable energy mix and is widely available and pollution free. Wind power can effectively control the impact of the increasing energy supply being placed on the environment. With the excessive consumption of energy around the world, more interest and investment has been injected into the research and utilization of renewable energy. Wind power is primed for large-scale development and commercial application. Various countries have been increasing their investments in technical research on wind power generation and its related technologies. The annual growth rate of the global wind power industry has reached 40%, with more than 100 countries stepping into the industry, making it an integral part of the global energy market.

  The pursuit for developing a green and low-carbon data center industry creates a great opportunity for wind power suppliers, creating a win-win cooperation between the data center industry and such suppliers. To give you an example, in 2013 Huawei built a data center in Ulanqab, an area that is surrounded by abundant wind power resources and provides wind power plants with a high installed capacity and low electricity prices. The solution has reduced costs and improved energy efficiency for the above data center.

  (2) PV power

  Thanks to technological advancements, photovoltaic (PV) power is seeing a considerable increase in cost-effectiveness (PV power prices are more or less at parity with coal prices thanks to large-scale PV plant development), technical strength, and public awareness. PV power is playing a crucial role in addressing climate change,

reducing energy use costs, and ensuring energy security. Distributed PV systems can be constructed near data centers to reduce power supply costs. To save land resources, PV systems can be even built on the rooftops of data center buildings. PV power is increasingly used to energize auxiliary facilities or secondary loads in data centers, such as lights, elevators, and monitoring systems. Multiple hybrid models, such as "PV + energy storage" and "PV + power grid," can uninterruptedly provide clean power for data centers to meet their electricity demand around the clock.

(3) Hydropower

As another type of clean and renewable energy resource, hydropower delivers numerous advantages in optimizing the electricity mix, ensuring safe operations, reducing power consumption, and improving the economic benefits of electricity. Data centers built in areas rich in hydropower resources can be powered by clean energy and cooled by local water. To reduce energy consumption and costs, a number of data centers in China are deployed in areas rich in hydropower resources. The Dongyuemiao Data Center, located in the vicinity of the Three Gorges dam of China, is fully powered by hydropower and cooled by the Yangtze river.

- **Dynamic microgrid**

  Using green energy resources such as wind, hydro, and PV power are crucial to the future power supply strategies for data centers, offering energy sustainability, and green and low-carbon development. However, green energy is unpredictable and is therefore unable to supply stable power over a long period of time, which can sometimes lead to a fluctuation or even failure of the power system.

  A microgrid is a small power generation and distribution system composed of distributed power supplies, energy storage and conversion equipment, loads, monitoring systems, and protection equipment. The microgrid allows data centers to consume local green energy when it is sufficient, avoid energy loss during transmission in the power grid, and improve energy utilization efficiency. It can be connected to the power grid through a single point and obtain the power from the grid when the energy supply is unstable. The microgrid adopts advanced control and uses a large number of power electronic devices. It connects distributed power supplies, energy storage equipment, and controllable loads together, so that it becomes a controllable load for the power grid system and can operate in either grid-tied or off-grid mode. In this way, both the microgrid and power grid can run safely and stably.

Up to now, there are multiple practices in deploying data center microgrids. For example, the renewable energy microgrid project of the Green Energy Center of Zhangbei Cloud Computing Base in China has a total installed capacity of 220 MW and will produce about 450 million kWh of electricity each year after being put into operation.

As the research of key microgrid technologies and the development of green and low-carbon data centers accelerate, microgrids will enter a period of rapid development.

(2) New energy storage

Energy storage technology has become an important way to reduce power costs in data centers by peak shaving. Data centers are hungry for power, where power costs account for 60% to 70% of a data center's OPEX. Power supply companies usually offer different electricity prices during peak and off-peak hours. Data centers can use energy storage systems to store power during off-peak hours and use the stored power during peak hours to reduce costs. According to the World Energy Outlook 2022 report released by the International Energy Agency (IEA), an increasing number of countries and regions have set renewable energy development goals and plans to accelerate their respective energy transition to green power. According to the French government's plan, the share of renewable energy in the power generation mix of France will increase to 40% by 2030, and the installed PV power capacity will increase tenfold and 50 offshore wind farms will be built by 2050. Japan's latest basic energy plan proposes that the share of renewable energy in power generation will increase to 36%–38% by 2030. The global renewable energy industry has officially entered the fast lane. With a higher market penetration rate of renewable energy, a larger demand for power system load balancing and a longer duration of energy storage will be created.

- **More lithium, less lead**

  As the demand for internal space capacity management and operational efficiency increases in data centers, data center reconstruction with increased power density is becoming an important pathway for data center upgrades. Lithium-ion batteries are rapidly becoming the next-generation energy storage equipment that is substituting lead-acid batteries in data centers thanks to their high energy density, output voltage, and safety. An increasing number of data centers have started to use lithium-ion batteries as power supply units. Compared with traditional lead-acid batteries, lithium-ion batteries offer multiple advantages.

  · Size: Lithium-ion batteries are small and light. Data center operators can directly place lithium-ion batteries in a higher position without using a reinforced floor.

  · Floor area: Lithium-ion batteries occupy only one third of the installation space for lead-acid batteries and therefore can better adapt to the environment of modular data centers. Lithium-ion batteries are easier to transport and install.

  · Service life: Lead-acid batteries have a short service life and usually have to be scrapped after three to six years of use, while lithium-ion batteries have a service life of 10 to 15 years. A longer service life means that battery replacement and maintenance costs in data centers will be significantly reduced.

  · Reliability: Mainstream data centers use long-life lithium iron phosphate cells and a four-level protection architecture to effectively ensure battery charging and discharging performance.

  · Management: Lithium-ion batteries can be combined with a more advanced battery monitoring system (BMS) to provide information such as battery O&M time and health status for data center O&M personnel. As the utility power supply keeps improving, energy storage batteries will find limited application scenarios in data centers. However, lithium-ion batteries will be more widely used as they can effectively reduce the O&M costs of data centers while facilitating safe and efficient management.

- **Hydrogen energy storage**

  Hydrogen energy complements the global carbon emission reduction strategies because it does not emit greenhouse gases or fine dust when being burned. Renewable energy sources such as wind and PV power are volatile and intermittent. Hydrogen energy overcomes these shortcomings, becoming a key supplement to the world's energy transition. Hydrogen energy storage is receiving more attention and seeing more applications around the world. More than 20 countries and regions

have released hydrogen energy strategies, and breakthroughs are being made in related technologies.

Electric energy storage methods mainly include pumped energy storage and lithium ion batteries. Whereas, hydrogen energy storage has advantages such as a long discharge time, high cost-effectiveness for large-scale storage, flexible storage and transportation, and zero environmental impacts. In addition, hydrogen energy storage can be used in a variety of scenarios. On the power supply side, hydrogen energy storage can reduce power curtailment and suppress fluctuations. On the power grid side, hydrogen energy storage can regulate the peak capacity of the power grid and relieve the congestion of transmission and transformation lines. As an engine of the digital economy, data centers are set to support the intelligent transformation of various industries. And the safeguarding of data centers is of strategic significance to this mission. If hydrogen energy is used to supply the power in data centers, hydrogen leakage may cause combustion and explosion, which will result in physical damage to the data centers. Therefore, one of the top priorities involves effectively managing hydrogen safety and ensuring zero accidents.

During the development of hydrogen energy, some enterprises have already applied this technology in data centers. For example, a data center uses up to 4 MW fuel cells as a substitute for diesel generators to provide backup power. Though still in its infancy, the ever-improving hydrogen energy storage will be deeply coupled with data centers.

(3) Liquid cooling

Besides IT equipment, the cooling system is the second biggest energy consumer in a data center. IT equipment in a data center continuously emits heat while running. When the power exceeds the rated range, servers may break down, causing service interruption and compromising the equipment service life. Therefore, a cooling system needs to be used to ensure the normal operation of IT equipment. To reduce the power usage effectiveness (PUE) of data centers, advanced cooling technologies are particularly important and have gradually become critical to data centers. Advanced cooling should be green, energy saving, innovative, modular, and integrated. In addition, intelligent approaches should be used in collaboration with the operating status of IT equipment to implement dynamical adaption and regulation.

Liquid cooling technology is applicable to high-power and high-density data centers. Data centers are exploring this highly potential technology. The average power density of data center racks increases year by year. Accordingly, the demand for liquid cooling technology is surging, and the market scale of liquid-cooled data centers continues to expand. Liquid cooling technology not only saves energy and reduces noise, but also improves the server density per unit space, boosting the computing efficiency and stability of data centers.

- **Full liquid cooling**

  At present, there are three technical pathways of full liquid cooling: cold plate, immersion, and spray. Cold plate liquid cooling involves indirectly transferring heat from heat-emitting components to the liquid coolant enclosed in the circulation pipeline through cold plates, and taking heat away through the coolant. In cold plate liquid cooling, the coolant is separated from the object to be cooled and is not in direct contact with electronic devices. The heat emitted from the object is transferred to the coolant through high-efficiency heat conduction components such as cold plates. Therefore, cold plate liquid cooling is also called indirect liquid cooling.

  Immersion cooling is a new heat dissipation technology that has attracted much attention across the industry in recent years. A specific coolant is used as the heat dissipation medium and IT equipment is directly immersed in the coolant, which dissipates the heat emitted from the running IT equipment through coolant circulation. The coolant exchanges heat with the external cooling source through the circulation process to release heat to the environment. With a special architecture, immersion cooling delivers the following unique advantages: First, the coolant used for immersion cooling is in direct contact with heat-emitting equipment and provides high

heat dissipation efficiency. Second, the coolant has high thermal conductivity and specific heat capacity, but little change in operating temperature. Third, energy efficiency is significantly improved as IT equipment with higher power density can be deployed. Immersion cooling features higher density, more energy saving, and better noise prevention performance than air cooling.

Spray cooling is a solution that deploys spray modules inside servers to spray an insulating liquid that cools heat-emitting components, and is harmless to humans, IT equipment, and the environment. Spray cooling features high component integration, heat dissipation efficiency, and energy saving but low noise. It is one of the most effective measures to reduce the cooling costs of IT systems and improve energy efficiency when high-power racks are deployed in data centers.

Multiple challenges still lie in liquid cooling. The deployment environment varies with the type of liquid cooling. Deploying liquid cooling systems in traditional equipment rooms would increase the deployment cost and difficulty. The compatibility of IT equipment and liquids and the friendliness of liquids to humans need to be considered when customers choose to deploy immersion or spray cooling. Cold plate liquid cooling does not require costly chillers. It reduces the total cost of ownership and improves the energy efficiency of data centers.

- **Air-liquid hybrid cooling**
  As liquid cooling is costly, data centers may choose a combination of air cooing and liquid cooling to reduce the CAPEX while achieving an optimal PUE. The combination of the two solutions helps customers reduce costs and achieve the PUE target. Air cooling and liquid cooling systems are located in different equipment rooms of the data center and are independent of each other.

Air-liquid hybrid cooling has become a new trend in the development of data center cooling technologies. In the future, a new landscape of "air cooling + liquid cooling" hybrid development will emerge in the data center market. Air cooling technology will not be completely replaced by liquid cooling technology. Customers can choose different data center cooling solutions based on their requirements. For data centers with low power per rack, customers still prefer air cooling. For high-density and large-scale computing scenarios such as supercomputing and energy survey, a combination of cold plate liquid cooling and immersion cooling can be flexibly selected with the cost factor considered, and liquid cooling can be a choice for energy-intensive components and equipment. Liquid cooling and air cooling technologies together will drive the future development of the industry.

- **Optimal PUE**

  Energy consumption and carbon emission indicators reflect the core competitiveness of data centers. They will be helpful to unlock the energy saving and emission reduction potential of data centers and raise the energy efficiency standards.

  An optimal PUE can be achieved through a combination of multiple factors in a data center. Data center characteristics vary with regions and industries. Low-carbon and energy-saving technologies should be used in the key systems of data centers throughout the life cycle from planning, design, and deployment, to management and O&M to achieve an optimal PUE.

- **Optimal water usage effectiveness (WUE)**

  Water resources are fundamental to the survival and development of our species, and strategic to maintain ecosystems and support socioeconomic development. Data centers are major water consumers. It is extremely important to strike a balance between PUE and WUE and achieve an optimal WUE.

  Data center PUE and WUE are closely related. Water evaporation is one of the most efficient heat exchange methods. High water consumption helps data centers achieve a better PUE. Therefore, we need to strike a balance between PUE and WUE and achieve an optimal WUE based on actual service requirements and geographical environments. For example, in areas with abundant water resources, an optimal PUE can be achieved by using water, while in areas with scarce water resources, the water consumption can be reduced by collecting rainwater, recycling waste water, and reducing the operating duration of chillers under wet conditions. It is estimated that the WUE will fall below 0.5 L/kW x h by 2030.

Flexible resources

Public clouds, industry clouds, and private clouds are widely used as digital and intelligent platforms in many industries. Large-granularity applications such as AI foundation models, metaverse, and digital twins are seeing explosive growth. This means that the cloud architecture will likely become a de facto standard for future data centers. Cloud operating systems will be added over hardware so distributed global data centers can provide large-scale, intensive, and scalable computing, storage, and networking resources on demand, all while ensuring the multi-tenancy security and performance SLAs for diversified applications government and enterprise customers require in many industries.

Next-generation cloud data center architectures will continue moving towards disaggregated pooling, flexible computing, and cross-region and cloud-edge synergy to achieve optimal return on investment.

(1) Disaggregated pooling

Resource pooling is an essential characteristic of cloud computing. It allows multiple tenants and applications to share physical resources to the greatest extent possible. However, the constraints of current technologies and data center architectures mean resources are often separated into large numbers of fragmented resource silos. This hinders large-scale and intensive sharing of resources. In the next 5 to 10 years, disaggregated pooling will become increasingly common in cloud data centers. Specifically, CPUs of different generations, memory on different nodes, storage (decoupled from compute), heterogeneous computing power, and DCN networks will be pooled respectively.

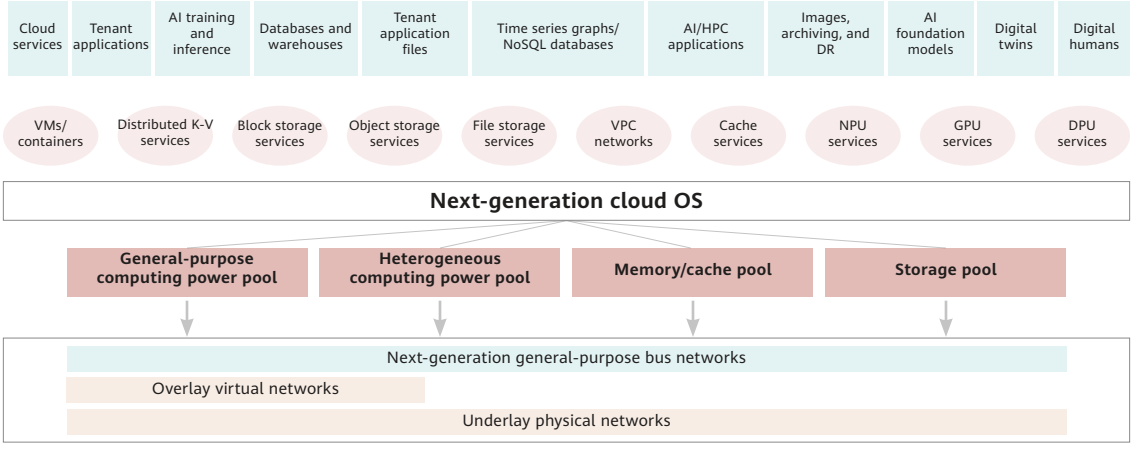
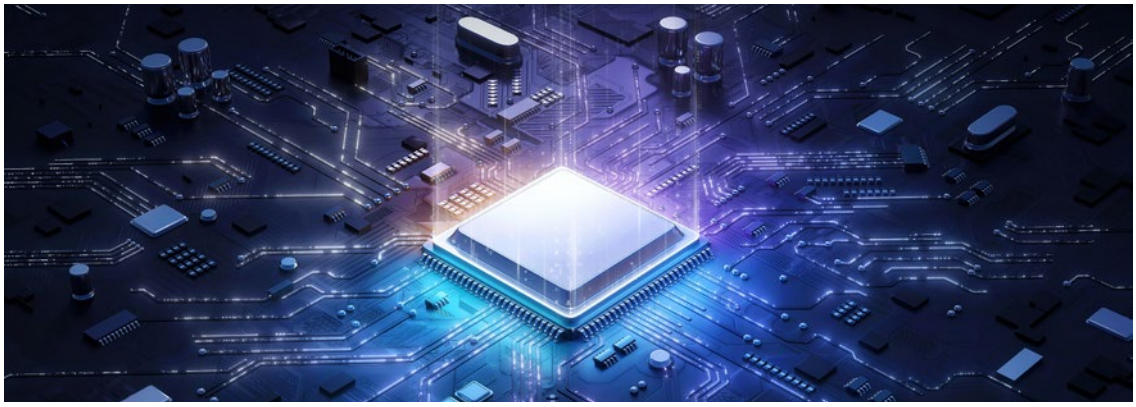


Figure 3-3 Disaggregated pooling in cloud data centers

- **CPU pooling (with multiple CPU generations)**
  Currently, computing power in a cloud data center is provided using a resource-centric model. When you choose a flavor for a VM or container, you are choosing a specific generation of CPU, such as Intel Sandy Bridge, Ivy Bridge, or Kunpeng 920/930. The resource pool for each generation of CPU is independent from each other. Cloud tenants like to select the latest generation of CPU. This leads to a lot of wasted compute resources using previous generations of CPUs, even though they are still within the useful life phase of the bathtub curve. To respond to this challenge, next-generation data centers will provision computing power in an application-centric model. This will abstract away certain CPU hardware differences for upper-layer compute services and resource scheduling layers. Additionally, it will leverage black-box real-time QoS detection to identify application QoS requirements and dynamically schedule CPU resources to achieve the optimal CPU overcommit ratio while meeting cloud tenants' SLA requirements on application performance.

- **Memory pooling across nodes**
  Cloud data centers allocate computing power from matching servers based on predefined flavors to minimize resource fragmentation. Each flavor contains information about how many CPUs and how much memory a VM or container should have. Memory cannot be overcommitted like CPUs. When CPU resources are insufficient, only application performance is affected. If memory resources are insufficient, application begin to run abnormally and may even fail. As a result, a large number of memory resources in VMs and containers are over-configured and underutilized.

  Theoretically, if some servers do not have sufficient memory resources, they can borrow idle memory resources from other servers in their computing cluster over the network. This is technically possible so long as the network can meet



the transport bandwidth and latency requirements. With unified bus and Remote Direct Memory Access (RDMA), bandwidths can reach hundreds of gigabits per second, latency can be as low as a few hundred nanoseconds, and tail latency can be 10-fold lower. It is technically possible to break through the physical boundaries between servers and centrally allocate memory resources to VMs and containers. However, cross-server remote direct memory access is one to two orders of magnitude slower than direct memory access through double data rate (DDR) memory channels. This can decrease application performance by 20% to 30%. Therefore, cross-server memory pooling is applicable only to cloud applications that can tolerate performance deterioration of up to 30%. It is not a good choice for performance-sensitive multi-tenant VMs or containers. Millisecond-level migration and instant memory allocation are still required to ensure application performance.

- **Heterogeneous storage and cache pooling**
  Unstructured data storage (such as block storage, object storage, and file storage) can use decentralized cross-AZ distributed key-value storage engines to create unified storage resource pools. However, semi-structured and structured data storage (such as SQL row-based and column-based databases, NoSQL databases, multi-dimensional data warehouses, graph databases, and time series databases) still integrate storage and compute to provide compute-side functions (data query, change, analysis, and processing) and storage-side functions (data persistence, availability assurance, parallel I/O read/write, and lossless elastic capacity management). Integrated storage and compute faces five unique challenges:

  · Different data processing and analytics tasks usually have different scaling requirements for compute and storage resources.

  · The cross-node data redundancy mechanisms of compute nodes and storage nodes conflict with each other. As a result, some databases, data warehouses, and big data clusters can only use servers with integrated storage and compute. They cannot utilize software-defined elastic storage or share compute resources with other compute services or tenant applications.

  · Data I/O paths often take multiple detours at both the compute side and the storage side, causing data-access performance bottlenecks.

  · Sharing data assets between different processing phases is difficult. Copying data and storing copies for redundancy purposes is expensive.

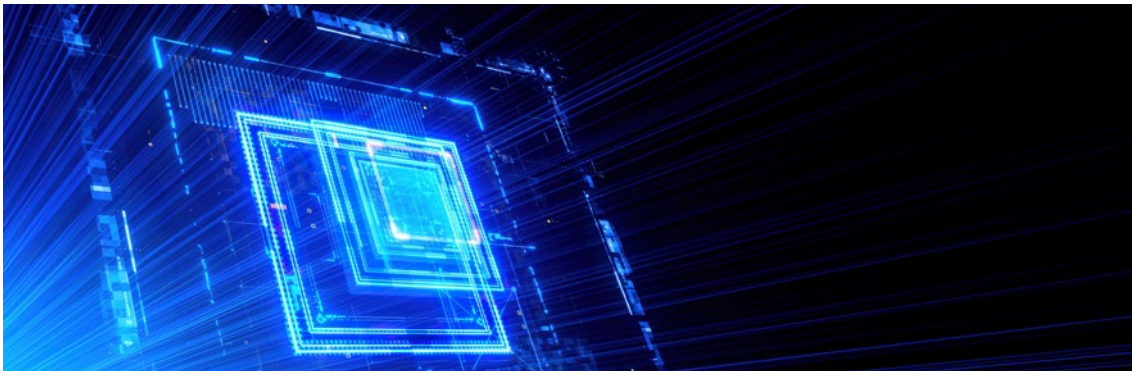
· The cross-AZ multi-active architecture has a complex processing logic to ensure data redundancy at the computing layer.

To address these challenges, next-generation cloud data centers will have to take multiple measures (such as data copy sharing across data compute engines, near-compute cache pooling, distributed calculus offloading for near-data processing, unified metadata management for heterogeneous compute engines, and intelligent tiered data storage) to create unified storage resource pools for structured, semi-structured, and unstructured databases.

- **Heterogeneous compute pooling**

  With the advent of AI foundation models, metaverse, and digital twins, cloud-based GPU/NPU heterogeneous computing power will gradually replace general-purpose CPUs and become the key computing power for AI training and inference, digital humans, and digital twin cities. Demand for such computing power will grow exponentially. However, in the primary/secondary compute architecture, GPUs and NPUs function as secondary PCI devices and are attached to a specified number of CPU cores for exclusive use by cloud servers or containers. Servers with a single GPU or NPU card or multiple GPU cards cannot meet the training requirements of foundation models. The PCI buses in servers and TCP/IP or RDMA networks across servers have to enable close collaboration between GPU clusters. This severely limits the linear acceleration of GPU and NPU clusters and impairs the cost-effectiveness of foundation model training.

  The advancement of networking technologies such as unified bus and RDMA enable NPU or GPU cards across clusters to be fully meshed to improve the cost-effectiveness of foundation model training. With software-defined GPU or NPU pooling, a physical GPU or ASIC acceleration chip can be divided into several or even dozens of isolated compute



units, and GPU or NPU chips on different physical servers can be aggregated for an operating system (on a physical machine or a VM) or a container to complete distributed tasks. CPU servers without GPU or NPU acceleration chips can also invoke GPU or ASIC acceleration cards on remote servers to complete AI computational tasks. CPUs can be decoupled from GPUs and heterogeneous computing power can be pooled to provide more scalable GPU and NPU resources.

- **DCN network pooling**

  A distributed software-defined overlay network can act over a physical switching network to pool multiple DCN networks. This allows VPC networks to be elastically provisioned for tens of millions of VMs and millions of tenants when there is a unified physical network with millions of interconnected physical servers. This can help overcome the bottlenecks that hinder horizontal expansion of traditional hardware routers and gateways, decouple logical network addresses from physical network addresses, and allow for flexible ACL policy setting for interworking. However, the use of software-defined overlay networks increases the complexity of multi-tenant, multi-application network technology stacks. Locating network connectivity faults can be a daunting task.

  Ethernet/IP network ports have been upgraded to hundreds of Gbit/s, the capacity of physical switches has been upgraded to Tbit/s, and user-mode Data Plane Development Kit (DPDK) and data processing unit (DPU) have been introduced to continuously optimize the throughput and latency of multi-tenant overlay networks. However, network transmission and routing between tenants or applications still uses the decades-old TCP/IP protocol stack. The link-layer Ethernet lacks efficient E2E flow control, the transport-layer TCP has a high processing overhead and a long transmission delay, and retransmission upon packet loss is inefficient. In terms of E2E QoS, these overlay networks cannot meet increasingly demanding requirements, like ultra-large bandwidth, extreme-low latency, and predictable network latency and packet loss between distributed concurrent processing units or microservices, when tightly coupled, large-granularity cloud applications are required for foundation model training, metaverse simulated rendering, and search-based recommendation. Overlay networks also restrict further efficiency improvements in pooling storage, memory, and heterogeneous compute resources. Although RDMA can overcome some of the challenges, the network cannot be expanded to 100,000 or millions of nodes, and end-to-end precise flow control is still far from meeting the requirements of cloud data centers.

Next-generation data centers will move away from this two-layer pooled architecture and towards a lightweight, RDMA-enabled, single-layer pooled architecture. This single-layer pooled architecture can support CPU/NPU uninterrupted processing and DPU offloading, and allow seamless scaling to millions of nodes. This architecture can minimize unnecessary overhead across multi-layer protocol stacks, better support compute and storage resource pooling, and increase the cost-effectiveness of tightly coupled, large-granularity cloud applications.

(2) Flexible computing

Elastic computing is a mainstream for provisioning computing power in cloud data centers. Cloud service providers predefine flavors of VMs or containers so that cloud tenants can select a flavor and a CPU overcommitment ratio based on their applications' performance requirements. Resources can then be billed based on the selected flavor and CPU overcommitment ratio. This resource provisioning method minimizes fragmented compute resources. However, it usually allocates excess resources, and average utilization rates of compute resource pools is only 20% - far lower than the average allocation rate of 80%. Nearly half of the physical computing power from tens of millions of cloud servers around the world remains idle.

To address these issues and dynamically adapt to applications' changing requirements for compute resources, cloud data centers will introduce flexible computing - a more flexible and intelligent way to allocate and provision computing power. In 2030, function-level resource allocation is expected to be possible

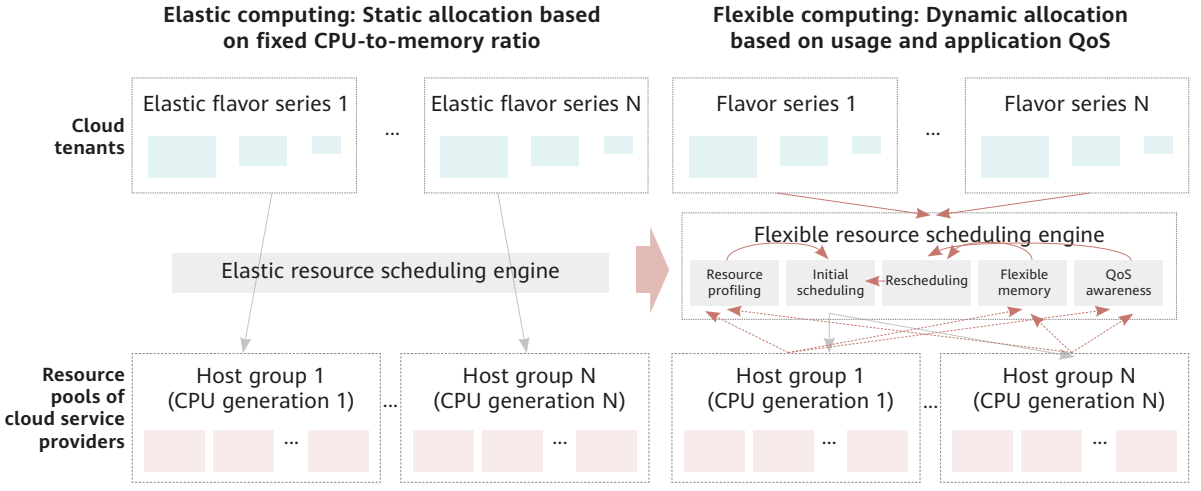


Figure 3-4 Comparison of elastic computing and flexible computing

in leading cloud data centers. This can greatly improve the utilization of compute resource pools and allow cloud tenants and developers to use dynamic computing power like other utilities, such as water and electricity. Cloud tenants and developers will then only have to pay for what they use and nothing more.

Flexible computing can scale computing power both horizontally and vertically to offer ultimate elasticity. It is much like "flexible manufacturing" where production can adapt to market changes. Flexible computing estimates fine-grained resource requirements, perceives QoS requirements, and provisions customizable computing power to adapt to subsequent changes to requirements. Flexible computing requires four key technologies: application-driven fine-grained profiling and initial resource allocation, application performance QoS degradation awareness with AI foundation models, flexible instance rescheduling, and dynamic overcommitment of flexible memory.

- **Application-driven fine-grained profiling and initial resource allocation**

  Flexible computing power is allocated based on application requirements and refined insights into resource requirements at the instance level and cluster level.

  To meet instance-level resource requirements, flexible computing removes the limitations on fixed CPU-to-memory ratios, regardless of the minimum granularity of compute resources or whether the compute resources are cloud native. Prior to instance provisioning, profiles are created based on past resource usage, and refined flavors are defined to best match service requirements. After an instance is provisioned, the host continuously monitors its resource usage, dynamically creates profiles for the instance, and adjusts the resource allocation policy to strike a balance between resource supply and demand. Additionally, flexible instances with different priorities are designed to meet cloud tenants' varying requirements for performance QoS, cost, and prioritized rescheduling upon reaching a certain degree of QoS degradation. This will help break elastic computing's restrictions on pooling resources from overcommitted and non-overcommitted clusters or from hosts with different CPU generations. With flexible computing, unified resource pools can be created from hosts with multiple generations of CPUs, instances with different priorities can co-locate on the same host, and preemptive scheduling can be used.

  To meet cluster-level resource requirements, flexible computing does not work like elastic computing, which relies on manual intervention and historical experience, and takes numerous rounds of trial and error to plan physical compute resources and define the scaling

policies for cloud servers and containers. Conversely, flexible computing uses queuing theory to calculate the minimum physical compute resources required by clusters in a given queuing probability based on the start time and end time of past tasks and multi-task concurrency characteristics. This reduces the cost wasted on trial-and-error and manually maintaining the physical compute resources. Flexible computing uses AI time series forecasting and time-frequency domain modeling tools to provide dynamic resource modeling and forecasting for VM or container clusters within several minutes or even seconds.

- **Application performance QoS degradation awareness with AI foundation models**

  Flexible computing also differs from elastic computing in that it not only perceives the dynamic resource requirements of workloads, but also quantifies the QoS requirements of workloads. A flexible computing scheduling system can initially allocate resources to flexible instances with different priorities based on the 95th percentile point of the CPU usage, average CPU usage and variance, and accumulated CPU usage of flexible hosts to control the performance conflict probability within a threshold. However, when multiple instances on a host compete for resources, there is still a chance that the QoS deteriorates to or even falls below the threshold. In this case, rescheduling will be necessary. This requires the flexible computing scheduling system be able to quantitatively assess the QoS of the workloads.

The most direct way to do this is to measure the application-layer performance of workloads in a white-box manner. However, most cloud applications run on multiple instances, instead of a single instance. Application-layer performance metrics cannot directly show each resource instance's contribution to QoS deterioration. This means QoS deterioration awareness is still required for each resource instance. Flexible computing can collect multi-dimensional performance metrics of all resource instances from the underlying host OS in non-intrusive black-box manner, including performance metrics for CPU, memory, storage I/O, network I/O, NUMA, L3 cache miss, and frontend and backend microinstruction stalling cycles.

Flexible computing can extract workload performance characteristics (such as if the workloads are CPU-intensive, memory-intensive, storage-intensive, network-intensive, or a combination of them) from the massive amounts of performance data collected with the help of AI foundation models that are interactively pretrained with self-supervised learning, fine-tuned with supervised learning, and optimized with reinforcement learning from human feedback (RLHF). A small number of typical workload training samples can then be added in supervised learning tasks to establish a fitting relationship between the resource-layer QoS and the application QoS degradation. Similar to GPT models that can be continuously trained online, a black-box workload performance QoS deterioration model can continuously enrich performance QoS features based on observable workload performance characteristics, so that it can better and better predict QoS deterioration for more types of unknown workloads.

- **Flexible instance rescheduling**

  When a flexible scheduling system detects application performance QoS of multiple VMs or containers on the same host degrading to a threshold, it triggers a rescheduling to relieve or eliminate the QoS deterioration. Rescheduling can be hot or cold migration that is not perceptible to services, or it can be a rescheduling that is perceived as a collaboration.

  Hot and cold migrations do not require any modification or adaptation to application-layer software, but hot VM migration incurs high overloads in compute and network resources. The duration of hot migrations is determined by the memory size and CPU idleness of the involved flexible instances and available network bandwidth and latency. A cross-host or cross-VM CRIU cold container migration requires that the snapshots of process-level CPU runtimes be persistently written into or read from shared SSDs and be loaded from the memory. Such migrations can interrupt services for a few hundred

milliseconds. This means hot and cold migration is preferred for the flexible instances whose QoS deteriorates by large amounts, but only when all of the following conditions are met: the physical servers support more than 100 Gbit/s of bandwidth, microsecond-level latency, and RDMA protocols; the CPU usage of VMs or containers is lower than 60%; and there is 16 GB of memory or less.

A rescheduling incurs lower resource overheads than a hot or cold migration, and the service layer and resource layer can cooperate to isolate overloaded instances and gracefully close unfinished tasks or web sessions to ensure a smooth service experience. However, slight modifications are required to application-layer task scheduling and web load balancing software. The best approach to rescheduling is to integrate QoS deterioration-driven rescheduling into the cloud-native framework to allow rescheduling without any modification or adaptation.

- **Dynamic overcommitment of flexible memory**

  Memory resources differ from CPU, storage, and network bandwidth resources in that they are exclusive to VMs and containers. In the future, memory will replace CPU as the biggest cost contributor - making up about two thirds of the total cost of a resource pool. Dynamically allocating memory resources across tenants and applications will become critical to improving the utilization of cloud computing power.

  The memory resources of a host and their guests are managed independently, so idle memory resources held by guests normally cannot be reclaimed for reuse. Memory ballooning has been introduced to allow the physical host to retrieve unused memory from guests and share it with others, however guests still need to notify the host that their memory is idle using an asynchronous notification system. As a result, guest memory may not be released promptly and the host memory may reach the upper usage limit. These are the clear drawbacks of memory ballooning.

  In the future, cloud data centers will turn to a flexible memory architecture to transform their independent layered memory page management architecture into a flat memory page management architecture. On the premise of not affecting the hypervisor-level isolation of memory pages for VMs and containers, metadata on idle memory pages will be synchronized in real time between guests and hosts to improve host awareness of guest memory requests and releases. This completely strips away the drawbacks of memory ballooning.

  In the future, cloud data centers with

the flat memory page management architecture will be able to obtain the physical memory page request and release history of all VMs and containers in a non-intrusive manner. They will be able to profile peak memory usage, average memory usage, and standard deviations, and dynamically overcommit memory to flexible instances with different priorities. When medium-priority flexible instances encounter physical memory page faults, they will be able to allocate idle physical memory pages from other servers through the RDMA network or allow read and write access to the memory swap area on SCMs and SSDs (which incurs 20% to 30% more performance overheads when compared with direct memory access). When high-priority flexible instances encounter physical memory page faults, the flexible instances with a medium or low priority can then be hot or cold migrated or a rescheduling can be activated to free up memory for higher-priority flexible instances.

(3) Cross-region and cloud-edge synergy

Disaggregated pooling in cloud data centers is restricted to the compute, storage, and network resources of a single geographical region. A typical pool has the capacity of a few million servers and the physical resource pools of multiple availability zones 50 to 100 km apart that are interconnected with

10-Tbit/s optical fibers. Construction costs, lease costs, electricity costs, PUE levels, CO2 emission factors, and expandable computing power vary from geographical region to geographical region, so computing power cost also varies. For example, computing power in the Ulanqab and Guiyang Regions of western China is 10% cheaper than that in the Regions in China's first-tier cities like Beijing, Shanghai, and Guangzhou. Horizontal cross-region collaboration will be required to break through the physical boundaries.

Demands to extend cloud resources from central areas to distributed edge sites closer to cloud tenants' access points are increasing as we see more mobile terminals and IoT technologies and cross-region deployment of enterprise services. Cloud data center infrastructure will need to further evolve from centralized to distributed deployment, and large-scale central areas will need to collaborate vertically with distributed edge sites.

- **Horizontal collaboration**

  To meet the requirements for horizontal collaboration, cloud data centers will evolve from a region-aware architecture to a regionless global computing architecture in the future.

  From the perspective of cloud service providers, multiple physical Regions and distributed edge nodes in a certain geographic area are integrated into a unified, all-domain logical resource pool. The all-domain resource scheduling engine is used to streamline all physical Regions and edge nodes in the logical resource pool. All resource requests can be allocated to tenants, thereby optimizing the input-output ratio of data centers, reducing energy consumption and carbon emissions, and solving the problems of unbalanced resource allocation across physical Regions and the imbalance between supply and demand of regional computing resources.

  From the perspective of cloud tenants and developers, cloud data centers provide a unique and logical entry for the development, deployment, and service routing of applications across all domains. In this way, cloud tenants and developers do not need to be concerned about the cloud service APIs or development framework entries of each independent Region, nor the service deployment, resource provisioning, service routing, data synchronization, wide area network (WAN) connection management, or network costs across multiple Regions and edge nodes. In this way, serverless and automatic scaling are realized in physical Regions, and a streamlined development experience parallel to that with standalone deployment can be achieved across physical Regions and across the cloud and the edge.

  Cross-region or cross-cloud-and-edge WANs have higher latency and bandwidth costs than non-blocking physical networks in a Region. Therefore, for the resource orchestration and scheduling layer, it is necessary to: 1) Define the access latency (at cold, hot, and warm layers) of application resource instances and data instances and the affinity between the instances. 2) Build a unified model for all-domain cloud resource orchestration and scheduling by taking into account the total cost of ownership (TCO) of Layer 1/Layer 2/Layer 3 of data centers, the dynamic bandwidth cost of WANs, and the basic information (such as the overall allocation rate, energy consumption, and carbon emission) related to the resource pools.

- **Vertical collaboration**

  At the core of lightweight edge/distributed cloud are deployment across geographic locations and the fact that users can hand over all or part of their edge infrastructure to cloud service providers for O&M. Such

deployment is characterized by all-domain collaboration:

(1) Service collaboration (unified tenant service applications): The applications are distributed in the primary geographic Region and multiple distributed edge geographic locations. Microservice collaboration including microservice discovery and communication across the edge and the cloud needs to be performed across different sites. Full-link governance capabilities such as routing, rate limiting, and circuit breakers need to be provided. Typical release processes including grayscale releases (canary releases) and blue-green deployment need to be supported.

(2) Service management collaboration: From the cloud, users can manage complex service models at the edge, and reduce O&M costs through centralized management.

(3) Data collaboration: Cloud data centers need to support data synchronization and sharing between distributed sites and central Regions, implement seamless connection between applications in different distributed cloud locations, and ensure data consistency between applications in distributed clouds.

(4) Resource collaboration: A unified distributed resource scheduling mechanism is constructed to select appropriate sites for resource allocation and disaster recovery (DR) orchestration based on distributed capabilities, locations, service running status, resource usage, and user habits and intentions.

In addition, in IoT scenarios with edge computing, cloud data centers need to provide collaborative management of cloud-edge-device resources. Nodes and devices need to be managed in a unified manner on the cloud. The functions of nodes and devices need to be abstracted, and data access between cloud-edge-device needs to be implemented through various protocols, with unified O&M on the cloud.

- **Global elastic network services**

  With our sights set on the target architecture of cloud data centers, over the next 5 to 10 years we expect a 100-fold increase in the WAN connection resource requirements for horizontal collaboration between heterogeneous clouds across different geographic Regions within the cloud and across different cloud vendors, in order to facilitate vertical collaboration between central Regions and distributed edge sites across clouds, and between cloud tenants' end users and distributed edge sites or nearby geographic Regions. However, considering the costs of long-haul transmission and routing devices, WANs, unlike DCNs in data centers, cannot serve as free resources for cloud tenants. WANs are scarce resources with a high cost per unit bandwidth. The key to solving these contradictions is to introduce all-domain elastic network services with transmission QoS/SLA guarantees and optimal cost-effectiveness, in order to support the skyrocketing WAN bandwidth interconnection requirements that stem from horizontal and vertical collaboration.

  "All-domain elastic network services" need to break free from the constraints of the physical exclusive mode of traditional optical fiber/MPLS leased line connections and the static WAN link capacity allocation and routing mode with a predefined maximum bandwidth. To this end, "all-domain elastic network services" need to support dynamic WAN link capacity allocation and routing capabilities for the application loads and data assets of cloud tenants. A key step in realizing this function is to support time-and-space characteristic sensing for WAN service interoperation and synchronous/asynchronous data replication traffic across Regions, across the cloud and the edge, and even across heterogeneous clouds. In doing so, a "multi-active redundancy" and "elasticity on-demand" type of mutual relationship is established between "WAN link capacity allocation and routing capabilities" and the open Internet plane. "All-domain elastic network services" need to go beyond the best-effort WAN transmission QoS guarantee of the open Internet, and provide end-to-end real-time latency optimization and congestion control capabilities for user experience-sensitive web sessions (remote API calling and web page access) and real-time media services, and provide "one-stop access to the nearest cloud" with ultimate cost-effectiveness and experience assurance for cloud tenants.

Peer-to-peer interconnection

(1) Hyper-convergence

Since Intel introduced the x86 architecture in 1978, various interconnection protocols have been developed to enable computers to offer different physical, transmission, and functional features. As shown in Figure 3-5, UltraPath Interconnect (UPI), NVLink, and Compute Express Link (CXL) work between processors; PCIe, CXL, NVLink, and SATA work between processors and peripherals and storage devices; and Ethernet and InfiniBand work between nodes. To implement the functions of different peripherals, the chip design must take into account the physical layer and controller for each specific type of interface. When the communication traffic flows across different protocol interfaces, the conversion between protocols generates extra hardware and software overheads, and also increases the power consumed during bridging.

This challenge calls for a hyper-converged interconnection architecture, which aims to break down physical boundaries between chips, reduce protocol conversion overheads, and eliminate communication software stack overheads, so as to reduce communication latency, increase communication bandwidth, and optimize interconnection utilization.

- **Streamlining intra-die protocols vertically**

  A unified interconnection protocol reduces protocol conversion and avoids level-by-level bandwidth convergence of the on-chip bus, PCIe bus, and network ports. As a result, the end-to-end interconnection bandwidth becomes the same as the processor port bandwidth.

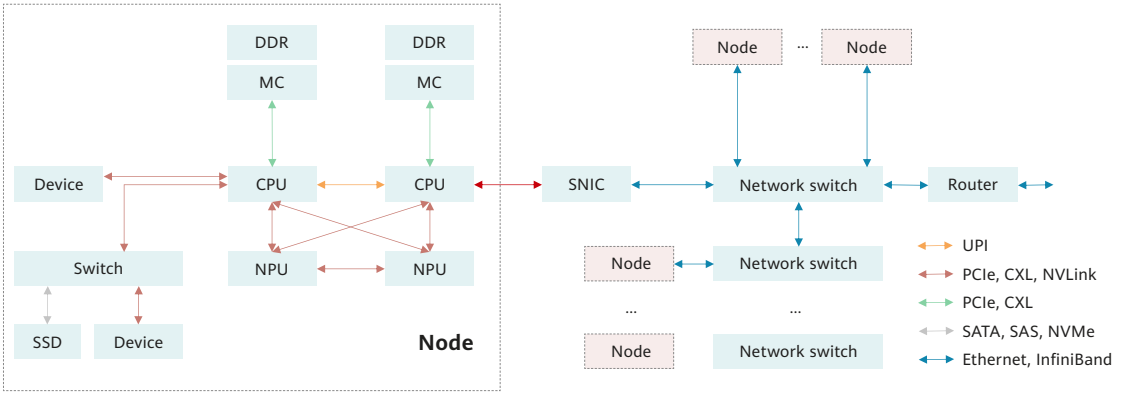


Figure 3-5 Multiple interconnection protocols in computing systems

- **Unifying link interfaces horizontally**

  The advanced memory management mechanism pushes memory semantics directly to software. Computing system components communicate with and invoke each other without any intermediates. As a result, data can be transferred between nodes with quicker memory access and lower communication overhead.

- **Building a data-centric architecture integrating storage, compute, and network**

  A single type of compute resources, a single node, or a system expansion pattern with a fixed ratio can hardly meet fast-changing application requirements. Furthermore, the interactive computing of soaring volumes of data poses big challenges to data centers' computing efficiency and interconnectivity. To streamline data processing and storage utilization, next-generation data centers must be data-centric and be built on a hyper-converged compute, storage, and network architecture.

  Employing a unified protocol stack for compute, communication, and storage services helps to overcome the limitations of legacy architectures in which the networks of general-purpose computing, high-performance computing, and storage are separated from each other. The unified protocol stack converges the three networks into one and drives the evolution of lossless networks towards a hyper-converged network architecture. It is estimated that by 2030, approximately 80% of large data centers will be operating over a hyper-converged Ethernet network.

  The evolution of the architecture is reflected in two places. (1) Storage-compute decoupling at the macro level: Compute and storage resources are deployed separately. They are connected through a high-throughput data bus and data is accessed using unified memory semantics. As a result, heterogeneous compute and storage resources such as CPUs and GPUs are decoupled in order to be scheduled in a more efficient way. (2) Storage-compute integration at the micro level: Near-data processing minimizes unnecessary data movement. Dedicated data processing computing power is deployed at the edges of data generation, on the data transmission networks, and in the data storage systems. The network, storage, and compute integration improves the efficiency of data processing.

(2) High performance

2030 will see yottabytes of data across all lines of business and industries. Data storage and compute resources must grow rapidly to meet the soaring demand for data. However, only 5% of the world's new data is utilized every year, and this figure is far from the desired

level for data value mining. Next-generation high-performance computing data centers must be delicately designed to make the most of data.

- **Scalable, massively parallel technology from chips to data centers**

  The computing power of chips is not growing as quickly as the volume of data is growing. According to a third-party research report, the AI computing demand has increased by a factor of 750 over two years, whereas the computing power of chips, driven by Moore's Law, has only increased by a factor of 2. Therefore, next-generation data centers must support the massively parallel technology to close the gap between computing power and computing demand. In a massively parallel system, each large dataset is split into small data blocks. Each computing chip in the data center only needs to process one small data block. The parallel computing middleware of a data center is divided into two layers: cross-node Spark, Flink, and Hadoop, and intra-node (chip-level) CUDA, OpenCL, and SysCL. Next-generation data centers will have unified, scalable, massively parallel computing middleware that works on a chip to data center basis.

- **High-speed peer-to-peer interconnection architecture**

  In a massively parallel system, computing chips need to communicate with each other in real time to exchange intermediate computing results. However, the computing power driven by Moore's Law has grown far beyond interconnection bandwidth and memory bandwidth. According to a third-party research report, over the past 20 years, computing power has increased by a factor of 90,000, whereas interconnection bandwidth and

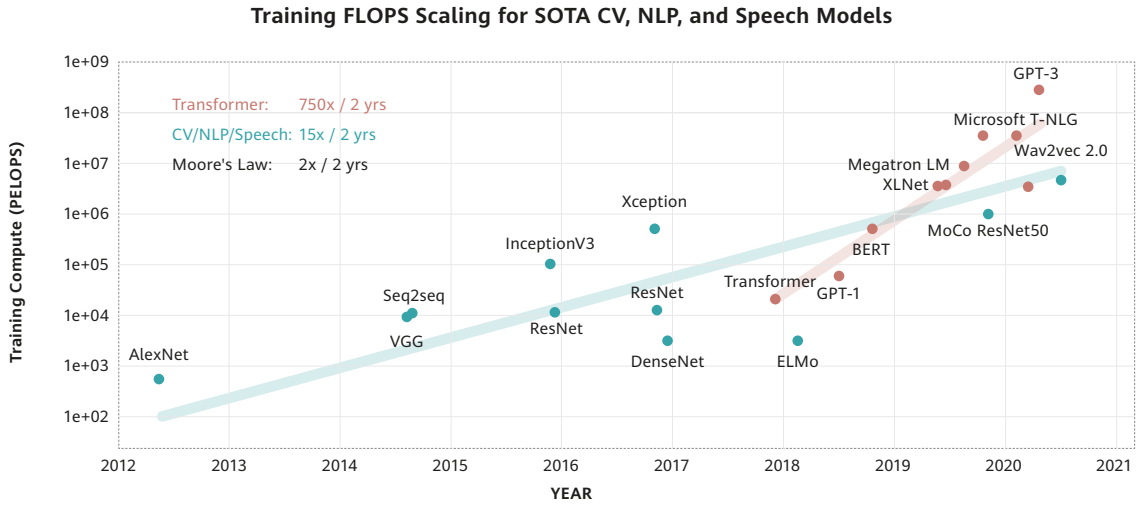


Figure 3-6 Different models' computing power demands

memory bandwidth have increased by a factor of 30. Next-generation data centers will employ a more efficient interconnection architecture to reduce the imbalance between computing power and interconnection bandwidth. Based on co-packaged on-board optics, optical switching, dynamic Torus, and optical interconnection, the high-speed interconnection architecture has been designed to handle the high bandwidth and low latency requirements of next-generation data centers. A new unified interconnection protocol is expected to eliminate data communication protocol conversion and implement peer-to-peer high-speed interconnection.

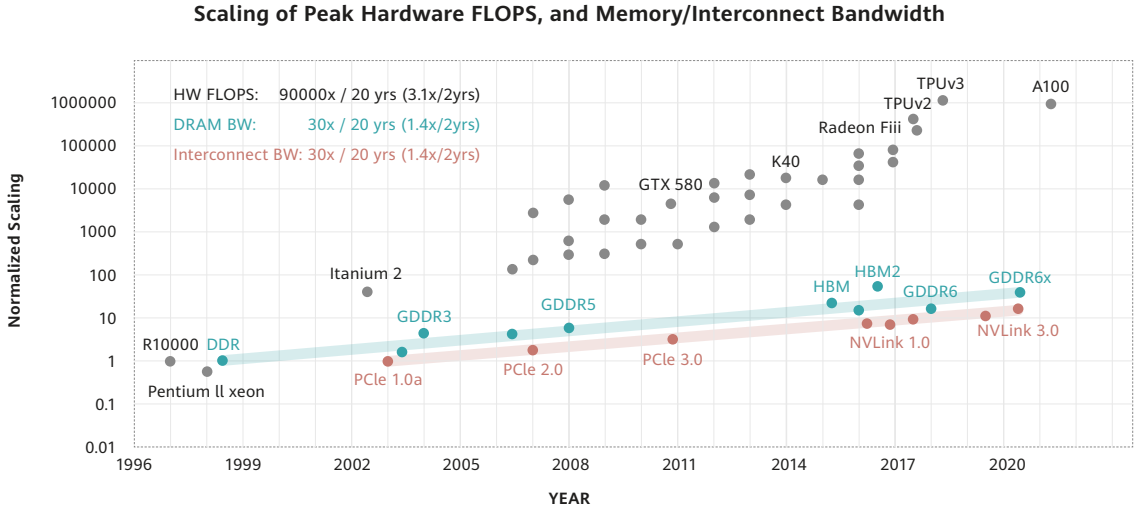


Figure 3-7 Computing power, interconnection bandwidth, and memory bandwidth

- **Lossless data center networks**

  ChatGPT is driving the emergency of AI foundation models with trillions of parameters, which far exceed the processing capability of a single GPU. A large number of GPUs execute AI computing tasks concurrently and share computing results between each other. They are interconnected over a lossless network with low latency and zero packet loss to build a large-scale computing cluster. Industry practices show that latency and packet loss issues lead to GPUs being underutilized in AI foundation models. Consequently, lossless data center networks have become a research hotspot. The industry has launched high-performance Ethernet products and chips that are specifically designed for AI.

  To establish a lossless network, hyper-converged switching technologies can be

introduced in data centers to achieve zero packet loss and 10-µs-level forwarding. To boost latency-sensitive applications such as supercomputing, network devices in data centers can participate in aggregating and synchronizing computing information, and this reduces communication latency and improves computing efficiency through computing-network collaboration.

As data centers have evolved from standalone units to a networked pattern, a lossless network across data centers has become indispensable. At present, telecom operators are exploring potential technologies for mutual sensing between computing power and networks. For latency-sensitive applications, networks can play a vital role in scheduling computing power to streamline communication with zero packet loss and deterministic latency.

- **Chip-level long instruction pipeline technology**
  To alleviate the problem of insufficient memory bandwidth, next-generation computing chips have been designed to reduce the memory access frequency. The long instruction pipeline technology divides the computing process into multiple phases, and the data in each phase is processed in parallel. With the chip-level parallelism, the data in the intermediate phases is not written back to the memory. This technology lowers the memory bandwidth consumption and reduces the imbalance between the computing power of chips and the memory bandwidth.

- **Distributed, multi-level cache systems**
  To achieve the Data Center 2030 vision, a distributed, multi-level cache system must be developed to further explore data locality and reduce long-distance communication between data centers.

  This cache system consists of multiple levels. Each cache level has its own capacity and speed. Distributed processing enables computing chips to access data more quickly, reduces their wait time, and automatically manages data based on the data access frequency and data importance. The next-generation cache system is expected to fully explore storage resources in a data center, and this will improve the overall throughput of the data center.

(3) Intrinsic optical capabilities

Powerful computing chips are witnessing an I/O bandwidth increase. The port rate of such chips is expected to reach Terabit-level or higher by 2030. According to Yole's prediction, 100% all-optical connectivity will be implemented in data centers by 2028.

As the speed of a single channel increases, serial communication with 100/200 Gbit/

s or higher speed creates challenges in power consumption, crosstalk, and heat dissipation. Against this backdrop, traditional optical-electrical conversion interfaces are unable to meet the requirements of increasing computing power. The proportion of co-packaged on-board optics in data center interconnect (DCI) will continue to increase. Compared with traditional solutions, the co-packaged on-board optics solution is expected to have 1/3 lower E2E power consumption, and will become a key technology to break through bandwidth bottlenecks and implement green development in data centers.

The network architecture of data centers will also change. The industry has started to research new optical cross-connect technologies to leverage the advantages of optical switching in bandwidth, port, power consumption, and latency. By doing so, the industry expects to meet two key system requirements in data centers: network scale and traffic bandwidth.

- **High-speed optical interface (1.6T/3.2T)**
  High-speed optical interfaces are used to connect devices in data centers. The optical interfaces include SR, FR, and LR interfaces, each differing in connection distance. Technical solutions vary depending on the transmission distance. The rate increase of the high-speed optical interfaces depends largely on the capacity of switches in data centers and serializer/deserializer (SerDes) technology development. The capacity of switches doubles every two years, with 200 Tbit/s or 400 Tbit/s switching capacity expected to become a reality by 2030. The single-port rate needs to increase to 1.6 Tbit/s or 3.2 Tbit/s accordingly.

  Optical connection technologies can be classified into direct detection and coherent detection technologies. Featuring low cost and power consumption, direct detection technologies are the most common for high-speed optical interfaces in data centers before the 800G era. As the rate increases, direct detection technologies are affected by issues such as dispersion and four-wave mixing (FWM), which shorten the transmission distance. Considering these factors, coherent detection technologies may replace direct detection technologies in data centers. In the 800G era, IEEE 802.3dJ will define the two types of detection technologies for 10 km scenarios. However, coherent detection technologies suffer from high power consumption and cost. In the future 1.6T/3.2T era, it is likely that direct detection and coherent detection technologies will coexist.

  Direct detection technologies will remain dominant in the 1.6T/3.2T era, and develop along scale-up and scale-out paths concurrently. As the rate of a single lane continues to increase, the increase of concurrent channels will require more

optical fibers or increasing use of the wavelength division multiplexing (WDM) technology, which is also developing. In the 800G era, the single-lane 100G technology will be inherited and the single-lane 200G technology will be developed. In the 1.6T/3.2T era, single-lane 100G and single-lane 200G technologies will be used for multiplexing, or single-lane 400G technology will be developed. For example, the 16 x 100G 1.6TSR solution was initiated in the IEEE 802.3dJ project. Some companies have expressed their expectations toward the 8 x 200G solution to construct 1.6T. As the solution uses the 8-wavelength multiplexing technology, it will face challenges such as dispersion and FWM, calling for research on new wavelength allocation solutions, dispersion management technologies, and low-power-consumption equalization technologies. For the single-lane 400G technology, high-bandwidth components, high-order modulation formats, and polarization multiplexing technologies can be used.

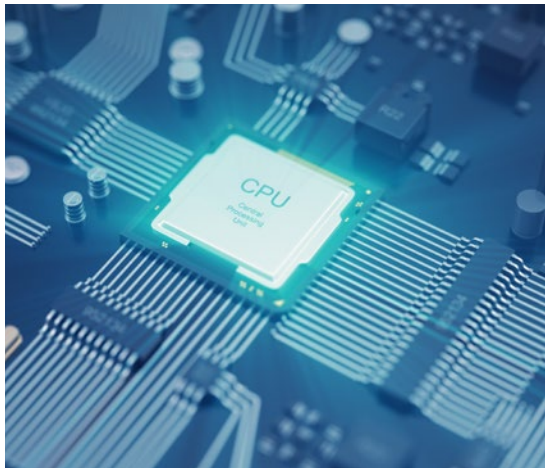
Traditionally, coherent detection technologies are used for long-haul optical transmission. Due to challenges such as dispersion and FWM, the transmission distance of direct detection technologies is continuously shortened. This has led to coherent detection technologies being increasingly deployed for DCI. Coherent detection technologies feature high transmission performance and can flexibly use the optical digital signal processor (oDSP) to compensate for dispersion. However, as mentioned already, the cost and power consumption are high. To combat this, many universities and enterprises have proposed the concept of coherent-lite. For example, low-cost light sources such as distributed feedback (DFB) gray light sources and quantum dot light sources are used to replace the distributed Bragg reflector (DBR) light sources for long-haul coherent transmission, and a light source pool is further used to share light sources. This helps to reduce the costs and power consumption. The optical domain polarization tracking scheme is used to simplify digital signal processing, and the segmented silicon photonic modulator is used to avoid digital-analog conversion (DAC) at the transmit end.

- **Co-packaged on-board optics**

  Reducing per-bit cost and power consumption has always been the goal of high-speed optical interface technology. Over the past decade, the capacity of switches has increased 80-fold, and the overall power consumption has fallen by 75%. In switches, the power consumption of application-specific integrated circuits (ASICs) has been reduced by 90%, and that of optical interfaces by 67%. Although the per-bit cost and power consumption of optical interfaces are also decreasing, such decrease is much slower than the power consumption reduction of ASICs in switches. The root cause is that optical interfaces depend on the SerDes technology, a digital-analog hybrid technology that evolves slower than ASIC in terms of energy efficiency. To further reduce power consumption, the SerDes circuit distance must be shortened or the quantity of SerDes circuits reduced. Therefore, many new technologies such as on-board optics (OBO) and co-packaged optics (CPO) are emerging in the system structure of optical interfaces, and CPO has become a hot topic in the industry.

  (1) The co-packaged on-board optics technology for data center switches — CPO

  Currently, there are two main technology paths: silicon photonics-based path and VCSEL-based path. (VCSEL: vertical cavity surface emitting laser)



  Silicon photonics technology has become the main path of the multi-channel integrated transceiver because of its high integration and compatibility with the complementary metal-oxide semiconductor (CMOS) process, which potentially reduces the cost. There are two solutions for the light sources of the CPO technology on the silicon photonics platform. One is the pluggable light source pool module technology. Considering the high failure rate of the light sources and to facilitate easy replacement in the future, multi-channel and high-power laser chips are encapsulated into a pluggable module and placed on the panel side. The chips are connected to the optical engine chips near the switching chip through polarization-maintaining optical fibers to provide a continuous laser source, a light source form widely used in the industry. For the other

solution, a few vendors have strong III-V/Si heterogeneous integration capabilities and can directly integrate light sources on silicon photonic engines. The 2:1 backup mode is implemented to improve the yield of light sources. This mode represents the second light source technology path. Currently, there are three technology paths for the high-speed modulator of the silicon photonics platform. The first is the relatively mature Mach-Zehnder (MZ) modulator technology. As the MZ size is large (hundreds of μm), after multi-channel integration, the optical engine size is large and the power consumption is relatively high. The second is the microring modulator. The microring modulator is small in size (dozens of μm) and has low power consumption (low drive voltage). However, the microring modulator requires a very stable operating wavelength tracking system. The third is an EA modulator using the Ge material, and its size is also dozens of μm. For the modulator, light absorption is enabled through the Franz-Keldysh effect.

Some vendors in the industry are also promoting the VCSEL-based CPO, largely due to its low power consumption (< 5 Pj/bit). This technology can largely meet the interconnection requirements within 100 m. In the future, VCSEL components will be upgraded to have fewer modes or a single mode. It is also expected that the interconnection length can reach the km level. Currently, the baud rate of mature VCSEL components is 25 GBd, and 50 GBd components are expected to be put into commercial use in the coming years. Although the bandwidth growth is slightly slower than the development of silicon photonics, VCSEL can use external multiplexers and demultiplexers to implement WDM to improve the single-fiber capacity. Arrayed VCSEL/PD components can also be used together with multi-core optical fibers (with a core spacing of about 40 μm) to implement large-capacity transmission.

(2) Co-packaged on-board optics technology for high-performance computing — optical I/O technology

The high-performance computing (HPC) cluster is a powerful computing platform connected by high-speed communication networks whose communication capability has become important support for the xPU cluster. Further improving the interconnection bandwidth has become a focus in the industry. Public awareness has started to grow around optical I/O technology, which places optical transceiver chips in a computing chip package, and therefore is also referred to as in-packaged optical connection technology (in-packaged optics). With this technology, the fan-out bandwidth of chips can be greatly improved, and the power consumption of optical

interconnection reduced, making the bandwidth density and power consumption equivalent to those of intra-board/intra-cabinet electrical interconnection. In addition, the interconnection distance (km level), which cannot be reached by electrical interconnection, is now realized, and a new technology path featuring low power consumption and large capacity is also provided for cluster system interconnection. The optical I/O technology is mainly enabled through silicon photonics, specifically a microring bus WDM technology with a low modulation rate (30–60 Gbit/s). For one thing, in the modulation rate range, there is a relatively optimal E2E power consumption level (about 5 pJ/bit). For another, a multi-channel integrated WDM bus is implemented by leveraging the narrowband working feature of microrings. Doing so can greatly expand edge interconnection bandwidth density, making it easy to reach 100 Gbit/s/mm or even Tbit/s/mm. Currently, this field mainly focuses on technology paths including the implementation of a dense WDM microring modulator, control of a multi-channel microring modulator, a multi-wavelength external light source technology, and an advanced packaging technology.

- **Optical cross-connect**

  In recent years, the industry and academia have widely studied new optical cross-connect (OXC) technologies. By leveraging the advantages of optical switching in bandwidth, port, power consumption, and latency, the industry expects to meet two key system requirements in data centers — network scale and traffic bandwidth. OXC is mainly classified into wavelength-level cross-connections and fiber-level port cross-connections. Up until 2030, MEMS OXC and sub-μs fast OXC technologies will be the research focus in data center scenarios.

  (1) MEMS OXC

  The micro-electro-mechanical system optical cross-connect (MEMS OXC) is an optical cross-connect system device based on the micro-electro-mechanical system technology. An array formed by a pair of optical collimators is used as an input/output (I/O) port and a pair of MEMS micromirror array chips are used to control light beams, ensuring that any input port can be connected to any output port. As such, the MEMS OXC features high integration, high rate, and low power consumption.

  (2) On-chip integrated optical switch

  Based on which key technology is applied, on-chip integrated fast optical switches are classified into five types: thermo-optic effect, free carrier effect, Pockels effect, Kerr effect, and silicon-based MEMS (Si-MEMS).

The thermo-optic effect indicates that the refractive index of the material is regulated by using the temperature-sensitive characteristics of the lattice material structure. Optical switches can have the advantages of an ultra-compact size of 100 μm, 10 mW level switching power consumption, and sub-microsecond (sub-μs) switching latency. The free carrier effect is a special effect based on silicon materials. The length of optical switches ranges from 300 μm to mm-level, and the switching latency reaches ns-level. Both the Pockels and Kerr effects are non-linear optical effects. The electro-optic response time of an optical switch with a non-linear optical effect ranges from ps to fs, with no extra loss generated. However, a higher drive voltage or a longer component is required. Leveraging the attraction/rejection behavior of the suspended waveguide structure by electrostatic force, the silicon waveguide micro MEMS system (Si-MEMS) directly changes the physical spacing between waveguides to change the optical path. The switching speed of the optical switches reaches the sub-μs level. Compared with other technologies, Si-MEMS can provide higher isolation and lower loss, and allow a more compact structure. However, the reliability and durability of switches are restricted by the mobile waveguide or metal electrode structure.

- **New fiber media**

  The development of next-generation DCI focuses on high rate, high density, low latency, low cost, and easy O&M. In this case, the application of new optical fibers will revolutionize the optical interconnection of data centers. With special and excellent fiber features, the hollow-core fibers and multi-core fibers will further promote optical interconnection with lower latency, higher density, and lower costs in data centers.

(1) Hollow-core fibers

Hollow-core fibers eliminate the limitations of traditional quartz optical fibers. Based on the anti-resonance mechanism and specific cladding structure design, hollow-core fibers can restrict light transmission to air fiber cores to change the transmission medium of light in optical fibers, eliminating the problems caused by intrinsic material limitations. Compared with solid-core fibers, hollow-core fibers have advantages such as low latency, low dispersion, and low nonlinearity. First, the transmission speed of light in air fiber cores is 1.5 times that in the glass medium, greatly shortening the communication latency between servers and GPUs in AI-enabled data centers. Second, because the transmission medium of hollow-core fibers is air, the material dispersion is low, which helps extend the transmission distance of the high-speed optical modules in data centers and reduce the optical interconnection cost. Third, in addition to low material dispersion, air has a smaller nonlinear refractive index coefficient and a lower nonlinear effect than glass materials such as silicon dioxide. This greatly suppresses signal distortion caused by optical interconnection in data centers and ensures higher communication and network quality.

(2) Multi-core fibers

In a multi-core fiber, multiple fiber cores share a cladding. Each fiber core is single-mode, and the crosstalk between fiber cores is small. This increases the density by several times compared with traditional single-mode fibers. In a multi-core optical fiber, multiple optical signals can be transmitted in fiber cores concurrently. The small crosstalk also greatly improves the communication capacity. Therefore, the application of the multi-core optical fibers will have a revolutionary impact on optical interconnection in data centers. Replacing multi-mode fibers with single-mode fibers, single-core fibers with multi-core fibers, and hot swapping modules with COBO/CPO modules will be the future cabling trend of data centers. Multi-core fibers have the potential to become an 800G+ interconnection solution in the future. They can greatly improve the optical transmission capacity and spectral efficiency, save cabling costs and pipe resources, and reduce energy consumption. In addition, they have multiple parallel physical channels, which are more likely to be used in next-generation data center cabling.

SysMoore

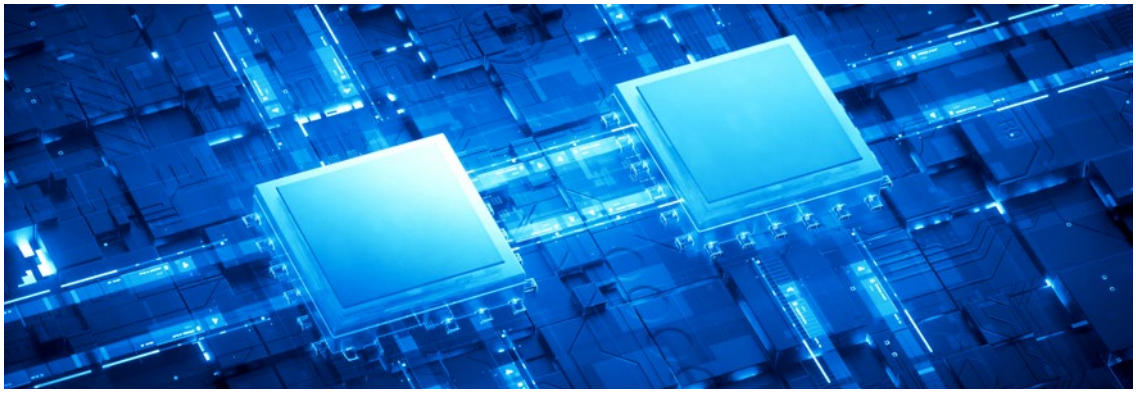
(1) Large chips and chiplets

Semiconductor integrated circuits are the cornerstone of the modern information industry. However, Moore's Law, which once set the pace for the development of integrated circuits for decades, is now reaching its limits in terms of physics and economics, and traditional silicon-based electronic technologies are approaching their upper limits. Therefore, new chip technologies are urgently needed to promote the development of the information industry.

- **Chiplet packaging technology**
  Traditional chips are restricted in terms of die size and die yield due to limitations in the wafer exposure (1X reticle: 25 mm x 32 mm), and this directly hinders improvements in chip performance and cost reduction. The 2.5D silicon/fan-out (FO) interposer + chiplet technology can effectively improve the die yield, reduce chip costs, and boost chip performance by stacking and integrating chiplets, and flexibly adapting them to various product specifications. In addition, the single-bit power consumption in 2.5D packaging is only half that of board-level interconnection solutions in traditional packaging.

  In response to industry development and ultra-large-scale chip requirements, the size of a 2.5D silicon/FO interposer is expected to exceed 4X reticle by 2025, and the size of the package substrate is expected to exceed 110 mm x 110 mm in the future. However, the application of larger 2.5D interposers and substrates leads to engineering challenges in terms of yield, delivery time, and reliability. Therefore, an integrated and innovative substrate architecture is a top priority.

- **3D large chip manufacturing**
  Compared with traditional 2D/2.5D advanced packaging and heterogeneous integration technologies, the 3D chip



technology has obvious advantages in terms of interconnection density, bandwidth, chip size, power consumption, and comprehensive chip performance. It is the core technology for SoC integration in key scenarios such as high-performance computing and AI. This technology will evolve from die-to-wafer (D2W) to wafer-to-wafer (W2W) and from micro-bumping to hybrid-bonding, and then to monolithic. It will be widely used in 3D memory on logic, logic on logic, and optical on logic, and will implement the stacking of more heterogeneous layers in the future. In terms of stacking, 3D chips adopt an ultra-high-density bonding technology that decreases the pitch to less than 10 μm. Compared with traditional 2.5D packaging, 3D chips have obvious advantages in bandwidth and power consumption. Specifically, the single-bit power consumption is expected to drop to 1/10. Further exploration is required to achieve smaller through-silicon vias (TSVs) in terms of materials and basic process technologies. Nevertheless, 3D stacking doubles the density of local power consumption and current, and this directly affects the overall power supply and heat dissipation paths of the system.

The chiplet technology based on chiplet integration will mature earlier than 3D large chip manufacturing. 3D large chips will emerge once the relevant processes and technologies mature.

(2) New computing power

The Dennard scaling law has come to an end in silicon-based semiconductors. Continuing or surpassing Moore's Law has become a major challenge in the computing field. Both academia and the industry are looking for new computing paradigms such as analog computing and non-silicon-based computing, to make computing more energy-efficient.

- **Quantum computing-accelerated engineering**

  Quantum computing hardware is currently in the high-speed engineering phase, where the number of quantum bits (qubits) is multiplying rapidly. It is estimated that quantum chips with more than 10,000 physical qubits will become available within the next five years. On today's noisy intermediate-scale quantum (NISQ) hardware, it is the most feasible direction to construct a hybrid computing system of classical and quantum computers. Quantum simulation, quantum algorithms for combinatorial optimization, and quantum machine learning are mainstream application scenarios in the industry. Quantum simulation offers a new computing paradigm for drug discovery and new material R&D. Quantum algorithms for combinatorial optimization utilize the parallel computing capability of quantum computing to more quickly and effectively solve problems such as logistics scheduling, route planning, and network

traffic allocation. Quantum machine learning will become a new approach for AI computing acceleration. In the next decade, the physical qubit scale of a single quantum chip needs to be continuously enlarged, the coherence time of qubits and the fidelity of quantum operations need to be enhanced, and the system scalability needs to be improved through quantum chip interconnection. In terms of software and algorithms, the quantum software stack needs to be refined, and the quantum algorithms need to be optimized based on application scenarios to reduce the quantum circuit depth and complexity, so as to gradually promote the commercial use of NISQ quantum computing. In addition, the fault tolerance design of quantum computing needs to be enhanced to make the quantum system more reliable. There is still a long way to go before the application of a general-purpose quantum computer.

- **Optoelectronic accelerators based on analog optical computing**

  Light travels fast and does not consume a lot of energy. Physical phenomena such as interference, scattering, and reflection of light can be expressed using mathematical models. By modulating, controlling, and detecting optical signals, specific computing tasks can be completed. In addition, photons are bosons and naturally have features such as wavelength division multiplexing, mode division multiplexing, and orbital angular momentum (OAM) multiplexing. Implementing multiple-dimensional parallelism by means of analog optical computing is a promising direction of optical computing development, in that it is expected to accelerate computing capabilities in scenarios such as optical signal processing, combinatorial optimization, and AI. To enable the large-scale application of optical computing, it is most important to heterogeneously integrate the active/passive components on chips, improve the efficiency of optical signal coupling, control insertion loss and noise, and meet the computing precision requirements of specific scenarios based on which an optoelectronic system needs to be built to accelerate specific computing tasks.

- **Large-scale application of non-silicon-based computing**

  Transistors based on 2D materials have the advantages of short channels, high mobility, and 2D/3D heterogeneous integration. They are expected to continue Moore's Law to 1 nm process. In addition, 2D materials that have an ultra-low dielectric constant may also be used for component isolation on integrated circuits. Such materials may be first applied in fields such as optoelectronics and sensing. Currently, 2D materials and components are in the initial research phases. In the next five years, the yield of industrial-grade wafers based on 2D materials needs to be increased, and after this, the electrode and component structures need to be refined to improve the comprehensive performance of 2D transistors. Carbon nanotubes (CNTs) have ultra-high carrier mobility and atomic-level thickness, and are known for their high performance and low power consumption. When the size is extremely reduced, CNT-based transistors are about 10 times more efficient than silicon-based transistors. CNT-based transistors are expected to be put into small-scale commercial use in biosensors and radio frequency circuits within five years. In the future, efforts should be made to continue upgrading the preparation process of CNT materials, so as to reduce surface pollution and impurities and refine material purity and CNT arrangement consistency. In addition, the contact resistance and interface state of components require optimization for higher injection efficiency. When the size of a carbon-based semiconductor device can be miniaturized to the size of an advanced silicon-based semiconductor device, it is expected to be used widely in scenarios requiring high performance and integration.

(3) New storage

Wide application of big data and AI has highlighted the importance of data-driven computing and the value of data. Two major challenges facing data storage systems relate to how quickly they can meet the data processing needs of the compute unit and how they can achieve low-cost, long-term retention (LTR) of data. To address the challenges and better leverage the value of data, new data storage turns to diverse storage media and data-centric architectures.

- **Diverse storage media**

  By the 2030s, the world will generate around 1 yottabyte (YB) of data per year, and 50 zettabytes (ZB) of valuable data will need to be stored (23 times more than in 2020). This will require high-capacity, energy-efficient, and cost-effective storage media along with resilient storage systems that are highly reliable, scalable, and durable, and possess data computing and analytics capabilities for quicker data access.

  In terms of future data lifecycle management, high-speed and high-performance storage media will be needed for hot data, medium-speed and high-capacity media for warm data, and low-speed and low-cost media for cold data.

  (1) DRAM is the mainstream choice for high-speed and high-performance media.

  Dynamic random-access memory (DRAM) is a type of volatile memory that delivers the best performance. Thanks to the advanced 1α process technology, 1α nm DRAM currently boasts the industry's highest bit density of 0.315 Gb/mm^2. Large capacitors limit the effective area in the DRAM cell, so technologies like 3D DRAM, wafer thinning, and hybrid bonding have been developed to increase storage density and reduce power consumption. In addition, new types of non-volatile memories (NVMs) have continued developing and seen remarkable progress with ferroelectric RAM (FeRAM), magnetoresistive RAM (MRAM), resistive RAM (ReRAM), phase-change memory (PCM), and oxide semiconductors.

  FeRAM opened the door to Mb-level products and 8 Gb FeRAM using the 1x nm DRAM process is fabricated. MRAM has seen stand-alone Gb-level and embedded Mb-level products and is being developed to replace SRAM/DRAM in cache applications. For PCM, products with 512 GB capacity using 3D XPoint technology are already available to meet persistent memory and storage class memory (SCM) requirements. ReRAM has already seen Mb-level stand-alone products commercially available and Mb-level embedded products ready for mass production, and is being researched to find new solutions for computing in memory. Oxide semiconductors like indium-gallium-zinc oxide (IGZO) can be used for 2-transistor-0-capacitor (2T0C) DRAM cell and potentially achieve high unit density beyond 4F^2 by monolithic stacking.

  In general, these new types of NVMs are non-volatile and energy-efficient, meaning they can slash power consumption in storage devices. Despite this, they are not as good as DRAM in terms of storage density and erase/write times, making DRAM the continued mainstream choice for high-performance storage media until 2030.

  (2) SSDs have significant advantages for

medium-speed and high-capacity media.

HDD using magnetic memory has historically been the primary choice for high-capacity storage media. As thin film media made of iron-platinum magnetic alloy, heat-assisted magnetic recording (HAMR), and microwave-assisted magnetic recording (MAMR) mature, the storage volume of 3.5-inch hard drives will increase from the current 30 terabytes (TB) to 80 TB. The cost per TB will not drop much though because of the adoption of laser and microwave technologies. Thanks to rapidly developing semiconductor manufacturing and innovative 3D-stacking technologies, 3D NAND has great potential for medium-speed and high-capacity media. Currently quad-level cell (QLC) has already achieved scale shipment and penta-level cell (PLC) applications are being planned. The progress of 3D NAND shows that adding more layers in a cell is a feasible direction for capacity breakthroughs. 232-layer NAND has already been announced and 1000-layer NAND is expected within the next decade, so there is still room for 3D NAND storage density to be improved. It is estimated that, SSDs will cost the same per TB as HDDs by 2030 while also delivering obvious advantages in latency and bandwidth, and up to 80% of data centers will use all-SSD flash storage.

(3) Tapes and optical discs are commonly the favored low-speed and low-cost media options.

The rapid pace of digitalization has led to an exponential increase in the amount of data being aggregated in data centers, while the advent of AI is facilitating the extraction of more value from data. Currently, most regulatory and policy frameworks also require data to be stored for at least 30 years. This has prompted data center operators to reevaluate the significance of "antiquated" tape and optical storage. Tapes offer a distinct advantage in terms of cost per TB, thanks to their straightforward production process and large available storage capacity. Furthermore, tapes can leverage the head and magnetic powder technologies used in hard drives to consistently improve capacity density and ensure sustainable evolution. Currently, an LTO-9 tape cartridge boasts a capacity of 18 TB, with future expansion projected to reach an astounding 576 TB. Data on old tapes is required to be copied onto new tapes every seven years, due to format compatibility issues and limited media lifespan. The new Archival Disc (AD) technology, a successor of Blu-ray Disc (BD) optical storage, also boosts the capacity of a single disc to 500 GB or even 1 TB, and delivers an impressive data retention period exceeding 50 years. Furthermore, the capacity density of 12 discs is currently comparable to that of a single current tape cartridge. As media materials and servo technologies further develop, the capacity of a single disc will reach 2 TB or even 4 TB in the future. Therefore, tapes and optical

discs play a crucial role in data centers for storing cold data. Tapes are favored for their cost-effectiveness, while optical discs excel in providing extended storage periods.

- **Data-centered architecture**

  Emerging data-intensive applications such as big data, AI, high-performance computing (HPC), and Internet of Things (IoT) are driving explosive growth in data volume, with a staggering compound annual growth rate of nearly 40%. More than 30% of data will eventually be hot data. In addition, as we reach the limits of Moore's Law and Dennard's scaling law, the annual growth in Central Processing Unit (CPU) performance has dropped to 3.5%. This sluggish growth of data processing capabilities will not be able to keep up with the rapid expansion of data, leading to an imbalance between the power of data storage and the pace of data growth.

  Under the traditional CPU-centric architecture, the uneven distribution of services across space and time results in a low utilization of local storage resources, with over 50% of local memory and storage sitting idle. Furthermore, data movement and repeated data format conversion occupy a lot of CPU resources, leading to low data processing efficiency.

  In order to improve data processing efficiency and storage resource utilization, a new data center architecture is required to help us shift from being "CPU-centric" to "data-centric" in four areas:

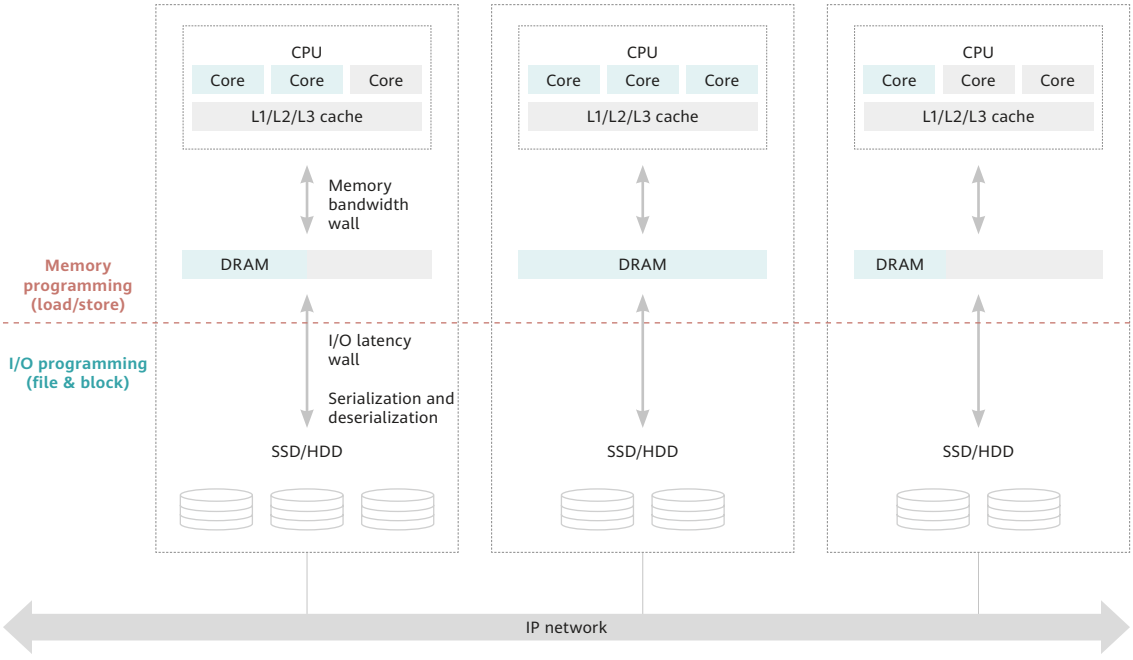


Figure 3-8 Traditional CPU-centric architecture

(1) Decoupled storage and compute in data centers

Compute and storage resources are deployed separately. They are connected through a high-throughput data bus and data is accessed using unified memory semantics. As a result, compute and storage resources are decoupled in order to be scheduled in a more efficient way. Storage-compute decoupling has already advanced beyond the traditional decoupling of CPUs from external storage devices such as SSDs and HDDs. It breaks the limits of compute and storage hardware resources, forming independent hardware resource pools (such as CPU pools, DPU pools, memory pools, flash pools, etc.) to enable the flexible expansion and sharing of different hardware. The decoupled storage-compute architecture has three characteristics: storage resource pooling, full-memory semantic access, and high-throughput peer-to-peer interconnection bus.

(2) Data and control separation within the storage system

The data plane and control plane are separated so that CPUs process only the tasks on the control plane, and tasks on the data plane are processed by heterogeneous

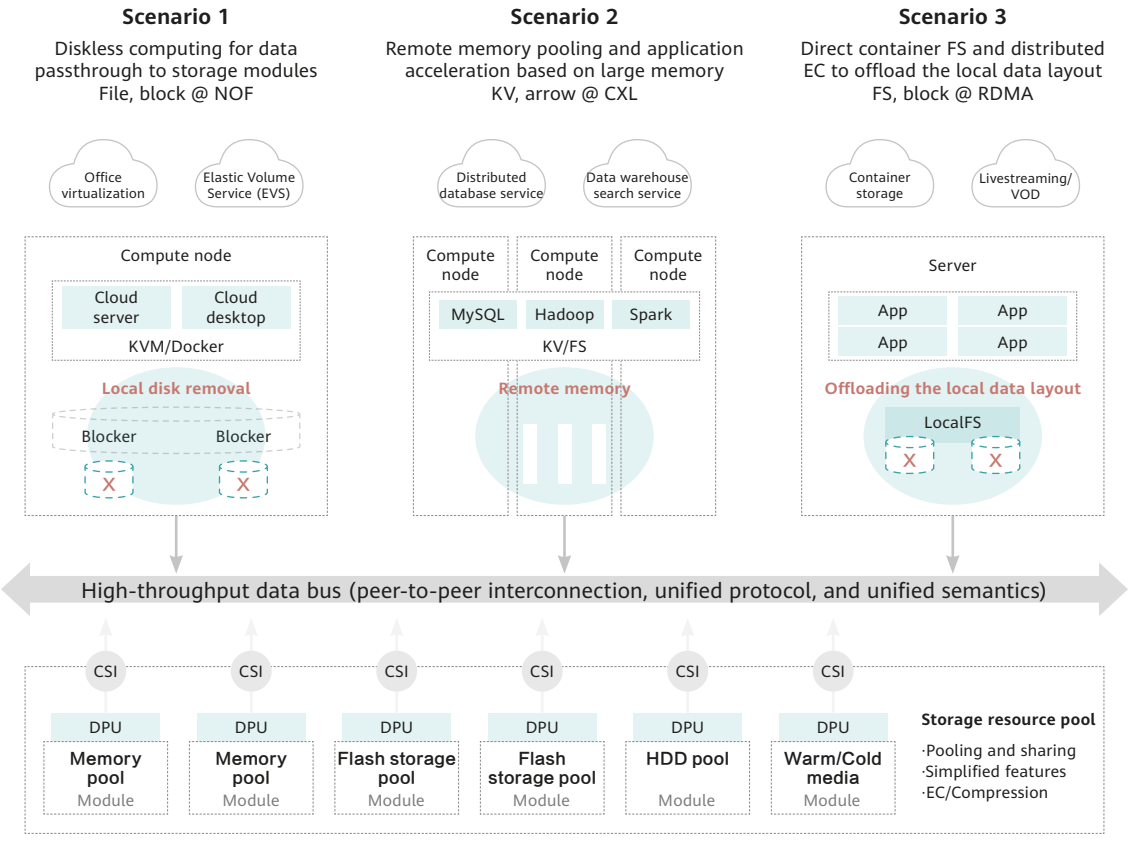


Figure 3-9 Decoupled storage-compute architecture

computing power such as data processing units (DPUs). This avoids repeated context switching and improves data read and write performance.

Traditionally, CPUs are the central hub of a storage system, and data read and write can only be performed with the support of CPUs. As a result, CPUs have become a system performance bottleneck because they cannot meet the ever-increasing performance requirements of emerging applications. With technologies such as I/O passthrough, data processing paths can be shortened from smart network interface cards (SmartNICs) and DPUs directly to drives, enabling fast data access from front-end cards to back-end media. In this way, CPU involvement in I/O paths will be reduced and the latency and throughput can reach record highs.

(3) Intelligent data fabric across data centers

The development of digital technologies generates a large number of demands for cross-region data mobility, in turn posing higher requirements on data availability and quality. However, regional restrictions and difficulties in data governance hinder the free flow of data and ultimately result in problems related to data gravity. Data fabric technology can be widely applied to different applications to dynamically and automatically coordinate distributed data sources and provide integrated and reliable data across multiple data platforms.

Based on AI and other technologies like knowledge graphs, intelligent data fabric can identify and connect data from different applications to discover service correlations between available data points. The edge, data centers, and the cloud frequently exchange data over

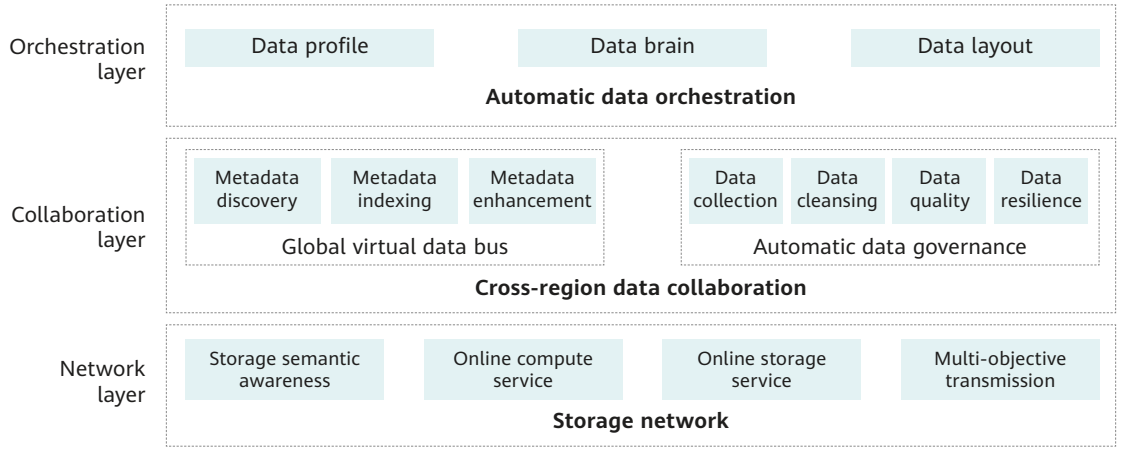


Figure 3-10 Intelligent data fabric framework

data networks. Intelligent data fabric can continuously analyze existing, discoverable, and inferable metadata assets to integrate cross-platform data and enable efficient data mobility and processing. To leverage intelligent data fabric, the challenges created by data gravity first need to be resolved. This will require technological breakthroughs in cross-region data collaboration, automatic data orchestration, and efficient storage networks.

(4) Application-oriented data acceleration

In a data-centric processing paradigm, data processing is performed in a specialized way rather than based on general-purpose computing power. Data used to be transferred directly to processors, but now computing power can be deployed near the data itself. Data is now processed with the most appropriate computing power at near-data sites, and nearby data processing is performed at the edge of data generation, during data mobility, and in data storage. More than 80% of all data is expected to be processed near memory or in memory by 2030. As a data carrier, data storage needs to provide both data access services and near-data processing acceleration services. Nearby data processing is performed mainly through three modes: diversified storage and compute convergence, data storage and network convergence, and data processing and network convergence.

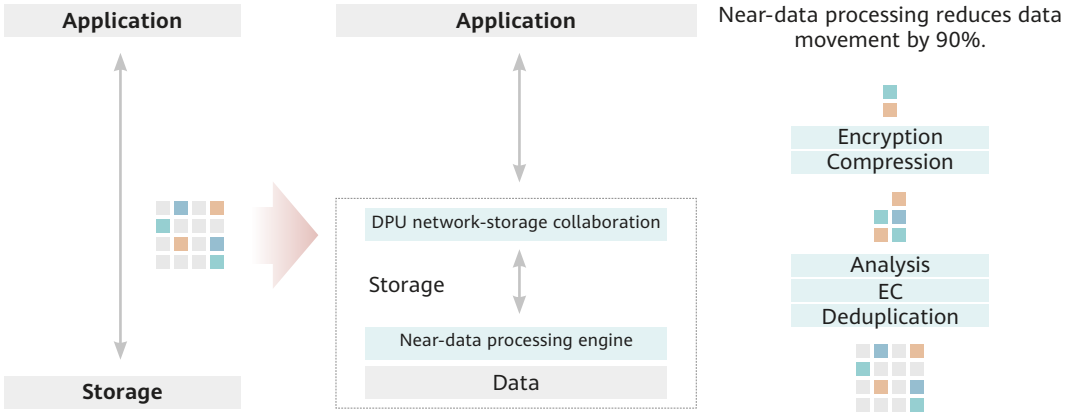


Figure 3-11 Application-oriented data acceleration

Reference Architecture for New Data Centers

04

In 2030, the functional positioning of data centers will change dramatically, as a result of the rapid increase in industry computing power requirements and the acceleration of innovation in data center technologies, such as computing, storage, networking, cloud, cooling, and green energy supply and storage. For example, data centers need to transform from enclosed, isolated facilities into infrastructures that are capable of participating in more extensive social-scale computing network collaboration; from coarse-grained resource management to more refined, efficient computing power supply; from data silos to a role that is capable of ensuring secure and reliable cross-domain transmission of large-scale data; from CPU-centric to data-centric computing architecture. These changes not only affect the current data center architecture, but also pose significant challenges for enterprise data center construction. We propose a new type of data center architecture with six key features.

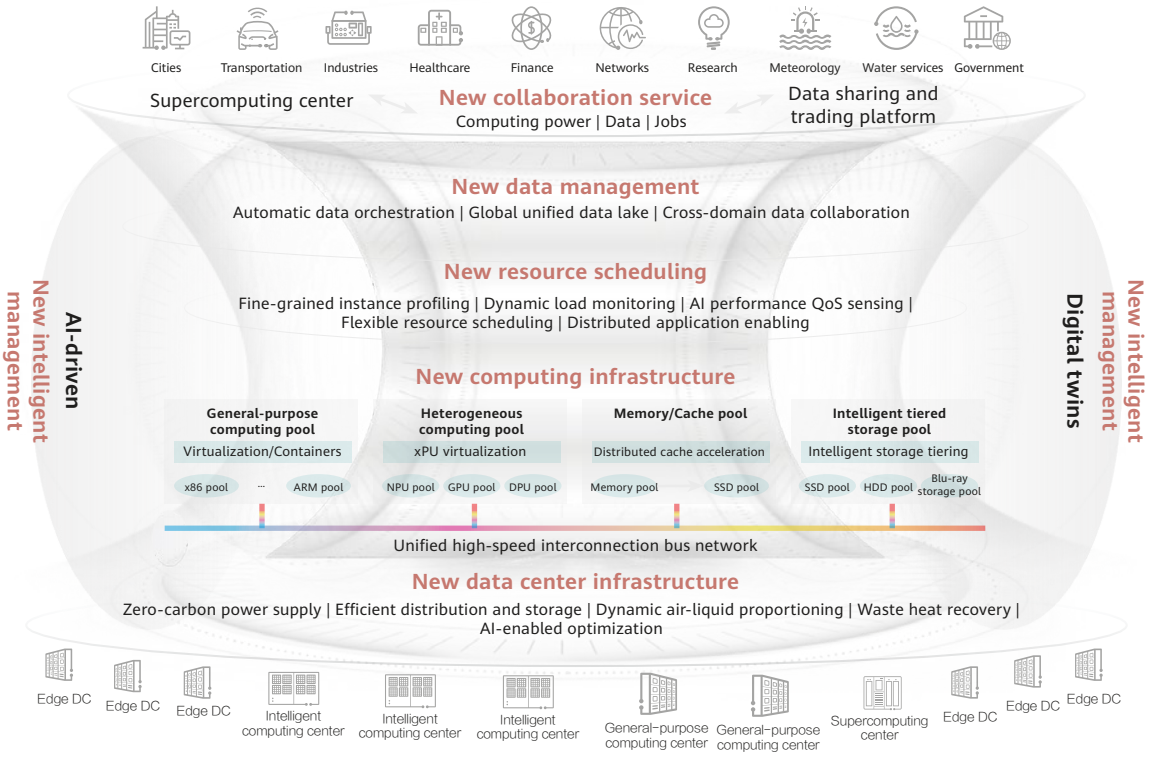


Figure 4-1 A reference architecture with six key features for new data centers

New data center infrastructure: Driving inclusive green growth with innovative power supply and cooling

As the data center scale continues to grow, the power consumption of data centers will continue to rise, which will result in multiple challenges in power supply and cooling. Challenges in power supply include a low proportion of green power, inefficient power grid utilization, high loss in power supply, and numerous diesel generators employed to ensure reliability. As a result, the effective power used for IT equipment is generally less than 80%. To cool data centers, compressors in cooling systems have to keep running for most of the time, resulting in low cooling efficiency. Moreover, the static cooling architecture cannot meet the requirements of rapid changes in computing power. In addition, the waste heat generated by data centers cannot be recycled. To address these challenges, we propose a new data center reference architecture that integrates key functions such as zero carbon, low energy consumption, and more flexible and elastic cooling in all weather conditions.

By 2030, new power supply systems will use long-term energy storage, hydrogen-fueled generators, and local PV to interact with virtual power plants to enable a synergy of generation-grid-load-storage. In this way, the power grid will fully utilize the surplus power reserves of data centers to meet changing load requirements. This will solve the random and intermittent issues of wind and PV power, improve the stability and utilization efficiency of a large proportion of clean energy that the power grid takes on, and ensure that almost all power used in data centers is green. Power supply systems will be further integrated to reduce energy loss. It is estimated that over 95% of the electricity supplied to a data center will be consumed by computing equipment.

By 2030, new cooling systems will use an architecture that is compatible with air cooling and liquid cooling to dynamically and flexibly schedule these two types of cooling, better meeting the rapidly increasing computing power requirements. By reducing the temperature difference in heat transfer and making full use of free cooling sources such as dry air and lake water based on local conditions, nearly 100% free cooling will be achieved, resulting in a twofold or threefold increase in the cooling energy efficiency. With the grade of waste heat improved and relevant businesses such as power generation from waste heat properly planned, it will become possible to fully utilize waste heat.

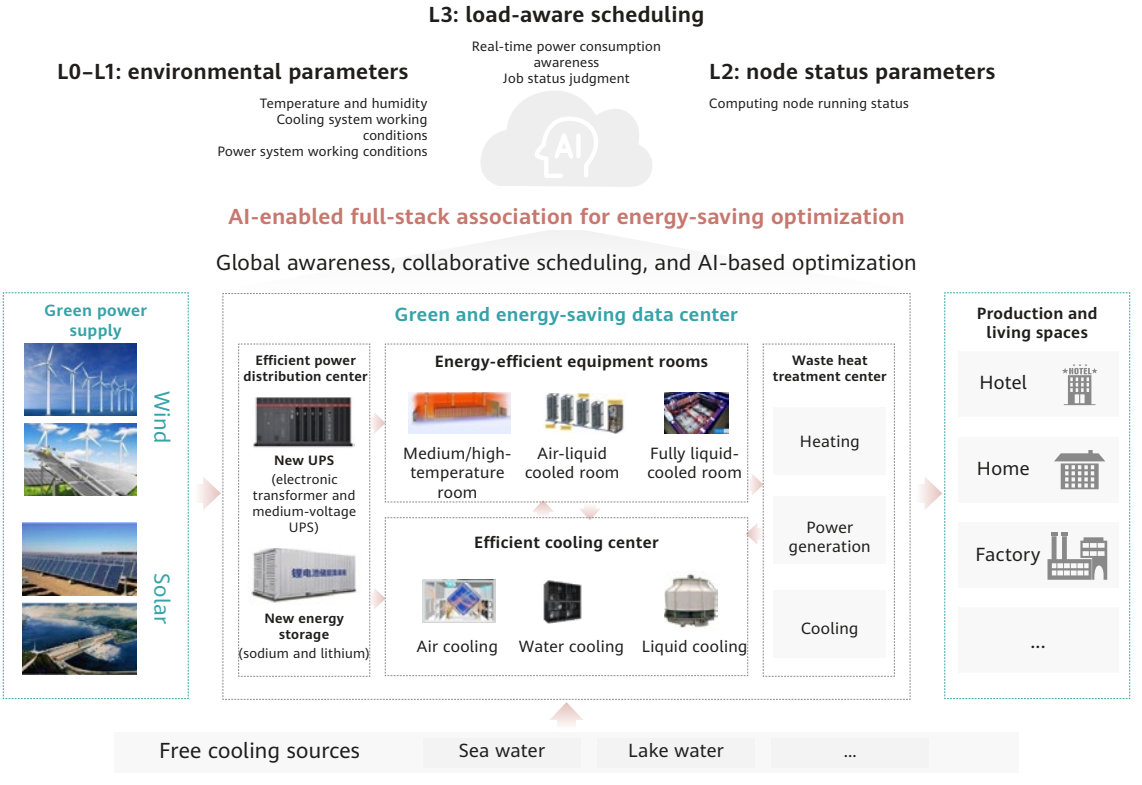


Figure 4-2 New reference architecture for data centers with green power supply and dynamic cooling

New computing infrastructure: Building a data-centric, diverse computing system

The majority of data centers still rely on the conventional multi-level hierarchical architecture, where each layer—compute, storage, and network—is a complete computer system consisting of components such as CPUs, memory, buses, and drives. However, this architecture has three pitfalls: memory, I/O, and computing. They result in slow data access and migration and impede large-scale distributed horizontal expansion.

By 2030, the next-generation data center computing architecture will have evolved from a CPU-centric multi-level hierarchical architecture to a data-centric peer-to-peer interconnection architecture with diverse computing power. This new architecture will be based on memory semantics. It will establish a unified, performant, programmable, and scalable interconnection network/bus (Unified Bus Fabric). It will prioritize data migration, conversion, and distribution, overcome the memory and I/O constraints, and unleash the computing power of CPUs and heterogeneous accelerators, so that computing and networks can be fully integrated to form an efficient supercomputer system.

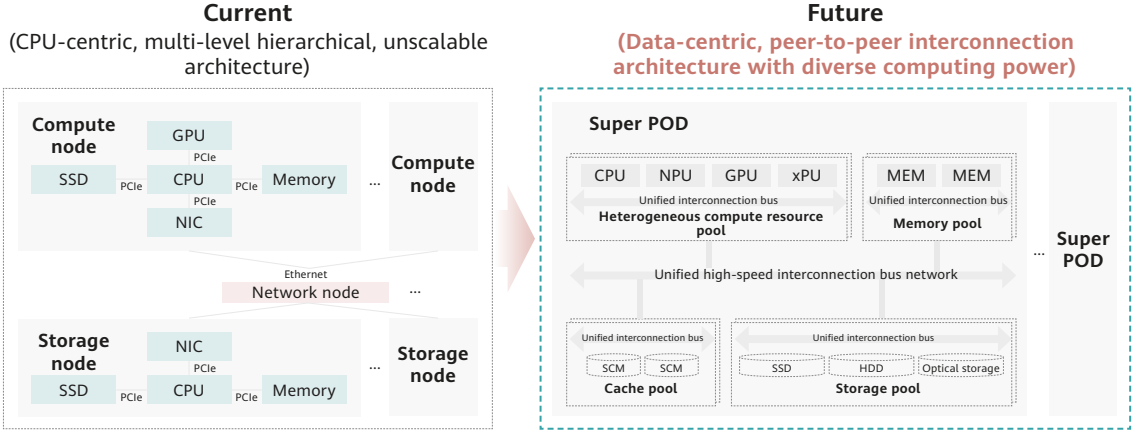


Figure 4-3 An architecture of the data-centric diverse computing system

New resource scheduling: Implementing application-centric, flexible scheduling

Just like every computer having an OS to schedule hardware resources including CPUs, memory, and drives, a data center also has its "OS" to provide distributed resource scheduling and coordination and implement data center–level elastic scaling. Data center OSs have evolved from the physical machine era to the current virtualization or cloud era, and are now advancing into the application-centric era.

In the era of physical machines, each server had an independent OS and ran only one application. As a result, the performance of a single server limited the deployment scale of applications. In the virtualization era, data center OSs manage resources by VM. Data center OSs leverage core virtualization technologies, such as software-defined networking (SDN), software-defined storage (SDS), and OpenStack, to present a high-performance server as multiple VMs. Then those VMs share the physical hardware resources of the host server while remaining logically independent of each other. Each VM accommodates an individual application, which does not interfere with applications on other VMs. This significantly improves server utilization and keeps data centers' operation costs low. However, operating and maintaining a cluster consisting of VMs can be quite challenging, especially when a fault occurs, as it is difficult to analyze the cause and locate the fault.

Over the next decade, application scenarios are expected to become increasingly diverse. At the same time, users are expecting to have direct access to resources, quick service start-up, unlimited service expansion, and seamless application migration. Applications will be the focus, which means that multiple data centers' compute, storage, and network resources will need to be consolidated and basic resources including CPUs, NPUs, GPUs, memory, and I/Os will need to be pooled and allocated to applications on demand.

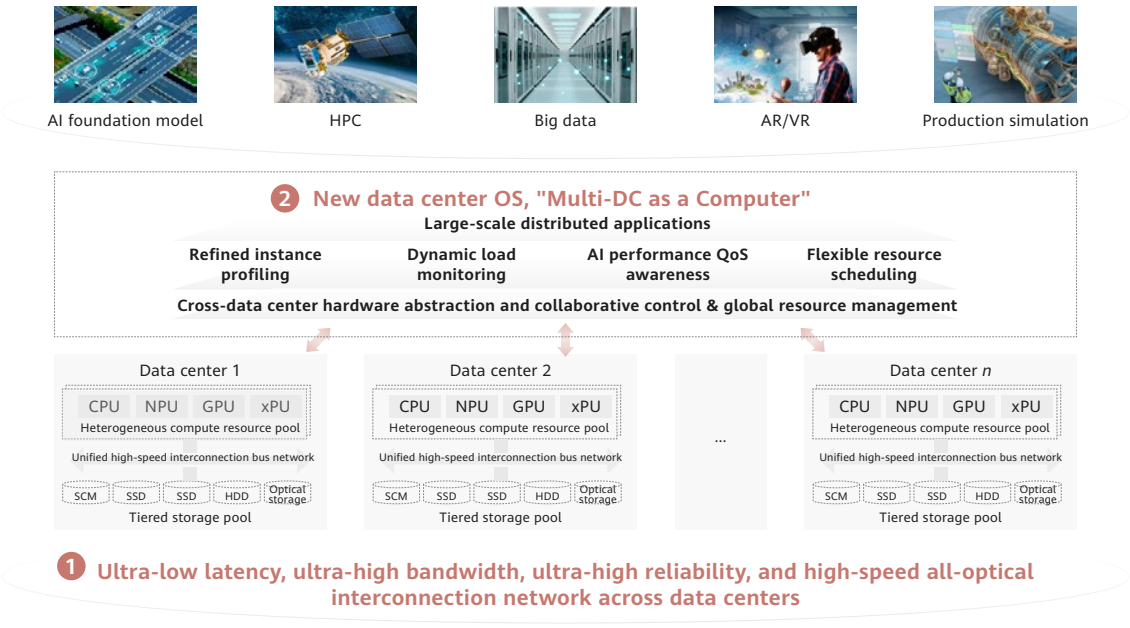


Figure 4-4 An architecture of the next-generation data center resource scheduling system

Meanwhile, the rapid rise of fields such as AI, scientific research, and the metaverse increases the demands on computing power. It is predicted that in the next three to five years, AI foundation models with trillions of parameters will emerge, and the computing power of a single data center will no longer be sufficient for AI training. Accordingly, the industry is looking into the use of clustering to overcome such performance limits, extend data centers beyond their physical limits, and thereby enable flexible scheduling of cross–data center cluster computing resources as well as agile application deployment.

To overcome the resource and platform limitations of a single data center and deploy large-scale distributed applications across data centers, an organization must: (1) establish a high-speed cross–data center network that allows for ultra-low latency, ultra-high bandwidth, and ultra-high reliability interconnection between data centers within a domain, and ensures the rapid scheduling and transfer of both data and tasks; (2) build an application-centric next-generation data center OS that abstracts and integrates cross–data center hardware resources and fully exploits hardware capabilities, while also providing refined and intelligent resource management functions, such as instance profiling, dynamic load monitoring, AI performance QoS awareness, and flexible resource scheduling, to improve global efficiency; and (3) provide deployment tools and running frameworks to deploy and operate large-scale distributed applications more efficiently.

New data management: Realizing instant visualization for data flow systems

Looking towards 2030, the demand for cross-domain data flows is becoming increasingly urgent. However, efficiency, security, collaboration, and management challenges stand in the way. First, there are myriads of data silos but a lack of global data views, resulting in low data utilization and difficult value mining. Second, the absence of tiered (hot, warm, and cold) data flow technology inhibits data flows between data centers. Third, cross-domain data collaboration is inefficient and cross-region unified metadata management is unavailable, making it impossible to analyze data in parallel. Fourth, inefficient data storage, high data storage costs, and slow data processing all have a negative impact on cross-domain query and analysis. To tackle these challenges, a logically unified data lake across domains, data centers, and storage forms is required. The data lake works with the data network brain to implement global data visualization, secure and efficient cross-domain data flows, and automatic, optimized storage tiering.

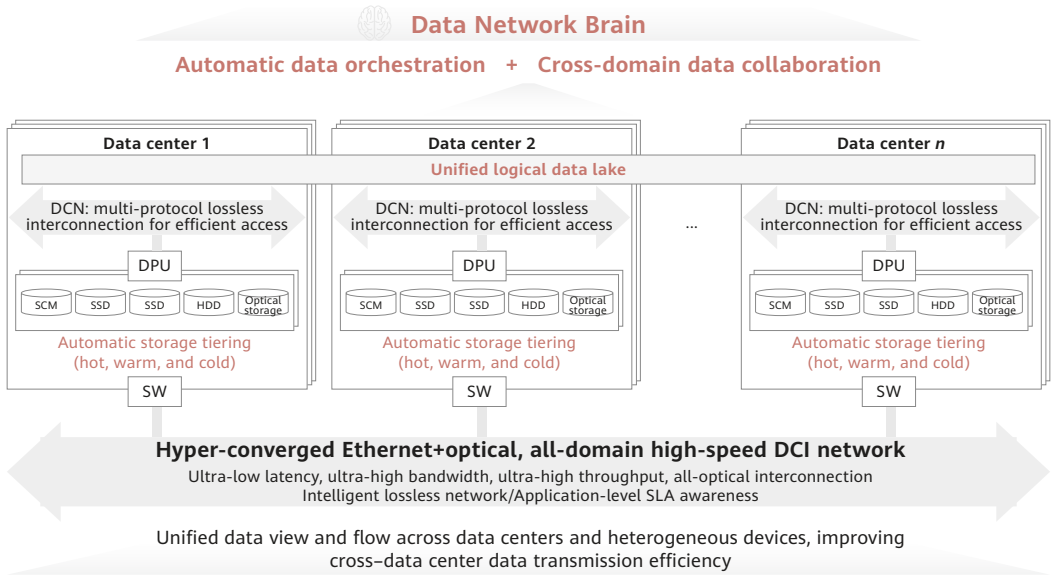


Figure 4-5 An architecture of the next-generation global data management system

New collaboration service: An open architecture to connect democratized computing power

In the future, diversified application scenarios will pose new requirements on the functional positioning of data centers. The data center landscape is evolving from one dominated by general-purpose data centers to one in which general-purpose computing centers, intelligent computing centers, supercomputing centers, and even data centers featuring optical computing and quantum computing can coexist. A system with collaboration between data centers and between the cloud and the edge will continue to grow. These application-driven, diversified data centers will work together to provide computing services. This will become an important form of computing power supply for data centers and will provide ongoing support for the development of the digital economy.

The new type of data center will no longer be an isolated data aggregation and processing center, but rather a part of the ubiquitous and inclusive computing service infrastructure. It will be an organic component of the entire social-scale computing network. It needs to be able to collaborate externally in order to participate in and comprehensively enable all fields of social production and life.

The new type of data center will need a more open and collaborative architecture, which will enable all data centers to open interfaces for collaboration between computing power, data, and operations while complying with unified standards. In addition, it will be able to quickly connect with external trusted sharing and trading platforms such as computing

power sharing and trading platforms, data sharing and trading platforms, and operation requirement distribution platforms. It will also be able to seamlessly participate in the division of labor and cooperate with democratized computing networks. The new data center will help to create an inclusive, open, and shared economic model which benefits everyone and which supports the rapid growth in demand for intelligent computing power across thousands of industries.

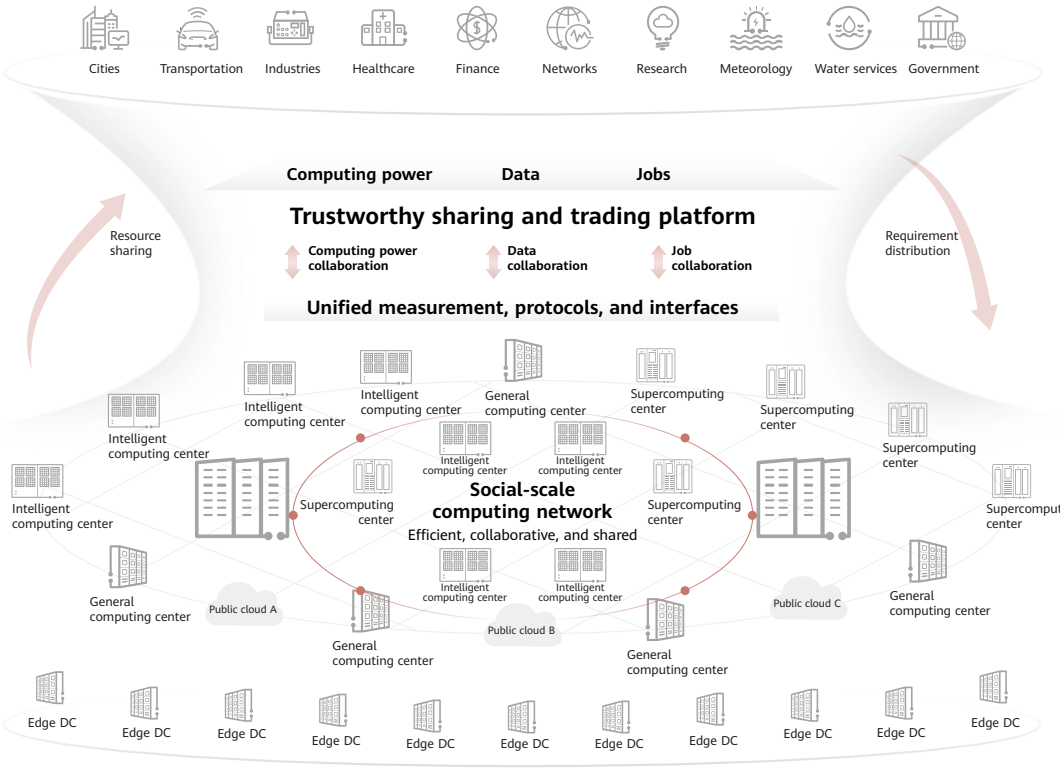


Figure 4-6 An open, shared data center collaboration architecture based on unified standards

New intelligent management: Enabling AI-driven, automatic data center O&M

Every operator aims to run their data centers securely, efficiently, and stably. By 2030, there will likely be more than one million IT hardware devices in hyperscale data centers, and hundreds of millions of application instances will be running in real time. As intelligent transformation progresses across industries, the data center scale will increase, the monitoring granularity of components will become more refined, the amount of monitoring data will increase, and new technologies and components will continue to

be introduced. Traditional data center O&M will face more severe challenges:

(1) Siloed management, many O&M tools, and poor collaboration make problem demarcation and locating difficult. Currently, data center management is scattered and problems are solved from an isolated perspective. Different devices such as infrastructure equipment rooms, computing, storage, and network devices have different monitoring, alarm, and log recording systems. In addition, there is no inter-system linkage. The integrated analysis capabilities of logs, data, and alarms are weak, making it difficult to locate and rectify faults.

(2) O&M data is scattered, making it difficult to fully unleash the value of the data. O&M data in the current data center is scattered, and lacks a centralized organization system and a unified data indicator system, and as a result, data integration and analysis difficult. In addition, data is difficult to obtain. The majority of data collection is done using manual methods, but this results in quality inconsistencies. Furthermore, analysis methods are limited, so the value of the data cannot be fully extracted.

(3) The level of automation and intelligence is low. According to a survey conducted by China's Academy of Information and Communications Technology (CAICT) in the

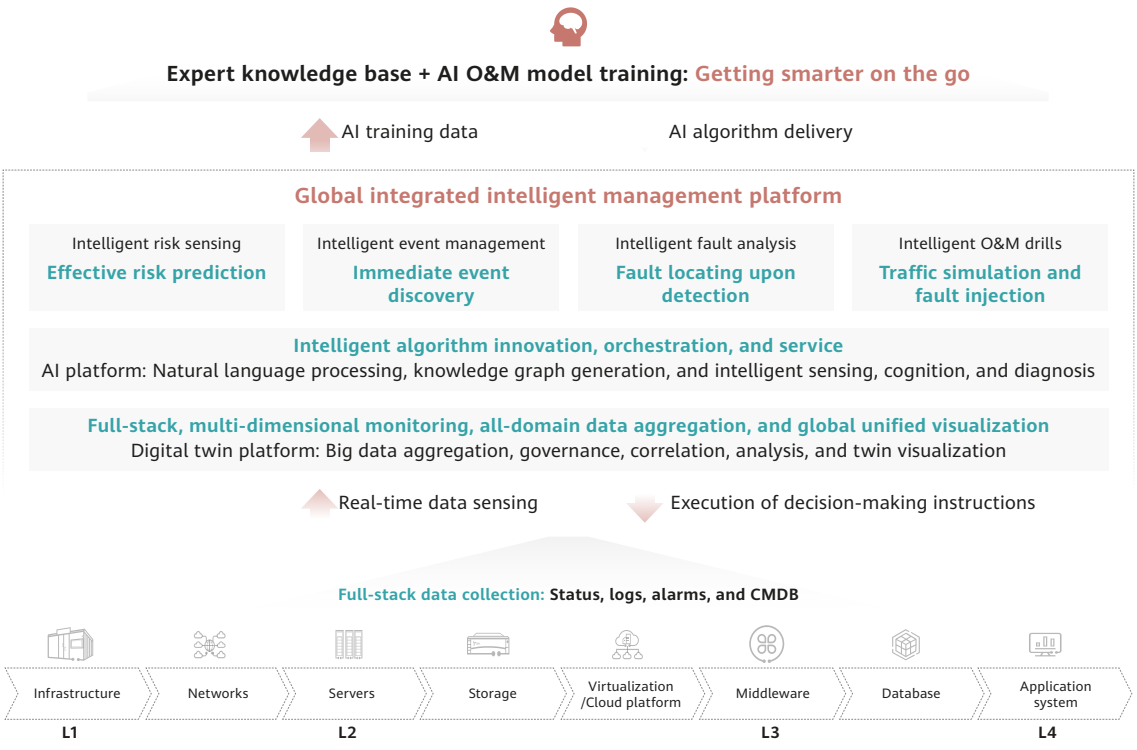
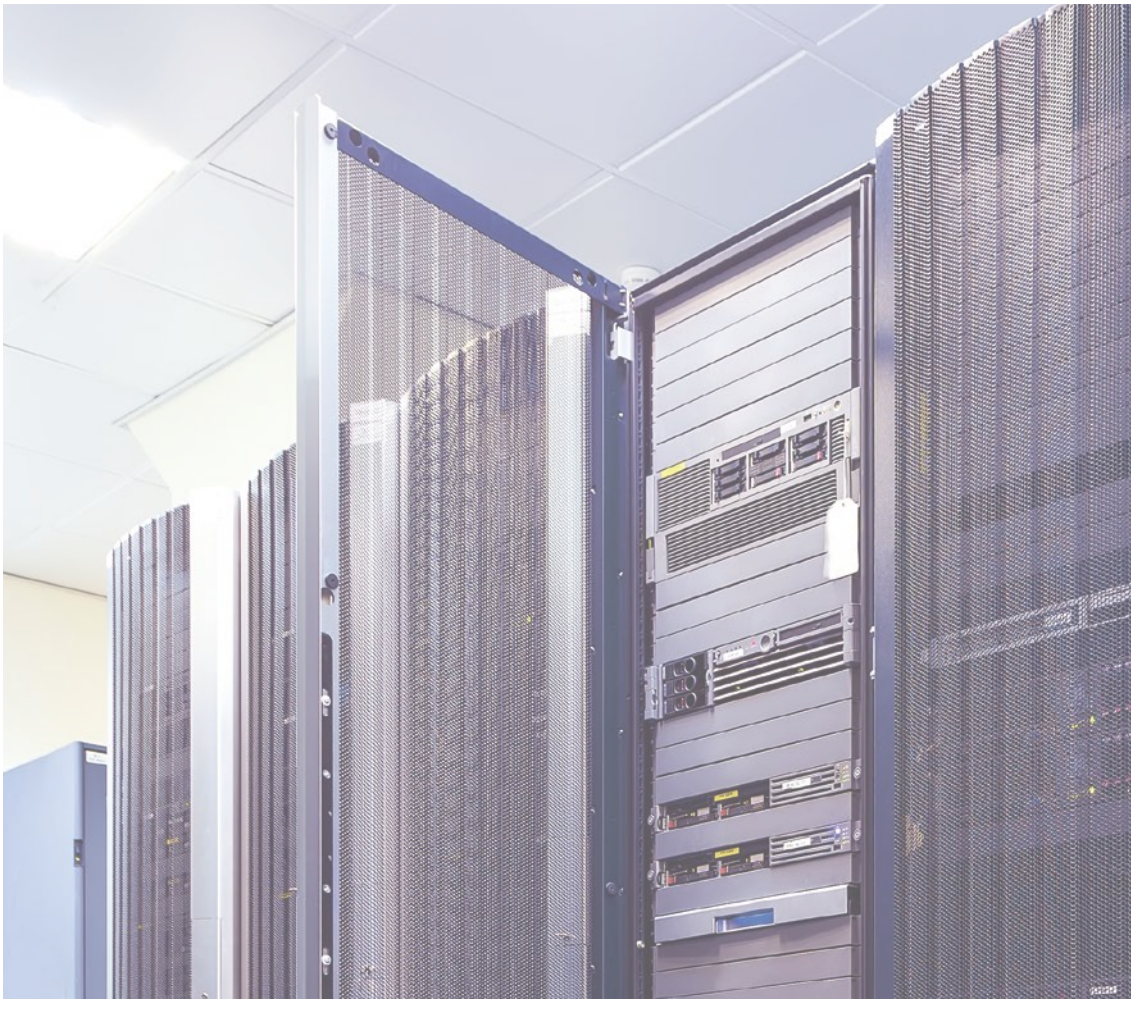


Figure 4-7 Reference architecture of the global integrated intelligent O&M system for new data centers

Research Report on the Development of Intelligent Data Center O&M (2023), most data centers in China still rely on manual O&M.

Looking ahead to 2030, data center management needs to evolve from being labor-intensive to being technology-intensive. Key technologies such as big data, AI, knowledge graphs, and digital twins must be leveraged to build a global integrated intelligent O&M system. This system will implement full-stack data collection, all-domain data aggregation, and a single-pane-of-glass visualization. With the help of expert knowledge bases and O&M foundation model training, risks can be detected earlier, problems can be solved faster, operation decisions can be made more accurately, and O&M can be managed and controlled more intelligently. The result will be data centers with a more advanced autonomous driving network (ADN).



Development and Call to Action

05

We are living in a time that is full of both challenges and opportunities.

Digital technologies, such as AI, 5G, and cloud, are changing our lives and penetrating various sectors at a faster pace. While envisioning the changes that technologies are bringing to our lives, we are constantly pondering over the impact of new technologies on the environment and ecology.

Data centers are the engine of the digital economy. Only by continuously improving the efficiency of data centers can we provide the power that constantly drives the development of the digital economy. The overall advancement of data centers can be comprehensively evaluated if we take energy efficiency, computing efficiency, data efficiency, transmission efficiency, and operation efficiency (5Es) into account. With all of the efficiency factors optimized, we will finally solve the structural contradiction between the rapid growth of computing power requirements and the sustainable development of data centers, ultimately creating greater value for a smart society.

In the coming decade, 5E-in-1 data centers will be diverse, ubiquitous, secure, intelligent, zero-carbon, energy-saving, and feature flexible resources, SysMoore, and peer-to-peer interconnection.

When the wind blows, let's ride the waves! Through architecture, system, theoretical, and engineering innovations, as well as the joint efforts of all stakeholders including industries, universities, research institutes, and customers, we believe that a sustainable intelligent world is fast approaching.

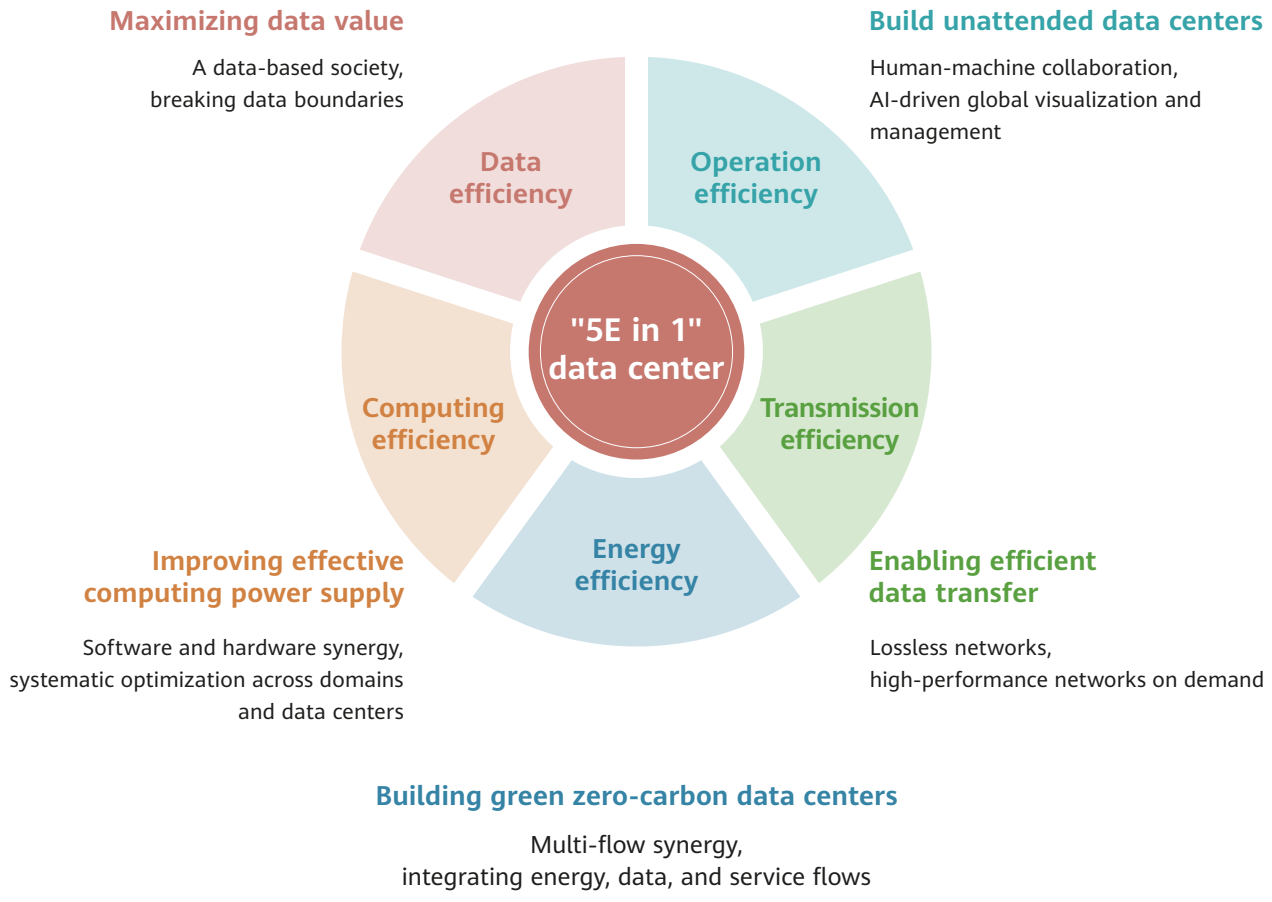


Figure 5-1 5E-in-1 data center

Appendix 1: Indicator system of key prediction data

Technical Feature	Indicator	Definition	2030 Prediction
Diversity and ubiquity	General-purpose computing power of a cluster	Effective computing power of a single cluster with software and hardware tuning	>70EFplops
	AI computing power of a cluster	Effective computing power of a single AI cluster	750 EFLOPS
	Cluster storage capacity	Effective storage capacity of a single cluster	Exabytes
	Percentage of data collaboratively processed by clouds and edges	Ratio of data requiring edge and data center processing to the total data volume	80%
	Digital access rate of enterprises' production devices	Ratio of enterprises' production devices that can be accessed through edge data centers to enterprises' total production devices, after being IoT-enabled and intelligentized	80%
Security and intelligence	Data security investment ratio	Proportion of data security investment to the total data center investment	20%
	System-level availability	System-level availability = System's annual MTBF/(System's annual MTBF + MTTR) (MTBF is short for mean time between failures, and MTTR is short for mean time to repair)	99.999%
	Disaster recovery (DR) coverage of important data	Percentage of important data and associated application systems for which DR is available	100%
	Automation level	L1: human assisted; L2: partially automated; L3: conditionally automated; L4: highly automated; L5: fully automated. (Automation includes automatic predictive troubleshooting and analysis, automatic emergency handling, and AI-powered energy efficiency management. L4 indicates operations approaching truly unmanned, and L5 indicates operations without any human involvement.)	L4
Zero carbon and energy conservation	Power usage effectiveness (PUE)	Total data center power consumption/IT equipment power consumption	1.0x
	Renewable energy factor (REF)	Renewable energy consumption/Total data center power consumption	80%
	Water usage effectiveness (WUE)	Water consumption/IT equipment power consumption	0.5 L/kWh
Flexible resources	DC-level resource pooling rate	Ratio of compute, storage, and network resources available for global scheduling to all resources in a single DC	80%
	Proportion of new cloud-native applications	Ratio of new cloud-native applications to all new applications	90%

Technical Feature	Indicator	Definition	2030 Prediction
Flexible resources	Resource allocation granularity	Granularity of compute, storage, and network resource allocation, scheduling, and billing	Function-level
Peer-to-peer interconnection	Penetration rate of hyper-converged interconnection bus technologies	Penetration rate of unified hyper-converged interconnection bus technologies	60%
	Penetration rate of hyper-converged Ethernet networks	Ratio of converged networks of general-purpose computing, high-performance computing, and storage to all networks of data centers	80%
	Penetration rate of optical + computing collaboration	Ratio of cluster computing power that supports computing power collaboration using spine layer on the all-optical direct connection AI parameter plane to the total computing power of the cluster	50%
	Penetration rate of optical + storage collaboration	Ratio of data transmitted in cross-WAN high-speed mode using all-optical direct connection SSD to the total transmitted data	50%
SysMoore	Percentage of all-flash storage	Ratio of all-flash storage capacity to the total capacity of data centers	80%
	Penetration rate of RDMA storage networks	Percentage of RDMA-based storage networks	80%
	Percentage of data processed near memory or in memory	Ratio of the amount of data processed near memory or in memory to the total amount of data processed	30%

Appendix 2: Abbreviations and acronyms

Abbreviation/Acronym	Full Spelling
3GPP	3rd Generation Partnership Project
5G	5th Generation of Mobile Communication
ABAC	Attribute-Based Access Control
AI	Artificial Intelligence
AIGC	AI-Generated Content
API	Application Programming Interface
AR	Augmented Reality
ARM	Advanced RISC Machine
ASIC	Application-Specific Integrated Circuit
BMS	Battery Management System
CDN	Content Delivery Network
CMDB	Configuration Management Database
CMOS	Complementary Metal-Oxide-Semiconductor
CPO	Co-Packaged Optics
CPU	Central Processing Unit
CRIU	Checkpoint/Restore In Userspace
CUDA	Compute Unified Device Architecture
CXL	Compute Express Link
DAC	Digital-to-Analog Conversion
DBR	Distributed Bragg Reflector
DC	Data Center
DCN	Data Center Network
DCOI	Data Center Optimization Initiative

Abbreviation/Acronym	Full Spelling
DDR	Double Data Rate
DFB	Distributed Feedback Bragg grating
DPDK	Data Plane Development Kit
DPU	Data Processing Unit
DRAM	Dynamic Random Access Memory
E2E	End-to-End
EA	Electronic Absorption
EB	Exabyte
EC	Erasure Code
EFLOPS	ExaFLOPS
ETH	Ethernet
ETSI	European Telecommunications Standards Institute
FeRAM	Ferroelectric Random Access Memory
FLOPS	Floating-point Operations per Second
FPGA	Field Programmable Gate Array
FS	FusionSphere OpenStack
GeSI	Global e-Sustainability Initiative
GPT	Generative Pre-trained Transformer
GPU	Graphical Processing Unit
HAMR	Heat Assisted Magnetic Recording
HDD	Hard Disk Drive
HPC	High-Performance Computing
HTTP	Hypertext Transfer Protocol

Appendix 2: Abbreviations and acronyms

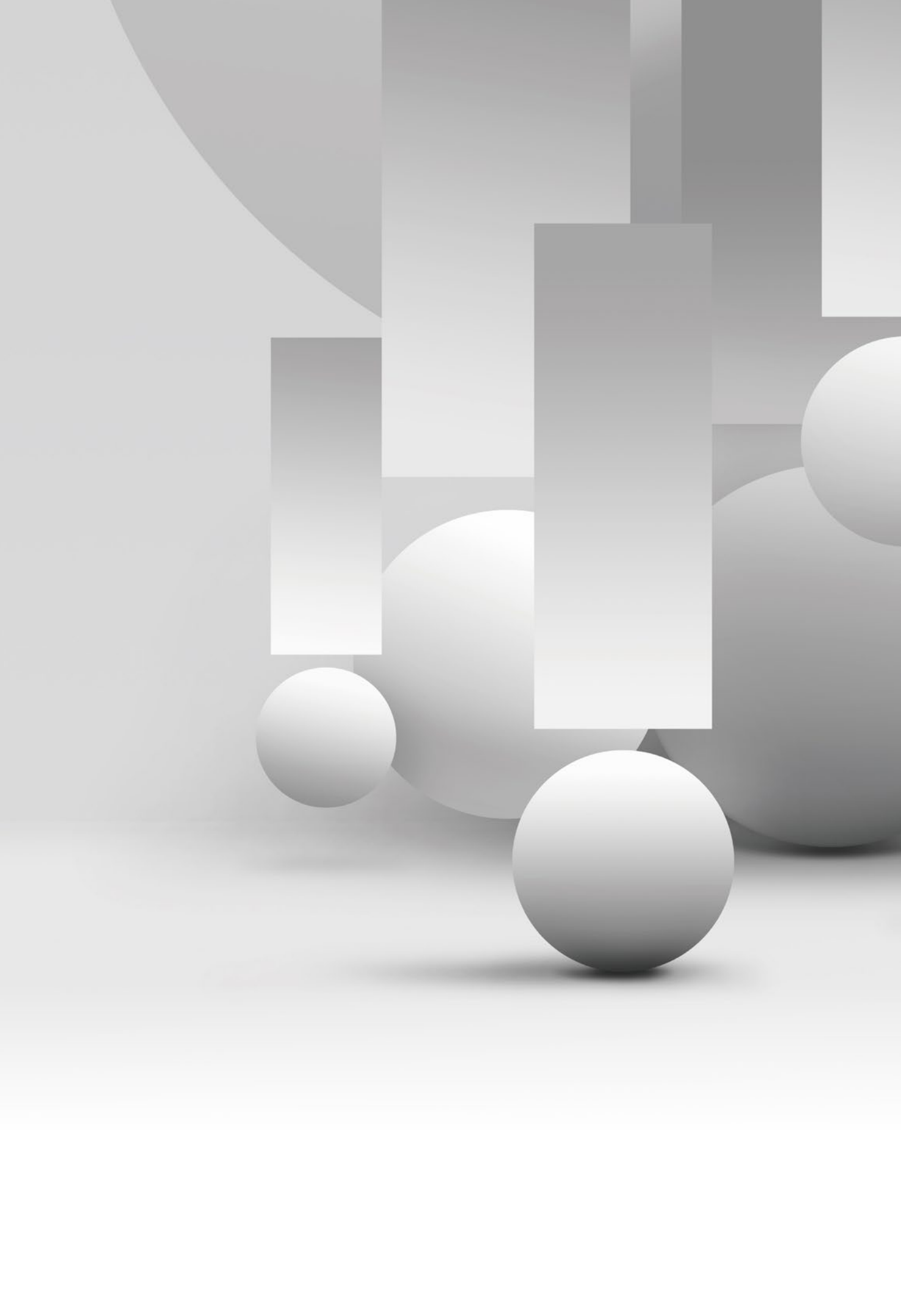
Abbreviation/Acronym	Full Spelling
HTTPs	Hypertext Transfer Protocol over Secure Sockets Layer
I/O	Input/Output
IB	InfiniBand
ICT	Information and Communications Technology
IDC	Internet Data Center
IGZO	Indium Gallium Zinc Oxide
IO	Input/Output
IoT	Internet of Things
ISP	Internet Service Provider
IT	Information Technology
K-V	Key-Value
KVM	Kernel-based Virtual Machine
KW	Kilowatt
LR	LongRange
MAMR	Microwave Assisted Magnetic Recording
MC	Main Control
MEC	Multi-access Edge Computing
MPLS	Multi-Protocol Label Switching
MRAM	Magnetic Random Access Memory
ms	Millisecond
MW	Megawatt
MZ	Mach-Zehnder modulator
NAND	Non-volatile Memory Device

Abbreviation/Acronym	Full Spelling
NG DCOS	Next Generation Data Center Operating System
NISQ	Noisy Intermediate-Scale Quantum
NOF	NVMe over Fabrics
NoSQL	Not only SQL
NPU	Neural Processing Unit
NUMA	Non-Uniform Memory Access
OBO	On Board Optics
oDSP	optical Digital Signal Processor
OS	Operating System
OXC	Optical Cross-Connect
PB	Petabyte
PCI	Peripheral Component Interconnect
PCIe	Peripheral Component Interconnect express
PCM	Phase Change Memory
PUE	Power Usage Effectiveness
QLC	Quad-Level Cell
QoS	Quality of Service
RBAC	Role-Based Access Control
RDMA	Remote Direct Memory Access
ReRAM	Resistive Random Access Memory
SATA	Serial Advanced Technology Attachment
SCM	Storage Class Memory
SDN	Software-Defined Networking

Appendix 2: Abbreviations and acronyms

Abbreviation/Acronym	Full Spelling
SDS	Software-Defined Storage
SLA	Service Level Agreement
SNIC	Standard Network Interface Card
SQL	Structured Query Language
SR	ShortRange
SSD	Solid-State Drive
swTPM	Software Trusted Platform Module
TB	Terabyte
TCO	Total Cost of Operation
TCP/IP	Transmission Control Protocol/Internet Protocol
TEE	Trusted Execution Environment
TOR	Top of Rack
TPM	Trusted Platform Module
UB	Unified Bus
UPI	UltraPath Interconnect
UPS	Uninterruptible Power Supply
VCSEL	Vertical Cavity Surface Emitting Laser
VM	Virtual Machine
VPC	Virtual Private Cloud
VR	Virtual Reality
WebGL	Web Graphics Library
WORM	Write Once Read Many
WUE	Water Usage Effectiveness

Abbreviation/Acronym	Full Spelling
xPU	A portfolio of architectures (CPU, GPU, FPGA and other accelerators)
XR	eXtended Reality
YB	Yottabyte
ZB	Zettabyte
ZFLOPS	ZettaFLOPS



HUAWEI

Version 2024

Intelligent Campus 2030

Building a Fully Connected, Intelligent World

Intelligent Campus

2030

Bringing Digital to Every Campus for Pervasive Intelligence

Foreword

Peter Nijkamp

A Beacon for Technological Progress

Campus ..... Latin roots

Campus is an ancient Latin term that originally refers to an open field outside the city where collective activities were organized that would be less suitable in inner-city locations (e.g. big festivities, military exercises). The concept of a campus as a dedicated people's meeting place became in the course of history particularly relevant for expanding university complexes which needed a lot of space. For example, Princeton University built already in the end of the eighteenth century in New Jersey an extra-city university complex, including teaching facilities, libraries, dormitories, and leisure facilities. This has in the course of years been followed by many campus initiatives all over the world. The concept of a modern campus has been more general. Rather than only referring to universities or colleges, it manifests itself nowadays often as an integrated centrifugal and centripetal knowledge hub. The rapid advent of contemporaneous ICT and digital technology has in the meantime created an unprecedented boost for advanced knowledge centers world-wide, in both the public and private domain. They are nowadays often called smart campuses seeking to share, generate, and create new applications through a synergy of physical and virtual proximity.

Intelligence ..... another Latin origin

Intelligence is another ancient Latin word which expresses the human ability to bring together or to synthesize knowledge and information; it is the cognitive potential to learn and understand new things by combining insights from different information sources. In contrast to smart behavior, which is based on learning and studying a demarcated knowledge issue, intelligent behavior is characterized by synergetic insights generated through sharp and brilliant brains and open minds. Intelligence

does not only solve complex problems, but is also able to prevent the occurrence of such problems. The emerging digital technology has also induced the rising popularity of interconnected data and knowledge networks, which are based on efficient information and knowledge sharing. Knowledge platforms and social media networks are a great illustration of the collective significance of modern multi-actor intelligence systems for industry and policy-making.

Intelligent campus ..... evolution

An intelligent campus is an integrated knowledge hub, in which both physical and virtual proximity plays a critical role. It is – in contrast to a traditional campus – not an island, but it forms an archipelago of virtually connected, actionable competency and innovation centers, which provide also open access to operators, users, and society at large. The influence radius of an intelligent campus far exceeds a local or regional territory, and in many cases this radius mirrors a global proximity, including academic, industrial and societal liaisons. The novel clever combination of hardware, software and humanware creates a new innovation arena with a global coverage, not only for industry but also for the achievement for the Sustainable Development Goals (SDGs), as formulated by the UN.

Intelligent Campus 2030 ..... the next frontier domain

In the rapidly evolving landscape of our modern knowledge society – as sketched out above – the Huawei Intelligent Campus 2030 study may be seen as a milestone that seeks to shed light on the transformative voyage that modern knowledge campuses all over the world are set to embark upon. It is therefore, desirable to offer an insightful exploration of future endeavors and to map out efforts driving a more promising trajectory of contemporaneous campus initiatives. It is remarkable that over the past years the release of the Huawei Intelligent World 2030 series has set the tone for a new perspective, culminating in the Intelligent Campus 2030 report. The report encapsulates global campus insights and practices as its building blocks, and may become a blueprint for future intelligent campuses worldwide. An intelligent campus is not a goal in itself. It seeks to develop and favor knowledge-oriented and digitally-based insights for academia, industry, policy makers, and society globally. Its technical features serve as a convergence trajectory for both scientific and innovation

consensus, accelerating the evolution of technology in campus environments globally. An intelligent campus is not a static idea; it is forward looking. In the pursuit of a connected and intelligent world, Intelligent Campus 2030 goes beyond conventional campuses, opening doors to an intelligent world where innovation shapes the very fabric of campus experiences. Intelligent campuses, as outlined in the Campus 2030 report, serve as foundational pillars for a sustainable and livable society. By seamlessly integrating connectivity and computing technologies within technological innovation institutions, the groundwork for high-quality society developments that are adaptive, sustainable, and digitally interconnected will be laid.

Beyond the intelligent campus ..... society and people across the globe

In the spirit of ongoing global collaboration, Intelligent Campus 2030 explores a future where global cooperation accelerates collective and shared progress for all societies, including:

- Establishing cross-institutional data connections that transcend geographical boundaries, enabling collaborative innovation and interdisciplinary advancements for multiple institutions.
- Spearheading digital inclusion initiatives to ensure that diversified campuses worldwide can benefit from them.
- Facilitating international exchanges so that academic, industrial, and policy-making partners the world over can define the future of intelligent campus together.

Intelligent Campus 2030, advocating for human-centric development, explores how intelligence seamlessly integrates into the human experience. It also focuses on how technologies influence our lives, in particular:

- Promoting ethical AI practices within intelligent campus technologies, emphasizing transparency, fairness, and accountability.

- Embedding accessibility features into intelligent technologies to ensure inclusivity for individuals with diverse abilities.
- Automatically iterating intelligent technologies within campuses and showing more care for campus users.

The Huawei Intelligent Campus 2030 report ..... connecting the future

As we extend our gaze towards the future, the Intelligent Campus 2030 report is beckoning us to envision a world where innovation, collaboration, and intelligence converge to redefine the very essence of campuses and industrial innovation. This report is a roadmap for institutions world-wide to navigate the digital world, embracing future-proof sustainable and inclusive practices, to be shared in a global knowledge society. The challenges ahead of us are formidable and require a common involvement of all stakeholders in the global knowledge domain. They can be synthesized in the form of a Huawei knowledge "flower power" symbolizing the benefits by sharing innovative knowledge among academic, industrial, policy-makers, and societal partners.

Fellow of the Royal Netherlands Academy of Sciences
Fellow of the Academia Europaea

Foreword

Ding Lieyun

Intelligent Buildings Help Reconstruct Intelligent Campuses

As humans, we have an innate desire to explore the future. This trait has been apparent throughout our history, from agricultural and industrial societies to the current smart society led by digital technologies. And despite the uncertainty of the future, digitalization, intelligence, and low carbon are trends that we can be certain will shape it. The intelligent world is akin to a blank canvas, brimming with endless possibilities. The tireless efforts of people from all walks of life add color to it, reshaping our production models and lifestyles.

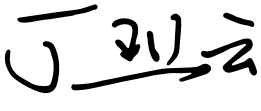
Campuses, where people live and work, are one of the basic units that make up a city. As such, they are integral to realizing carbon reduction goals. They also serve as the foothold for building a fully connected, intelligent world. As campuses continue to evolve and expand, they have transformed into a blend of physical and digital spaces that represent technological innovation.

Campuses and the buildings within them are set to undergo significant changes. The construction of intelligent buildings will drive the high-quality development of the building and construction industries. Such buildings are designed to be environmentally friendly and low-carbon, utilizing intelligent construction technologies to achieve industrialized construction. This will also pave the way for the transformation and evolution of campuses, for which we need more effective, sustainable, and efficient ways of growth. New technologies and demands are driving the constant upgrade and innovation of campus service scenarios and business models. This is driving campuses to become a self-evolving system that is eco-friendly, smart, and sustainable.

In the future, campuses will be revolutionized in terms of their planning, construction, O&M, operations, and experience. The campus consultation and top-level design are based on the "4A" architecture (business architecture, application architecture, information architecture, and technology architecture), while AI-powered advanced construction will be popularized. In terms of campus O&M and operations, people, vehicles, things, and events on a campus will be sensed from multiple dimensions to enable on-demand data processing and intelligent display. Sensors will be associated with human behaviors and big data analytics will be used to reduce carbon emissions and save energy. Personnel experience will be prioritized. Campuses will pay close attention to the immersive interaction experience between people and people, people and machines, and people and virtual space, while also attaching great importance to the impact of campus environments on people's physical and psychological health, with the purpose of delivering a superior experience to them.

As a leading global provider of information and communications technology (ICT) infrastructure and smart devices, Huawei is committed to pioneering the future through technological innovation. Huawei's Intelligent World 2030 report, released in September 2021, provides a comprehensive overview of the upcoming trends in ICTs and their applications for the next decade. The report aims to assist a range of industries in identifying new opportunities and discovering new sources of value. Intelligent Campus 2030 is part of the Intelligent World 2030 series, and outlines the vision and blueprint for future intelligent campuses by combining the expertise of industry professionals. In the report, Huawei proposes six key technical features, including intelligent twins, harmonized sensing and communications, ubiquitous intelligent connectivity, intent-driven ultra-broadband, security and resilience, and all-domain zero carbon.

As outlined in the report, the future of intelligent campuses is full of endless possibilities. We believe that it will have a significant impact on the development of the building and campus industries, guiding them as they stride toward the intelligent world.



Academician of Chinese Academy of Engineering

Foreword

David Wang

Stepping Closer to an Intelligent World

To remain competitive in the digital economy, major players worldwide plan to boost investment in data, software, computing, and networks over the following decade. The ongoing evolution of information and communications technologies (ICTs) constantly enhances our understanding of the digital world, propelling us toward a reality saturated with AI. Campuses, as fundamental components of cities, play a crucial role in this evolution. They are currently a focus of development in major cities worldwide, aiming to offer people more efficient, convenient, and comfortable ways of life.

Campus development occurs in two stages. The first stage involves conventional campuses, where countless extra-low voltage (ELV) systems are vertically and independently constructed. This stage, the industry standard before 2018, is characterized by the coexistence of narrowband, IP, and non-IP networks. The second stage, now widely endorsed by industry experts including standards organizations, associations, and design institutes, has a digital platform at its core. This stage converges IT/OT data, establishes IP-based networks, and integrates multiple networks into one.

Huawei predicts that intelligent campuses will become a self-evolving system based on the campus operating system (OS), relying on significant innovation in ICTs such as 5G-A, Net5.5G, F5G-A, cloud computing, big data, and digital platforms. The positioning of intelligent campuses will also undergo profound changes. Campuses will transition from being closed, isolated, and autonomous entities to intelligent and connected social spaces which serve the local community. They will shift from independent subsystem resource management to more refined and efficient resource supply for campus services. There will be higher demand for upgrading on-site manual data processing across domains and systems to in-depth data convergence based on the assumption that large-scale data flows will be much more secure.

The new stage of intelligent campuses is full of possibilities. As a world-leading ICT infrastructure and smart device provider, Huawei is committed to leading the future through technological innovation by maintaining heavy investment in R&D. Over the past three years, Huawei engaged in extensive discussion with hundreds of scholars, customers, partners, and research institutes. Through collaboration, Huawei gathered insights from industry experts and its own team to envision the next generation of intelligent campuses and launched the report Intelligent Campus 2030.

The report discusses the future of intelligent campuses, characterized by being digital, converged, resilient, people-centric, and green. It sheds light on the trends, vision, and potential scenarios for intelligent campuses, and outlines six technical features for future campuses. It systematically expounds technical challenges and directions for innovation, including intelligent twins, spatial interaction, ubiquitous intelligent connectivity, intent-driven ultra-broadband, security and resilience, and all-domain zero carbon. The report also proposes a reference architecture powered by key technologies for future intelligent campus construction. The purpose of this report is to assist in the construction and advancement of global intelligent campuses and to speed up progress toward an intelligent world.

Johann Wolfgang von Goethe, a German polymath, scientist, and writer, once said, "Ambition and love are the wings to great deeds." We believe that intelligent campuses 2030 are within reach, and we are excited to collaborate with industry experts and partners to explore, innovate, and make our dreams come true together. We look to build intelligent campuses with a human touch by using state-of-the-art technologies.

Executive Director & Chairman of the
ICT Infrastructure Managing Board, Huawei

Introduction

The digital economy has become instrumental in restructuring global resources. As digitalization and intelligence are sweeping across the globe, cutting-edge technologies like AI, digital twins, Net5.5G, F5G-A, 5G-A, and 6G have brought breakthrough after breakthrough. The economy is taking a new shape and new competition is emerging. An intelligent world is taking form at an incredible pace.

Climate change has also taken the world by storm. In 2022, global carbon emissions reached a new all-time high of 36.8 billion metric tons. Global boiling caused by ceaseless carbon emissions places a new sense of urgency on sustainable development worldwide, making green energy and low-carbon development paramount. At the same time, consumption upgrades and widespread automation will trigger a new round of requirement revolution, entailing dramatic changes in people's jobs and lifestyles.

The campus is a basic unit in the making of a city. It is the main place where people live and work. It acts as an important carrier to boost the digital economy, and a key point to realize green and low-carbon transformation. Intelligent campuses will undergo sustainable development and evolution. Profound and complex changes in today's world will transition campuses from connectivity of everything to intelligent connectivity of everything. In recent years, a great deal of research has gone into the creation of intelligent campuses. At the moment, intelligent campus construction is well underway.

How do you envision the intelligent campus as 2030 approaches? What will its characteristics be? What is the core driving force for its development? What exciting changes will technologies like AI, Net5.5G, F5G-A, 5G-A, and digital twins bring to future intelligent campuses?

Huawei has conducted extensive discussion with industry scholars, partners, and research institutes to deliver the Intelligent Campus 2030 report — which concludes our thoughts on the development of next-generation intelligent campuses — based on a wide breadth of industry experience. The report provides insight into three driving factors for intelligent campus development over the next decade, and proposes that intelligent campuses in the future will be characterized by being digital, converged, people-centric, resilient, and green. Additionally, it depicts ten typical scenarios of future intelligent campuses and systematically sheds light on the key technical features that support campus development.

Huawei is committed to bringing digital to every campus for pervasive intelligence.

Trends and Visions

01

Global sustainable development, digital economic growth, technological innovation, and the debut of the intelligent world have brought more important responsibilities and entirely new missions to campuses.

1.1 Driving Forces

The advent of the digital economy marks the beginning of the Fourth Industrial Revolution. Campuses will embrace new development opportunities as they bolster social economic development. According to the 14th Five-Year Plan on Digital Economy Development, by 2025, the digital economy will have expanded on all fronts, with the added value of the core industries expected to account for 10% of the GDP. This number will rise to 30% by 2030. Intelligent campuses will improve coordinated management, refined operations, and online and offline resource sharing through digital infrastructure construction. Smart brains, digital platforms, and virtual campuses will become a reality on intelligent campuses, which will then be reshaped into high-standard digital economy campuses.

The aim to satisfy core requirements of users guides campuses to transform toward data interconnection and all-domain intelligent transformation. Society has changed, and so have production methods and people's lifestyles. Consumers used to purchase products based primarily on utility. In the future, they will pay more attention to personalized and tech-assisted experiences. The main target customers will be those who pursue high quality with an appetite for additional or unique experiences. Multi-layer and diversified consumption requirements will encourage every campus to interconnect data between various business types, explore user requirements, and establish linkage. This will result in multiple new business types and scenarios. In addition, quality consumption requirements will propel data convergence

between campuses and cities to provide more personalized and targeted services.

Automation has added to labor market inequality, as many production and clerical workers saw their jobs disappear or their wages plummet. Automation has encouraged campuses to become unstaffed and intelligent through the creation of new business types and demands. Machines will replace humans in repetitive low-skill tasks, but investment in novel technologies will create a large number of positions that involve human-machine interaction. For campuses, unattended business operations like unstaffed deliveries, factories, restaurants, and logistics, will become the mainstream. For users, the transition from conventional office to human-machine collaboration requires campuses to provide corresponding environments and digital tools so that users can quickly adapt to new positions.

1.2 Development Trends

Driven by the rapid progress of the digital economy, faster innovation of forward-looking technologies, and revolution of social requirements, intelligent campuses will be characterized by being digital, converged, people-centric, resilient, and green.

1.2.1 Trend 1 Digital

Campuses That Are All-Intelligent and Fully-Connected Are Becoming a Reality

1.2.1.1 Infrastructure Featuring 10 Gbps Connections and Digital Platforms Will Become the Core Foundation for Campuses

The current stovepipe infrastructure can no longer support the digital economy or low-carbon development on campuses. Campuses in 2030 should deploy high-quality and moderately advanced IoT sensing infrastructure. By using networks that offer 10 Gbps connections as well as digital platforms, the infrastructure will ensure everything can sense and is intelligently connected.

The novel IoT sensing infrastructure based on Net5.5G, F5G-A, or 5G-A facilitates the connectivity of everything on campuses. According to Huawei's Intelligent World 2030 report, the total number of global connections will exceed 200 billion by 2030, marking the advent of an era with 10 Gbps connections. Next-generation connectivity will contribute to precise positioning, full coverage of high-speed networks, and borderless information transmission between campuses and cities. Future campuses will deploy ubiquitous IoT sensing devices to collect and monitor the running status data of the infrastructure, as well as gain operations insights from numerous IoT devices that generate massive amounts of data. As a result, campuses will be able to implement automatic operations, decision-making support, as well as prediction and warning.

The centralized foundation with the digital platform as the core helps build digital and intelligent campuses. Future campuses will provide converged and streamlined aggregation and data services based on the digital platform to enable cross-system data convergence, multi-application scenario linkage, and agile innovation.

1.2.1.2 Intelligent Buildings That Can Sense and Think Will Become the Core Element of Campuses

Intelligent buildings have become an essential part of the planning and layout for industry development across various countries. The development of comprehensive intelligent campuses is going to be a huge area of investment. The construction industry will intelligently transform and upgrade using digital and intelligent technologies. It will establish real-

time connections with campuses from start to finish and deploy IoT sensing devices to help campuses fully understand user demands as well as provide more personalized intelligent services.

As an important part of future campus construction, intelligent buildings will evolve from individual intelligence to swarm intelligence. Looking ahead, buildings will no longer be just simple, separate objects. Instead, they will become intelligent and stay online in real time with comprehensive sensing and proper collaboration. The buildings will be able to perform self-learning, self-diagnosis, and autonomous decision-making and execution. In addition, campus buildings will advance to the point of intelligent connectivity of everything for architectural complexes, bolstering comprehensive digital and intelligent campus construction. Unattended operations, automatic control, and adaptive learning can be developed by integrating digital technologies into robots and other intelligent devices, which boosts construction and operations efficiency.

Intelligent buildings will focus on user demands. Advanced technologies like IoT, wireless positioning, and mobile intelligent robots continuously enhance building sensing to achieve comprehensive awareness and real-time interconnection. This enables buildings to autonomously sense and respond to user behavior and preferences in real time, which meets refined user requirements.

1.2.1.3 Intelligent Operations with Data-Based Governance Will Become the Mainstream

Data is a key production factor for digital economy development. Intelligent campuses need to attach importance to data value mining. Multi-system data linkage supports lean operations and improves campus management services. Refined data operations will not only become the underlying support for intelligent campus construction and operations, but will also guarantee them. Proactive data value mining ensures sustainable campus operations.

Decision-oriented real-time monitoring, situational awareness, and closed-loop analysis and decision-making based on global data analysis facilitate precise and efficient management decisions. It has become a must for intelligent campuses to build decision-making platforms to analyze the overall situation in real time. The platform predicts campus business risks alongside operational exceptions by learning and analyzing a multitude of data. In addition, it dynamically displays campus operation and comprehensive business statuses, assisting managers in precise and efficient decision-making.

Based on multi-system data linkage, the campus digital platform converges data, boosts service development, and improves quality and efficiency. Campus operators will interconnect data from multiple systems to shift service functions from isolated to collaborative, with shorter time and less labor input, achieving highly efficient operations.

1.2.2 Trend 2 Converged

Campuses Will Feature Virtual-Real Fusion and Boundless Function Integration

1.2.2.1 Campuses Will Employ Virtual-Real Fusion in Diversified Scenarios

Virtual-real fusion will be a new trend for future intelligent campuses. It will emphasize experiences in the virtual world while focusing on digital overlays in the physical world. Digital technologies will recreate the relationship between people and campus spaces, stimulate new production and living requirements, drive new scenario-based applications to provide novel services, and evolve digital, physical, and human spaces on intelligent campuses. In the

future, intelligent campuses will seamlessly blend the physical and digital worlds based on replication through digital twins, the creativity of digital native, and the possibilities provided by virtual-real fusion. This will result in digital campuses that can deliver a more intelligent and convenient experience.

Replication through digital twins serves as the fundamental building block of campuses in the metaverse, as it enables precise replication and restoration of the physical world. It will be widely utilized in industrial production, operations management, and information dissemination on campuses. In the future, as lidar and remote sensing image technologies continue to mature, they will push the iteration and upgrade of unattended factories, metaverse stadiums, and immersive exhibition halls.

The creativity of digital native is the engine of intelligent campuses in the metaverse. It will spawn new scenarios related to campus planning, design, and service management in the digital economy. By using models created by digital native, campus designers can simulate future campus operations in virtual environments and provide optimized operational strategies during the planning and design phase. Digital native is expected to revolutionize office service process management and predictive maintenance on next-generation campuses.

The possibilities provided by virtual-real fusion enable us to envision the most intelligent campus possible in the metaverse. It breaks down the boundaries between the virtual and physical worlds, creating new versions of digital culture, entertainment, and sports. Technologies such as holographic projection, 360° panoramic screens, and virtual reality (VR), will facilitate the implementation of immersive art exhibitions, interaction with virtual idols, stadium augmented reality (AR), and more.

1.2.2.2 Campuses Will Be Integrated with Cities Through Borderless Convergence

As urbanization accelerates, the melding of campuses and cities is inevitable. Future intelligent campuses will deeply integrate their spaces and functions into the development of cities.

Intelligent campuses will focus on converging spatial functions along with promoting integrated function development for surrounding regions and cities. Campuses will become borderless and offer open spaces and diverse functions. Together with cities, they will grow side by side. This can improve their overall competitiveness and keep development sustainable. Intelligent campuses will have to break down their physical borders. This will allow employees to feel more connected to their cities, attract and retain more talent, and promote faster innovation and development.

Intelligent campuses will prioritize the convergence of city data. By linking campus operations with city services, campuses can precisely identify user service needs, expand their service scope, and improve service satisfaction and operational efficiency. They will be able to offer city-level services utilizing a systematic approach and a constantly evolving digital platform, which will provide users with personalized services that exceed their expectations. In addition, digital technologies will be deeply integrated into city applications, and digital platforms will improve campus operations management and lead to refined and diversified operations.

1.2.3 Trend 3

People-Centric

Human-Centricity Is the Key to Making Future Campuses Viable

1.2.3.1 More Proactive, Smarter, and More Personalized Campus Services Create a Smooth Experience for Users

Future campuses will be people-centric. They will proactively sense user needs, and provide personalized services for different types of users like office employees, decision-makers, and operators. Additionally, future campuses will be able to predict user behavior and associate different scenarios with each other, delivering an immaculate experience for the people who live and work there.

Future campuses will be capable of learning employee habits and sensing their needs in advance. This allows them to provide a comfortable and convenient service, and create a novel office experience for existing users and young generations. Future campus construction should take into account the youth's idea of a convenient experience. Campuses should be able to sense their behavior and environments. Sensed information can then be associated with the central building control system to intelligently shift the parameters of the office environment, such as temperature and lighting. This creates a cozy office that provides an optimal experience while consuming the least amount of energy.

It is the job of campus operators to undertake operations management in an efficient and refined way. To this end, industrial, service, and management resources will be integrated into one platform to reduce costs and streamline logistics. Routine service operations and personnel activities within the campus should be sensed in real time and precisely analyzed and predicted through the use of AI, IoT, and other technologies. Alarms can then be automatically generated regarding any exceptions, and service tickets automatically dispatched. All these practices will bring campus operators an efficient and intelligent operations experience that surrounds human-machine collaboration.

Managers of future campuses will act as leaders in campus digital transformation, and will pay close attention to sustainable development. Campuses will build a decision-making center powered by real-time situation analysis. Data

across all domains will be consolidated and displayed in a visualized way so that managers have direct access to current campus information and can implement policies that are informed and targeted. Future intelligent campuses should capitalize on AI, digital twins, and big data to evolve toward smart decision-focused operations, which is necessary for scheduling resources centrally in the management field.

1.2.3.2 Campuses Will Be Able to Better Understand Users and Provide a More Human Touch

While striving to maximize productivity, intelligent campuses should always put users first, and show consideration for everyone on the campus to improve their experience. **Intelligent campuses will serve not only as places for face-to-face communication and interaction, but also as spaces that meet people's emotional needs, strengthen social connections between people, and help them reach their full potential.**

Being social and fulfilling one's potential are important emotional and human-centric values for campus users. They hope campuses can build a virtual social platform that delivers an immersive experience and integrates online and offline interaction. Users want campuses to offer personalized services that cater to their needs and goals through proactive sensing and analysis. By building an innovation platform that can facilitate in-depth communication and cultural exchange, campuses can provide personalized services which are more proactive and inclusive.

1.2.4 Trend 4

Resilient

Security and Resilience Are the Core Assurance for Campus Management and Operations

1.2.4.1 Single-Point Defense Will Give Way to Proactive and Three-Dimensional Security Protection

Campuses will transition from being closed and separated to open and integrated, which will require higher levels of campus security. As such, campuses should develop three-dimensional proactive security across physical, data, and network fields to put a stop to any potential threats.

Campus security management will integrate with AI and machine learning to improve device analysis and decision-making as well as predict and prevent risks proactively. According to Memoori, an independent analyst company focused on intelligent buildings, more than half of network security cameras will have onboard AI by 2028. For example, the multi-modal image recognition of AI cameras can automatically locate people of interest and suspicious behavior, helping determine security risk levels based on analysis algorithms, and related resources can then be automatically scheduled to solve problems. AI algorithms enable devices to automatically adjust their operating modes and make accurate decisions based on real-time situations.

In addition, the construction of intelligent campuses should focus on data security, with close attention paid to data classification and grading, security assessment, and cross-border security. Relevant policies are making these systems more detailed and complete. Supporting data security specifications and standards systems will be improved as well. However, there has been no targeted security policy for the intelligent campus domain. In the future, specifications on data risk assessment and cross-border data transfer must be refined. Critical information infrastructure and hardware deployment on intelligent campuses must strictly comply with corresponding laws and regulations.

Campus digitalization is based on network security, which is the core requirement of building intelligent campuses. Campus digital transformation heavily relies on high-performing networks as its foundation to cope with new network security threats. The widespread use of digital technologies has eliminated physical network boundaries on campuses, rendering traditional border protection and security architectures ineffective. Future network security protection systems for campuses will be able to proactively detect, promptly monitor, and intelligently analyze network threats, automatically responding to them and processing them within seconds. Hybrid network security solutions that combine diverse technologies and analytical data will be the primary tools for managing network security in the future.

1.2.4.2 Situational Awareness and Agile Recovery Are a Must for Campuses to Become Highly Resilient

The world today is evolving at formidable speeds. Such immense change naturally leads to risks, and black swan events frequently occur. This raises higher requirements when it comes to the strong adaptability and recovery of campuses. **In the years to come, intelligent campuses must be able to instantly respond to risks and threats, effectively tackle various changes and impacts, maintain the operation of campus services as usual, and quickly recover from any shocks.**

By leveraging AI, intelligent vision, IoT, and big data, intelligent campuses can comprehensively sense the operational security within the campus. They can promptly detect potential risks, assess and analyze them, as well as monitor and warn of them. These technologies improve the agile recovery of campuses in the case of uncertain, complex, and dangerous environments. Campuses will continuously update and optimize the risk database to cover all risk types, their likelihood of occurring, the extent of their impact, and those who may be held liable. Models will be designed based on intelligent algorithms to empower monitoring and evaluation, the generation of alarms, and the prevention and control of risks. AI, digital twins, and 3D holographic technology can be used to accurately simulate situations when a campus is attacked. Multi-modal data analysis helps decision-makers conduct emergency recovery experiments in a simulated environment. They can then make more precise decisions and emergency scheduling can be more intelligent in the event of a real attack. Campuses will use unmanned aerial vehicles (UAVs) and other methods to quickly identify emergency locations, as well as significantly strengthen their capabilities in addressing highly complex and dynamic shocks. Robots that are able to sense and make decisions can be intelligently scheduled for emergency repair and rescue, making rescue in complicated and high-risk scenarios more precise, efficient, and secure.

1.2.5 Trend 5

Green

Campuses Will Eventually Become Zero-Carbon to Help Achieve Sustainable Development

Sustainable development is crucial for the future of humanity. Campuses play a critical role during green and low-carbon transformation, and using digital technologies to promote sustainable development and transform energy structures has become an important trend.

As climate change worsens, global boiling caused by carbon emissions makes sustainable development imperative. As a major source of global carbon emissions, campuses are a key link in addressing climate change. Figure 1-1 shows that global carbon emissions grew in 2022, reaching a new all-time high of 36.8 billion metric tons. Against this backdrop, improving campus resource output and carbon productivity will become a key entry point for emission reduction strategies of various countries.

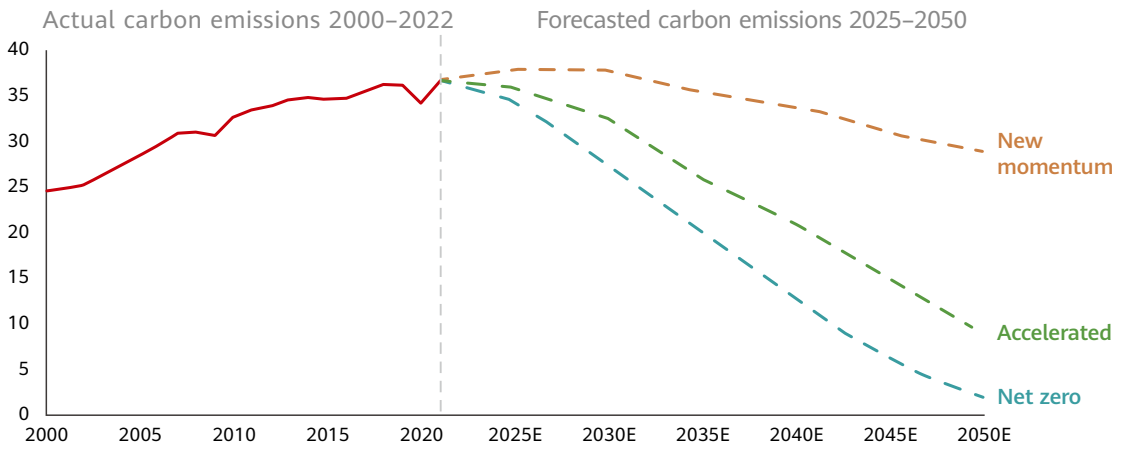


Figure 1-1 Historical and forecasted global carbon emissions (unit: 100 million metric tons)

The transformation of global energy structures is now at a critical juncture, prompting campuses to consolidate energy management systems and increasing renewable energy utilization. According to bp's Energy Outlook, the usage of nuclear energy and renewable energy such as hydropower and PV will reach 72% and 81% respectively by 2050, as shown in Figure 1-2.

1.2.5.1 Zero-Carbon Campuses Will Shift from Optional to Mandatory

Zero-carbon/Near-zero-carbon campuses are vital to China's dual carbon strategy. In compliance with policies for green, clean, and cyclic development, campuses are encouraged to explore zero-carbon transformation throughout

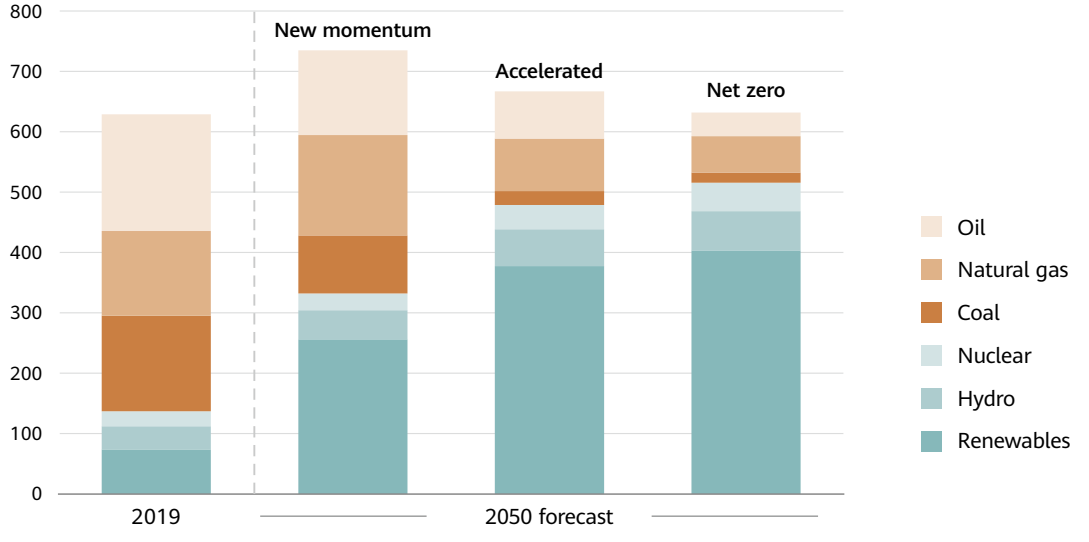


Figure 1-2 Global primary energy fuel sources (unit: TW)

the lifecycle, such as construction, O&M, and management. Campuses must strike a balance between social development and the dual carbon goals, explore sustainable and low-carbon construction paths, and properly implement the dynamic zero-carbon goal.

Many regions have issued preferential and incentive policies to encourage carbon emission reduction, providing strong support for the construction of low-carbon/near-zero-carbon/zero-carbon campuses as well as the implementation of new green building regulations. For example, City A released a policy regarding the development of the green energy industry in August 2023. The policy evaluates and recognizes zero-carbon campuses and green low-carbon factories based on their new energy projects and energy saving reconstruction. Cash rewards will be offered to top-level zero-carbon campuses that meet the criteria.

Policies regarding carbon emission control when it comes to low-carbon/near-zero-carbon/zero-carbon campuses and the energy-saving management of green buildings are becoming increasingly stringent. For instance, in June 2023, a city's local government fined key sewage treatment plants around CNY2.4 million because their carbon emissions exceeded the permitted limit. In the future, more enterprises will be included in the Emissions Trading Scheme (ETS) as the carbon trading market matures. Carbon quotas, which are mainly distributed free of charge, will gradually transition to paid allocation and carbon taxes will be introduced. Campus carbon emissions are mostly attributed to buildings. As such, ultra-low energy consumption buildings will become necessary to zero-carbon campus construction. Intelligent campuses are expected to construct ultra-low energy buildings by 2030.

1.2.5.2 Green Energy Structures, Efficient Energy Management, and Sustainable Production and Lifestyles Will Become Key

The transformation toward zero-carbon/near-zero-carbon campuses will begin from three aspects: energy structure, energy management, and zero-carbon lifestyles. Digital technologies will comprehensively enable and optimize zero-carbon transformation measures, helping campuses achieve the goal of carbon neutrality.

A higher proportion of renewable energy alongside the application of green technologies promotes campus energy structure transformation. Campuses will further increase the use of renewable energy such as PV, wind power, and hydropower to optimize the campus energy structure from the source. Moving forward, AI will play a greater role in predicting power generation capacity and demands for renewable energy, power grid operation and optimization, and distributed resource management. In addition, smart microgrids as well as energy storage facilities (hydrogen/ammonia/electrochemical) will be deployed to comprehensively facilitate the low-carbon transformation of campus energy structures.

Carbon planning, emission reduction, and trading comprehensively reduce the carbon footprint on campuses and efficiently manage

campus energy. The concept of carbon neutrality must be embedded into the entire process of campus planning, construction, management, and operations. A zero-carbon management platform can be set up to carry out carbon accounting, achieving the goals of carbon peak and neutrality. Negative emissions technologies, the purchase of carbon sink products, alongside carbon capture, utilization, and sealing (CCUS) will strengthen the carbon absorption of campuses. In particular, CCUS captures carbon dioxide during industrial production to provide carbon emission reduction solutions for energy-consuming enterprises. In addition, a carbon asset management platform will be established to provide green finance and carbon inclusive services. The platform encourages campus enterprises to actively participate in the carbon market and achieve carbon neutrality by purchasing carbon sink products.

Green and low-carbon jobs and lifestyles fundamentally reduce campus energy demands. Based on green energy supply and efficient energy operations, future campuses will convey the value proposition of energy saving and emission reduction to campus users. They will advocate green lifestyles such as low-carbon commuting, paperless office, and bring your own utensils (BYOU), and help people develop awareness surrounding the topic. The transformation of campus production to zero-carbon is a complex and multifaceted project. It should start with upgrading the production infrastructure to low-carbon, electrifying devices, and powering them with sustainable energy. In regard to people's lives, future campuses will build a digital carbon footprint bonus point system to help users track their carbon footprints in real time. The system will be gradually improved to guide green work and living habits. Meanwhile, campuses will advocate environmentally-friendly commuting modes of using fewer automobiles. Green travel can be made a reality by exploring the trial run of autonomous shuttles on open campuses.

1.3 Vision

1.3.1 Intelligent Campus Definition

Oriented toward 2030, we clearly define intelligent campuses as an all-intelligent, people-centric, green, and low-carbon self-evolving system that seamlessly integrates physical, digital, and human spaces.

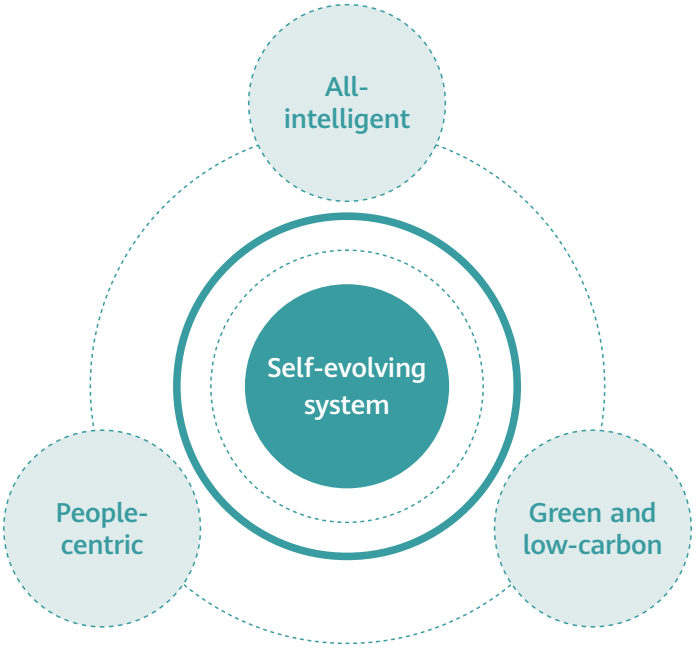


Figure 1-3 Definition of intelligent campus 2030

All-intelligent: Full intelligence will vitalize campuses. Physical, digital, and human spaces will be closely integrated through digital technologies. IoT sensors serve as the eyes of intelligent campuses. They sense any change happening on the campus and collect data. Meanwhile, the digital platform acts as the campus brain; it supports situational awareness, risk prediction, and precise decision-making to facilitate intelligent management and operations. Campus service applications serve as the hands of the campus, while campus networks act as the nervous system. Together, they can enable self-growth and iteration within campuses.

People-centric: Digital technologies can transform campuses into spaces that provide people-centric services and emotional value. An intelligent campus will shift its focus from functions to the people who work and live on it, becoming more people-centric. The evolution and inclusion of ICTs will bring more care and consideration for people on campuses. They will be able to better understand user needs and provide personalized smart services proactively.

Green and low-carbon: The intelligent campus will become a green, low-carbon, and highly resilient self-evolving system. Campuses embrace digital technologies to promote green energy structures, efficient energy management, as well as sustainable production and lifestyles. Such technologies transform campuses toward being green, low-carbon, and zero-carbon, achieving sustainable development.

1.3.2 Intelligent Campus Vision

For Huawei, its vision regarding intelligent campuses is to Bring Digital to Every Campus for Pervasive Intelligence.

Bringing digital to every campus: Huawei is committed to bringing digital technologies such as digital platforms, high-speed networks, cloud computing, and smart terminals to every campus.

Achieving pervasive intelligence: In addition to making novel smart management methods, operations modes, and experiences possible, digital technology helps campuses achieve self-growth and iteration.

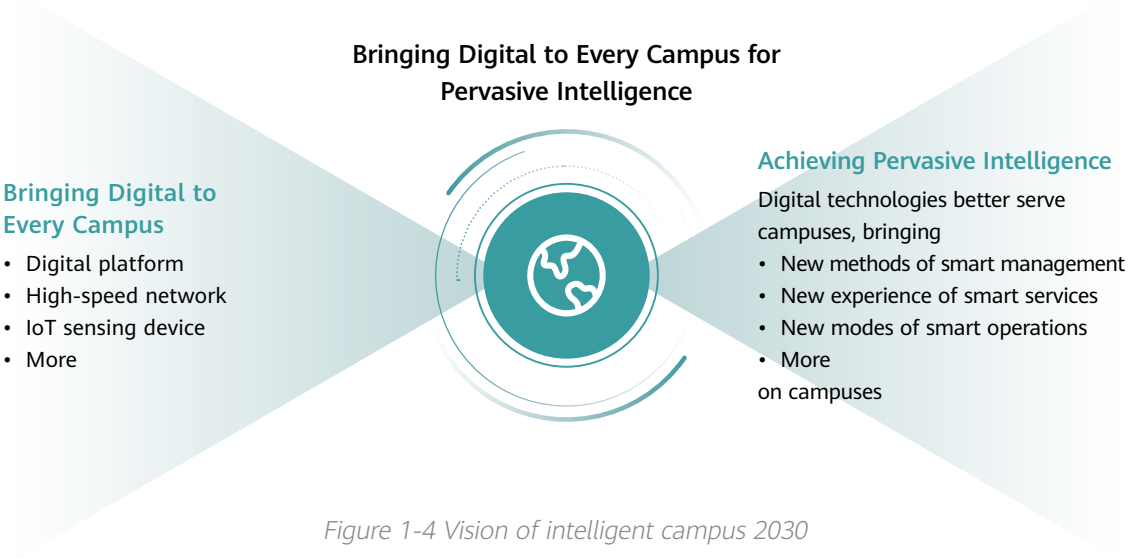


Figure 1-4 Vision of intelligent campus 2030

Scenarios

02

Ever-changing digital technologies and higher campus service requirements constantly iterate and upgrade service scenarios, as well as stimulating new scenarios and applications.

Typical scenarios in 2030 must follow the development trends of forward-looking industries. A number of key technologies are closely related to intelligent campuses and will be widely applied in future scenarios. These include cognitive intelligence (generative AI), digital twins (intelligent platform), spatial interaction (immersive experience and human-machine collaboration), and more, which will affect and reshape campus office, production, and lifestyles. In particular, ubiquitous intelligent connectivity (IoT sensing and edge intelligence) and digital twins (intelligent platform and digital modeling) will enable campus operations management and decision-making.

In addition, typical scenarios in the future will reflect the development trends of intelligent campuses — they will be digital, converged, people-centric, resilient, and green. There are a lot of issues to be tackled. For example, how can campuses stay people-centric, better understand users, and provide them with more proactive and personalized services? How can campuses perform intelligent operations to replace manual governance with data-based approaches, and become zero-carbon by promoting green energy structures, efficiency energy management, and sustainable production and lifestyles?

Based on these questions, we have selected 10 typical intelligent campus scenarios. As digital technologies advance, such scenarios will be further developed based on future campus applications and practices.

2.1 Scenario 1: Holographic AIOC

2.1.1 Definition

Holographic visualization for global display, all-round intelligent situational awareness, and autonomous, precise analysis and decision-making combine to make campus operations visible, manageable, and controllable.

2.1.2 Description

Faster campus digitalization and constantly refined governance make holographic AI + IOC (AIOC) an important method to build digital and intelligent campuses. It has intelligent sensing, cognition, and decision-making capabilities. Campus digitalization will evolve from pure data mining and integration toward insights into campus management and operational decisions. IOC is no longer a simple data visualization tool. Instead, it will become a smart brain that assists campuses in intelligent all-domain management as well as efficient and precise decision-making by collecting and analyzing massive amounts of service data. Using the digital platform, light field holographic rendering, intelligent human-machine collaboration, and other state-of-the-art technologies, the holographic AIOC breaks campus information barriers and displays real-time campus operating status in holographic visualization mode, facilitating more dynamic and intelligent campus management. The holographic AIOC also gradually becomes autonomous with the aim to replace manual decision-making, allowing campuses to visualize, manage, and control operations.

The holographic AIOC applies technologies such as the campus digital platform and digital twins to aggregate, monitor, and analyze campus-wide operating data in real time. It visualizes the overall campus status, intelligently senses campus situations and operations, and supports intelligent command and emergency dispatch. This ensures smarter scenarios, more intelligent operations, and more efficient management on campuses. Typically, a campus involves an abundance of data. The holographic AIOC comprehensively aggregates and analyzes campus data in regard to various scenarios such as security, transportation, access, energy, and assets. It precisely maps the physical space to the digital through light field holographic rendering and digital twin models. This reproduces the campus operating status to the maximum extent, as well as implements multi-dimensional situational awareness and holographic display. The holographic AIOC adapts to decision-making requirements across different campus scenarios

and integrates multiple types of data into the holographic model. This brings an excellent visual experience and supports real-time interaction. Thanks to AI, the holographic AIOC enables campuses to perform knowledge inference based on massive amounts of data and build knowledge networks so as to deduce the in-depth logic behind the data. This contributes to intelligent insights and cognition, campus operating status sensing, as well as precise analysis and collaborative command across campuses. The holographic AIOC also performs intelligent command and emergency dispatch according to the overall situation; it delivers execution instructions in human-machine interaction mode. High bandwidth, low-latency 5G-A/6G networks, and intelligent analysis of machine learning integrate sensing terminals into emergency scenarios. The holographic AIOC quickly analyzes risks based on the real-time backhaul of on-site voice and HD videos and schedules optimal resources through human-machine collaboration, efficiently managing campus emergency responses from start to finish.

Capitalizing on the AI foundation model and campus digital platform, the holographic AIOC will gradually develop autonomous prediction and intelligent analysis for efficient decision-making and analysis, which will assist campus managers during governance. Dynamic data self-learning enabled on the campus digital platform, machine learning, algorithms, and models equip the holographic AIOC with autonomous knowledge acquisition and iterative evolution. On this basis, the holographic AIOC can independently predict development trends and operational risks, as well as identify campus operations or operational management exceptions. It supports campuses in making decisions on investment attraction by autonomously learning data about regional industry development and campus economic operations. Regarding service ticket management, the holographic AIOC accurately identifies rule violations such as trespassing and parking violations, and automatically handles relevant service tickets to implement proactive problem awareness, timely warning, and efficient handling.

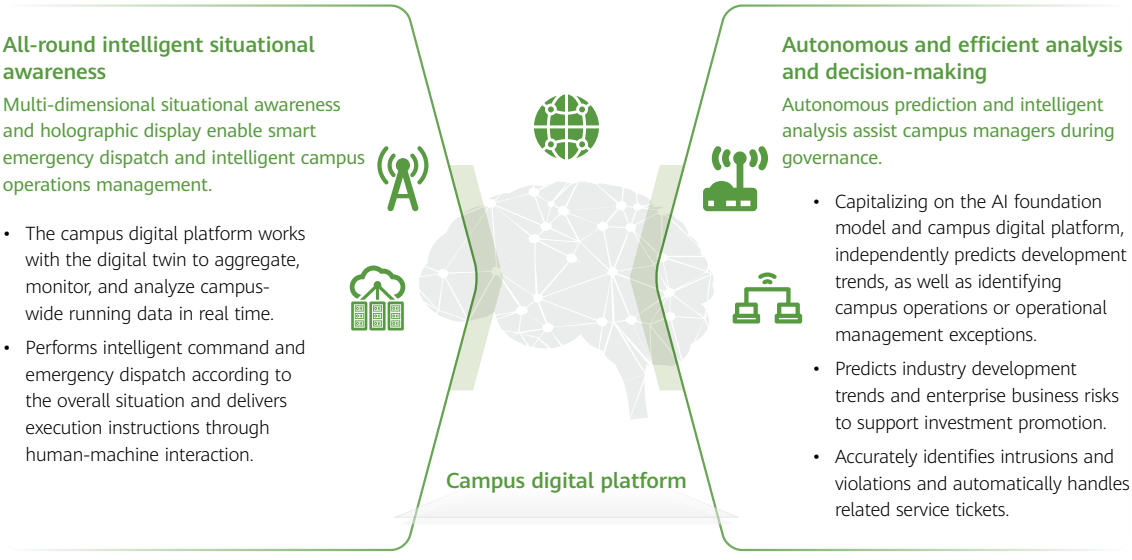


Figure 2-1 Holographic AIOC scenarios

2.2 Scenario 2: Transcendent Smart Office

2.2.1 Definition

Generative AI will completely change the way people work on campuses. High-quality simplified office networks and intelligent twin technologies will enable employees to work anytime, anywhere, and with significantly higher efficiency.

2.2.2 Description

Transcendent smart office will reshape the way people work and inspire greater creativity in them than ever before. The campus office mode will gradually evolve from hybrid offices to transcendent smart office that can take place anytime and anywhere. This technology will extend its reach to anyone with office needs, regardless of location. Time-consuming repetitive tasks, low statistical work efficiency, and similar issues lower the working efficiency of employees, who anticipate personalized office services and a healthy and convenient office experience. In the future, campuses will utilize simplified office networks, AI-based collaboration, and digital twins to greatly enhance the campus office space and work itself. These technologies will create a new way of connecting office groups and make campus offices more flexible and creative.

The smart office mode will integrate online and offline office resources and offer intelligent office assistant services. This will create an efficient and comfortable office experience that features human-machine collaboration for office groups. AI technologies based on high-quality office networks and large models will provide intelligent assistant services. These assistants are capable of understanding personalized requirements and are highly professional, which ultimately enhances office collaboration efficiency. In terms of office services, the super AI office assistant

will accurately understand employee schedules and work habits, provide considerate and customized work plan suggestions, automatically generate briefings and weekly reports, and automatically collect, analyze, and process complex data. The AI office assistant also offers conference services such as preparing and sending notifications before a conference, real-time multi-language translation during a conference, automatic presentation of reports, and automatic generation of role-based meeting minutes after a conference. In terms of environment services, the AI office assistant will cover a wider range of campus offices. It can intelligently optimize office environments based on behavior and environment awareness. For example, it dynamically adjusts parameters such as lighting and temperature based on changes in weather, indoor environments, and office personnel working habits. This ensures the most comfortable and personalized office environment with the lowest power consumption.

Transcendent smart office removes barriers between physical space and the digital world, providing campus users with cross-region, immersive, and interactive office services that are available anytime and anywhere. High-quality office networks alongside digital twins will result in an immersive office experience, which is characterized by holographic conferences and virtual human secretaries. Holographic conferences are expected to integrate virtual conference spaces, virtual participants, and offline participants to create more immersive remote collaboration and interaction. For example, holographic conference participants may be able to align space with remote users through simple means, creating an immersive portal to the virtual conference space. A virtual human secretary is expected to manage schedules and travel arrangements through holographic images and virtual roaming, eliminating the need to do so manually. Additionally, during busy periods, campus managers can attend activities and conferences online through the use of digital twins, meeting the requirements for business opportunity negotiation and daily communication.

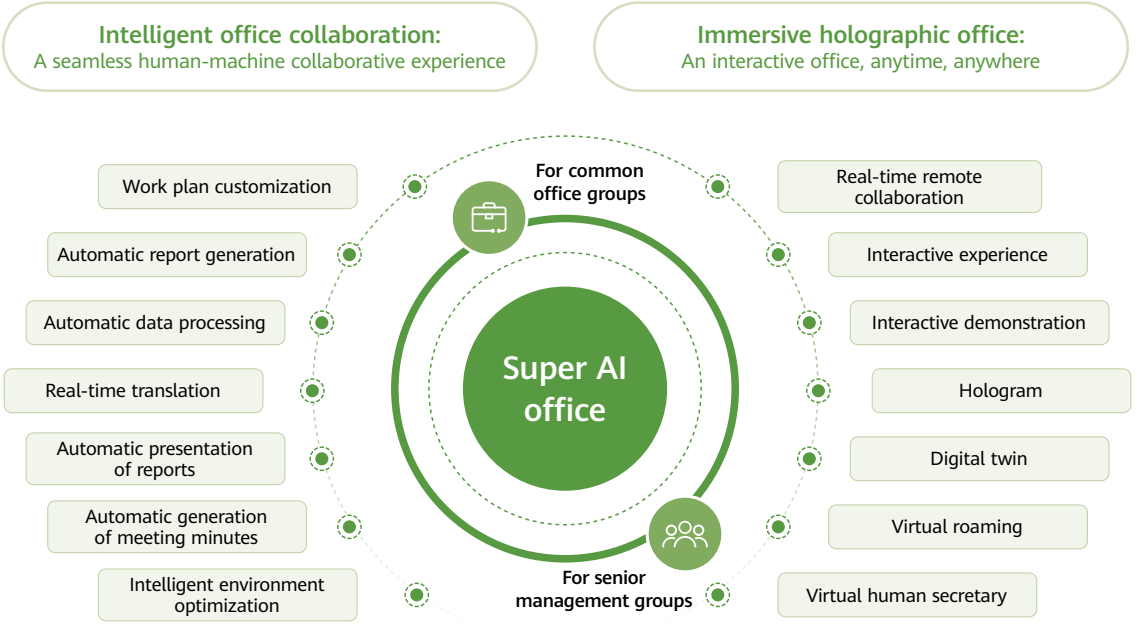


Figure 2-2 Transcendent smart office scenarios

2.3 Scenario 3: Multi-Region Access Based on Intelligent Sensing

2.3.1 Definition

Intelligent campuses of the future will provide users with a premium, efficient, and convenient access experience across multiple associated regions and scenarios.

2.3.2 Description

As technologies advance and transportation vehicles evolve, new access scenarios will unfold on intelligent campuses. Campus travel will evolve from the current 'two points, one line' approach to a more complex multi-region and multi-scene linkage approach. It will transform from simply meeting basic needs to a premium, convenient, and streamlined experience. Moving forward, we must cater to the more personalized travel needs of office personnel and visitors traveling to and from a campus. By utilizing technologies like AI, big data, IoT, and V2X, we can achieve intelligent access in all scenarios with cross-city and cross-campus data convergence, resulting in convenient, personalized, and interactive access experiences.

Access based on multi-region linkage, including the exterior of a campus, parking lots, office areas, and public areas, will bring a highly efficient and customized daily travel experience for campus office employees. Campuses will utilize AI algorithms, big data, and intelligent multimodal self-learning technologies to provide them with pre-travel planning, intelligent parking, and seamless access services. The campus service platform will offer personalized travel plans based on users' travel plans before they arrive on campuses. Public transport users can access live information regarding bus and metro congestion directly from their mobile devices. Drivers can receive intelligent push notifications, including optimal route recommendations, to keep abreast of road conditions. Upon arrival at the campus parking lot, office personnel can enjoy diverse intelligent parking services, including seamless entry, intelligent guidance to available parking spaces, quick parking, active parking route planning, automatic allocation of shared parking spaces, reverse vehicle search, and seamless departure. Upon arriving

at a campus, office personnel can smoothly enter and move within the campus through personalized digital identities and intelligent access control. For example, they can easily book and take elevators, while high-precision target recognition gates release access in seconds. This makes commuting easier for everyone, all while ensuring campus security.

Visitor access based on multi-scenario linkage, including visitor reservation, reception, navigation, and management, provides visitors with a convenient, immersive, and interactive experience. V2X and mmWave sensing technologies will also offer them immersive AR navigation, automatic parking, and unstaffed shuttle bus services. Digital twins and AI technologies are combined to generate an immersive 3D map. This map allows visitors to intuitively and clearly obtain a virtual view of roads and destinations on their mobile devices before they visit a campus, upgrading the campus navigation experience. Upon arrival at the parking lot, license plates can be accurately identified, and automatic parking services can be provided based on mmWave sensing technology. This allows for precise vehicle control, autonomous obstacle avoidance, and autonomous parking. Once visitors exit the campus parking lot, unstaffed shuttle buses equipped with lidar, ultrasonic radar, and HD cameras will automatically take them to their desired office locations. The buses use precise positioning, automatic planning, and a vehicle-pedestrian perception system to navigate. Before visitors leave the campus, the buses again transport them to their designated exits. Visitors who choose to drive manually can also use these buses to reach the parking lot and find their car, making departure seamless.

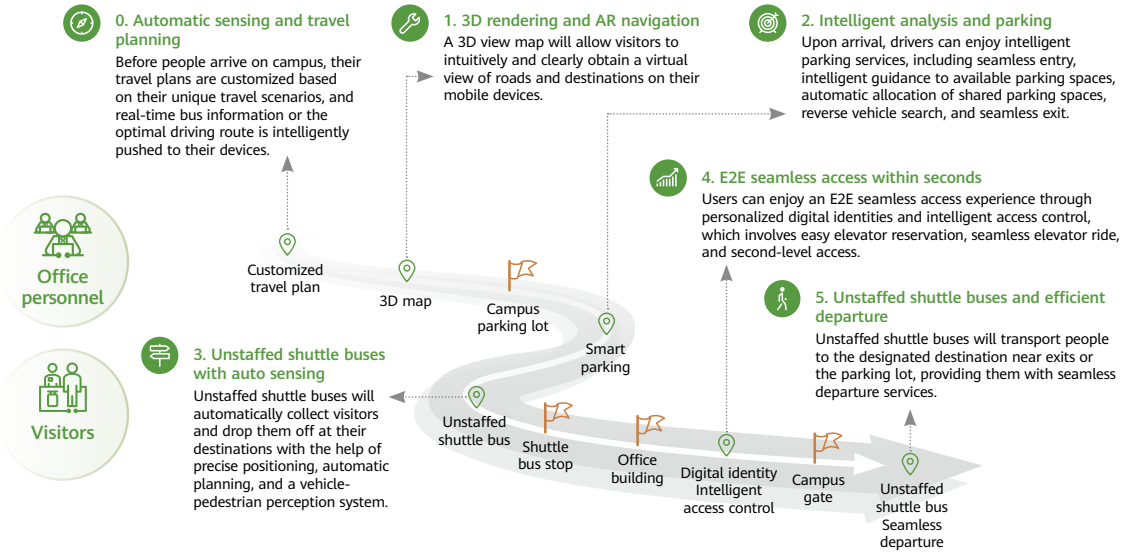


Figure 2-3 Scenarios of multi-region access based on intelligent sensing

2.4 Scenario 4: Fully-Automated Asset Operations

2.4.1 Definition

The campus digital platform will enable all-domain asset visualization, remote management, and clean data, thereby improving asset operational efficiency and unlocking value from assets.

2.4.2 Description

The combined development of campuses and technologies such as IoT, blockchain, and big data will undoubtedly give birth to fully automated asset management and refined asset operations. Current campus asset management heavily relies on manual processes, leading to inefficient asset stocktaking, numerous idle assets, and costly O&M. In the future, static asset management will transition to dynamic, and user-centric value management for assets will be prioritized. Assets will be managed throughout their lifecycles, taking user requirements into account. All-domain assets will be visualized and can be managed remotely, generating clean data at the same time. This can enhance asset operational efficiency and maximize asset value.

Technologies behind the visualization and remote management of assets include IoT, AI, and blockchain, alongside holographic, visible, and real-time asset monitoring and exception warning systems. Holographic and visible asset operation monitoring is based on the campus digital platform and IoT. All-domain interconnected assets, along with multimodal data from IoT sensing and intelligent vision, facilitate intelligent asset stocktaking and real-time operating status monitoring. On top of this, 3D rendering models holographically visualize such statuses. In addition, relying on remote and real-time monitoring, warnings will automatically and promptly be generated against any asset exceptions, with service tickets automatically dispatched to handle the exceptions within seconds.

AI-powered full-lifecycle operations analysis, predictive maintenance, and decision-making for assets will result in intelligent operations and higher utilization. AI foundation models will enable campuses to implement predictive asset maintenance. Developed based on historical and real-time operational data, these models can predict potential faults and automatically formulate and execute maintenance plans, transforming traditional passive maintenance into proactive maintenance. This will reduce the failure rate and increase the reliability of campus assets.

Moreover, intelligent operations analysis will help campus operators make more informed decisions and unlock value from assets. Indicators such as asset utilization, the proportion of idle assets, and asset ROI, can be automatically analyzed based on their operating status and performance data. On this basis, managers and operational personnel can make accurate decisions to mine asset data value and activate seldom-used or invalid assets, thereby enhancing the full-lifecycle utilization of assets.

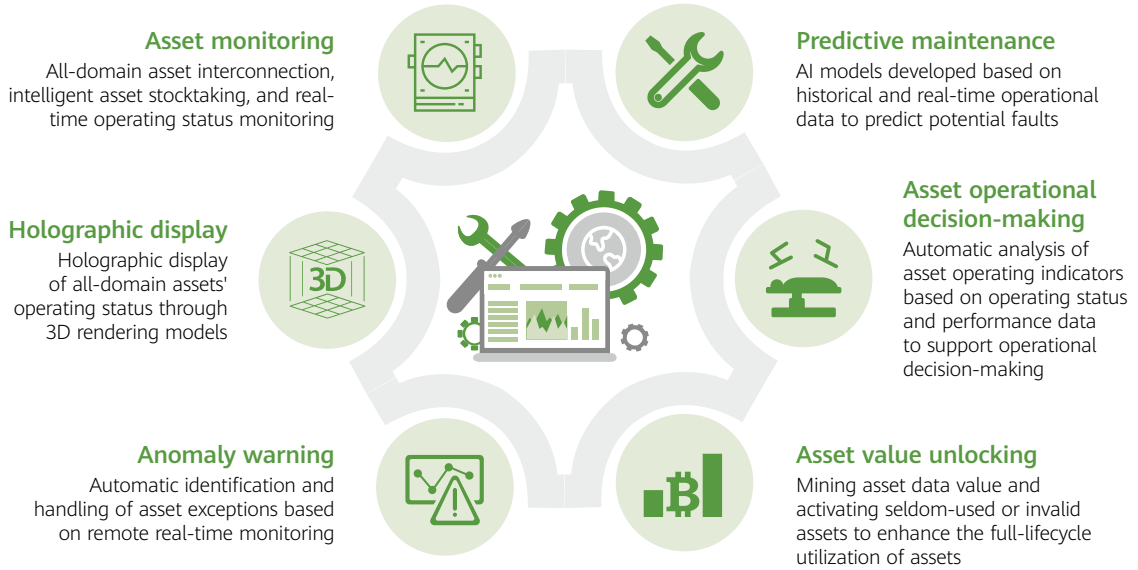


Figure 2-4 Fully-automated asset operations scenarios

2.5 Scenario 5: Multi-Element Linked Logistics Scheduling

2.5.1 Definition

Future intelligent campuses will create a three-dimensional and highly efficient logistics management system that links ground and low-altitude areas, facilitating collaboration among people, vehicles, goods, and yards.

2.5.2 Description

As low-altitude wireless coverage, generative AI, IoT sensing, and visual sensing technologies become more integrated and widely used in the logistics industry, campus logistics is expected to move toward intelligent and precise

logistics scheduling based on "computing + data + models". Next-generation campuses are expected to use AI foundation models as the core for logistics scheduling. This will enable automatic matching of logistics requirements and resources. By comprehensively perceiving, digitally connecting, and deeply integrating people, vehicles, goods, and yards, they can centrally manage logistics, information, and service flows, realizing real-time logistics scheduling and precise decision-making.

Intelligent scheduling throughout campuses can be achieved through linkage between data from multiple systems. Adaptive vehicle and warehousing management and visualized goods tracing will be utilized to enable vehicle-goods collaboration and ensure secure and efficient operations. Technologies such as AI and IoT are utilized to obtain real-time information about goods. Demand forecasting guides the entire process of adaptive management on campuses, resulting in improved transportation reliability and reduced operating costs. The logistics vehicle management platform and the campus security system will interwork for vehicle management and scheduling. Planning for vehicle entry time, precise parking policies, and vehicle scheduling policies can be automatically optimized through comprehensive analysis of vehicle reservations, transportation routes and paths, and vehicle arrival time. This will realize efficient vehicle-goods collaboration and zero waiting times when goods are moved to and from campuses. As for intelligent warehousing, equipment modeling will be standardized, and AI algorithms combined to optimize operations and ensure streamlined goods transportation and efficient production. Goods management involves intelligent matching with waybills, loading docks, and vehicles to track goods throughout the process. Visualized video tracing will be utilized in abnormal cases to quickly locate incident locations and handle alarms and events within seconds. This effectively solves problems such as untimely detection and handling, poor goods management, and disorderly personnel entry and exit. Campus logistics management

platforms will be interconnected with access, visitor, and attendance management systems to accurately detect and verify personnel entry and exit as well as goods warehousing to ensure safe operations and implement the automatic management of people, vehicles, goods, and yards. City-level logistics vehicle scheduling based on big data and AI algorithms is suitable for large-scale campuses with a wide business scope, high transportation timeliness requirements, and numerous random goods orders. Campus logistics platforms will connect to city-level transportation platforms. These platforms will automatically optimize goods transportation routes based on road topology, historical traffic data, and real-time traffic data. This will effectively improve the efficiency of vehicle scheduling and management, while also reducing transportation costs.

Low-altitude full wireless coverage, IoT, big data analysis, and AI algorithms will help build a three-dimensional and cyclic intelligent low-altitude logistics network and an intelligent logistics system that integrates various types of logistics devices, such as drones, high/low-speed unmanned vehicles, and robots. After receiving a delivery task, a drone will take off from the warehousing center carrying the requested goods. Based on GPS and intelligent vision, it will fly along the routes planned by the intelligent scheduling system and place the goods in the designated area. Next, intelligent devices working on the ground, such as high/low-speed vehicles and robots, will respond swiftly. They will continue the delivery task and quickly deliver the goods to the destination, such as a designated floor in an office building, a restaurant, or a warehouse in the production workshop. The entire precise and collaborative process links the ground with low-altitude operations. It can meet logistics and delivery requirements in different scenarios, including office, public, and production areas, speeding up delivery and supporting efficient collaboration.

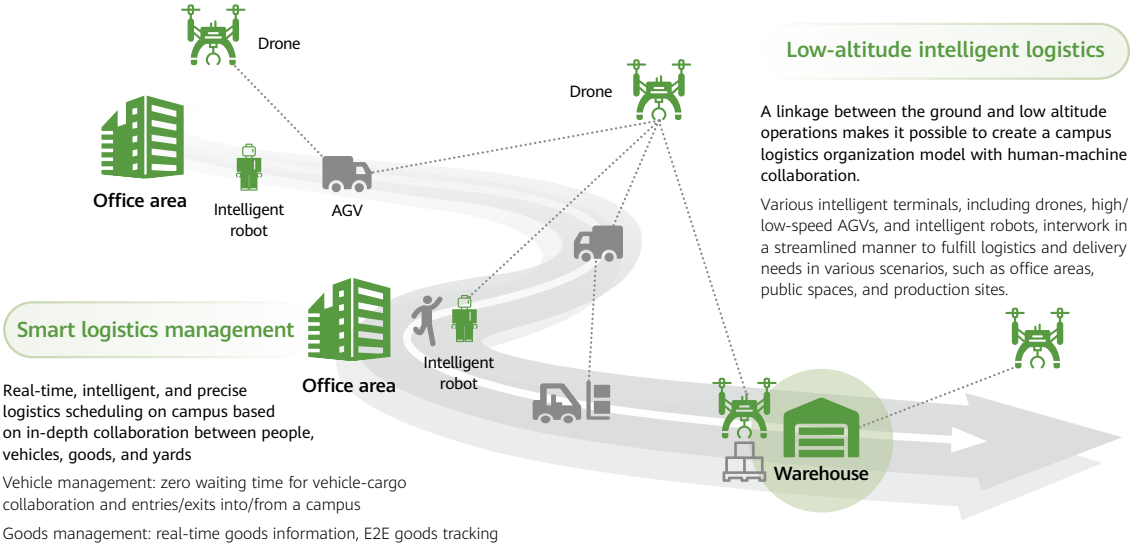


Figure 2-5 Possible logistics scenarios with multi-element linkage

2.6 Scenario 6: 10 Gbps Highly Reliable Production

2.6.1 Definition

Powered by industrial intelligent computing edge, deterministic production networks will reshape industrial control and the sensing system, ensuring stable production and accelerating innovation.

2.6.2 Description

Under a new round of technological revolution, edge intelligence, industrial Internet, AI, and other booming technologies are reshaping the way we produce and manufacture things, driving production and manufacturing to be fully intelligent. Intelligent campuses gather a large number of production enterprises, playing a vital role in the in-depth integration of the digital economy and the real economy, and are at the forefront of promoting the development of high-quality intelligent manufacturing. Campuses continuously evolve their digital technology application and platform capabilities to enable the traditional production and manufacturing field to create new productive forces. With the popularization of 5G-A/Net5.5G/F5G-A, future campus production will be based on 10-gigabit deterministic networks and gain support from industrial intelligent computing infrastructure. This will reshape industrial control and the sensing system, leading to stable production and faster innovation in R&D.

Production devices are connected through an intelligent deterministic production network

that features lossless roaming. Service data and instruction data are transmitted to production devices through the network, enhancing production stability and reliability. Deterministic networks support high-precision remote control in production, with real-time data collection from the production site and instant delivery of production control instructions to achieve precise control over production devices such as intelligent robots. This significantly improves human-machine collaboration efficiency and ensures stable production. Additionally, the intelligent deterministic network with lossless roaming supports automated guided vehicles (AGVs) to transport goods without human intervention. With AGVs, 80% of raw material, supply chain, and vehicle fleet management workflows can be automated, vastly improving productivity and reliability.

Industrial intelligent computing infrastructures such as edge computing centers and intelligent computing centers will reshape the production and manufacturing system, contributing to high-quality manufacturing and innovative R&D. Powered by edge computing, industrial production will witness higher quality and efficiency. Automatic optic inspection (AOI) based on edge computing technologies can implement real-time and automatic detection at any location in the production line. Through computer vision and deep learning, AOI can automatically perceive and accurately identify production defects to constantly improve quality control, boosting productivity and providing insights into operations. Simultaneously, intelligent computing centers for production will speed up innovation in R&D. Relying on strong computing power, AI foundation models, and intelligent algorithms, high-quality and flexible intelligent computing centers can deeply analyze massive amounts of user and product data. This supports trial production and virtual simulation tests to improve R&D efficiency.

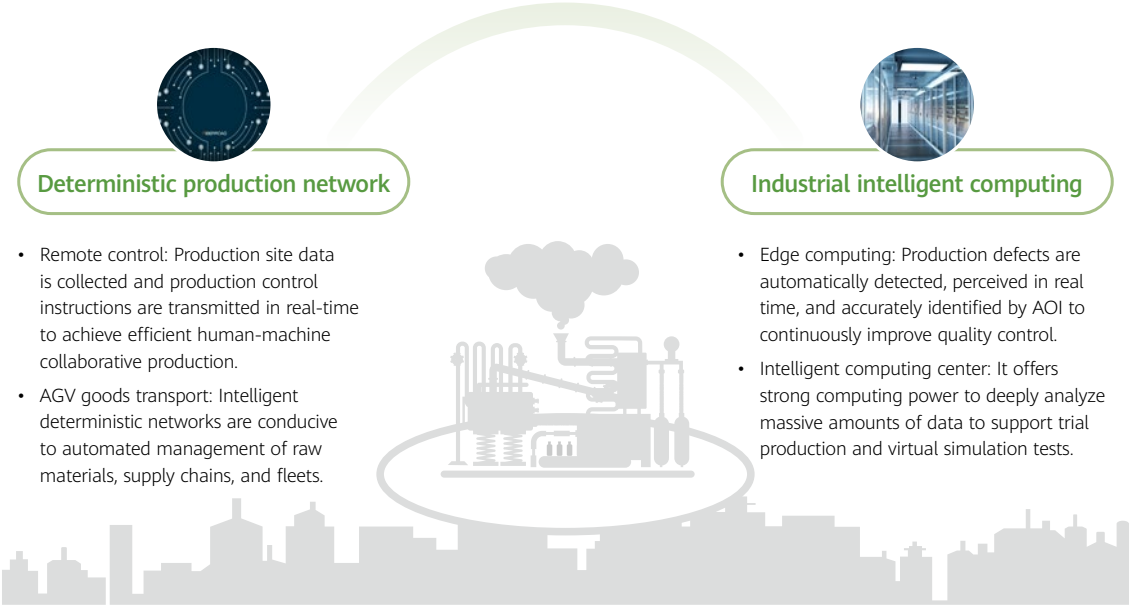


Figure 2-6 10 Gbps highly reliable production scenarios

2.7 Scenario 7: Digital Health Services

2.7.1 Definition

Quality health-related resources will be utilized to provide holographic and virtual health services and intelligent first-aid services. This will enable campuses to deliver a proactive and targeted health management experience.

2.7.2 Description

As people become more health literate and aware, campus health services are becoming more personalized, convenient, and accurate. Advancements in digital health technologies such as 5G-A, Net5.5G, and F5G-A, wearable devices, and AI will significantly improve accessibility and diversity of campus health services. Intelligent campuses will be able to offer full-lifecycle health management services featuring accurate sensing, prediction, and intervention tailored for individual users.

10GE high-speed networks will support campuses in close association with quality health resources. Therefore, they can provide holographic and virtual health services, bringing users attentive and desirable health management experiences. Technologies like digital twins, VR, holographic projection, and generative AI running on 10GE high-speed networks will allow campuses to seamlessly integrate high-quality health resources into their management systems. Health experts' digital twins will appear to users as if the experts are physically present, offering a range of health management services such as psychological counseling, chronic disease consultation, and fitness guidance. This provides campus users with a seamless, engaging, and personalized online consultation and interaction experience.

The campus health platform will be associated with intelligent warning devices to offer first-aid services, facilitating campus health risk detection and emergency rescue. 10GE networks are expected to be widely used on campuses, with high bandwidth, high-density coverage, and ultra-low latency. Intelligent warning devices that utilize big data, machine learning, and other technologies will monitor the status of campus users in real time and connect to the campus digital platform when they detect any exceptions. This will facilitate resource scheduling, second-level emergency response, and emergency rescue.

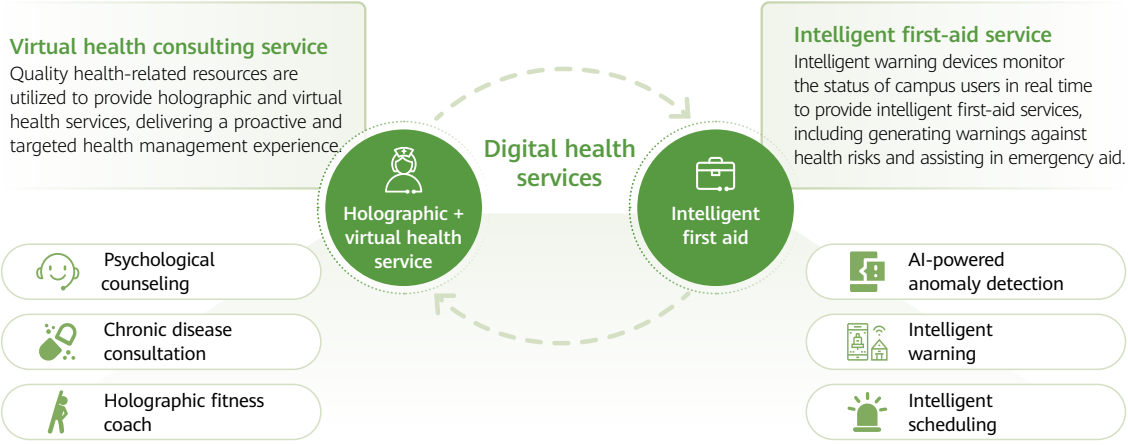


Figure 2-7 Digital health service scenarios

2.8 Scenario 8: Ultra-Immersive Interaction

2.8.1 Definition

Fueled by 10GE ultra-high-speed networks, future campuses will be able to break the boundaries between dimensions through spatial interaction, providing an immersive interactive experience across time, space, and media.

2.8.2 Description

By integrating AI, IoT, XR, and naked-eye 3D technologies, immersive experiences on campuses will be extended to more fields. With the help of 10GE networks, this will foster the creation of innovative and highly interactive applications that offer ultra-immersive experiences. AI will integrate with technologies like digital twins and XR to provide ultra-immersive experiences across time, space, and media. Venues, exhibition halls, and public entertainment spaces will play a crucial role in the development of interactive experience spaces with virtual-real integration and symbiosis.

Spatial interaction will break the boundaries between dimensions. Virtual-real interaction and human-machine interaction will spawn ultra-immersive interactive services and bring campus users a brand-new virtual-real integration experience. Applications like virtual digital humans and special holographic effects will help build activity sites that support virtual-real interaction. Virtual digital hosts created through target image and motion capture, computer graphics (CG) modeling, generative AI, and natural language processing (NLP) will be able to perform manual tasks like activity introduction, on-site interaction, and sign language generation. Holographic

projection, near-eye display, VR interaction, and immersive space projection will be in conjunction with on-site stage design and environments to create realistic simulations. On top of this, campus users can obtain an immersive service experience as their interaction with virtual roles will trigger real-time responses given by offline intelligent robots. Virtual roles, which are created using human-computer interaction technology, can detect users' behaviors and postures, track their locations, and identify user-given service instructions. Online and offline linkage activates offline service robots to respond in real time, accurately meeting users' needs while also providing them with a new and immersive human-machine interaction experience.

With digital twins, an ultra-immersive entertainment space with virtual-real integration can be created to provide immersive exhibition, gaming, and other entertainment services, leading to a futuristic experience for campus users. Immersive exhibitions are grounded in AR, naked-eye 3D, and interactive sensing technologies, which help create an immersive exhibition space for digital arts and cultural tourism. Immersive audio and videos are seamlessly integrated with exhibition halls on campuses, creating a magnificent and immersive experience for visitors, with realistic visual and thrilling sound effects that maximize sensory stimulation. When it comes to immersive games, digital modeling of gamers replicates their postures and behaviors in the virtual world. Therefore, they can interact with both offline and virtual roles, breaking the limitations of traditional immersive experiences that only allow for in-place interaction, ultimately maximizing the immersive gaming experience.

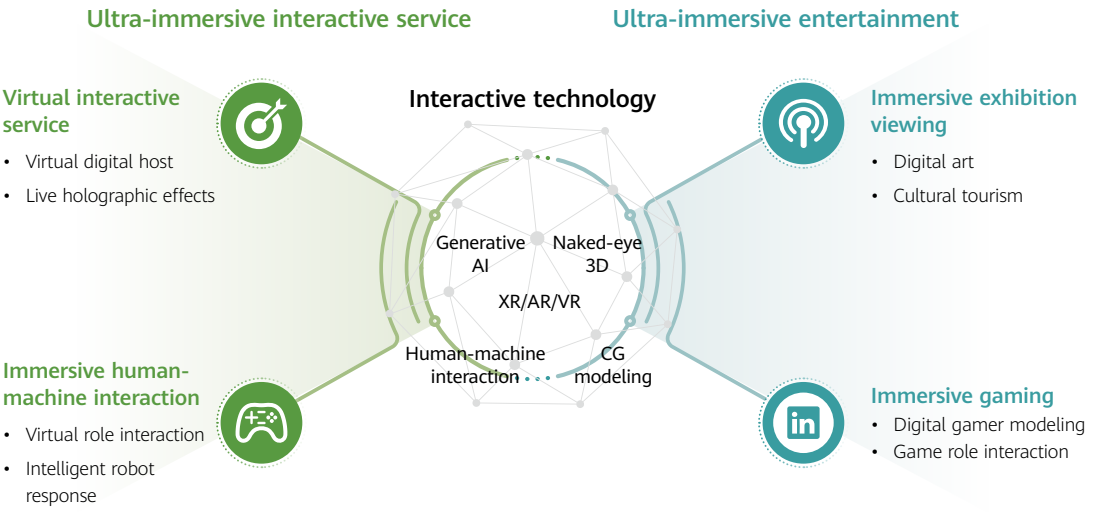


Figure 2-8 Ultra-immersive interaction scenarios

2.9 Scenario 9: Metaverse Life

2.9.1 Definition

By utilizing 10GE high-speed networks and intelligent twins, we can construct an incredibly immersive virtual world that offers unique and engaging entertainment experiences.

2.9.2 Description

Core metaverse technologies, such as digital twins, 5G-A/6G, AI visual and holographic interaction, XR and sensing, and emotional interaction are maturing. This will enable campus activities to expand beyond work and production to culture, entertainment, and consumption. Metaverse applications will gradually permeate the business ecosystems of campuses. Metaverse technology will drive the collaboration between the real and virtual world, creating new ecosystems for consumption on campuses. This will provide the next generation of campus consumers with immersive, interactive, and experiential services.

The metaverse in the field of consumption aims to create an immersive offline space that blends virtual and physical elements. This space will offer real-life entertainment experiences, including AR versions of street views, viral sightseeing spots, and navigation, to cater to the preferences of young consumers. The ultimate goal is to provide a futuristic digital consumption experience. Anime and comics, puzzle games, social media, and so on put forward diversified consumer demands. Campuses will create immersive spaces that integrate cultural and

social exchange, entertainment, and shopping. A range of real-life entertainment consumption experiences will make leisure time in the metaverse more interactive and engaging. AR street views connect virtual human performances with campus buildings, presenting real-time images of virtual humans performing under various beautiful lights. Users can also scan the code using their mobile devices to enjoy fun experiences like AR roaming, AR art exhibitions, AR light shows, and AR store information. AR viral sightseeing spots combine popular consumption activities and store promotions into plot-based games to engage and attract young users. This increases the frequency of interaction between customers and malls, attracts more customers, and enhances the shopping experience on top of the loyalty of consumers. AR navigation is expected to gradually achieve full coverage across indoor and outdoor areas. AR provides street navigation, indoor shopping guides, and precise vehicle guidance services, and automatically recommends corresponding shops and services according to consumer preferences. In this way, consumers can easily find physical stores, which promotes sales conversion.

The metaverse will provide an immersive online entertainment experience. Within this virtual space, consumers can enjoy unique shopping experiences, including AR shopping, try-on, and interaction. This can effectively attract consumers to shopping malls and brands, as well as promote new products. On the metaverse virtual consumption platform, consumers can create their own virtual digital identities, immerse themselves in virtual consumption experiences, and communicate and interact within the digital space. Before consumers arrive at the mall, the metaverse virtual space can recreate product usage scenarios based on consumer preferences and environmental parameters. This allows users to get a feel for the product in the virtual world. Upon arrival at the mall, consumers can try on clothes and makeup, and get professional advice from the brand's stylist in the virtual space. Consumers can connect with friends from afar, create digital identities, shop together in the virtual space, and share their real experiences trying on new products. This can boost consumption power and improve campus business.

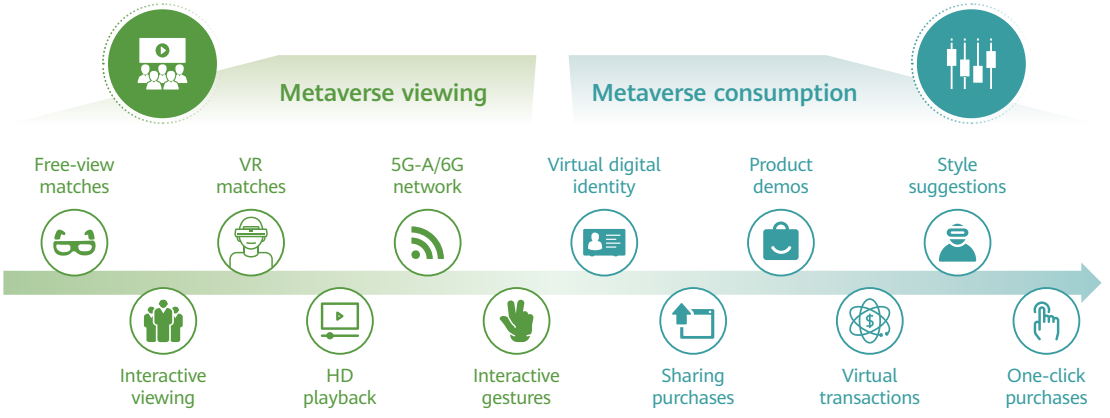


Figure 2-9 Metaverse life scenarios

2.10 Scenario 10: Smart Energy Management

2.10.1 Definition

By integrating digital and energy technologies, campuses can achieve full-lifecycle zero-carbon management, optimize energy operations, and make accurate decisions, ultimately leading to the development of zero-carbon intelligent campuses.

2.10.2 Description

Zero-carbon intelligent buildings with lower energy consumption and lower carbon emissions have become the norm for campuses. To create zero-carbon buildings, advanced technologies like intelligent algorithms, multimodal learning, and big data prediction need to be used. This will enable intelligent and eco-friendly management, which encompasses intelligent perception, risk warning, and energy consumption optimization. With these technologies, carbon emissions are reduced and managed throughout the lifecycle of campuses. Campuses will facilitate the achievement of carbon neutrality and carbon peaking. An integrated energy system for campuses features generation-grid-load-storage integration, multi-energy complementation, and dynamic balancing between supply and demand. Combining AI, intelligent algorithms, big data prediction, generation-grid-load-storage coordination, and green power supply, a reliable AI-based energy prediction system will optimize the supply and procurement decisions of clean energy to ensure the quality of campus energy systems while helping achieve a balance between low carbon use and economic efficiency.

In terms of full-lifecycle zero-carbon management, the energy supply and usage efficiency of campus buildings will be improved through digital technology-enabled design and renewable energy deployment. Digital technology will enable design features such as natural lighting, heat insulation, ground source heat pumps, and the use of renewable energy such as hydro, wind, photoelectric, nuclear, biomass energy, and solar, so as to achieve 100% renewable energy supply. For example, users can predict future energy supply and consumption trends, as well as calculate real-time proportions of new energy utilization and carbon dioxide emission reduction, based on the overall area and equipment used for renewable energy deployment on campuses. Additionally, zero-carbon intelligent buildings utilize advanced green and low-carbon technologies to reduce carbon emissions and enhance the operational efficiency of campus buildings. Merging BIM and AI generative design technology helps achieve refined management of carbon emissions during construction and building use. For instance, we can use environmental data to make energy-saving and design recommendations, such as building orientation and material selection. Energy monitoring centers and carbon big data platforms can instantly detect indoor temperature, humidity, illumination, carbon dioxide, and outdoor wind and rainwater. They can also intelligently manage and coordinate the usage of electricity, water, heating, cooling, and renewable energy within and between buildings. This optimizes energy consumption and makes it controllable through visualized data. To achieve net zero emissions on campuses, carbon capture, absorption, and trading can be used to offset the carbon emissions of campuses.

An AI-based energy management system accurately predicts campus energy load through coupling of different types of energy systems

and multi-dimensional data analysis, resulting in refined energy operations and management. AI-based energy prediction systems will use AI model training to accurately predict energy demand. The systems will consider factors like campus area, population, facilities, energy consumption habits, and consumption history. At the same time, in line with relevant laws and regulations, commodity market trading dynamics, and the unstable supply of clean energy such as PV and wind power, they will automatically develop energy purchase and supply strategies and assist campus management in energy procurement decision-making. A reliable AI-based energy prediction system will maximize the use of clean energy to replace traditional energy sources in the energy supply structure while ensuring a stable campus energy supply. This enables the construction of sustainable low-carbon campuses. The integrated energy system, based on AI technology, can automatically optimize energy storage and usage on campuses by analyzing energy demand and supply. This can lead to a significant reduction in carbon footprints. To reduce energy loss and improve the resilience and flexibility of the energy supply system, the reactive power compensation system

driven by data and various energy, heat, and cooling systems will be further optimized. For example, AI algorithms enable smart microgrids on campuses. They help efficiently utilize clean energy and solve the problem of unstable distributed energy, thus balancing loads for comprehensive power grids. AI algorithms can also provide energy system operation optimization policies during energy use. Closed-loop control of terminal devices can be used to reduce the energy consumption of campus enterprises, improving energy system security and cost-effectiveness. For example, peak shaving and energy consumption optimization strategies can be tailored to the characteristics of enterprise power loads, which can reduce the demand for power grid and electricity expenditure.

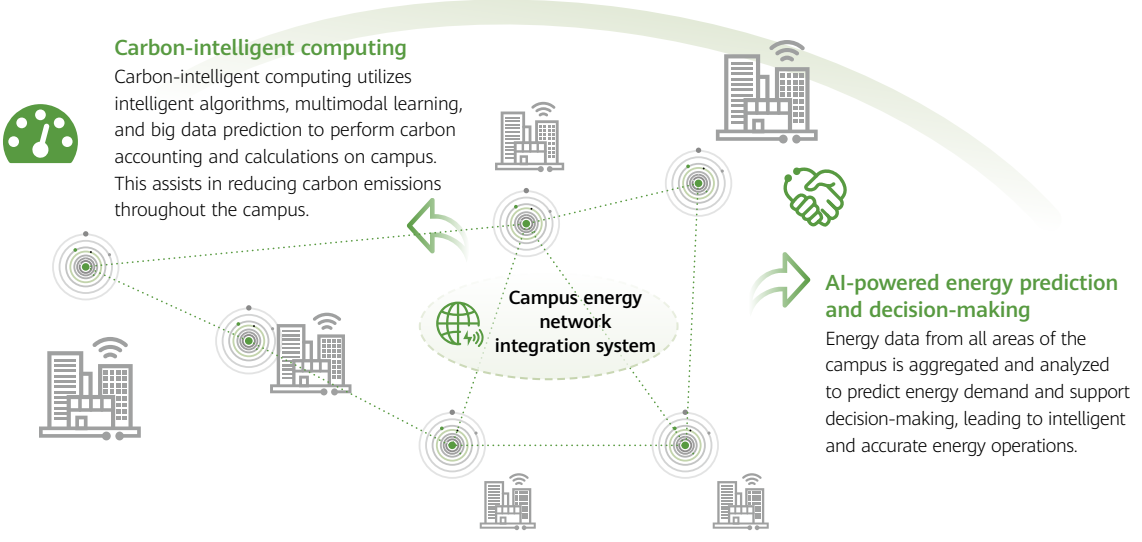


Figure 2-10 Smart energy management scenarios

Key Technical Features and Reference Architecture

03

3.1 Key Technical Features

In the future, an intelligent campus will be a self-evolving system rather than a stacking of disconnected ones. These novel intelligent campuses use ICTs to reconstruct the multi-dimensional data of people, devices, things, and space within the physical campus.

Huawei believes that future intelligent campuses will boast six technical features: intelligent twins, spatial interaction, ubiquitous intelligent connectivity, intent-driven ultra-broadband, security and resilience, and all-domain zero carbon.

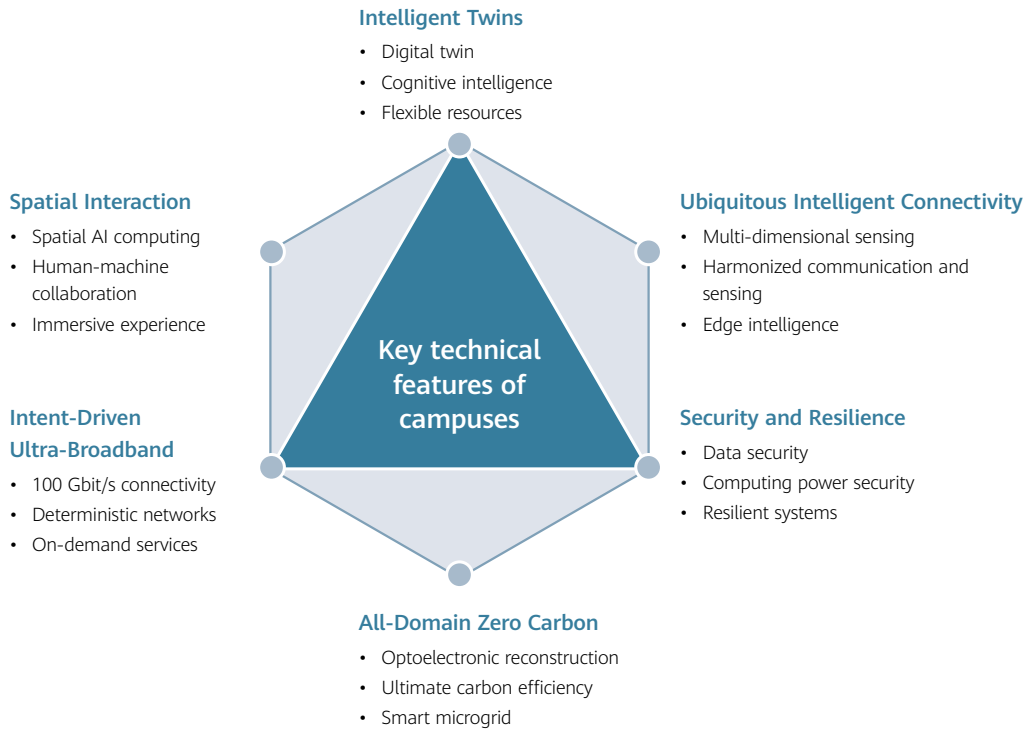


Figure 3-1 Key technical features

3.1.1 Intelligent Twins

The future intelligent campus is a highly digital space. Using a centralized intelligent platform, a digital twin can be established between physical and digital spaces. It allows us to simulate the real world and predict the near future with high precision. The intelligent platform for future campuses needs to be able to flexibly process massive amounts of data as well as schedule ever growing resources. The platform should employ an array of AI technologies that are automated, autonomous, and generative, and make use of knowledge computing. In doing so, it can gain an intelligent cognition of the physical space it simulates. It must also provide flexible compute, storage, and transport resources.

Digital Twin

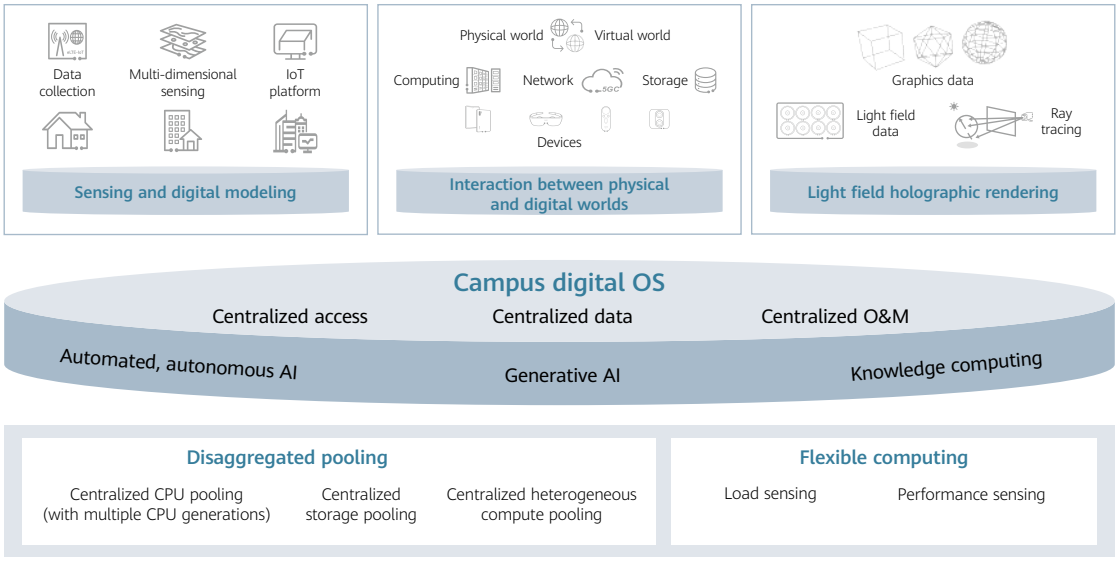


Figure 3-2 Digital twin

The digital twin of an intelligent campus creates virtual models of physical entities to perform data modeling, improve the digital space, and develop related business services. It relies on multi-dimensional sensing, digital modeling, interaction between the physical and digital worlds, light field holographic rendering, intelligent platforms, and more to map the physical space to the digital space.

1) Multi-dimensional sensing and digital modeling

- Multi-dimensional sensing: Massive amounts of data in the physical world of the campus is collected and stored, including that of videos, access control systems, facilities and devices, environments, and conferences. Multi-dimensional data processing and convergence require high-resolution sensing, object location, imaging, and environment reconstruction. The amount of data generated during this process is even larger. Screening, preprocessing, modeling, and simulation of such massive amounts of data rely on powerful computing and integration across multiple disciplines, such as AI, cognitive science, and control science.
- Digital modeling will require 100 times more computing power. Managing such an enormous amount of multi-dimensional data and transforming it into a 3D model is a big challenge. 3D modeling requires huge computing power as it is based on images and video streams from different angles, data collected by array cameras and depth cameras, and multi-dimensional campus sensing.

2) Interaction between the physical and digital worlds

Device-cloud synergy allows data to be processed and transmitted in real time, supporting the collaboration, linkage, and synchronization between the digital and physical worlds. A large amount of state queries and message transmission require that people and things interact with each other at latencies of less than 5–10 milliseconds, the bandwidth reach hundreds of Mbit/s per user, and the required computing power tens of TFLOPS.

3) Light field holographic rendering

Future campuses will adopt holographic rendering to build a digital twin display system

which provides users with the same experience as they have in the physical world under balanced 3D lights. AI content generation requires massive AI computing power to precisely map the high-concurrency rendering needed. It is estimated that after the light field holographic rendering technology is implemented in 2030, the demand for AI computing power will increase by 64 times, and one user will require more than 10 TFLOPS of computing power.

4) Intelligent platform

As the digital OS for future campuses, the campus digital platform provides the following:

- Connectivity services: A wide variety of intelligent and dumb terminals on campuses can seamlessly connect to the network. Together with tail access and platform intelligence, they support campus business services.
- Data services: Centralized data services include access, integration, and sharing of multiple data sources, unified workflow, consistent user experience, multi-device interconnection, multi-application collaboration, and multi-data convergence.
- Centralized O&M: Centralized platform O&M for multiple tenants, services, and applications, multi-service innovations, multi-application development, as well as multi-operations collaboration and security.

Cognitive Intelligence

On future campuses, a multitude of data will be aggregated on an intelligent platform that carries the digital twin. The daily traffic volume will be accurate to each building, floor, room, or even location. The volume of the data will be 100 times what we are used to today. Cognitive intelligence, under the scope of automated, autonomous AI, knowledge computing, and generative AI, is essential to supporting campus digitalization. Huawei predicts that AI will be present within all services of newly built campuses by 2030.

1) Automated, autonomous AI

The broad application of license plate recognition (LPR), target recognition, AR/VR, and other novel technologies in campus service scenarios like security, attendance, and conferences shows that AI has gradually become essential to operations of intelligent campuses. Breakthroughs have been made in the following key technologies, helping AI evolve from supervised learning to an automated, autonomous mode.

- Self-supervision and self-feedback: Training signals can be incorporated online in a self-supervised fashion, so that feedback is available during inference, not just during the training phase.
- Online continuous learning: At present, a model's learned representations are formed without constraints. The representations that result from different training sessions may be radically different even if they are of the same model structure. Models need to overcome catastrophic forgetting so that learning can be carried out continuously, and training and inference can converge into a single process.
- Multi-task natural interaction: Models manually designed for different tasks need to be replaced by models that can learn to encode for different tasks and switch between different modalities in context and on demand.

With automated, autonomous technologies, AI can employ new methods — such as transfer learning, few-shot/zero-shot learning, self-

supervised/weakly-supervised/semi-supervised/unsupervised learning, and active learning — to overcome our dependence on manual training, design, and iteration of models. This will eventually allow AI to reach autonomy. AI autonomy will make models more homogenized, with the same model serving multiple purposes. The scaling of data and the prevalence of online learning will lead to more centralized model production. Industry applications in multiple domains will converge to a handful of or even a single ultra-large model. Models will learn to pick up and train on new data as they operate. Increasingly simplified deep learning will gradually negate the need for manual intervention. Therefore, AI will gradually take over basic campus services such as security, access control, fault diagnosis, and facility energy saving as well as lead other services like handling, operation, events, and authentication. Bringing intelligence to infrastructure will take a huge burden off the workforce and lead to high-quality services.

2) Knowledge computing

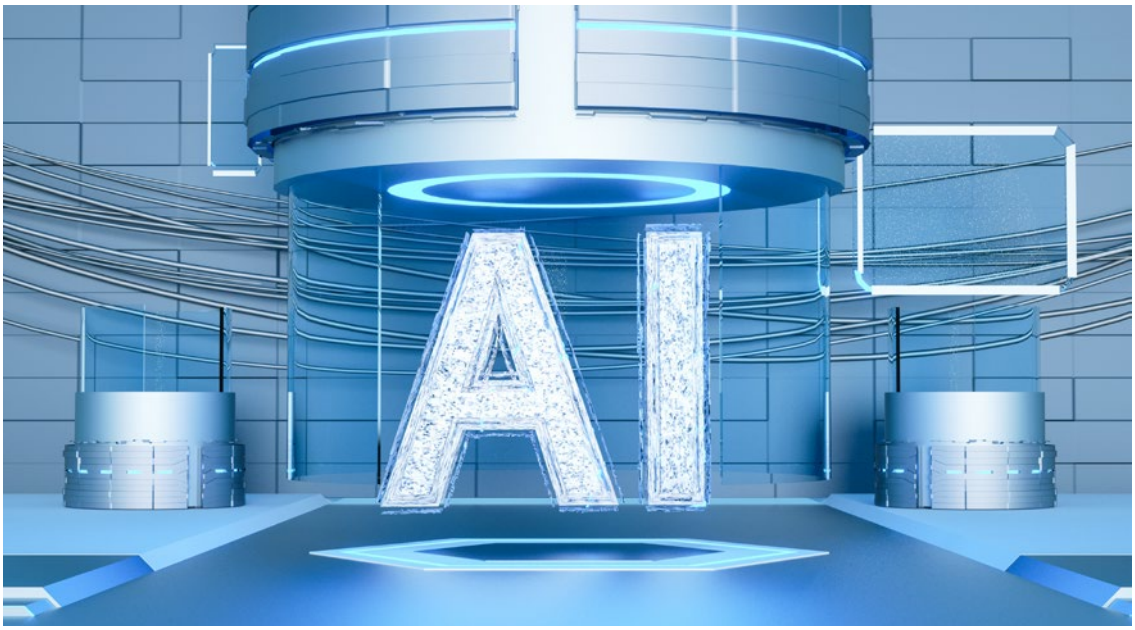
The industrial application of AI requires the ability to make high-quality decisions based on expert domain knowledge across multiple disciplines. A complete technical system is needed for knowledge extraction, modeling, management, and application.

Knowledge computing is used for both cognition and perception. It will require breakthroughs in algorithms regarding the massive retrieval of sparse information, capture of dynamic-length knowledge, knowledge attention, and large-scale graph computing. The training schema for cognitive intelligence will require advances in high-frequency knowledge retrieval during training and inference as well as feature enhancement based on knowledge combination. In terms of computing, it will be necessary to solve a number of problems regarding the training and inference for high-frequency random retrieval, high-speed data communication, and some graph computing puzzles such as random walk and structural sampling.

Knowledge computing will have the following high-order cognition features:

- Knowledge extraction: The data source will not only include text and structured features, but also complex and multi-level knowledge. This includes areas of research like multi-modal knowledge alignment, extraction and fusion, complex-task knowledge extraction, and cross-domain knowledge extraction.
- Knowledge modeling: We will move on from developing scenario-specific, atomized, automated, and large-scale knowledge graphs to integrating these scenario-specific graphs into general knowledge graphs.
- Knowledge application: It will go from simple query and predictions to high-order cognitive tasks such as causal reasoning, long-distance reasoning, and knowledge transfer.

Next-generation AI knowledge computing will gradually develop high-order cognition, which will enable applications across a wide variety of industry campuses. It is predicted that humans will be replaced in some scenarios and decision-making/service robots that can perform logical inference will be available in 2030.



3) Generative AI

Generative AI powers automated content production. It allows computers to abstract the underlying patterns related to a certain input (such as text, audio files, and images) and use it to generate expected content. Generative AI is used in identity protection and audio synthesis, among other fields.

Generative AI creates data that is similar to training data, rather than simply replicating it, so it can incorporate human creativity into processes of design and creation. In the development of generative AI applications, the key objective is to create generation models that are capable of evolving and dynamically improving over time. The field of generative AI is facing the following challenges:

- Some generative models (such as generative adversarial networks, or GANs) are unstable, and it is difficult to control their behavior. For example, generated images may not be sufficiently accurate; they may not produce the desired output; and the cause cannot be located.
- Current generative AI algorithms still require a large amount of training data and cannot create new things. To address this, algorithms capable of self-updating and evolving are needed.

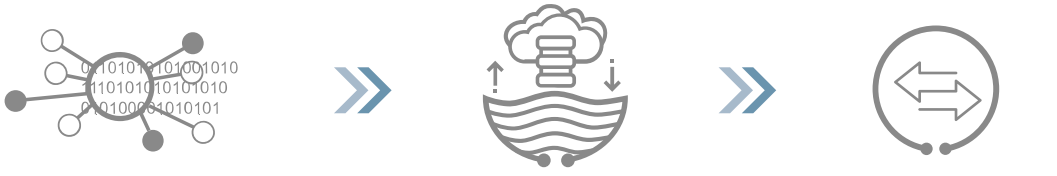


Figure 3-3 Generative AI

- Malicious actors can use generative AI for spoofing identities and can exploit vulnerabilities in AI tools to conduct remote attacks, resulting in serious threats to online information security. This includes the likes of data breaches, model tampering, and spam.

Generative AI will gradually have the following characteristics:

- Model self-evolution: Model self-update and self-evolution resolve specific issues.
- AI from cognition to creation: AI can incorporate human creativity into design and creation, including when it comes to artistic creation, auxiliary artistic creation, and auxiliary content generation.
- AI vision and holographic interaction: Future chips support generative AI engines to create new content in the virtual world and provide an immersive experience.

Generative AI alongside holographic rendering conferences and transcendent smart office will significantly improve the management and services of future campuses. Currently, generative AI requires a large amount of training data as inputs. It is estimated that in 2030, AI will be able to proactively search for training data to form a novel self-evolving generative AI model.

Flexible Resources

Resources like public, industry, and private clouds are widely used as digital and intelligent platforms for campuses. Large-granularity applications such as AI foundation models, the metaverse, and digital twins are seeing explosive growth. This means that the cloud architecture will likely become a de facto standard for future intelligent campuses. It can provide large-scale, intensive, and scalable compute, storage, and transport resources on demand, all while ensuring the multi-tenancy security and performance SLAs for diversified campus service applications that public and private sectors require.

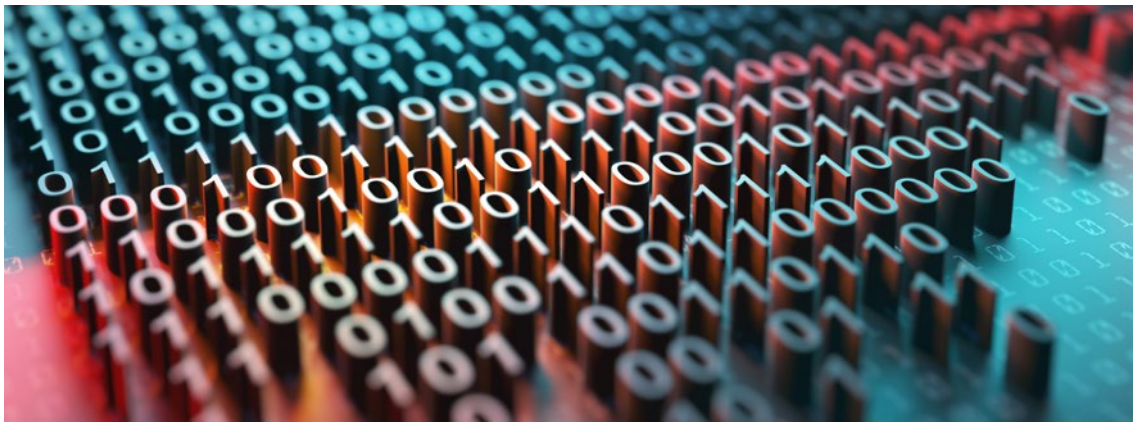
Data foundations for digital OSs of future intelligent campuses will continue moving toward disaggregated pooling and flexible computing to establish intelligent twins.

1) Disaggregated pooling

Resource pooling is an essential part of the campus OS data foundation. It allows multiple tenants and applications to share physical resources like compute, storage, and transport to the greatest extent possible. Over the next 5 to 10 years, disaggregated pooling will become increasingly common in data foundations for campus digital OSs. Specifically, CPUs of different generations, storage (decoupled from compute), and heterogeneous computing power will be centrally pooled.

- Centralized CPU pooling (with multiple CPU generations): Computing power that is used in a future campus digital OS data foundation will be provided using an application-centric model rather than a resource-centric model. This will resolve certain CPU hardware differences for upper-layer compute services and resource scheduling layers. Additionally, it will leverage black-box real-time QoS detection to identify application QoS requirements and dynamically schedule CPU resources. This will achieve the optimal dynamic CPU reuse while also meeting cloud tenant SLA requirements on application performance.

- Centralized heterogeneous storage and cache pooling: Unstructured data storage like block, object, and file storage can use decentralized cross-AZ distributed storage engines to create centralized storage resource pools. However, semi-structured and structured data storage still integrates storage and compute to provide compute-side functions (data query, change, analysis, and processing) and storage-side functions (data persistence, availability assurance, parallel I/O read/write, and lossless elastic capacity management). Data foundations for digital OSs of future campuses will have to take multiple measures — such as data copy sharing across data compute engines, near-compute cache pooling, distributed calculus offloading for near-data processing, centralized metadata management for heterogeneous compute engines, and intelligent tiered data storage — to create centralized storage resource pools for structured, semi-structured, and unstructured databases.
- Centralized heterogeneous compute pooling: With the advent of AI foundation models, the metaverse, and digital twins, cloud-based GPU/NPU heterogeneous computing power will gradually replace general-purpose CPUs. It will become the number one computing power for AI foundation model training and inference, as well as rendering and simulation of digital humans and digital twin campuses. With software-defined GPU or NPU pooling, a physical GPU or ASIC acceleration chip can be divided into several or even dozens of isolated compute units. GPU or NPU chips on different physical servers can be aggregated for a digital OS data foundation (on a physical machine or a VM) or a container to complete distributed tasks. CPU servers without GPU or NPU acceleration chips can also invoke GPU or ASIC acceleration cards on remote servers to complete AI computational tasks. CPUs can be decoupled from GPUs and heterogeneous computing power can be pooled to provide more scalable GPU and NPU resources.



2) Flexible computing

To dynamically adapt to the changing requirements of applications for compute resources, data foundations for digital OSs of future campuses will introduce flexible computing, a more flexible and intelligent way to allocate and provision computing power. This can greatly improve the utilization of compute resource pools and allow cloud tenants and developers to use dynamic computing power like other utilities do, such as water and electricity. Cloud tenants and developers will then only have to pay for what they use and nothing more.

Flexible computing also differs from elastic computing in that it not only perceives the dynamic resource requirements of workloads, but also quantifies their QoS requirements. Flexible computing can collect multi-dimensional performance metrics across all resource instances from the underlying host OS in a non-intrusive black-box manner. This includes performance metrics for CPU, memory, storage I/O, network I/O, and more. Flexible computing can extract workload performance characteristics (such as if the workloads are CPU-intensive, memory-intensive, storage-intensive, network-intensive, or a combination) from countless services collected. This is accrued with the help of AI models that are continuously iterated and optimized online. A small number of typical workload training samples can then be added in supervised learning tasks to enable flexible computing that can be observed in black-box mode.

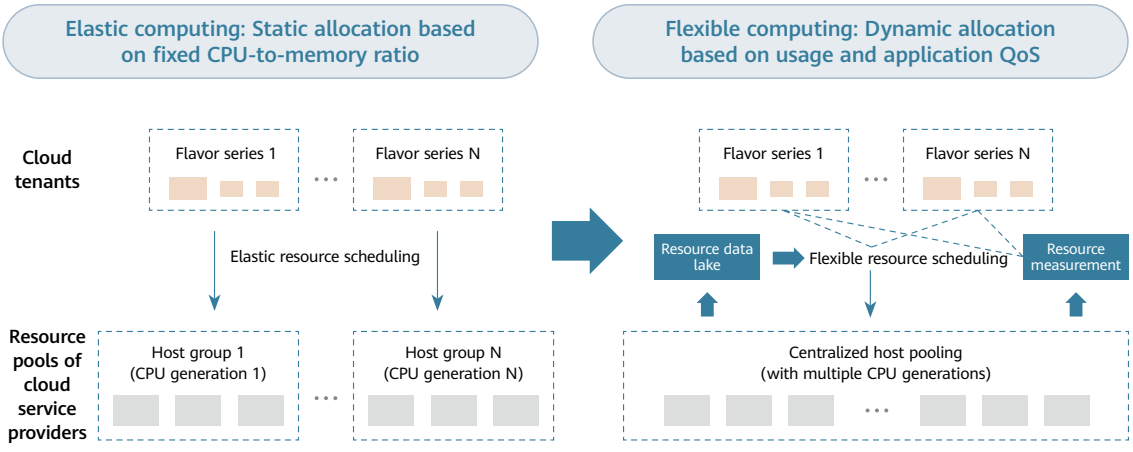


Figure 3-4 Flexible computing

3.1.2 Spatial Interaction

Human-machine interaction is expected to renovate campus office, increase production, and improve people's everyday experiences. In the future, spatial interaction will mainly be of two kinds: virtual-reality and human-machine. Future intelligent campuses will provide smart experiences, including the likes of naked-eye 3D display and XR, which are more intelligent, personalized, and immersive. In addition, human-machine interaction on the campus will shift its focus from machines to people. Facilities will be highly interconnected and automated. Machines will transform from mere tools to partners that can understand, learn, and make decisions. All this will take us to a new phase of intelligent human-machine collaboration. Immersive experience and intelligent human-machine collaboration require low latency, ubiquitous access, and powerful intelligence. These factors drive the development of human-machine interaction networks and spatial AI computing.

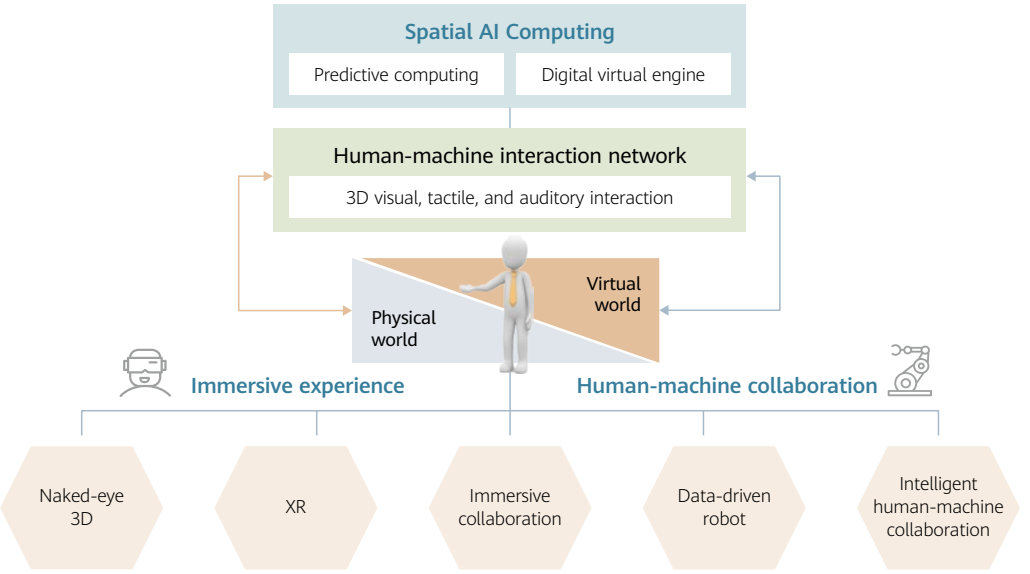


Figure 3-5 Spatial interaction

Human-Machine Collaboration

Human-machine collaboration employs AI and machine learning to enable collaboration between people and machines, as well as between people and facilities. On a campus, human-machine collaboration processes and analyzes data in real time through computing with cloud-edge-device synergy, thus increasing data processing efficiency. Human-machine collaboration also places higher requirements on networks, data-driven robots, and intelligence.

1) Human-machine collaboration network

- Campus time-sensitive networking (TSN): Traditional robots use hard programmable logic controllers (PLC) and human-machine interfaces (HMI) on each access side to automatically control robots and devices. To process services in a distributed manner, each access side needs specific hard PLCs. However, this approach makes management challenging, O&M complex, and expansion inflexible. The TSN technology is set to become

prevalent in various scenarios, including smart factories, campus devices, and intelligent service robots on campuses. The latency and jitter of asynchronous TSN are predicted to be as low as 10 µs in 2030. Campus TSN provides a protocol suite developed by IEEE 802.1 to enable deterministic minimum delay on the non-deterministic Ethernet. TSN provides a set of universal time-sensitive mechanisms for the data link layer of the Ethernet protocols. It also guarantees real-time, deterministic, and reliable data transmission on the standard Ethernet, thereby improving data transmission efficiency.

- Tactile Internet: In future industrial networks and IoT, the tactile Internet has the highest requirements on networks for bidirectional closed-loop management from perception to execution. Typical scenarios of the tactile network include self-driving vehicles, industrial automation, telemedicine, and VR/AR. Compared with sight and hearing networks, the biggest changes in tactile networks lie in ultra-responsive and ultra-reliable connectivity. The E2E latency for ultra-responsive connectivity is 1 ms. This means that the network latency must remain at the 100-µs level, with 99.99999% availability.

2) Data-driven robot

Robots will be intelligent individuals on a campus, which will use the accumulated data to create a closed-loop data flow. This will continuously drive robot optimization, improve campus intelligence, and ultimately, enable collaborative management and independent coordination for thousands of intelligent robots.

Data-driven robots will mainly include the following technical features:

- Intelligent data foundation: Obtains data (such as images, sounds, point clouds, maps, and motions) from the environment, helps migrate spatio-temporal data to the cloud, and builds an intelligent data foundation for intelligent campuses. The main technical points include raw data, mapping algorithms, and data service platforms (data and algorithm management). For example, environmental sensing–based mapping supports real-time map creation for scenarios across more than 1 million square meters, helping build digital campus scenarios.
- Robot simulation: Significantly quickens the simulation and R&D of robots on intelligent campuses alongside the generation of

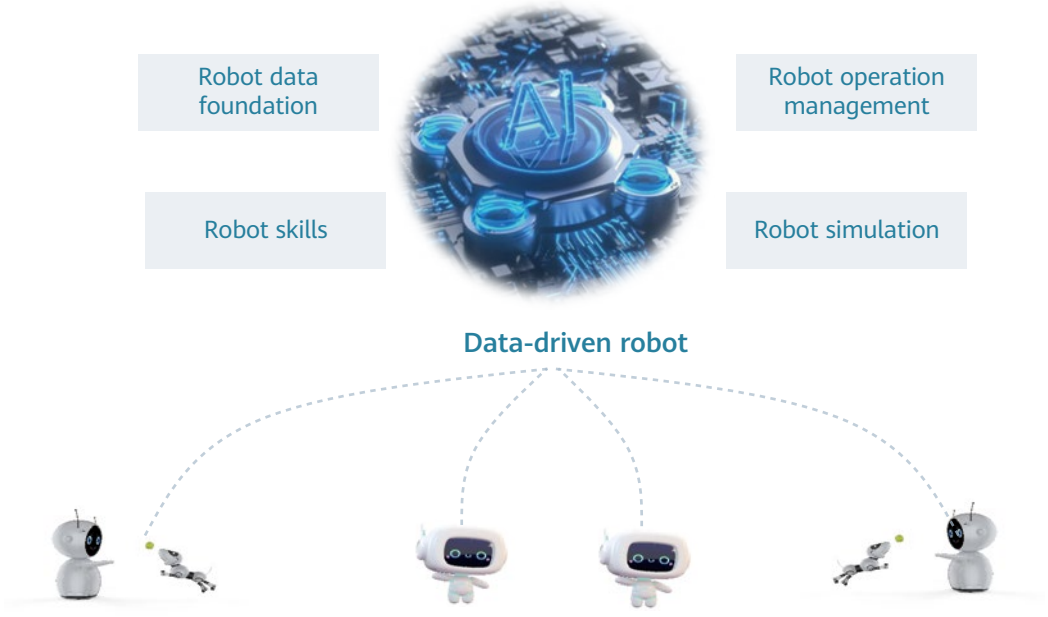


Figure 3-6 Data-driven robot

different scenarios. Improves the applicability of robots. The main technical points include 3D objects, worlds, and robot modeling in the simulation environment, cloud native robot simulator and its physical engine, and simulation-based data generation. The simulation service will shorten the setup period from weeks to days, accelerate world creation, and provide an out-of-the-box simulation world plus a real-time simulation experience.

- Robot skills: Uses the real data platform and robot simulation to obtain robot data. The data drives the development of relevant skills, thus addressing corner cases at a low cost. The main technical points include data-driven skill development pipelines for robot skill ecosystem construction, plus a lifelong learning and development mode for preset and new skills.
- Robot operation management: Monitors robot running status and performs centralized management to reduce labor costs. Migrates running data to the cloud to facilitate subsequent intelligent analysis and cloud brain construction. The main technical points include the edge-cloud synergy framework and monitoring and upload of robot running data. Finally, manages more than 10,000 robots and performs real-time monitoring and global dispatching management.

3) Intelligent human-machine collaboration

Human-machine interaction studies the interaction between people and physical systems as well as people and digital systems. The interaction process here is led by people. The application of intelligent systems has made interaction shift to collaboration.

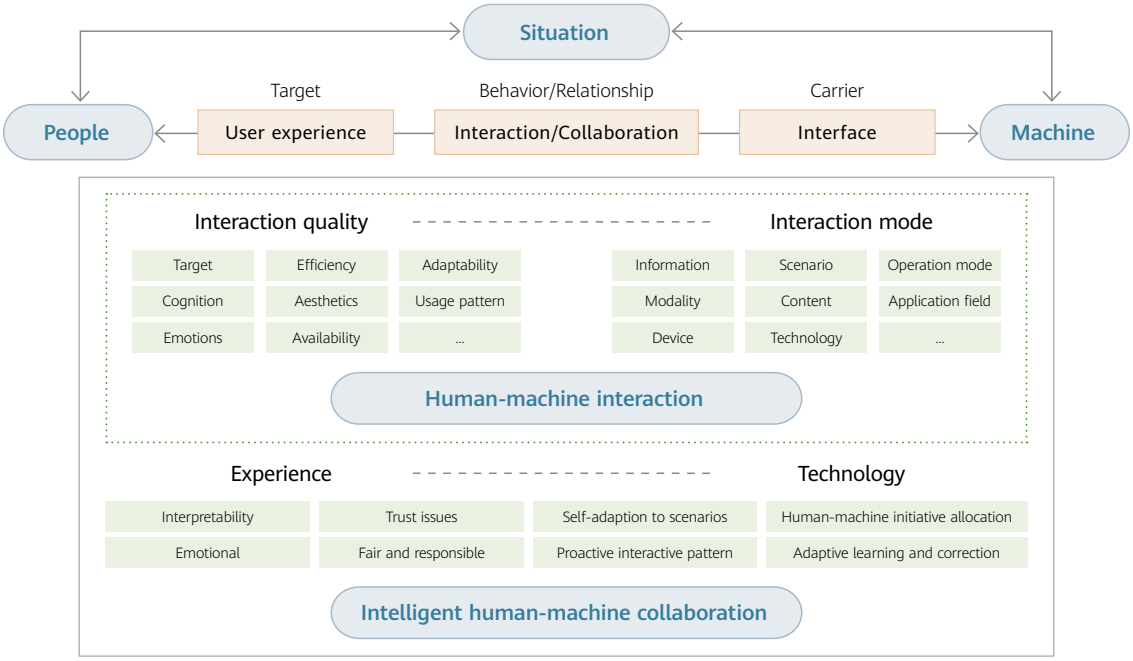


Figure 3-7 Intelligent human-machine collaboration

AI and its application to robotics have introduced autonomy to machines. Intelligent human-machine collaboration involves four technologies, as seen below.

- Human-machine agency allocation: Both humans and machines have their own degree of agency, including the boundaries of behavior and decision-making. Intelligent human-machine collaboration fully considers human agency so that the two entities can better collaborate.

- Adaptive learning and correction: When interacting with people, intelligent systems can learn and correct themselves based on user requirements and preferences. The systems will use reinforcement learning (RL) to achieve adaptive learning and correction.

- Scenario adaptation: Scenario sensing and understanding, exception handling, and target and behavior planning are the prerequisites for AI-based scenario inference and understanding. In the future, an open and scalable platform for building and researching multi-modal collaboration systems using cloud-based foundation models will become mainstream. The platform will have a time-based programming mode for parallel coordinated computing while providing tools and AI components for data visualization, processing, and learning.

- Proactive interaction mode: AI can enable machines to obtain and process massive amounts of information regarding users and scenarios. They can use this data to predict user intent and detect potential problems in advance. This marks a shift in role of machines from pure tools to partners that can understand, learn, and make decisions, which supports proactive responses and changes.

Immersive Experience

Cutting-edge office equipment, as well as VR, AR, MR, naked-eye 3D, and other technologies, are poised to revolutionize the way we work over the coming years. Experience interaction networks, XR, naked-eye 3D, and immersive collaboration will create a brand-new immersive experience, and enable participants to be engaged in virtual environments, which improves collaboration efficiency and promotes innovation.

1) Experience interaction network

As XR/3D becomes more popular, experience interaction networks will provide network access that features low latency, high bandwidth, and high reliability. Additionally, they will support the self-discovery and self-networking of a large number of office devices on campuses.

- Continuous improvement in networks: XR large-scale real-time rendering and 3D reconstruction will be transmitted to the cloud for processing through the intent-driven ultra-broadband network. To support a smooth and immersive XR experience, gigabit bandwidth and 10 ms-level latency are the minimum requirements. 8K multi-view naked-eye 3D and 8K 120 fps XR will put new pressure on networks. 10GE access and millisecond-level latency are necessary for services to be popularized.
- Device self-networking connection improvement: Multi-machine collaborative office will be promoted from homes to campuses. Distributed networking based on office devices with different capabilities offers a more cohesive office environment. For instance, mobile computers as well as projection screens, whiteboards, and MR devices in conference rooms will automatically create collaborative conferences. There will be more novel I/O devices connected to computers for a better user experience. Distributed self-networking of office I/O devices must be supported by device discovery and connection, distributed networking, and adaptive transmission technologies.

The network discovery technology of distributed networking is used to discover peripheral distributed devices through different media, such as Wi-Fi, Bluetooth, and Ethernet. Distributed networking enables distributed devices with different capabilities and characteristics to form a network. The adaptive transmission technology of distributed networking ensures that proper transmission technologies are provided for services based on network loads, device capabilities, on-site conditions, and power consumption requirements, helping reduce transmission overheads.

2) XR and naked-eye 3D

XR and naked-eye 3D display can provide more realistic, extensive, and personalized experiences to users. This technology will not only reform how people consume content for entertainment, but also create new business opportunities and social benefits. The latest technologies contributing to XR and naked-eye 3D display include near-eye display, perception and interaction, network transmission, and AI. Ongoing technological breakthroughs will help improve experience as well as the ecosystem. By 2030, the number of XR users on campuses is expected to reach 1 billion.

Naked-eye 3D display: The implementation of naked-eye 3D display involves three major phases: the digitalization of 3D objects, network transmission, and optical or computational reconstruction and display. There are two types of naked-eye 3D display technology: light field display (through lenslets) and the use of spatial light modulators (SLMs).

- Light field display: Leverages the binocular parallax to create 3D visual effects. It uses parallax barriers, lenticular lenses, and directional backlight, all of which create a fairly rigid point from which one can spectate. Their large-scale adoption would require the real-time capture of user location and dynamic adjustment. An alternative approach would be to use SLMs.

- SLM: An interferometric method is used to store all the amplitude and phase information of light waves as they scatter on the surface of a 3D object in a recording medium. When a hologram is irradiated with the same visible light, the original object's light wave can be reproduced thanks to diffraction, providing users with a lifelike visual representation.

XR display and interaction technology: Currently, XR is still at a stage of partial immersion. Today, typical XR involves 2K monocular resolution, 100°–120° FOV, 100 Mbit/s bitrate, and 20 ms motion-to-photon (MTP) latency. If all content is rendered on the cloud, 20 ms of MTP latency is the threshold above which users start to report feelings of dizziness. XR will reach full immersion by 2030, by which time it will be supported by 8K monocular resolution, 200° FOV, and a gigabit-level bitrate. There will also be technical breakthroughs in display and interaction modes.

3) Immersive collaboration system

Centered around experience, immersive collaboration has been evolving from offline to online, from single-space to multi-space, and from pen, paper, and discussion to real-time digital collaboration using ICTs. As a typical example of immersive collaboration systems, immersive conference technologies will create a brand-new experience that features 3D immersive audiovisual environments and panoramic smart collaboration. The current evolution of these technologies reflects the rising importance of immersive collaboration systems.

Regarding service data flows in immersive conference application scenarios, there is a huge focus on the following aspects:

- Audio, video, and AI data collection: Collecting conference audio and videos will involve the evolution of: multi-camera video array collection, audio array collection, and multimodal data collection. Multi-camera modules, instead of a single camera, can be set up to capture different angles and perspectives during a conference. First, one conference camera uses the multi-camera module architecture and integrates the panoramic video stitching algorithm, extending the conventional narrow FOV of 80° to a panoramic FOV of 360°. Second, multiple cameras are cascaded to form an all-scenario video collection array that covers the 3D space. This, coupled with the update and iteration of the 3D optical collection chips, will advance methods of video collection while keeping the array layout cost low. Conference audio collection strengthens sound pickup quality through linear, ring, area, and cascaded arrays, and performs audio enhancement based on AI. It supports AI voice enhancement and noise suppression, which are applicable to short-distance and long-distance sound pick-up respectively. Additionally, voice and infrared sensors, cameras, and lidars track and recognize people and things.

- Audio, video, and data encoding and decoding: The 2D encoding and decoding technology for multimedia video conferences is being upgraded from H.261/H.263/H.264/H.265 to H.266, and will develop toward 3D video encoding and decoding standards. The encoding and decoding resolution is evolving toward 4K/8K 60-120 fps, with audio encoding and decoding technology standards upgrading from mono- and dual-channel to multi-channel. Thanks to device-cloud synergy, the cloud helps process massive amounts of audio and video encoding and decoding data.

Conference audio, video, and collaborative data communications use wired, Wi-Fi/BT, and NearLink networks. Device-cloud synergy and multi-device collaboration are going to require high bandwidth, numerous connections, low latency, and the interference-free transmission of application data as a default for collaborative office devices and services in the future. NearLink implements multi-level cascading of audio and video collection devices with a bandwidth of hundreds of Mbit/s and a latency of milliseconds. In addition, AI-based network detection technology provides super error concealment (SEC) and jitter-free algorithms to ensure smooth video conferences even if the packet loss rate exceeds 30%.

- Immersive audiovisual experience and AI collaboration and interaction: Audio and video display and data interaction are migrating from small-sized LCD screens to ultra-large LED screens. They are also evolving toward OneWall zone-based collaboration and interaction as well as naked-eye 3D display. The light field and 3D technologies, stereo sound localization with multi-speaker arrays, and smart office assistants using multi-modal AI sensors come together to provide multi-dimensional immersion.

Spatial AI Computing

Enterprise office access has made a transition from relying on intelligent devices to becoming intelligent collaboration spaces thanks to new AI and sensing technology. In the intelligent collaboration space, cloud, edge, AI, and devices interlink to identify user intent, perform predictive computing on space and data, and provide users with faster and more accurate services through real-time collaboration of campus digital virtual engine.

1) Predictive computing

As AR/XR and AI technologies advance, human-machine collaboration during enterprise office will embrace intergenerational changes in terminal forms, human-machine interaction, and applications. The AI platform and intelligent devices' spatial computing for enterprises will form a new intelligent collaboration space. In such a space, predictive computing alongside real-time interactive computing will be implemented using AI, high-performance computers, and real-time data processing technologies. The computing technologies can be applied to a wide range of fields such as VR and videoconferencing to provide a smoother user experience. The computing power required per user is 1 TFLOPS.

Unlimited I/O expansion, distributed collaborative computing, virtual-physical office, device-cloud distributed applications, and tactile Internet bring both major challenges and huge opportunities to network latency, network bandwidth, and openness. Networks and application devices provide intelligent collaboration spaces for enterprises and employees.

2) Digital virtual engine for campuses

3D tools like powerful rendering and physical simulation will be more broadly applicable, like in a digital twin intelligent campus. The technologies are also known as campus digital virtual engines. This engine can display real-time data regarding the intelligent campus. It affects campus management, and can even directly control campus devices. It improves operational efficiency, resource utilization, and the quality of the campus environment.

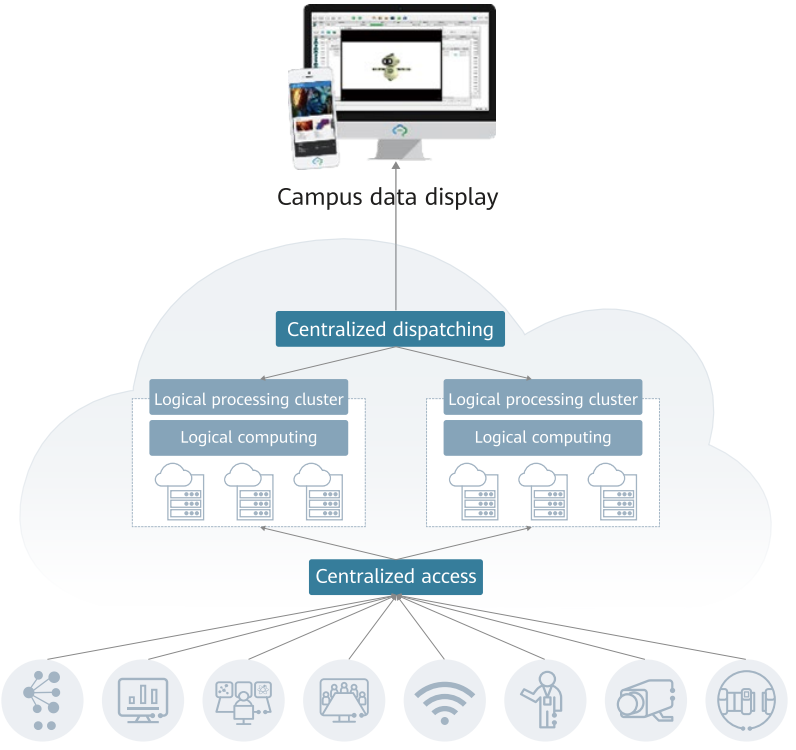


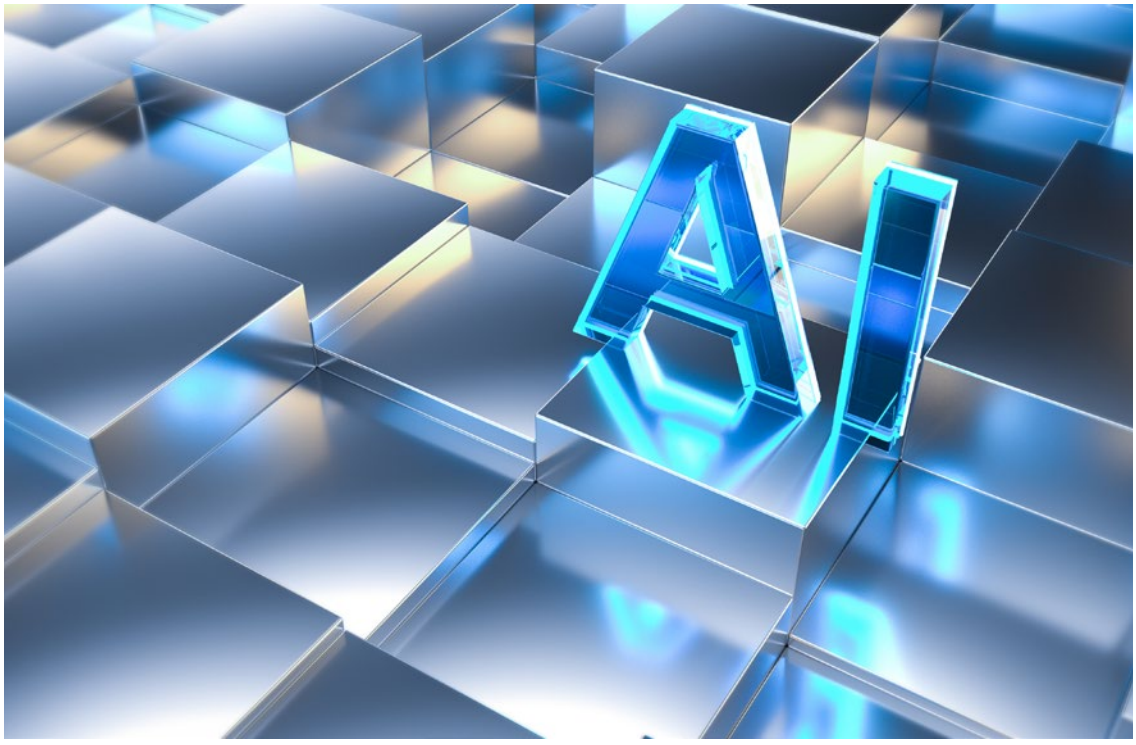
Figure 3-8 Campus digital virtual engine

Each physical device on the digital twin intelligent campus can be regarded as an agent. An agent can upload data to the campus digital virtual engine on the cloud for real-time display. It can also be controlled by the engine to change the physical status. IoT and robotics are booming in the market. As a result, a wide variety of agents that take many forms and do many things are likely to be seen on future intelligent campuses.

Unlimited computing power on the cloud plus high-speed communications buses are crucial to displaying the data of numerous agents in a cloud-based campus digital virtual engine/virtual space. When the campus digital virtual engine becomes distributed, 10,000 agents can be connected to the same space for real-time interaction.

- Access of numerous agents and real-time data upload: An agent manages its own engine server based on its home access, and the digital virtual engine servers of multiple campuses share statuses. The total number of connected agents increases to more than 10,000.

- Real-time display and control of massive amounts of data: Provides interest-based data acquisition during data display, feeds back data based on requirements, and offers interfaces to view and control a multitude of data.



3.1.3 Ubiquitous Intelligent Connectivity

Countless sensors and IoT devices in the physical space, which are used for data collection, aggregation, and processing, are the prerequisites for constructing their digital counterparts. Based on the outlook for future intelligent campuses, technical requirements for ubiquitous intelligent connectivity come from multi-dimensional sensing, harmonized communication and sensing, and edge intelligence, which use data to power intelligent services across industries.

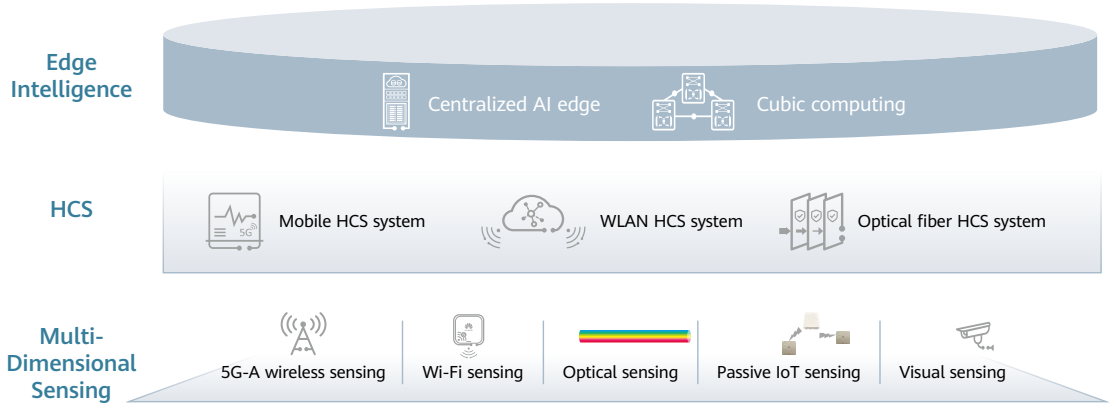


Figure 3-9 Ubiquitous intelligent connectivity

Multi-Dimensional Sensing

Future intelligent campuses should be able to sense. By combining the latest sensing technologies like 5G-A wireless sensing, Wi-Fi sensing, and optical sensing with traditional passive IoT sensing and visual sensing, they will build a highly accurate digital sensing network that provides real-time visibility and efficient operations. This network holistically senses and identifies people, devices, things, events, spaces, and environments within the campus from many angles.

1) 5G-A wireless sensing

5G-A wireless sensing is particularly useful in scenarios like connected vehicles and drones. With 3GPP Release 16, precise positioning can already achieve meter-level accuracy, and future releases are expected to hone this accuracy to a centimeter level. As wireless networks move toward higher frequency bands like millimeter wave and terahertz, wireless sensing will be applied in smart cities, weather forecasts, environmental monitoring, medical imaging, and more.

2) Wi-Fi sensing

IEEE 802.11bf defines standards of Wi-Fi sensing, which detects and identifies surrounding environments and events while generating data based on unlicensed frequency bands. It covers functionalities such as high-precision positioning, posture and gesture recognition, breath detection, emotion recognition, and perimeter security for indoor, outdoor, in-vehicle, warehouse, and freight yard scenarios, among others, on campuses.

It can be classified into the following types by sensing capabilities:

- Coarse-grained sensing: Simple hardware extracts obvious signal features like Doppler frequency shifts to identify events, including person detection, intrusion detection, motion detection, and more.
- Fine-grained sensing: High-precision detection methods like mmWave radars precisely detect and recognize subtle statuses such as personnel signs and gestures.

It can be classified into the following types by sensing scope:

- Single-site detection and sensing: A single device implements fixed-point detection and sensing within a specified area. Examples include off-bed monitoring and fall identification.
- Multi-site networking sensing: Joint detection by multiple sites and information combination and deduplication cover the entire continuous area. Specific results include shopping mall/supermarket personnel statistics, trajectory tracking, and indoor environment imaging.

3) Optical sensing

Optical sensing technology can be used to sense changes in faults, vibrations, stress, temperature, humidity, gas, sound, and lighting. It can support device/pipeline fault diagnosis, and environment and facility stress monitoring, as well as provide service features such as temperature, humidity, ventilation, and lighting adjustments.

- Fiber-based resource sensing: Technologies like fiber topology, E2E optical power visualization, and meter-level optical path fault diagnosis are coupled with software algorithms. When a network fault occurs, the O&M personnel can quickly demarcate it and identify whether the fault occurred in an optical fiber or a network device.
- Fiber-based vibration sensing: The optical sensing system employs optics and AI to ensure the response latency can reach the millisecond level at a distance of 100 kilometers. It can effectively identify external sound and vibration signals from optical cables deployed along the perimeters, warn of intrusions at perimeter fences, and promptly identify threats. The system works with video security devices to construct a multi-level intrusion detection system.
- Fiber-based environment sensing: Distributed temperature/humidity sensors, laser gas sensors, fiber-based distributed sound sensors, and zone-based illumination sensors monitor temperature/humidity, CO_2 density in active/rest zones, full-area sound intensity, and full-building light distribution to provide 3D spatial data. This leads to precise temperature/humidity adjustment, intelligent ventilation, proactive noise suppression, and automatic lighting control.

4) Passive IoT sensing

Current active IoT devices still rely on batteries or other power supplies. Therefore, passive IoT collection with low power consumption is becoming popular, which requires the following three technologies:

- IoT devices with ultra-low power consumption: The simplified RF architecture and crystal oscillator-free receivers/transceivers reduce IoT RF power consumption by one or two orders of magnitude.
- Wireless directional energy transmission: Directive wireless energy transmission and efficient air interface energy collection and conversion technologies replace the built-in batteries of IoT devices.

- Joint scheduling of energy supply and collection: Coordinated directional energy transmission and device data collection ensure continuous, stable, and reliable collection control for battery-free IoT devices.

5) Visual sensing

Visual sensing uses cameras to collect information about the environment, space, people, and events, and then generate image data. The technology is constantly evolving to minimize its impact on the environment and people, simplify installation and deployment, as well as improve sensing precision.

- Mini front-end device: The front-end device is small and easy to deploy across various scenarios. It is more harmonious with the environment.
- Simplified front-end device: With the support of high bandwidth, low latency, and edge pooling, cameras are designed to be simple. They collect only original signals and directly send them to the backend.
- Multi-dimensional front-end device: Supports converged access of multiple visual sensing and image devices, including mobile devices like mobile phones and digital cameras, and fixed devices like network cameras and monitoring cameras.
- Multi-channel sensing: Sensing based on diverse technologies that complement each other, such as visible light, thermal infrared, millimeter wave, and sound wave, achieves a high precision and a low false alarm rate.

Harmonized Communication and Sensing

Harmonized communication and sensing (HCS) refers to the extension of communications technologies into the realm of sensing. There are three categories of HCS systems: mobile HCS system, WLAN HCS system, and optical fiber HCS system.

1) Mobile HCS system

From 1G to 5G, communications and sensing have remained independent from each other. For example, a 4G communications system is only responsible for communications, and a radar system is only responsible for functions like speed measurement, sensing, and imaging. This separation wastes wireless spectrum and hardware resources, and the separation of functions often results in high latency for information processing.

As 5G-A and 6G gain popularity, the communications spectrum will expand to include millimeter wave, terahertz, and visible light. This means the communications spectrum will soon

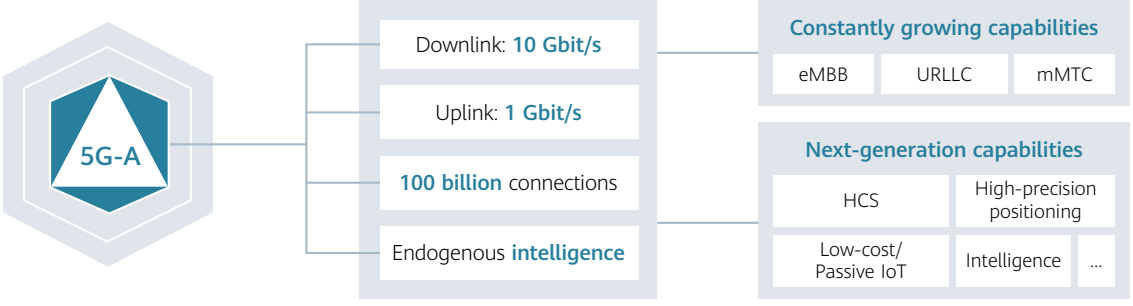


Figure 3-10 Mobile HCS system

overlap with the spectrum previously reserved for sensing systems. HCS facilitates unified scheduling of communications and sensing resources.

Communications and sensing resources can be multiplexed through time, space, and code division, so that sensing functions can be added to base stations as required at lower deployment costs. High-isolation antennas are full-duplex and have 5G TDD enabled, meaning that they can support a co-frequency co-time full duplex (CCFD) sensing mode. They can co-cover communications and sensing while also delivering optimal communications performance. An integrated architecture for communications and sensing can help provide E2E sensing services across different industries and ensure data security.

Ultra-wideband and multi-antenna capabilities can achieve centimeter-level sensing. Multi-site collaboration of cellular systems can add 3D multi-angle sensing with no blind spots to single-site sensing. In addition, the target recognition algorithm based on machine learning can improve the recognition resolution using rich multi-domain (time, frequency, space, and code) 5G NR information.

2) WLAN HCS system

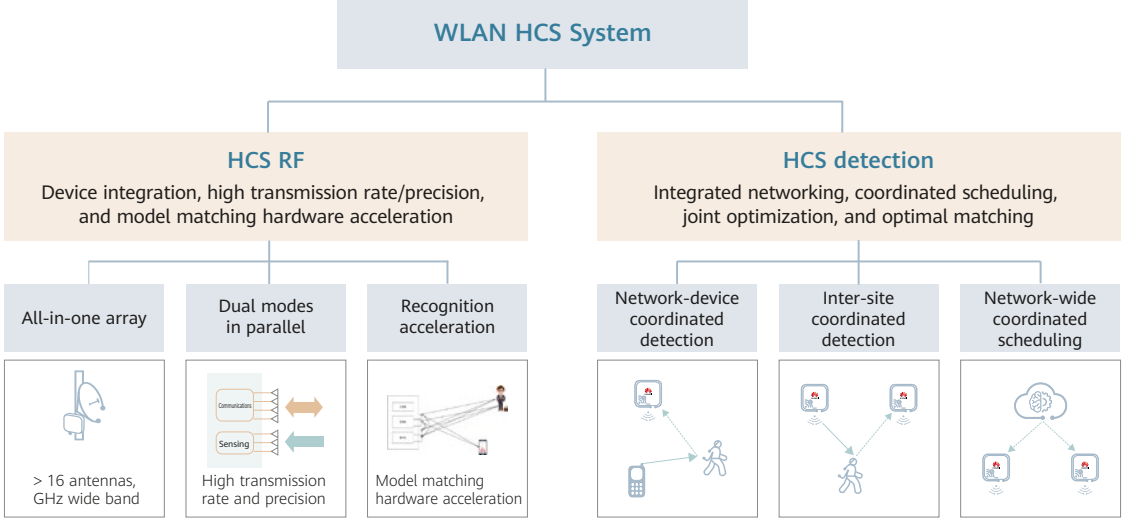


Figure 3-11 WLAN HCS system

The WLAN HCS system leverages existing hardware, chips, radio frequency ports, and operating frequency bands of communications sites to integrate communications, sensing, signal processing, and channel resource scheduling. This realizes a high sensing precision while maintaining the original communications bandwidth, latency, transmission, and energy consumption.

The HCS system achieves high sensing precision mainly by utilizing the HCS radio frequency and collaborative detection.

The HCS radio frequencies can be categorized into microwave and mmWave frequency bands depending on the operating frequency band.

- Microwave frequency band: It uses the sub-7 GHz spectrum and device resources. To achieve decimeter-level precise positioning and meet coarse-grained detection requirements (such as presence detection and intrusion detection), it uses isolated omnidirectional antennas with fewer than 8 transmitting or receiving antenna elements and a frequency bandwidth of less than 160 MHz.
- mmWave frequency band: It uses more than 16 antennas and a high detection bandwidth (x GHz) that allow for centimeter-level precise positioning and millimeter-level precise sensing detection. This level of precision meets the requirements for various sensing applications, such as detecting movement, identifying and monitoring behaviors, vital signs, and trajectories, and environmental imaging.

HCS collaborative detection uses the following technologies to coordinate the behaviors of sites and devices across the entire network and ensure the best balance between communications quality and sensing precisions:

- Network-device coordinated detection: Information obtained through sensing and detection is collected in parallel with regular network protocol interaction and data transmission.
- Inter-site coordinated detection: Collaborative detection between neighboring APs compensates for limitations in sensing quality caused by terminal traffic locations. It also provides network-wide location anchors for sensing and detection.
- Network-wide coordinated scheduling of sensing and communications: Network-wide communications traffic, site resources, and device resources are taken into account to ensure an optimal balance between communications quality and sensing precision.

3) Optical fiber HCS system

The optical fiber HCS system integrates sensing technology with optical fiber communications technology to sense the environments, objects, and events, providing more intelligent and efficient data services and applications.

Optical fiber communications technology is used to build high-speed campus networks, which facilitate high-quality transmission with:

- High bandwidth: This technology supports the direct transmission of front-end devices' original sensor signals at speeds of 100 Mbit/s to 1 Gbit/s, and to 10 Gbit/s. It also supports ultra-high-speed interconnection and collaborative interaction with more original information.
- Low latency: The technology ensures that the latency of synchronization between each front-end device is less than one frame. This allows for efficient and collaborative analysis of events while also generating complete tracks.

- Zero packet loss: Front-end devices' original sensor signals are transmitted without any losses, and with key data features preserved.

The multimodal information optical sensing technology collects a wealth of environmental information to provide the necessary data for services.

- Environmental sensing technology: Ambient optical sensors and infrared sensors sense and measure ambient parameters such as light and temperature to support smart office, environmental monitoring, and other application scenarios.
- Optical fiber sensing technology: Optical fibers function as sensors to detect various physical quantities and parameters in the environment, such as vibrations and stress, and to sense postures and movements during immersive experiences.

Edge Intelligence

The edge intelligence data system deployed on campuses intelligently processes multi-dimensional sensing data at the edge through cubic computing and centralized processing at the edge.

1) Cubic computing

Computing and storage infrastructures are distributed across different locations on the cloud, edge, and devices. Such infrastructures can be horizontally or vertically coordinated to complement each other and enable cubic computing. This addresses problems such as poor service experience, uneven distribution of computing, low utilization of computing resources, and information silos.

- Edge computing: To be able to apply edge computing on a large scale, certain challenges in areas like centralized processing at the edge, edge computing networks, edge security, edge standards, and open ecosystems must first be confronted. Edge computing enhances privacy and security for data collection, storage, processing, and transmission, as well as standardizes edge computing systems, software and hardware frameworks, interfaces, and protocols. It provides a solution to common intelligence problems that other industries are also facing. For instance, it meets requirements for edge acceleration, offloading, and performance breakthroughs.
- Multi-device collaboration: Over time, multi-device collaboration systems will gradually evolve from simple cooperation and connection systems into autonomous swarm intelligence. The multi-device collaboration technology aims to improve the problem-solving capabilities, overall performance, and robustness of multi-device systems. Multi-device collaboration can take various forms, such as task sharing, result sharing, and intelligent agents. Effective multi-device collaboration also requires solving problems related to cooperation and conflict resolution, global optimization, and consistent interaction and collaboration.
- Cloud-edge-device synergy: AI and emerging data-intensive applications on intelligent campuses are developing rapidly. The need to improve application experiences, for example by reducing latency or bandwidth costs, or by enhancing data privacy protection, is driving the development of cloud-edge-device synergy. Cloud-edge-device synergy uses multiple computing and storage devices at different locations on the cloud, edge, and devices to develop an integrated computing architecture and support task, intelligence, data, and network collaboration.

2) Centralized processing at the edge

The future campus service system will be able to centrally process multi-dimensional information gathered by front-end devices, which will also be accessed by various service subsystems. It will deliver a three-dimensional protection and sensing management experience, simplify the current management systems, and enhance operational and management efficiency.

- Resource pooling and sharing: Data from multiple sources is aggregated to set up shared resource pools, where resources can be invoked by each service system on demand.
- Collaborative analysis by intelligent algorithms: Intelligent algorithms carry out multi-dimensional analysis of more raw data to support more accurate prediction, identification, and intelligent handling of events. The intelligent and proactive management system becomes smarter the more it is used thanks to AI-powered multimodal self-learning and distributed collaboration.
- Centralized data processing: Data collected by each front-end sensor is centrally processed to facilitate unified service management and monitoring along with collaborative intelligent analysis.

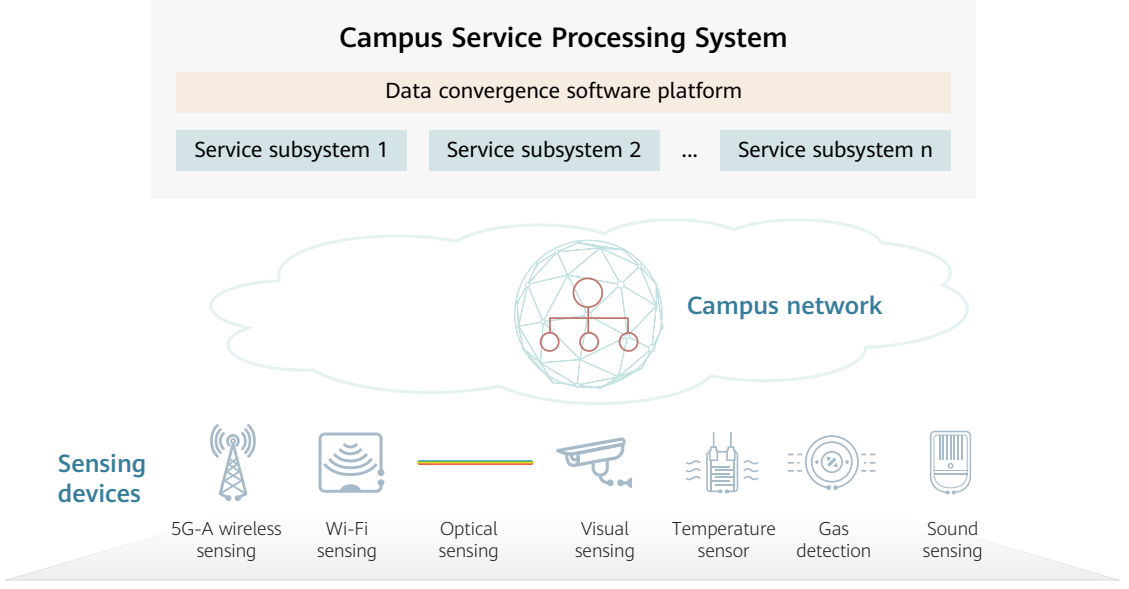


Figure 3-12 Centralized processing at the edge

3.1.4 Intent-Driven Ultra-Broadband

Network infrastructure is the foundation of informatization, and also the foundation of the future world that everyone will live in. The coming decade will see continuous improvement in network performance. Network ports will be upgraded from 400G to 800G or even 1.6T, and single-fiber capacity will exceed 100T. Huawei predicts that by 2030, there will be over 200 billion connections worldwide, and we will enter the era of 100 Gbit/s connectivity. Deterministic networks and on-demand network services should be provided to meet office, production, and life needs in typical scenarios, including holographic AIOC, transcendent smart office, 10 Gbps highly reliable production, metaverse life, and smart energy management. They can enhance the application experience of enterprise users.

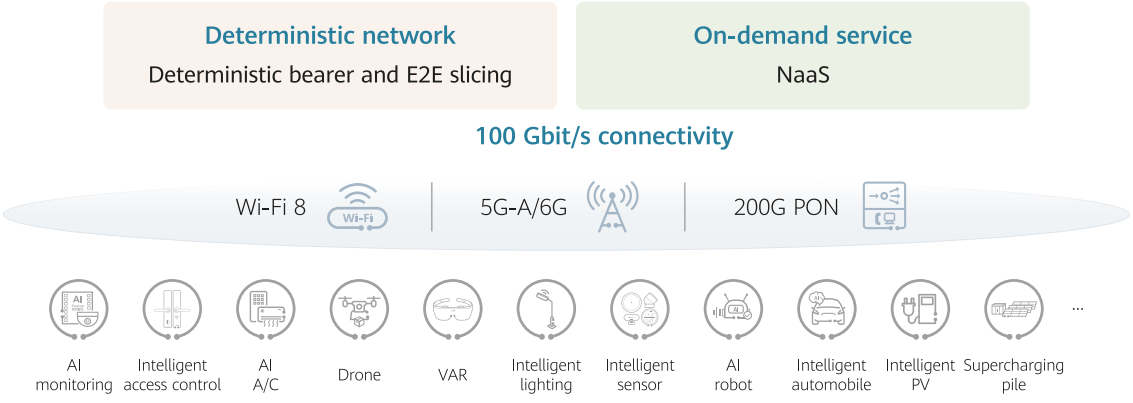


Figure 3-13 Intent-driven ultra-broadband

100 Gbit/s Connectivity

To make 10-gigabit campus networks possible, certain technologies are needed to support 100 Gbit/s connectivity. The first is next-generation Wi-Fi 8 technologies that support mmWave and high-density MIMO. 200G passive optical network (PON) technology will likely be used for optical access. The coherent detection technology typically used for wavelength division multiplexing (WDM) will be used in the PON field, which will significantly improve receiver sensitivity and support modulation formats with higher spectral rates to achieve higher data rates. Last, mobile 5.5/6G network research needs to focus on the flexible use of the sub-100 GHz spectrum bands and continuous evolution of massive MIMO.

1) Wi-Fi 8 wireless technology

With the application of holographic AIOC, transcendent smart office, and 10 Gbps highly reliable production, devices are requiring higher bandwidth, evolving from 100 Mbit/s to x Gbit/s. By 2030, campus networks must support a concurrent experience rate of 10 Gbit/s for multiple devices and a peak rate of WLAN APs at 100 Gbit/s. The current sub-7 GHz spectrums are unable to support the 100 Gbit/s bandwidth target, which is leading to the evolution of devices and devices to the mmWave spectrum.

While mmWave has abundant spectrum resources and offers significant advantages in terms of working bandwidth, its widespread

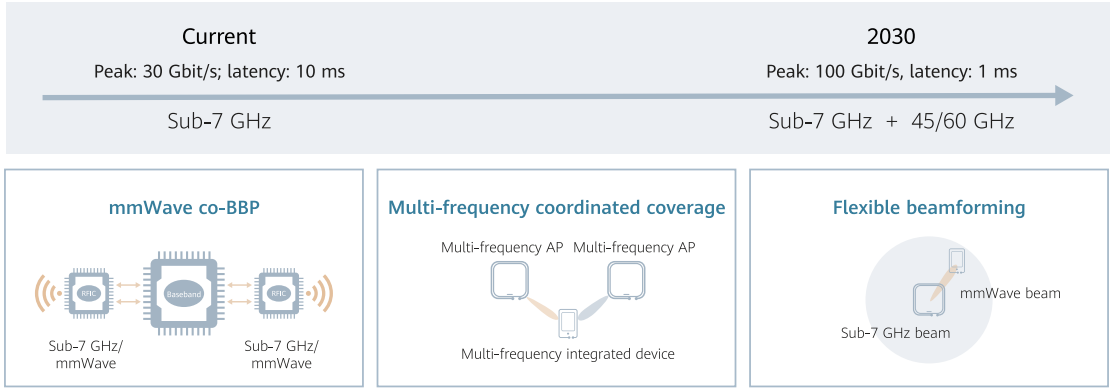


Figure 3-14 Three technical solutions

commercial use is impeded by three challenges: costs in integrating mmWave into network devices, reliability of mmWave links, and coverage continuity. These challenges can be addressed through the following technical solutions:

- Multi-frequency co-BBP: With a flexible baseband architecture, the same baseband can support both sub-7 GHz and mmWave RF. RF ports, the number of antennas, and frequency bandwidth can be configured and adjusted as needed through shared baseband to reduce hardware costs and power consumption of APs and devices that integrate mmWave.

- Multi-frequency coordinated coverage: Multiple high- and low-frequency links, inter-frequency joint coding, and dual-frequency collaborative switchover ensure services are not interrupted when signals are suddenly blocked. This is especially important because the weak diffraction, reflection, and

penetration capabilities of mmWave signals make services susceptible to interruptions when they are blocked.

- Flexible beamforming: To enable co-deployment and coordinated coverage, mmWave and sub-7 GHz RFs must maintain continuous coverage over typical inter-site distances. Due to the greater air interface attenuation of mmWave, traditional omnidirectional antennas are insufficient for coordinated coverage. Instead, low-cost, high-gain hybrid beamforming arrays are necessary, along with a low-overhead fast beam training alignment algorithm to quickly deploy and track the location of mmWave devices.

2) 200G PON

GPON/10GPON features high bandwidth (1 Gbit/s and 10 Gbit/s), long transmission distances (up to 40 km), low energy consumption (passive ODN), and low costs (P2MP traffic statistical multiplexing). Therefore, they are well-suited for service requirements in the data interconnection era and have been widely adopted. However, as new services (such as AR/VR/holography) and application scenarios emerge, users have higher requirements for latency and jitters while they also require bandwidth to continuously evolve from 10 Gbit/s to 50 Gbit/s and then to 200 Gbit/s. For instance, vertical industries require a centralized bearer for multiple service networks (such as data collection, video security, and production control) to ensure reliability and service security isolation, as well as to meet industry specifications and graded SLA requirements. The key characteristics involved are as follows:

- 200G high speed: The innovation of system mechanisms introduced the 200G coherent detection technology of WDM, pushing TDMA to evolve toward TDMA+FDMA/WDMA and realize 200G over a single fiber.

- ms-level latency and μs-level jitters: The new TDMA+FDMA architecture alongside the single-fiber multi-burst technology conduces to ms-level latency and μs-level jitters.

- Hard isolation between multiple services: Different timeslots, wavelengths, or modulation formats are divided for different services through time domain isolation and L1 subcarrier isolation technology. This provides physical hard isolation in multi-service integrated bearer scenarios.

3) 5G-A/6G mobile network

In the future, mobile network research needs to focus on flexible use of the sub-100 GHz spectrum bands and continuous evolution of massive MIMO. 3GPP Release 16 has defined two frequency ranges, FR1 and FR2, for 5G new radio (NR), covering all spectrum bands for International Mobile Telecommunications (IMT) between 450 MHz and 52.6 GHz. Research for Release 17 is still underway, and one important focus of this research has been the use of spectrum above 52.6 GHz for 5G NR, which has three characteristics:

- Increasing capacity: Improvements in air interface efficiency and networking innovation, and application of new communications technologies, such as electromagnetic information theory and holographic MIMO, are expected to increase the capacity of wireless networks by 10 to 20 times between 2025 and 2030. Engineering technology innovation and advancements in algorithms enable more precise channel estimation, beamforming control, and interference suppression, contributing to higher spectral efficiency. Additionally, AI-powered air interfaces make PHY and MAC allocation more adaptable and efficient, catering to the personalized needs of individual users.

- Three-dimensional collaboration: User experience at the cell center and edge varies greatly. As we move toward the future, ubiquitous 1 Gbit/s high bandwidth will be necessary to support 8K XR and 6DOF holographic services. Limited antenna space and higher capacity demands necessitate the use of multiple distributed systems at a single site. To achieve optimal convergence of the sub-100 GHz spectrums, different frequency bands, duplex mode, and uplink and downlink spectrum resources can be flexibly used. Distributed massive MIMO offers low-complexity distributed processing with near-optimal performance, and is easy to deploy in all scenarios (DRAN/CRAN, macro and micro sites) to ensure a consistent user experience. With flexible spectrum technology, we can determine the spectrum to be used, duplex mode, uplink and downlink, space division, and control/service resource allocation mode based on service characteristics, network load, user location, and rate.

- 100 Gbit/s uplink: For super uplinks, interference must be eliminated in all scenarios with different slot configurations, including under the Uplink DMIMO large-scale joint reception architecture, to increase the uplink capacity of a single cell from 10 Gbit/s to 100 Gbit/s. DMRS pilot capacity expansion, four-dimensional joint optimization, and ABF antenna innovation can increase the number of uplink layers and expand the capacity.

Deterministic Networks

Communications networks need to provide a deterministic experience through deterministic bearer and E2E slicing, so as to meet the service needs of holographic AIOC, transcendent smart office, metaverse life, and 10 Gbps highly reliable production.

1) Deterministic bearer

Campus services are becoming more diversified. Over the course of this decade, the campus traffic model will be disrupted with a focus of connected services shifting from traffic to both traffic and service latency. They will reposition from top-down content traffic that serves consumption and entertainment to bottom-up data traffic that serves intelligence across the whole industry. Intelligent robots will generate massive amounts of data, and this data will need to be processed in data centers. The network needs to serve an intensive layout with cloudified services at the center, and it can perform real-time and deterministic scheduling at the network layer based on service attributes.

- Wireless access: Real-time wireless access services require high instantaneous rates over the air interface. However, due to the spectrum constraints caused by the multiplexing of multiple pieces of UE on a single carrier, it is difficult to guarantee real-time performance. Moving forward, multicarrier aggregation technologies need to be developed so that carrier configuration is decoupled from transmission, improving the bandwidth of services under latency constraints on multi-band carriers.
- Wired access: For Ethernet networks, the current best-effort forwarding mechanism needs to be changed, protocols at the PHY and MAC layers need to be improved, and new technologies such as TSNs and deterministic IPs need to be integrated to ensure on-demand, end-to-end latency. PON upstream transmission, which relies on time division multiplexing (TDM), will evolve toward a reliance on FDMA to guarantee low latency.

2) E2E slicing

E2E slicing is a network virtualization technology with SLA assurance. Through network slicing, different logical or physical networks can be isolated from the network infrastructure to meet the SLA requirements of different industries and services. Types of slicing include wireless slicing, IPv6+ slicing, and cross-domain slicing management and service. This technology offers tailored private networks that operate independently and are separated from each other, catering to different industries and delivering superior network services for specific sectors.

- Wireless slicing: It can be further classified into hard slicing and soft slicing. Hard slicing is achieved through resource isolation, such as through static resource block (RB) reservation and carrier isolation for specific slices. Soft slicing is achieved through resource preemption, such as QoS-based scheduling and dynamic RB reservation. Currently, the bitrates of different network slices can be guaranteed based on priorities. The next step in the development of network slicing is to explore the most appropriate wireless

protocols for the PHY, MAC, Radio Link Control (RLC), and Packet Data Convergence Protocol (PDCP) layers. For example, we could have a PHY layer with a low-latency coding scheme for slices that support ultra-reliable low-latency communication (URLLC) services, or a MAC layer with an optimized hybrid automatic repeat request (HARQ) mechanism.

- IPv6+ slicing: Transport network slicing is achieved through physical isolation or logical isolation. The physical isolation technology uses Flexible Ethernet (FlexE) at the MAC layer to isolate services by scheduling timeslots. Logical isolation technologies mainly include SRv6, traffic engineering (TE), and virtual private network 6 (VPN6) at the IP layer. Logical service isolation is implemented through labeling and network equipment resource reservation. Further research is needed in the industry to explore the integration of technologies such as congestion management mechanisms, latency-oriented scheduling algorithms, and highly reliable redundant links for FlexE, TSN, and deterministic networking (DetNet). This can deliver bounded latency and zero packet loss for physical slicing, as well as low-granularity FlexE interfaces.

- Cross-domain slicing management and service: Moving toward 2030, the SLA awareness, precision measurement and scheduling of slicing need to be further researched in the industry to achieve automated closed-loop slicing control. 3GPP has defined an E2E network slicing management function (NSMF), which streamlines network slice subnet management functions (NSSMFs) to enable end-to-end automatic slicing. This can facilitate elastic slice service provisioning and capacity expansion or reduction.

On-Demand Services

1) Network as a service

As services are migrated to clouds, enterprise applications are being deployed across multiple zones. The communications network that connects services needs to span across campuses, wide areas, and clouds. Software engineers and application developers want to exclude network complexity, decouple service logic from network topology, and focus on application construction and new function development to improve the experiences of enterprise users when it comes to applications. As a result, business abstraction will occur at the network-as-a-service (NaaS) layer, and applications will only need to invoke network service–layer interfaces. The network will automatically establish connections and provide service assurance based on the application's definition.

- NaaS layer: The NaaS layer is a network layer that streamlines service connections for network developers and operators. It conceals network complexity and centrally manages connections using unified policies to provide traffic management, including service routing, load balancing, service discovery, observability, and security. It lies between application services and network connections and can be a mixture of physical or virtual network connections. Ensuring reliable, efficient, and secure communications between application services in this environment is a technical challenge.

- Global network routing: Global network routing utilizes independent path computation and transmission assurance technologies to build optimal transmission path networks based on network and PoP resources provided by basic network operators, cloud providers, and Software-Defined Cloud Interconnect (SDCI). The path with the lowest latency or transmission cost is selected as needed, based on real-time network status, to achieve global optimal application-oriented scheduling. Global network routing is situated beneath the NaaS layer and offers network computing and automatic configuration capabilities.

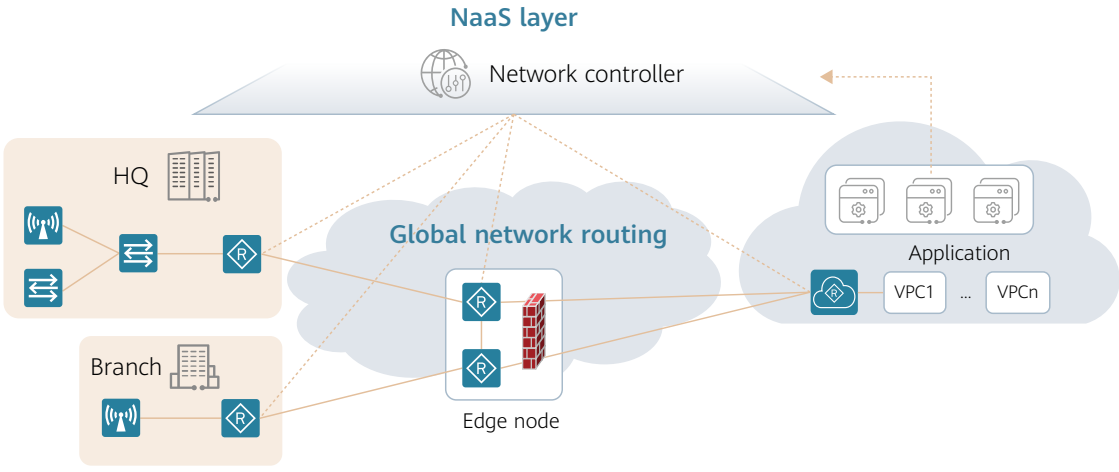


Figure 3-15 NaaS layer

3.1.5 Security and Resilience

The increasing adoption of cloud platforms worldwide means that the geographical and legal boundaries of security have been blurred. As a result, security and resilience assessment of systems, resources, and devices has become necessary. Moving forward, it is essential to ensure campus service security by tapping into data security, computing power security, and resilient systems.

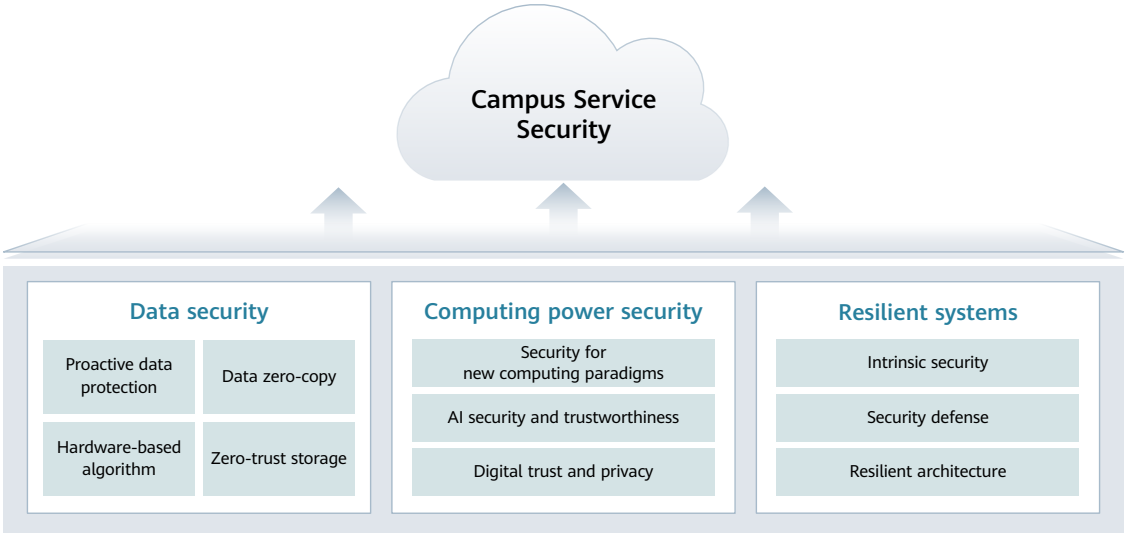


Figure 3-16 Security and resilience

Data Security

Production data can be regarded as core assets for future campus services. Therefore, it should be protected in multiple layers, from multiple dimensions, and across multiple products. We can use technologies like proactive data protection, zero-copy, zero-trust storage, and hardware-based algorithm to identify, detect, respond to, protect, and recover campus data. This creates a comprehensive awareness of data security at all times. Data access behavior, data information entropy, internal data correlation, and data distribution within a certain period of time can be collected to create data security profiles. Research on the sampling theory of mass data, converged processing of heterogeneous data, hardware-based AI detection algorithm acceleration, and causal analysis and inference will enhance threat detection accuracy and the ability to predict unknown data threats. This will eventually lead to accurate and swift threat detection, real-time processing, and dynamic assessment.

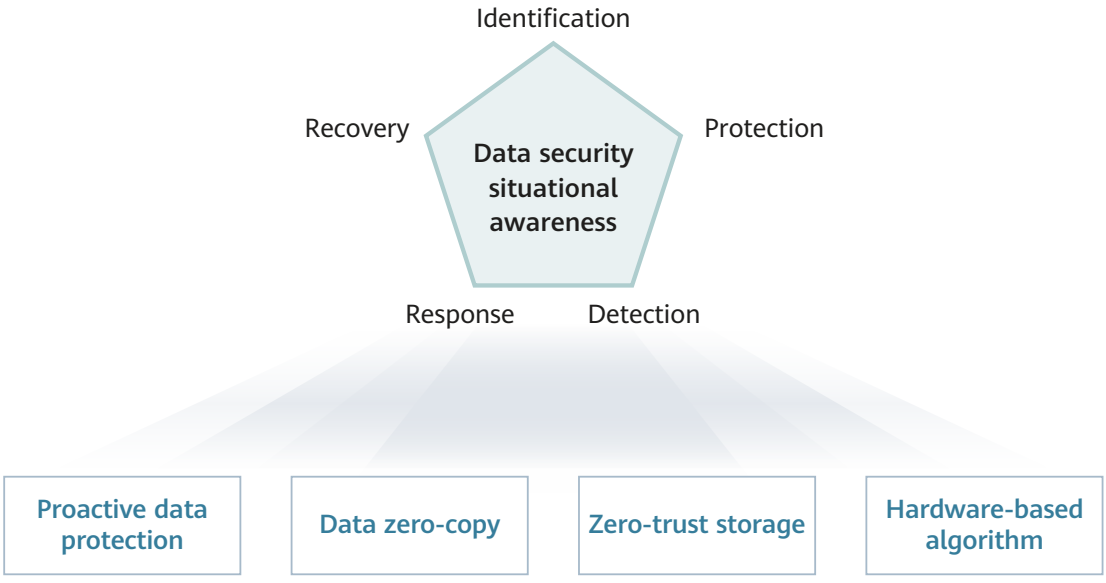


Figure 3-17 Data security

1) Proactive data protection

Studies on data security attacks and defense have revealed that the current security system for campus service data is unable to defend against virus attacks, including ransomware. To address this issue, it is necessary to enhance the campus data protection system by implementing technologies such as data timeline travel, native anti-tampering, and multi-dimensional linkage response.

- Data timeline travel: If campus data is damaged due to internal or external attacks, the data infrastructure must be able to restore the damaged data to a historical point quickly to ensure zero data losses. I/O-level data recovery, root cause determination, and other technologies make it possible to precisely locate data on the timeline and automatically recover it. The finest-grained data replay function can trace the attack sources, supporting adjustment and optimization of data security policies.
- Native anti-tampering: Data anti-tampering mainly relies on system-level data access control. However, campus service cloudification has extended the security boundary of campus data, enlarging the attack surface of a typical system, while it is difficult to guarantee data anti-tampering. Over the next decade, system-level data access control and the physical anti-tampering attribute of media are expected to be combined.
- Multi-dimensional linkage response: This technology allows for cross-device collaboration among network, security, endpoint detection and response (EDR), and storage devices. This implements multi-dimensional closed-loop threat processing and prevents threats from spreading. AI security analysis, causal analysis, and inference are expected to make autonomous decision-making and response more intelligent, thus enabling quick and accurate responses.

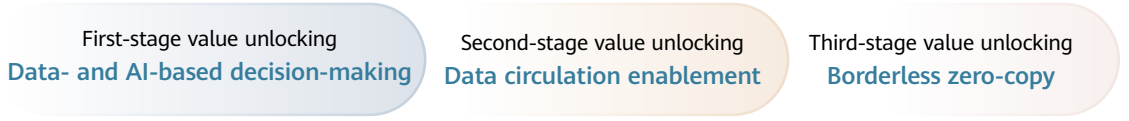


Figure 3-18 Data zero-copy

2) Data zero-copy

The value of data elements is unlocked in three stages. During the first stage, data supports service system operation, and promotes service digitalization and intelligent decision-making. The second stage involves data circulation, which enables quality data from different sources to be converged and aggregated in new services and scenarios. Data sharing and access control efficiency can be enhanced by capitalizing on technologies like cryptography-based access control, self-protecting data, efficient and transparent audit, and zero-copy data access (ZCDA). This makes for efficient data circulation and usage while maintaining secure data sovereignty. The third stage involves borderless zero-copy. ZCDA breaks data boundaries for data sharing and eliminating data silos.

- Cryptography-based access control: Cryptography is used to ensure data confidentiality. Users who do not comply with access control policies cannot decrypt data. Encryption technologies will be able to implement policy judgment and randomized processing on a ciphertext that is leaving a trusted domain so as to prevent any data that doesn't comply with predetermined access control policies from leaving the trusted domain.

- Self-protecting data: Data security has evolved from system-centric management and control to data-centric full-lifecycle security protection. As protected objects have been changed, the data capsule solution is expected to encapsulate access policies, use control policies, and encrypted data to ensure that data owners can control data and implement secure data transfer.

- Efficient and transparent audit: Blockchain is the mainstream technology used to audit data trustworthiness currently. Moving toward 2030, we need to use efficient and transparent audit technology to develop an audit solution to prevent campus services from being tampered with and efficiently store data. This will also cater to data read and write timeliness requirements in the production process.

- ZCDA: Due to the differences between application data models of different campus services, most applications are siloed with independent data copies. We can solve this problem in two steps. First, application data models will be deployed at the data storage layer and automatically generated based on the same data to eliminate data silos. Second, fine-grained access control and trusted network transmission based on chip certification will be used to implement efficient data access across trusted domains, ensuring secure data sharing.

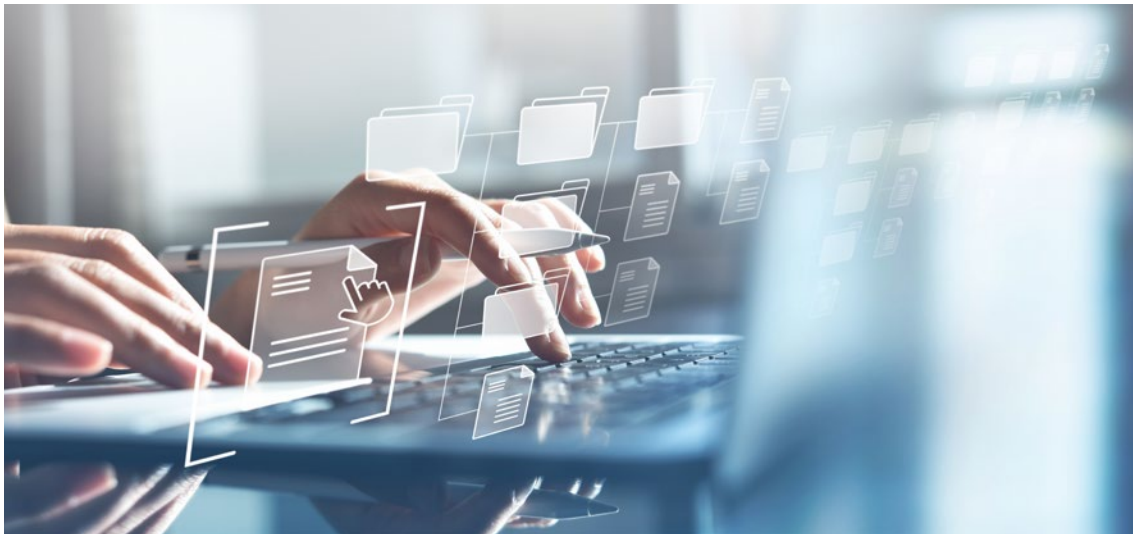
3) Zero-trust storage

Zero-trust storage addresses various security concerns, including data leakage, integrity damage, and data unavailability on campuses. Zero-trust storage treats all data access and operations as unauthenticated, and implements minimum authorization for access subjects, data, and data operations using technologies like mandatory data access control and full-path data encryption. Continuous verification and dynamic authorization are used to implement the finest-grained data access control.

- Mandatory data access control: Fine-grained data access control follows the principle of minimum authorization. It maps the characteristics of data access subjects, data attributes, and fine-grained data processing actions to ensure that only subjects can access and use the finest-grained data set under certain conditions. To ensure the accuracy and consistency of complex policies, technologies like formal verification, automatic policy generation, and compliance audit are utilized. These technologies address issues such as large-scale formal verification performance, automatic policy generation mechanisms, and complex rule matching.
- Full-path data encryption: In the data security system based on campus service boundaries, the security assumption of full-path data may cause data leakage. Encryption can be adopted for full-path data processing, including memory, storage I/O, network I/O, and cache. Native data security capability can be shared through centralized key management to prevent data leakage.

4) Hardware-based algorithm

By 2030, many valuable applications will be developed using campus data, and numerous AI models will be introduced to empower campus services. However, this will also bring about security challenges concerning both AI models and data. Therefore, the key focus will be on improving the security and robustness of AI models.



- Dedicated hardware acceleration: For computing-intensive tasks in the data security field, algorithm characteristics will be deployed on dedicated hardware, and the computing efficiency is expected to be improved through optimized instruction sets and heterogeneous computing resource scheduling, which will minimize the impact on system service performance. For example, AI-based threat detection will move from general-purpose computing processors to dedicated security hardware, improving detection throughput and accuracy.

- Trusted and secure hardware: Creates a heterogeneous trusted computing environment and implements secure computing, access, and audit under the heterogeneous computing architecture through software and hardware collaboration for post-quantum algorithms. When dealing with sensitive information and defending against complex attacks, hardware modules are often combined with software security measures to create a complete campus data security solution.

Computing Power Security

Intelligent twins and spatial interaction technologies will develop rapidly over the coming years with widespread adoption across different campuses. Built-in computing power security will offer a range of security mechanisms for products and solutions. It also works toward enhancing campus computing security capabilities in areas such as security for new computing paradigms, digital trust and privacy, and AI security and trustworthiness.

1) Security for new computing paradigms

In campus data centers, computing power is shifted to the memory, rendering traditional memory encryption mechanisms and hardware-based privacy computing technologies ineffective. Even when data is encrypted and processed at the application layer, it remains in plaintext, making it vulnerable to theft by privileged users and processes. To prepare for the future of campus services, which includes holographic AIOC, generative AI, ultra-immersive interaction, and metaverse scenarios, the main objectives

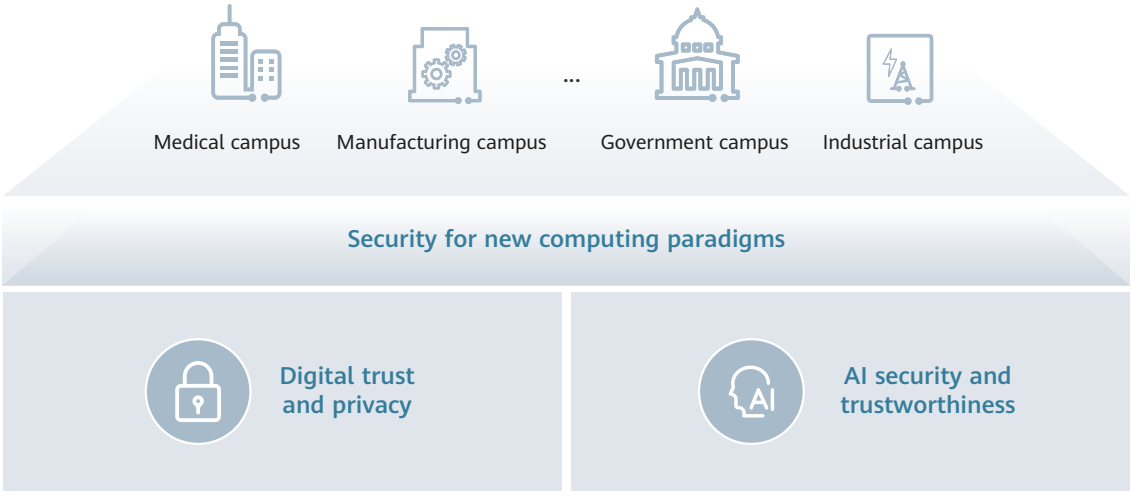


Figure 3-19 Computing power security

are to ensure the security of diverse computing power and efficient and secure collaboration between computing power and peripheral devices. The mainstream approach to achieving these objectives involves combining security with the in-network computing architecture, the diversified computing architecture, and the data-centric peer-to-peer computing architecture.

- Security + In-network computing architecture: After the zero-trust architecture breaks the security boundary, a finer-grained permission and access control mechanism should be adopted to support dynamic identity authentication and resource access policies. That means software will consume a large amount of CPU resources. However, an in-network computing architecture that uses hardware acceleration mechanisms for regular expressions can make policy execution 10–15 times more efficient.

- Security + Diversified computing architecture: To efficiently adapt to scenarios that require high computing power, such as large models and big data, CPU-centric confidential computing technology will evolve into data-centric heterogeneous confidential computing technology. This new technology is compatible with existing large model software frameworks and supports diverse computing power devices, such as GPUs, DPUs, and NPUs, to accelerate and collaborate with security computing power for confidential computing. In this way, these three issues can be resolved:
  - » The security computing power will be fully compatible with the common computing power ecosystem and can be flexibly configured.
  - » Security computing will extend from the trusted execution environment (TEE) of CPUs to diversified computing power devices.
  - » The security computing power can be scheduled flexibly and the computing power resources can be centrally managed.

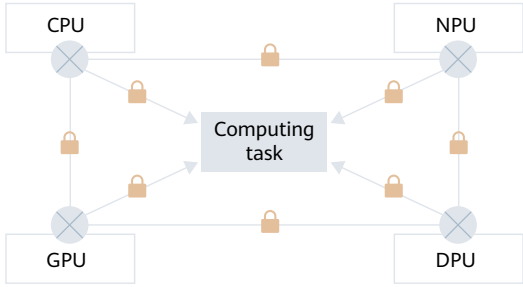


Figure 3-20 Security and diversified computing architecture

- Security + Data-centric peer-to-peer computing architecture: In a data-centric peer-to-peer computing architecture, high-performance SCM will connect with the memory bus in the system. This prevents residual data in the memory from a privacy breach after a power-off.

2) Digital trust and privacy

Data security computing has emerged to ensure the privacy and security of campus service data during computation. The main technical directions of data security computing include: TEE based on hardware security, homomorphic encryption and secure multi-party computation, and multi-party computation built on the sharing of secret slices between multiple parties. These technologies provide efficient protection for data privacy and strictly prevent privacy leakage and damage.

- TEE: The privacy computing environment based on hardware TEE can prove the completeness of data and its own innocence, avoiding security vulnerabilities. TEE is a core technology for computing security and is expected to be used in more than half of all computing scenarios by 2030.

- Homomorphic encryption and secure multi-party computation: They are considered to be the most ideal privacy computing technologies because it is possible to verify their technical security level mathematically. Approximation computing is maturing, and homomorphic encryption and secure multi-party computation have already been applied in image identification, the sharing of health data, and other specific domains.

- Multi-party computation built on the sharing of secret slices between multiple parties: TEE technology can greatly improve the performance of multi-party computation, while being used to enable the sharing of secret slices between multiple parties and eliminate data silos. In addition to that, security can be proved mathematically based on TEEs.

3) AI security and trustworthiness

AI systems require a large amount of data for training and inference. To fully protect the privacy of this data, it is critical to use appropriate protection technologies. The development of AI will accelerate intelligence in services such as smart office, asset operations, digital twins, and ultra-immersive interaction. Additionally, technologies for the protection of AI models and training data, AI attack detection and defense, enhanced AI security, and AI regulation will be constantly improved.

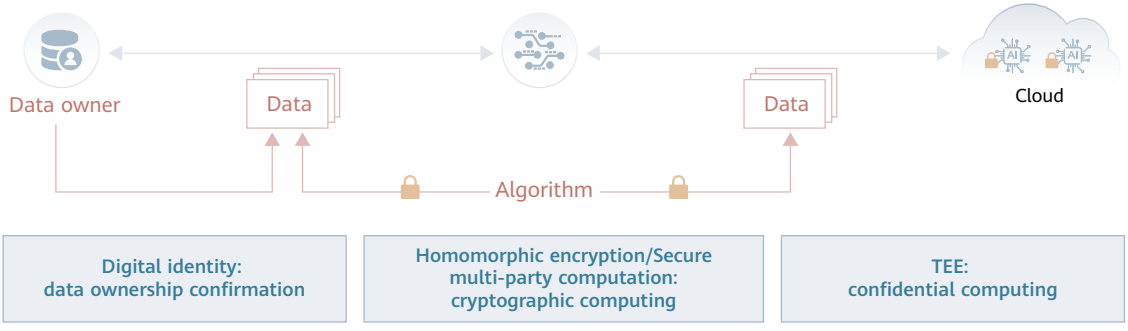


Figure 3-21 Digital trust and privacy

- Protection of AI models and training data: Encryption, mandatory access control, security isolation, and other mechanisms must be implemented to ensure security of AI models and training data throughout the data lifecycle, from collection and training, to inference. The memory encryption algorithms and architecture design for a memory hardware encryption engine on NPUs provide high-bandwidth, real-time, and encrypted memory data processing capabilities.

- AI attack detection and defense: Adversarial sample detection models should be implemented to better identify physical and digital evasions and other attacks on AI models, block attack paths, and prevent misjudgment when AI models are attacked.

- Enhanced AI security: This can avoid damage to AI models caused by unknown attacks, which can be achieved by enhancing model robustness, verifiability, and explainability. Adversarial training can improve the attack defense capability of AI models, and regularization of models can improve the generalization capability of adversarial samples.

- Compliant with AI regulatory requirements: AI models should also be continuously monitored and audited to comply with AI regulations. Blockchain, data capsule, and other related technologies can be used to ensure reliable audit results and the real-time tracking of issues.

Resilient Systems

Security refers to a system's ability to avoid attacks, while resilience refers to its ability to minimize damage when attacked. Resilience can be broken down into four sub-capability goals: anticipate, withstand, recover, and adapt. These goals can be achieved by implementing a series of technical control over the system.

To achieve campus service resilience, a deterministic model is needed to narrow down the attack surface, improve defense success rates, reduce the uncertainty of threats, and strive for more time for defense. The fundamental technologies of the campus resilient system architecture include intrinsic security, three-dimensional defense, and resilient architecture.

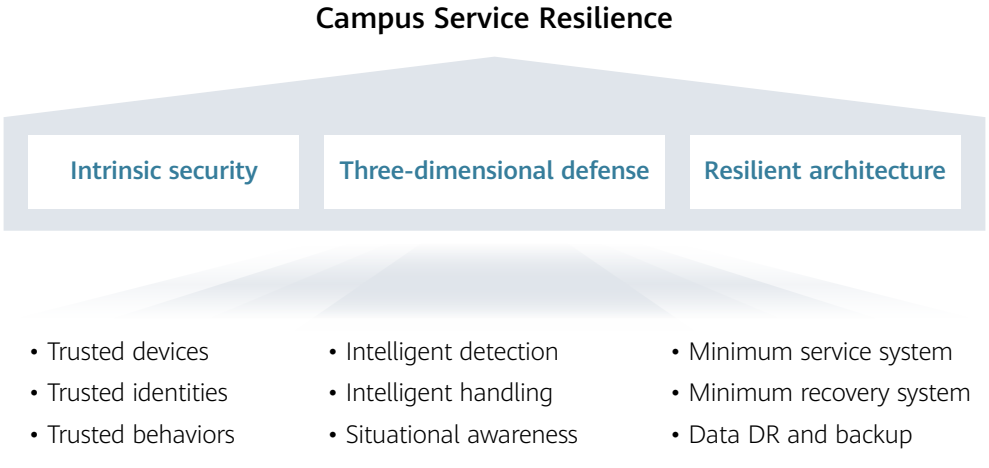


Figure 3-22 Resilient systems

1) Intrinsic security

A trusted environment can be established based on trusted devices, which involve trusted computing, trusted storage, trusted networks, and trusted applications. This will enable trusted device startup, trusted measurement, and remote attestation. Zero-trust access identity management is based on a trusted identity, which includes entity authentication, certification authorization, and mandatory access control. With trusted behaviors, default trust based on entity attributes will give way to verifiable trust based on service behaviors. This will help establish a deterministic model for campus services, strengthen the campus security bottom line, and narrow down the attack surface.

2) Three-dimensional defense

To ensure security for campus services, a three-dimensional and systematic approach is needed. Strenuous defense against endless threats will give way to deterministic assurance, and the key is to develop multi-layer defense capabilities in this process. The first line of defense for security is passive measures, which are based on the physical environment (campus security management regulations), network borders, devices and subsystems, application data, and detection of unknown threats. The second line of defense for security is proactive measures, which include attack simulation, proactive deception, and blocklist and trustlist creation. To increase the probability of successful defense, reduce threat uncertainty, and increase time for defense, a third line of defense should be built based on intelligent detection, intelligent handling, and situational awareness.

3) Resilient architecture

Intrinsic security and three-dimensional defense fall short in addressing the full spectrum of security concerns, particularly advanced persistent threats (APTs). While defense can fail, mission-critical services should be provided with deterministic assurance instead of constantly confronting threats. Resilience functions as the main measurement dimension, which involves using a minimum service system to downgrade management after an attack and ensure core services. Additionally, a minimum recovery system is used to automatically restore core services on the campus after severe damage to the system. Data is backed up using disaster recovery and backup methods to ensure that system service data can be restored to the last saved update.

3.1.6 All-Domain Zero Carbon

As carbon emissions are most concentrated on campuses in city environments, campuses should take the lead in reducing their carbon footprint. Campuses will evolve toward being low-carbon from multiple perspectives: First, continuous innovation in ICTs will lead to the optoelectronic reconstruction of basic campus devices, resulting in a simplified zero-carbon campus. Second, the use of energy routers, energy cloud brains, PV generators, and new energy components will bring about ultimate carbon efficiency. Third, generation-grid-load-storage synergy of energy systems, which is supported by smart microgrid, will bring a new green power supply and demand order with carbon as the trading unit.

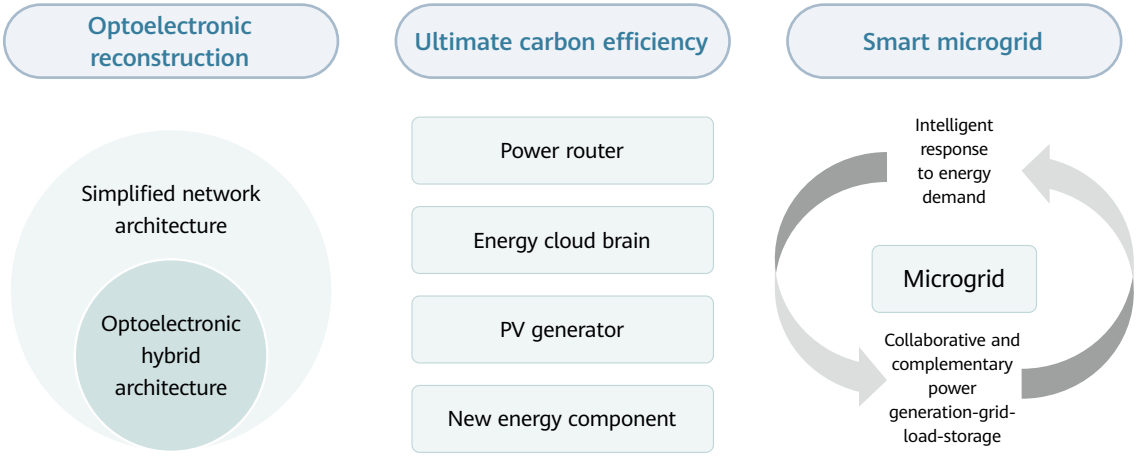


Figure 3-23 All-domain zero carbon

Optoelectronic Reconstruction

Driven by novel AI and big data analysis applications, campuses have higher requirements for data transmission bandwidth and urgently need to improve data transmission energy efficiency. As we approach the end of Moore's law and new technologies such as quantum computing have yet to mature, there has been a bottleneck in the continued improvement of computing, storage, and network energy efficiency. This makes it challenging to create a green and low-carbon network through basic technical innovation and expand the network capacity dozens of times while maintaining the same energy consumption in the next decade. We can introduce a campus network with a simplified network architecture and a new optoelectronic hybrid architecture to make the network green and low-carbon.

1) Simplified network architecture

Traditional networks are divided by services, with multiple coexisting networks leading to complex O&M. This model is increasingly difficult to adapt to the development of automated and intelligent networks. In the future, campus networks need to be reconstructed with an aim to achieve zero-carbon goals. A simplified three-layer network

architecture consisting of intranet, cloud network, and computing network will have to be built.

- Campus intranet: The campus intranet will change from two dimensions: network architecture and terminal access media. The number of physical network layers, ELV rooms, and remote power supply modules can be reduced by adopting passive and optoelectronic hybrid cable technologies. This will simplify the network architecture. Campus intranet can gradually evolve from one network to one device through simplified O&M logic and fewer O&M nodes. Under the trend of Ethernet-based bus, single-pair Ethernet (SPE) communications will be sped up and be simultaneously accessed by multiple applications in a serial manner. This allows it to be widely used for meters, switches, HMI, audio and video sensors, industrial edge, and executors.

- Campus cloud network: The campus cloud network is overlaid on the intranet using E2E slicing technology. It enables agile and open virtual networks that provide SLA assurance, and supports connectivity between the cloud and devices at the tenant level. It can be used for multiple purposes to increase network utilization and save network energy.

- Campus computing network: The computing network is used for connecting data and computing power at the service level, and provides computing power routing services and trustworthiness assurance for data processing. It is constructed based on distributed and open protocols. Through flexible scheduling of data, the computing network enables green, centralized multi-level computing power infrastructure that has a reasonable layout.

The three networks are interdependent. The computing network depends on the cloud network to enable agile building of virtual pipes and open interfaces that can be provisioned on demand, so as to provide real-time, elastic connections between data and computing power. The computing network also needs the support of the campus intranet to enable its most important features: low latency and high bandwidth.

2) New optoelectronic integration architecture

- Optical media: Traditional copper lines often suffer from signal attenuation and distance limitations, whereas fiber networks enable high-speed and long-distance transmission. If we replace traditional network cables with optical fibers, campus networks can obtain higher transmission speeds and stability. Fiber lines have lower transmission power and consume less energy compared to traditional copper lines, resulting in reduced energy consumption and carbon emissions on campuses. They also have longer lifespans, require less maintenance and replacement, and generate less waste, contributing to a circular economy.

- Optoelectronic integration architecture: In the next decade, the development of new products, such as optical input/output chips and CPO, will improve network electronic components' high-speed processing capabilities and reduce their power consumption. Coherent optical technologies will be applied to extend the transmission distance of high-speed ports on datacom equipment. With the development of 5G-A/6G, new types of antennas that directly connect to optical fibers will be used to reduce the weight and power consumption of base stations.

Optoelectronic integration is the way forward for structured improvement of equipment energy efficiency. CPO chips based on optical buses are expected to be commercially used. Some academic institutions are researching optical cell switching technology that could potentially replace electrical switching networks. Equipment-level optoelectronic integrated products using optical buses and optical cell switching technology are expected to be developed by 2030.

Optoelectronic integration technology at the network, equipment, and chip levels can continuously improve the energy efficiency of communications equipment, and meet the green campus network objective of increasing network capacity without increasing energy consumption.

Ultimate Carbon Efficiency

Future intelligent zero-carbon campuses will utilize cutting-edge digital technologies that rely on energy routers and energy cloud brains and combine PV generators with other new energy components to achieve coordinated scheduling of multiple energies and cascade utilization of them. This will reduce the levelized cost of electricity (LCOE). The traditional energy management approach, which mainly involves monitoring and O&M, will be replaced by closed-loop energy management that emphasizes execution, decision-making, and operations. This shift will enhance overall energy utilization and contribute to the development of intelligent zero-carbon campuses.

1) Energy router

There are many types of energy supply and consumption systems on campuses, each of which includes numerous devices. For example, the power supply system includes distributed PV devices, wind power devices, and electrochemical energy storage facilities; the heating system includes combined cooling, heating and power (CCHP) supply devices,

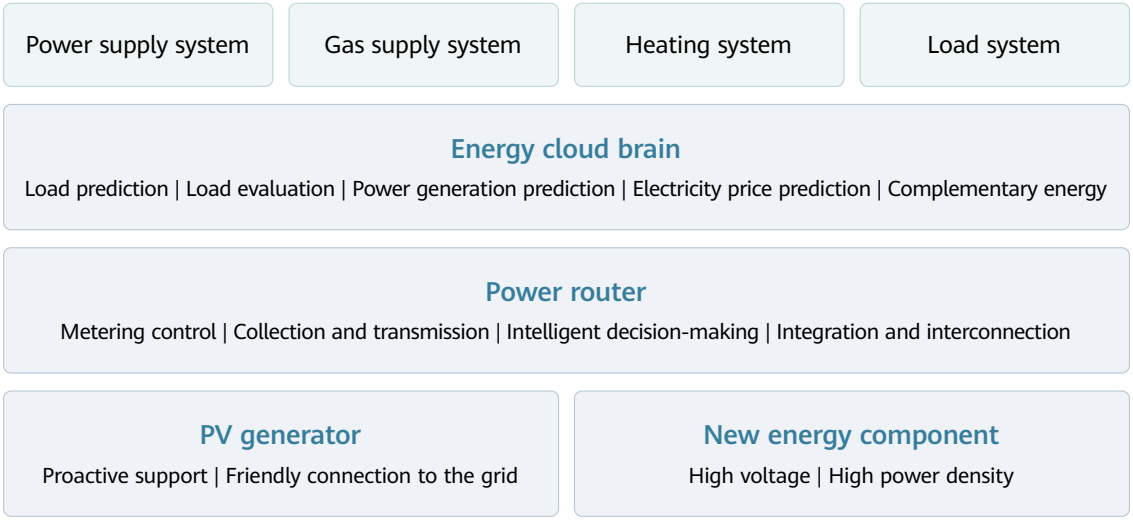


Figure 3-24 Ultimate carbon efficiency

heat pumps, and heat storage facilities; the gas supply system includes gas supply stations and hydrogen energy storage facilities; and the power load system includes factories, buildings, electric vehicles, and street lamps. Energy routers capitalize on power electronic conversion and control technologies to provide diverse electrical interfaces for various devices, and adopt standardized protocols to let different devices be integrated and interconnected. As a result, energy routers become core devices and energy hubs in microgrids. When combined with digital technologies such as 5G, energy routers will be capable of communications and intelligent decision-making support in addition to their basic metering and control functions. Energy routers can collect and transmit information such as device running status and energy usage in real time to implement unified data collection and classified metering, in addition to actively or independently managing the energy flow direction and power as instructed by users or dispatching centers.

The energy consumption data that energy routers collect and summarize in real time can be used to accurately capture and track carbon footprints. Measurement factor rules can be preset to accurately measure and monitor a campus' total carbon emissions in real time, ensuring that the data is reliable and trustworthy. This lays a foundation for IOC-based visualized carbon management, carbon quota management, and carbon asset trading.

2) Energy cloud brain

Massive energy consumption data of various loads on campuses is collected in real time and classified for measurement. Energy consumption models of different loads are trained, which, together with big data analysis and edge computing, helps intelligently control and optimize energy consumption behaviors such as air conditioning and lighting on campuses, achieving energy saving.

AI algorithms are used to predict the output of various distributed energy supplies and the load of various energy consumption systems on a campus. This data is then combined with various other factors such as meteorological prediction, power price changes, and requirements of power users to calculate a global intelligent dispatching solution. When the power supply exceeds the demand, any spare power can be reused multiple times (level by level) through energy transformation technologies and various energy storage devices. In addition to directly storing the spare power through electrochemical energy storage, the power can be converted into hydrogen energy through new technologies such as Power2Gas. This has two benefits: One is storing energy using hydrogen storage facilities, and the other is reducing gas procurement costs through a hybrid gas supply. When the power supply is less than the demand, the stored power can be released, or redundant heat energy and hydrogen energy can be transformed into power. By combining various energy sources such as electricity, heat, and gas, campuses can enhance their ability to be self-sufficient in energy supply.

3) PV generator

In a campus power grid, fossil fuel power plants and hydropower plants typically use conventional synchronous generators. These synchronous

generators utilize mechanical structures to provide stable voltage and frequency, thus facilitating frequency regulation and voltage control. However, as asynchronous generators gradually displace synchronous generators in power grids, the way power systems work will change fundamentally. In response, renewables-based power systems will need to simulate the technical indicators of synchronous generators, in order to proactively support the grid's frequency and voltage fluctuations. The goal will be to help power grids become safer and more reliable.

PV power generation technologies for campuses combine power electronics, energy storage, and digital technologies to simulate the electromechanical transients of synchronous generators. When connected to power grids, PV generators have many of the same external characteristics as synchronous generators, such as inertia, damping, primary frequency regulation, and reactive voltage control. As a result, PV generators can offer technical specifications that are similar to the synchronous generators used in fossil fuel power plants. PV power generation technologies for campuses can proactively support the operations of renewables-based power systems and make them more grid-friendly. PV power generation technologies for campuses provide a solid technical foundation for incorporating renewables into power grids, strengthen the resilience of the campus power grid, and ensure that it is safer and more reliable.

4) New energy component

As more campuses climb onboard the green and low-carbon initiative, their LCOE of green electricity will continue to become a key focus for investment models. By 2030, the LCOE of PV plants is expected to plummet, possibly even down to US$0.01 per kWh of electricity.

- Green electricity generation systems will be able to support higher voltage. As input voltage increases, so does output voltage. This in turn can reduce line loss in direct current systems and loss in low-voltage transformer winding, significantly increasing the systems' efficiency. In addition, solar inverters and transformers will become more compact, translating into a huge reduction in transportation and O&M workloads. Green electricity system maintenance will also be automated. Thanks to these trends, by 2030, the green electricity generation system's voltage will reach 1500 V or even higher, further slashing LCOE.

- Solar inverters will deliver higher power density because of advanced materials like SiC and GaN, better heat dissipation in chips, and topology technologies. These materials and technologies increase solar inverters' voltage, operating temperature and frequency, and reduce loss. By 2030, solar inverters will see their power density grows by over 70%.

- Intelligent green power plants: By 2030, AI is expected to be used in 90% of green power plants. As digital and green electricity technologies converge, they will make O&M, production, and asset management simpler, more intelligent, and more efficient. AI will handle the tasks that used to be performed by highly-trained experts, and support autonomous and collaborative optimization inside green power plants. Intelligent tracking algorithms make it possible for components, trackers, and solar inverters to work in tandem to improve efficiency. Fault location will be more precise and O&M times can be reduced from months to minutes.

Smart Microgrid

To reach carbon neutrality targets, campuses will increasingly utilize renewable energy sources such as solar, wind, and hydrogen power. However, these sources are unable to supply stable power over a long period of time. To address this issue, a smart microgrid with collaborative generation-grid-load-storage can be built to integrate power generation, transmission, distribution, storage, and usage based on digital and power electronics technologies. Combined with a socialized intelligent energy demand response, this smart microgrid will improve resource utilization on campuses.

1) Collaborative and complementary power generation-grid-load-storage

A microgrid is a small power generation and distribution system composed of distributed power supplies, energy storage and conversion equipment, loads, monitoring systems, and protection equipment. The microgrid allows campuses to consume local green energy when it is sufficient, avoid energy loss during transmission in the power grid, and improve energy utilization efficiency. It can be connected to the power grid through a single point and obtain power from the grid when the energy supply is unstable. The microgrid will make renewable energy more observable, measurable, controllable, and adjustable, increasing new energy consumption and creating new energy access systems that are less vulnerable. It will also enhance the group control and regulation capabilities of massive terminal systems, allowing power generation units and users to interact with each other in real time.

As the proportion of power that comes from renewable sources increases, the campus microgrid featuring collaborative "generation-grid-load-storage" can further reduce carbon emissions and work toward achieving the zero-carbon goal.

2) Intelligent response to energy demand

By utilizing a software platform, an intelligent response to energy demand can effectively bring together power equipment, energy storage equipment, and controllable loads that are dispersed throughout different campuses without the need to alter the physical network architecture. This approach allows for flexible scheduling, which enables efficient interaction with the microgrid. As a result, power supply and demand can be balanced across different spaces and times, leading to improved power grid security and increased consumption of renewable energy.

Intelligent decision-making for scheduling is a crucial aspect of responding intelligently to energy demand. By utilizing AI and big data, it is possible to reasonably and efficiently schedule power and energy storage equipment, which can help campuses accurately predict and dynamically optimize their energy usage. This approach allows campuses to function as a virtual power plant, supplying power and consuming surplus power, which ultimately leads to a balanced operation of the campus power grid and improved energy utilization.

3.2 Reference Architecture

By 2030, the positioning and architecture of intelligent campuses will have changed significantly. These changes will be driven by diverse digital and intelligent needs, various national and international policies, and ongoing innovation in fields related to intelligent twins, spatial interaction, ubiquitous intelligent connectivity, intent-driven ultra-broadband, security and resilience, and all-domain zero carbon. Campuses will transition from being closed, isolated, and autonomous entities to intelligent and connected social spaces which serve the local community. They will shift from independent resource management to more refined and efficient resource management of both internal and external applications. There will be higher demand for on-site cross-domain data processing and convergence based on the assumption that large-scale data flows will be much more secure. However, these changes have presented some challenges for the construction of intelligent campuses. To overcome the challenges and facilitate the evolution to intelligent campuses by 2030, Huawei proposes a reference architecture with six new features.

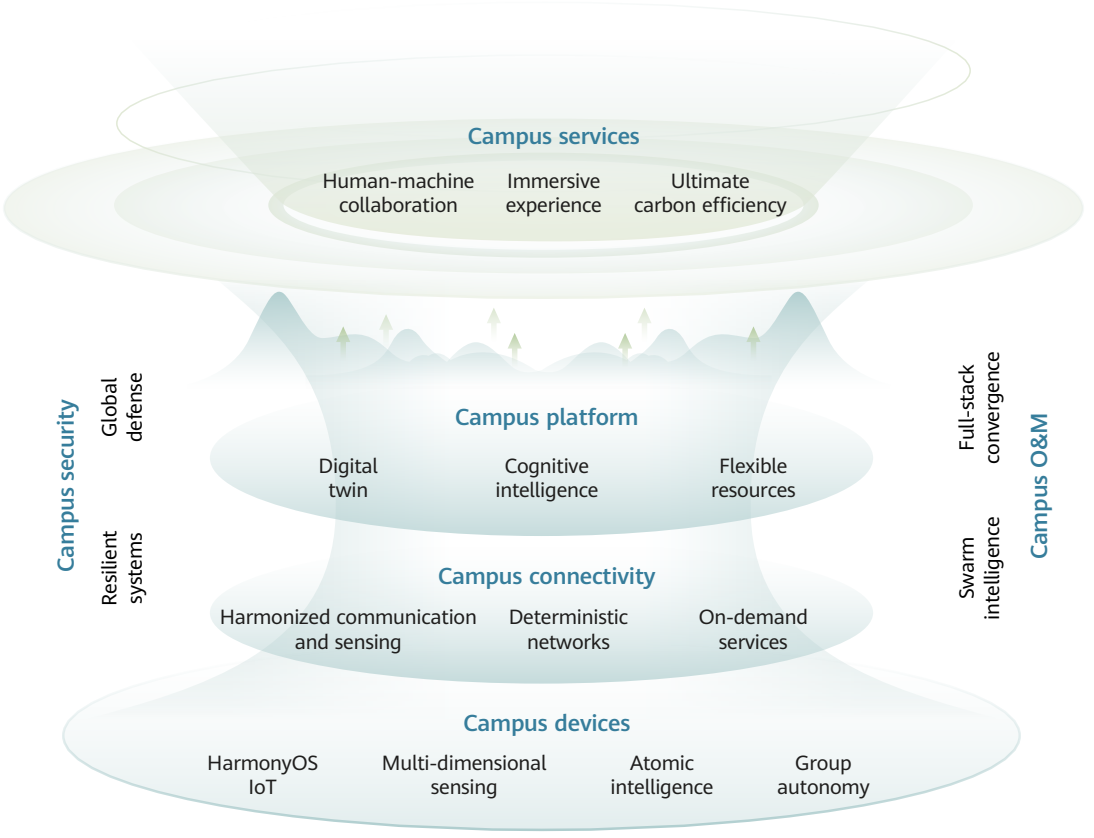


Figure 3-25 Reference architecture

3.2.1 Campus Service: People-Oriented, Digital, Green, and Low-Carbon Experiences

The intelligent campuses of the future will leverage big data, cloud computing, AI, and IoT to better serve people and to optimize the physical and virtual relationships between people, intelligent campuses, and smart cities. Future development directions may include unstaffed on-site operations, immersive experiences, digital services, and green and low carbon features.

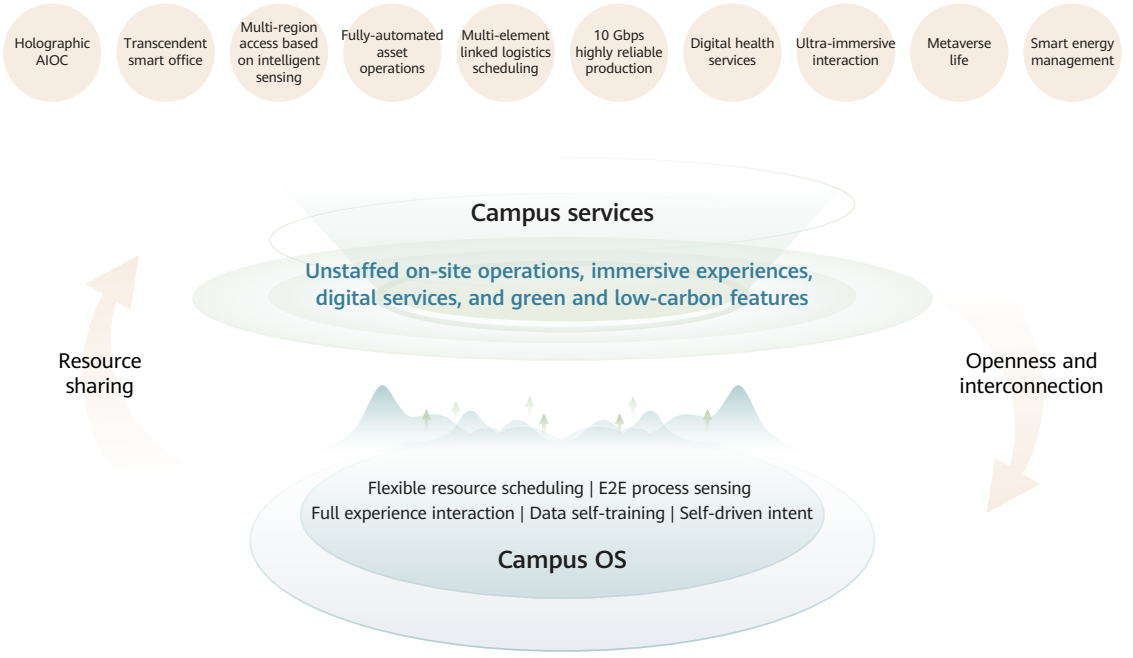


Figure 3-26 Campus services

Unstaffed On-Site Operations

Next-generation campuses will deploy robots to provide universal support for campus operations, BA, and community services. For example, robots will be responsible for basic, repetitive, and dangerous work such as monitoring the campus environment, cleaning, security, and logistics transportation. AI-based intelligent environment-monitoring robots can automatically optimize office and production environments. Personalized digital identity and intelligent access control systems provide an efficient and comfortable travel experience on campuses.

Immersive Experiences

Generative AI, IoT, and virtual interaction technologies such as converged holographic projection, naked-eye 3D display, and immersive spatial projection will be deployed on a large scale. There will be immersive, highly interactive, physical-virtual experiences both online and offline, such as holographic conferences and virtual workshops, negotiations, and exhibitions. These are likely to significantly increase collaboration efficiency.

Digital Services

Next-generation campuses will be able to aggregate the basic running data of service departments, buildings, and factories on campuses, streamline data resources within the campus and between campuses, and build a digital twin model for campus governance and different service scenarios. AIGC has committed to forecasting trends and performing risk assessments with a human touch to assist campus managers with their decision-making.

Green and Low-Carbon Features

Building information management (BIM) will be integrated with AI derivative design to represent the entire campus construction process using a three-dimensional model and implement refined management of carbon emissions on campuses. In addition, the AI-based energy forecasting system will optimize the supply and procurement decisions related to sustainable energy to achieve a balance between low carbon energy and cost-efficiency. This will ensure campus energy systems continue to function smoothly.

New campus services require a more open architecture to flexibly schedule all service resources, monitor the service process from start to finish, facilitate virtual-physical collaborations and interactions, and implement data self-training. They also need to perform centralized planning based on unified standards and future plans. This allows next-generation campuses to be integrated into smart cities and enables them to work toward building an open and mutually beneficial economic model. This facilitates the exploration of composite, intelligent, and green campus operations to support rapid digitalization across a wide range of industry campuses.

3.2.2 Campus Platform: Digital Platform with Intelligent Twins, Flexible Resources, and Citizen Development

The campus platform was designed to help campuses go digital. It consists of four parts: platform foundation, platform kernel, platform services, and development enablement.

- Platform foundation: This provides efficient computing power for campus services and stores massive amounts of data.
- Platform kernel: This offers the connection bus between campus devices and subsystems and flexibly schedules resources.
- Platform services: This layer includes data, twin, and AI services that support campus digitalization.
- Development enablement: This allows digital campus services to be developed by all.

The platform kernel, platform services, and development enablement constitute the campus digital OS, which manages and coordinates software and hardware resources on the platform foundation, standardizes people-thing-event connectivity models and service interfaces on campuses, and builds a campus digital twin. The campus digital OS supports efficient construction of intelligent applications, as well as digital innovation and sustainable development across different campuses.

The campus platform in 2030 will feature intelligent twins, flexible resources, and citizen development.

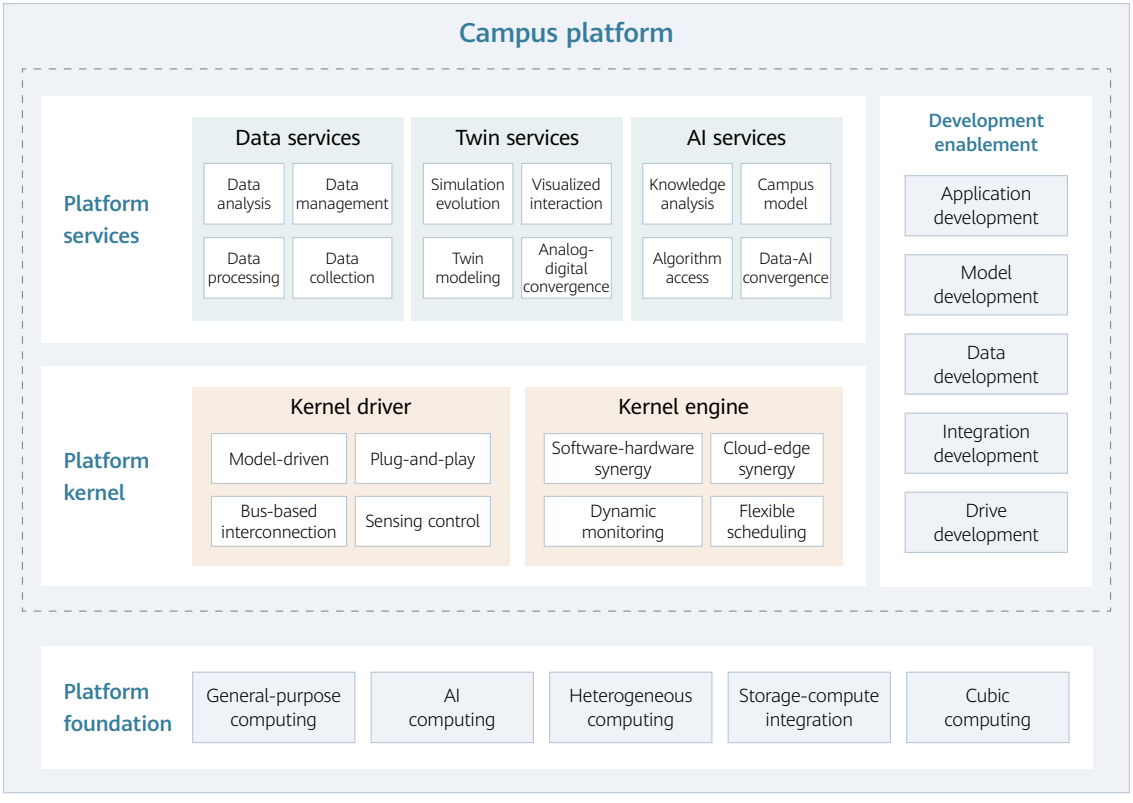


Figure 3-27 Campus platform

Intelligent Twins

- Intelligent platform service: The campus AI foundation model targets L1 scenarios. It obtains and analyzes a multitude of campus data to accumulate scenario-based knowledge as well as provide its capabilities for intelligent applications, meeting requirements of various business types on campuses.

- Platform twin service: Interconnection buses for people, events, and things, as well as IT, OT, and CT systems on campuses will be established. Together with standardized digital models and centralized service interfaces, they precisely sense and control the physical campus world in real time, and share campus-wide data. On-demand connectivity services, data services with intrinsic security, and twin services for the interaction between physical and digital worlds will also be available to build a highly digital and intelligent campus digital twin.

Flexible Resources

- Campus platform foundation: This provides large-scale AI computing power, massive storage, parallel computing, alongside device-edge-cloud cubic computing and serialized computing power. It also supports CPU/GPU/NPU/FPGA computing power pools.

- Platform kernel engine: Dynamic monitoring, on-demand loading, and flexible scheduling efficiently allocate and schedule compute resources. In addition, hotspot prediction, compilation optimization, and software-hardware synergy maximize computing efficiency.

Citizen Development

Development enablement: As digital transformation advances and becomes more widespread, campus users are looking for more agile and efficient ways to develop applications and AI models, harness the power of data, and integrate different systems. The traditional development mode that relies on professional IT personnel can no longer support campus service innovation. The low-threshold and intelligent development production line has been designed to optimize the work of service personnel. It includes no-code, composable, and generative development, as well as robot process automation. It helps campuses go digital faster by delivering a superior experience through intelligent, simplified, and innovative citizen development.

3.2.3 Campus Connectivity: Information Highway That Provides Flexible Convergence, Intelligent Experiences, and Agile Services

Campus connectivity is required to facilitate the provision of services, upload digital campus information, and deliver intelligent commands. Increasing campus digitalization and intelligentization will lead to a 100-fold increase in the number of sensing devices. A wide variety of intelligent terminals and cloud-edge synergy applications require more high-quality connections. Campus connectivity needs to support flexible access by numerous devices and subsystems to deliver deterministic SLAs for various campus applications and ensure premium application experience. In addition, service-oriented capabilities will become essential as they support changes to the businesses available on campuses and flexible expansion. Automatic and intelligent methods can be used to simplify O&M.

Flexible Convergence

- Ultra-broadband converged coverage: The campus network should support ultra-high bandwidth access, such as 100 Gbit/s Wi-Fi, 100 Gbit/s optical fiber, and 100 Gbit/s mobile network. The IPv6 address space allows the access by millions of terminals. Wireless, wired, satellite, and other communications technologies achieve multi-dimensional coverage of space, air, and land, providing universal connectivity on campuses.

- HCS network: Communications networks will be able to sense environments thanks to WLAN, optical, and 5G-A wireless sensing. For example, fiber networks will be able to sense

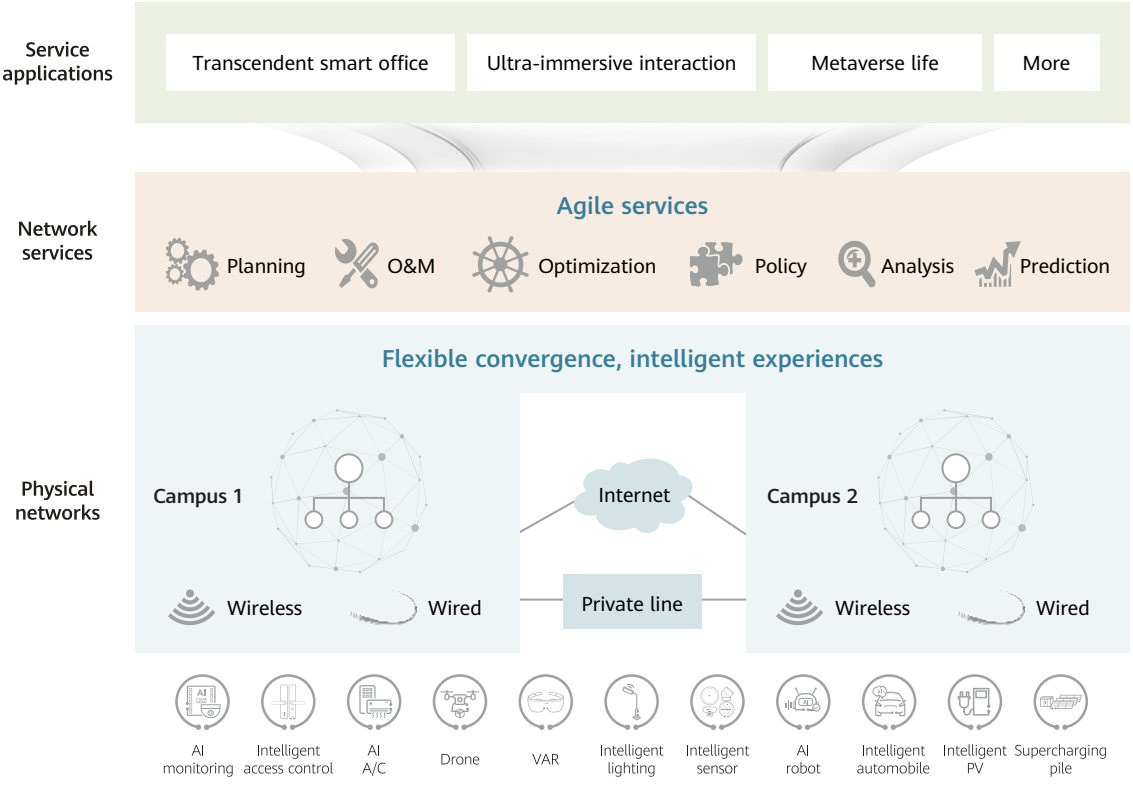


Figure 3-28 Campus connectivity

vibrations or stress, spectrum networks will be able to monitor water quality and identify gas leaks, and wireless networks will be able to sense location or motion. Joint scheduling of communications and sensing resources provides more methods to implement campus digital twins and facilitate intelligent decision-making.

Intelligent Experiences

Different services require different bearer network SLAs. E2E network slicing can ensure deterministic SLAs for various services and deliver a premium experience.

- Consistent office experiences: Touch-free device access and centralized policies that allow access from different locations improve mobility and security.
- Audio and video service experiences: The network burst bandwidth supports zero wait times, zero frame freezing, and zero latency.
- Auxiliary production service experiences: Continuous wireless coverage and zero packet losses during handover enables seamless roaming for mobile services like AGVs. The bandwidth reaching 10 Gbit/s helps VR, AR, the metaverse, and other applications evolve from auditory and visual digitalization to digital simulation of multiple dimensions, including hearing, sight, smell, and touch. The latency below 10 ms eliminates lag and dizziness during fast movement.
- Superior core production service experiences: The campus connectivity network that converges IT, OT, and CT offers 99.9999% reliability alongside deterministic latency and jitters (an OT-specific industrial control system < 100 ms, a PLC control period of 2–5 ms, and machine motion control < 1 ms), which ensures efficient automated production.

Agile Services

- Software-defined networking (SDN): On-demand connectivity generation allows for fast access and expansion of new devices and subsystems on campuses. Minute-level link provisioning enables fast cloudification of new campus services. On-demand adjustment of network resources guarantees experiences of critical services such as campus office conferences and automated production.
- Autonomous driving network (ADN): An ADN visually displays the digital health of networks, predicts network deterioration and faults, links with the alarm monitoring system for joint troubleshooting and ticket dispatching, and supports network self-optimization/self-healing based on policy control. The campus network reaches L4. Based on preset policies, it can proactively make decisions based on monitored key events.
- Network as a service (NaaS): Atomized network functions along with open network capabilities through service-based interfaces support flexible orchestration. Application requirements will interact with network capabilities and policies as well as user attributes to quickly meet diversified and customized service requirements and improve user experience from start to finish, quickly innovating more services.

3.2.4 Campus Device: Plug-and-Play, Collaborative Autonomy, and Agile Access

Future campuses will embrace full-connectivity, full-sensing, and full-computing. Digital all-campus devices will have to tackle numerous challenges, such as how to implement standardized fast access and interactions for countless devices, what to do when manual orchestration of massive amounts of device data fails to enable experience-based interactions, and how to cut cabling costs while providing services anytime and anywhere for IP-based devices. Huawei has designed a range of HarmonyOS-powered, autonomous, and wireless devices for the intelligent campuses of the future.

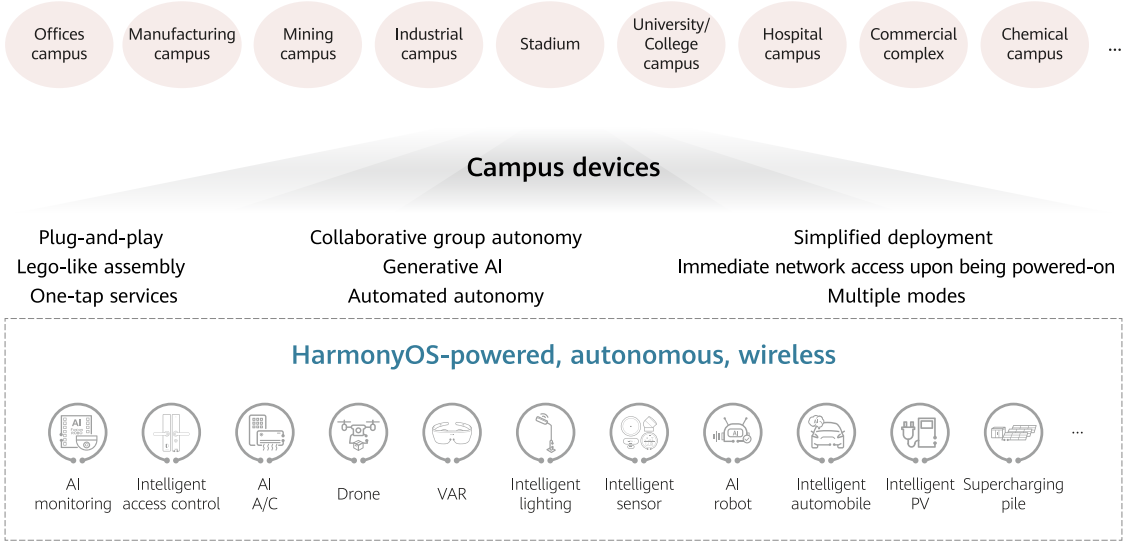


Figure 3-29 Campus device

HarmonyOS-Powered

HarmonyOS is equipped with unified thing models which can facilitate the implementation of plug-and-play devices and digital OSs on campuses. HarmonyOS can enable resource sharing between devices by automatically coordinating the surrounding resources for Lego-like assembly. For example, cameras can work with temperature and humidity sensors to extend their sensing capabilities. In addition, HarmonyOS-powered hardware collaboration supports one-tap services like tap-to-print using mobile phones.

Autonomous

Combining the current AI capabilities with breakthroughs in key technologies like intelligent cognition can make campus devices more autonomous, which means they will be capable of self-supervision and self-feedback through cloud-edge synergies. With autonomous AI, devices will achieve collaborative group autonomy. For example, lighting, air circulation, and air conditioning systems can deploy temperature, humidity, and light sensors to automate their functions in each area. Devices will be able to provide personalized services by tracking users' habits and implementing cloud-edge training and learning. Also, printers will automatically print the conference schedule for the current day. Moreover, devices will be able to automatically detect issues and attempt to adjust and repair themselves. For example, a real-time sorting system based on quality levels can be implemented by the conveyor belt monitoring system in coal mines.

Wireless

Devices will become IP-based, and they will collect more data and have more refined control requirements. To eliminate the need for expensive cabling and complex construction, some devices will use wireless technologies to accelerate their transition to being IP-based. They can immediately access the network after being powered-on, and can provide services anytime and anywhere. Campus devices can use a variety of technologies like Wi-Fi, passive RFID, RedCap, and NearLink to go wireless and agilely access the campus digital OS.

3.2.5 Campus Security: Ongoing Verification, Dynamic Authorization, and Global Defense

Complex and diversified networks brought by campus digitalization have fundamentally reworked the conventional information security management framework. This requires enterprises to tackle the challenges with a fresh mindset. For example, they need to transform the focus of protection from networks to data, break conventional thinking of clearing vulnerabilities with best efforts, and set up deterministic security assurance for campus services. Only in this way can information systems remain deterministic service statuses despite risks, vulnerability exposures, and defense failures.

Campus security must adapt to the ever-changing cloud, network, edge, and device requirements of campus services. Capabilities such as centralized policy management, dynamic resource scheduling, and on-demand service adjustment are essential for on-demand, elastic, and agile use of security services. To ensure service security for campuses in 2030, an all-round campus security system adopting the zero trust architecture (ZTA) will be set up based on three core principles — ongoing verification, dynamic authorization, and global defense.

Global Defense

Network-security collaboration quickly handles threats. The campus security brain evaluates device risks, user behavior anomalies, traffic threats, and application authentication behavior to create a complete trust chain. It also generates handling policies for users or devices with low trust scores and associates with network or security devices for rapid threat handling. This provides a security cockpit that features ongoing dynamic evaluation, visualized situational awareness, and global collaborative handling for campuses.

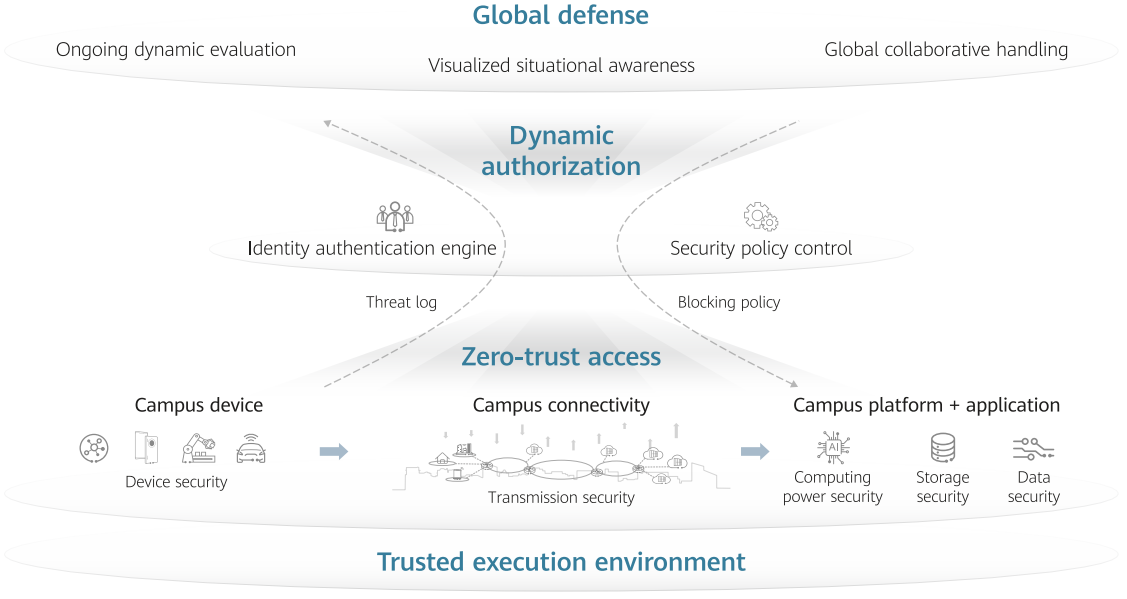


Figure 3-30 Campus security ZTA

Dynamic Authorization

Refined access control allocates permissions on demand. This principle refines permissions of campus service access targets to the application, function, and data levels. Only the required applications, functions, or data can be opened to access subjects, meeting the least privilege principle and significantly preventing potential attacks. Additionally, security control policies dynamically determine permissions based on access subjects, target objects, and environment attributes (device status, network risks, and user behavior) to implement refined and dynamic control over applications, functions, and data. The identity authentication engine is responsible for centralized personnel identity management and authentication, including the management of users, organizations, user tokens, and application tokens, as well as user identity verification. The security policy control engine performs dynamic and refined authentication on data service access requests. When the security level of a user changes, this engine promptly updates the access permission of the user.

Zero-Trust Access

Zero trust sets up a defense line for identity security. Identities of people, terminals, devices, and applications within the campus are centrally managed to build an identity-centric access control mechanism. The identity of the access subject, network environment, and terminal status are dynamically considered during authentication. Violations and exceptions during access are continuously monitored to ensure that users and terminals on the network are always trustworthy.

TEE

Infrastructure with intrinsic trustworthiness is the main environment to transfer campus service data. This is a secure and isolated execution environment where sensitive data and key operations can be protected from external attackers or malware. Typically, the TEE consists of hardware and software components and provides a secure execution environment to ensure data confidentiality, integrity, and availability. It ensures terminal, transmission, computing power, storage, and data security.

3.2.6 Campus O&M: Swarm Intelligence, Full-Stack Convergence, and Openness and Collaboration

Devices and service applications will become increasingly diversified as campus digital transformation thrives, and this will raise new requirements for flexible ICT and IoT adaptation and bring huge challenges for E2E campus O&M. Future campuses will require big data, AI, and other advanced technologies to perform automated and intelligent O&M as well as to improve O&M management. Future campus O&M should feature full-stack convergence, swarm intelligence, proactive assurance, and an open ecosystem.

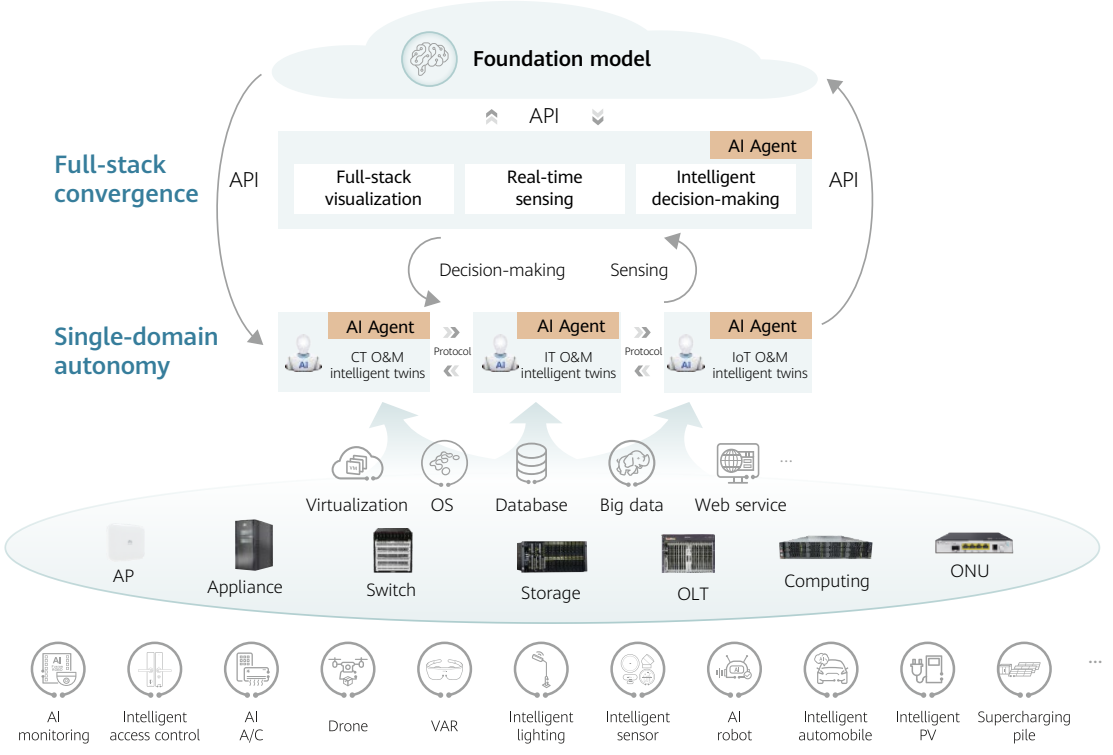


Figure 3-31 Campus O&M

Full-Stack Convergence

Future O&M will be oriented to full-stack ICTs and IoT technologies from the perspective of E2E services. The system should be able to display a complete service overview, including the status of devices, servers, storage, networks, virtualization, OSs, containers, middleware, and applications, and the relationships between them.

Swarm Intelligence

Every managed object, control unit, and application management component will be intelligent. Managed objects will use AI to iteratively provide real-time updates on their statuses and optimize them locally. Control units will intelligently close managed objects in the management domain, open management capabilities through interfaces, and collaborate with each other. Application management components will integrate the information from and capabilities of managed objects and control units to implement intelligent O&M using AI foundation models from the perspective of E2E services.

Proactive Assurance

Using AI foundation models, the system will proactively monitor services in real time from start to finish, analyze and evaluate their service statuses, and automatically provide service assurance or optimization suggestions.

Open Ecosystem

The O&M system should be able to provide abundant flexible external interfaces. It will be able to automatically adapt to services across various scenarios by configuring AI foundation models. They will also set open interface standards for various industries to better support industry development.

Call to Action

04

The world will be brimming with challenges and aspirations by 2030. As the core of the transition from the digital economy to intelligent economy, intelligent campuses are bound to develop even more rapidly.

To campus operators:

Campuses are crucial for people's work and lives, which are now embracing transformation, iteration, and rapid development. Driven by new policies, technologies, and requirements, campus service scenarios and business models will be constantly upgraded and innovated. Future campuses are moving toward smart spaces which are digital, converged, people-centric, resilient, and green. Compared with conventional campuses, both the external environments and internal conditions of intelligent campuses have undergone complex and profound changes, bringing a series of new requirements and challenges to intelligent campus planning, construction, and operations.

Intelligent campus construction is a systematic project that requires overall planning, ordered construction, step-by-step implementation, and value-based operations to achieve project success and maximize benefits throughout the lifecycle. A consistent blueprint can avoid unnecessary social cost increase and resource waste during the process. Thus, planning, construction, and operations of intelligent campuses are interdependent elements that complement each other. Specifically, the planning phase considers construction and operations; the construction phase involves operators; the operations phase implements the planned content.

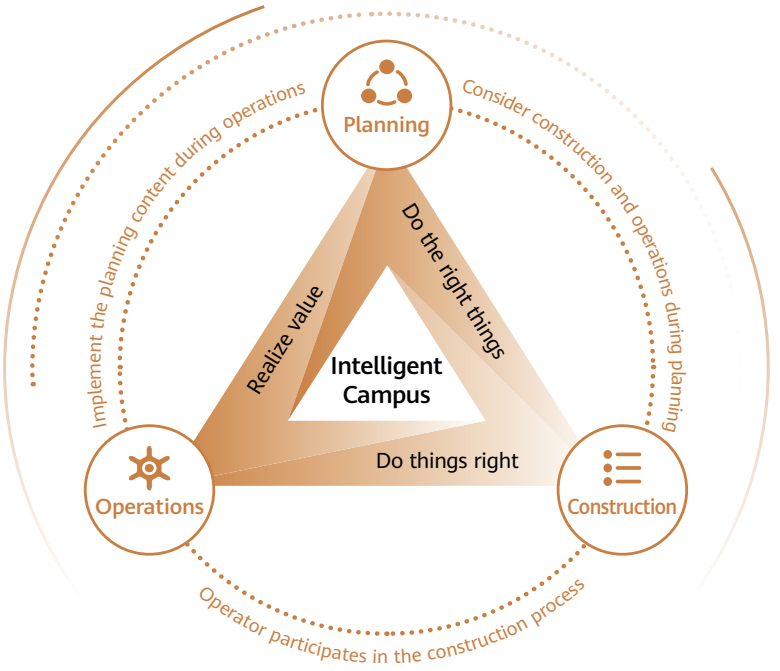


Figure 4-1 Integrated planning, construction, and operations

- During the planning phase, Huawei gains full insight into industry development, technical trends and applications, and smart requirements. We execute systematic and forward-looking intelligent campus planning, and deliver the architecture blueprint along with implementation roadmap planning to ensure we do the right things.

- Regarding construction, Huawei builds a four-layer architecture for the intelligent campus solution based on its powerful solutions and products. Our professional delivery services such as integration design, verification, and implementation ensure everything goes smoothly.

- In the operations phase, Huawei offers operations planning and design as well as service, data, and platform operations. Thanks to the above services, intelligent campuses are able to function well and be continuously optimized, gradually showcasing the benefits of digital transformation.

To industry partners:

By 2030, the intelligent campus industry will serve all manners of businesses, facilitate China's digital economy development, and boost the real economy. Openness and sharing, technological innovation, ongoing value creation for users, ecosystem support, and standards leadership will be the key to prosperous intelligent campuses.

Intelligent campuses present a comprehensive ICT solution oriented to campus management and operations. In addition to ICT infrastructure capabilities, they require industry partners to provide campus application software, devices and subsystems, technology integration services, and consulting and top-level design services. We therefore have created the following initiatives for partners:

- **Campus application software development partners (ISVs):** Campus application software is the direct entry for smart experiences of campus users. Campus applications will combine with platform capabilities, which will disrupt the traditional approach to application development that emphasizes single-point functions and waterfall development. Campus applications will leverage comprehensive data on people, events, things, and environments within the campus. By utilizing big data and AI technologies, we will adopt an agile, iterative development approach to deliver excellent user experiences. We realize that campus management and service requirements vary significantly between different business forms and scenarios. To remain competitive, campus application software must provide deep insight into campus value scenarios, provide best practices for campus management in different business forms, and enable efficient, agile, and cost-effective development to ensure continuous operations of campus services. To achieve business success, application software development partners must embrace platform capabilities, innovate for campus value scenarios, build application product systems with best practices, and deliver an ultimate smart experience to users that creates value.

- **Campus device and subsystem partners (IHVs):** Campus devices and subsystems are an important part of campus digital infrastructure in both the intelligent campus era and the smart campus era. These devices and subsystems serve as sensing and feedback points for information exchange between the digital and physical spaces on campuses. As intelligent campuses continue to develop, various intelligent devices and subsystems, such as mobile charging stations and logistics robots, are emerging to enhance campus operations. Innovative campus service scenarios require open campus devices and subsystems that offer standard data models and easy-to-integrate data interfaces for platforms and applications. This enables quick, closed-loop management of information acquisition, processing, application, and feedback. During this process, campus devices and subsystems will also receive support from intelligent scenarios, giving them more opportunities for usage. Therefore, campus device and subsystem partners should be open to implementing simplified, IP-based system interconnection and standardization of thing model data. In this way, they can achieve business success by expanding the industry space.

- **Campus technology integration partners (SIs):** Building an intelligent campus is a complex and comprehensive project that requires strong element integration. It encompasses various components such as campus applications, digital platforms, connections, devices, and subsystems. The composition needs to be implemented by professional technical integration services. Compared with the system integration (single system integration and engineering integration) required by weak-current intelligence, technical integration for intelligent campuses requires more "soft" integration. "Soft" integration has two aspects. The first is software integration, which allows intelligent campuses to combine platforms with subsystems, applications, and other platforms. Technology integration partners must be skilled in using multiple software integration platforms and developing software interfaces to achieve this. The second is IT-based integration. Intelligent campuses use IT architectures that differ significantly from traditional weak-current systems in terms of data transfer. As a result, technology integration partners must possess the ability to design comprehensive solutions for intelligent campuses. This includes designing technology architectures (TA), service/data flows (IA), and themed/specialized libraries required for campus applications. These requirements drive the technology integration from engineering to IT. Technology integration partners can achieve significant business returns by offering high-value professional services such as IT integration and data governance. Therefore, campus technology integration partners need to accelerate the development of professional and IT-based technology integration capabilities to ensure the overall construction.

- **Campus consulting and top-level design partners:** The goal of building intelligent campuses is to transform campus management and services using ICTs. The resulting benefits will include a low-carbon footprint, enhanced security, improved user experience, lower costs, increased efficiency, and innovative

models. For certain large-scale intelligent campus projects, consulting and top-level design are necessary to break down campus service objectives into operational scenarios, then break down the service architecture of those scenarios into technical solutions. These solutions are then applied to operational modes. The consulting and top-level designs are the links between campus service solutions and technical solutions, and ensure the implementation of key construction indicators. Campus consulting and top-level partners should have a range of methods, from campus service analysis to technology presentation, and use professional top-level design to guide the construction and operations of intelligent campuses. They possess the most comprehensive professional service capabilities in the entire industry ecosystem.

We hereby call on all industry parties to guide development and build a prosperous ecosystem. While prioritizing the ideal of adding value, we can integrate ICTs with construction to build a new future for intelligent campuses. Huawei also calls for joint participation in the development of intelligent campus standards, promotion of national and industry standards, and creation of a favorable industry environment. We hope that all industry ecosystem partners will attach importance to campus talent cultivation and development, and play an active role in building intelligent campuses that are industrialized, standardized, and healthy.

Now, the intelligent campus 2030 is already taking shape. Looking ahead, Huawei will continue to be customer-centric and demand-driven, and employ next-generation digital technologies to bolster the development of intelligent campuses.

A new era is arriving sooner than expected. Let's seize opportunities to cooperate with each other, explore and innovate, and build more intelligent campuses together.

Appendix 1: Technical Indicator Forecast

Technical Feature	Indicator	Indicator Definition	Forecast for 2030
Intelligent twins	Intelligent platform utilization	Percentage of campuses in China that use intelligent platforms	30%
	Light field holographic rendering computing power	A single user's demand for light field holographic rendering computing power in digital twins	10 teraFLOPS
	AI penetration rate	AI penetration rate of new campus services	100%
Ubiquitous intelligent connectivity	Optically-sensed gas density precision	Gas density sensing precision of the HCS system that combines communications with IoT sensors	< 10 ppm
	Optically-sensed illuminance precision	Illuminance sensing precision of the HCS system that combines communications with IoT sensors	< 0.1 lux
	Optically-sensed humidity precision	Humidity sensing precision of the HCS system that combines communications with IoT sensors	< 2%RH
	Optically-sensed temperature precision	Temperature sensing precision of the HCS system that combines communications with IoT sensors	<0.2°C
	Human body imaging precision of radar	Precision of detecting human body position and contour based on multi-radar collaboration	Centimeter-level imaging
Spatial interaction	Number of XR users	User quantity of XR, which is an important means to support immersive interaction	1 billion
	TSN jitters and latency	Necessary jitters and latency from sensing to collaboration of human-machine interaction supported by asynchronous TSN	10 μs–100 μs
	Miniature holographic bandwidth	Bandwidth capacity required to display a 10-inch miniature hologram	12.6 Gbit/s
	Simulation computing power	Single-user computing power required to shorten network latency through computing to simulate real user behaviors	1 TeraFLOPS
	Area of the map drawn in real time	The capability of the data foundation service to draw a spatial map in real time	1 million square meters
	Agent density of real-time interaction	Number of agents that support real-time interaction in the same physical space	10,000

Technical Feature	Indicator	Indicator Definition	Forecast for 2030
Intent-driven ultra-broadband	Wireless access bandwidth on campuses	Peak rate of an AP and cell capacity of the mobile network	100 Gbit/s
	Deterministic latency	Wireless air interface latency of an AP	1 ms with reliability of 99.9999%
	Multi-modal wireless coverage	The medium coverage distance of multi-modal Wi-Fi with microwatt-level power consumption	2–10 km
	10GE Wi-Fi penetration rate	Global utilization of 10GE Wi-Fi on enterprise campuses	84%
	Penetration rate of 5G industry private network	Global utilization of 5G industry private network on enterprise campuses	35%
Security and resilience	Proportion of investment in data security	Proportion of data security investment to the total IT resource investment	20%
All-domain zero carbon	Campus PV LCOE	LCOE of PV power generation on campuses	US$0.01 per kWh
	Green electricity voltage level on campuses	Voltage level of the green electricity generation system on large campuses	1500 V

Appendix 2: Abbreviations and Acronyms

Abbreviation/Acronym	Full Spelling
2D	2 Dimensions
3D	3 Dimensions
3GPP	3rd Generation Partnership Project
5G-A	5.5th Generation Fixed Network
5G	5th Generation of Mobile Communication
5G NR	5G New Radio
5G-A	5G-Advanced
6G	6th Generation of Mobile Communication
ABE	Attribute-Based Encryption
ABF	Analog Beamforming
ADN	Autonomous Driving Network
AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AIGC	AI-Generated Content
AND	Autonomous Driving Network
AP	Access Point
AR	Augmented Reality
ASIC	Application-Specific Integrated Circuit
BIM	Building Information Management
BLE	Bluetooth Low Energy
BOS	Balance-of-System
BYOU	Bring Your Own Utensils
CCFD	Co-Frequency Co-Time Full Duplex
CH4	Methane
CPU	Central Processing Unit
CRUD	Create, Read, Update, and Delete
D2W	Die-to-Wafer
DetNet	Deterministic Networking
DMRS	Demodulation Reference Signal
DoF	Degrees of Freedom
DR	Disaster Recovery
E2E	End to End
EDR	Endpoint Detection and Response
ELV	Extra-Low Voltage

Abbreviation/Acronym	Full Spelling
eMBB	Enhanced Mobile Broadband
EMS	Element Management System
F5G-A	5.5th Generation Fixed Network
FDM	Frequency Division Multiplexing
FDMA	Frequency Division Multiple Access
FlexE	Flexible Ethernet
FLOPS	Floating-Point Operations per Second
FO	Fan-Out
FPGA	Field Programmable Gate Array
FPS	Frames per Second
FR1/FR2	Frequency Range_1/Frequency Range_2
GaN	Gallium Nitride
GAN	Generative Adversarial Network
GPT	Generative Pre-Trained Transformer
GPU	Graphical Processing Unit
HARQ	Hybrid Automatic Repeat Request
HCS	Harmonized Communication and Sensing
ICT	Information and Communications Technology
IMT	International Mobile Telecommunications
I/O	Input and Output
IoT	Internet of Things
ISP	Image Sensor Processor
IT	Information Technology
LCD	Liquid Crystal Display
LCOE	Levelized Cost of Electricity
LED	Light Emitting Diode
LPR	License Plate Recognition
MAC	Media Access Control
Massive MIMO	Massive Multiple-Input Multiple-Output
Mbit/s	Megabits per Second
MIC	Microphone
mMTC	Massive Machine-Type Communications
MPP	Maximum Power Point
MR	Mixed Reality
MTP	Motion-to-Photon
NaaS	Network-as-a-Service

Abbreviation/Acronym	Full Spelling
NE	Network Element
NPU	Neural Processing Unit
NR	New Radio
NSMF	Network Slice Management Function
NSSMF	Network Slice Subnet Management Function
ODN	Optical Distribution Network
ONT	Optical Network Terminal
OS	Operating System
OT	Operational Technology
P2MP	Point-to-Multipoint Networking
P2P	Point-to-Point
PCS	Power Conversion System
PDCP	Packet Data Convergence Protocol
PEC	Privacy-Enhancing Computation
PHY	Physical Layer
PLC	Programmable Logic Controller
PON	Passive Optical Network
PPD	Pixel per Degree
PUE	Power Usage Effectiveness
QoS	Quality of Service
RAM	Random Access Memory
RB	Resource Block
RDMA	Remote Direct Memory Access
RF	Radio Frequency
RFID	Radio Frequency Identification
RL	Reinforcement Learning
RLC	Radio Link Control
RLHF	Reinforcement Learning from Human Feedback
RoT	Root of Trust
SDCI	Software-Defined Cloud Interconnect
SDGs	Sustainable Development Goals
SDK	Software Development Kit
SDN	Software-Defined Networking
SEC	Super Error Concealment
SIEM	Security Information and Event Management
Sim2Real2Sim	Simulation-to-Real-to-Simulation

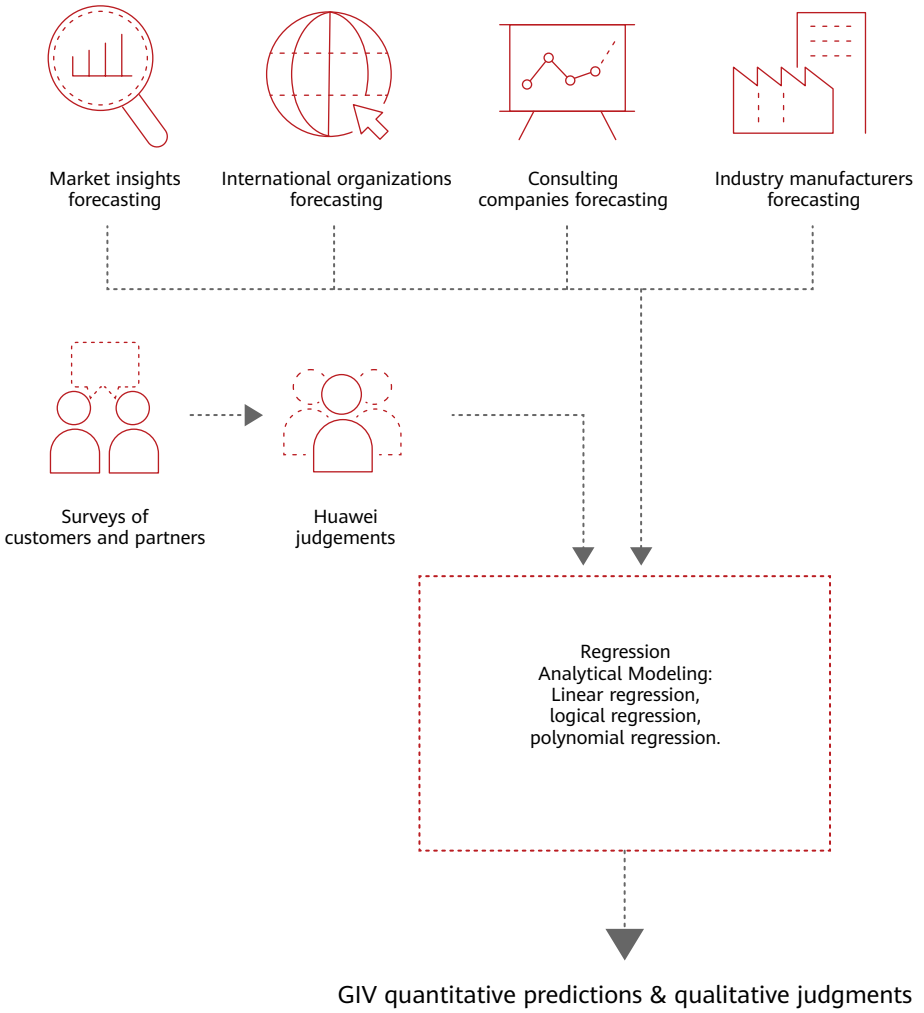
Abbreviation/Acronym	Full Spelling
SLA	Service Level Agreement
SLM	Spatial Light Modulator
SOAR	Security Orchestration, Automation, and Response
SOC	Security Operations Center
SPE	Single-Pair Ethernet
SQL	Structured Query Language
SRv6 Slice-ID	SRv6 Slice Identifier
TCC	Time Coordinated Computing
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TE	Traffic Engineering
TEE	Trusted Execution Environment
TFLOPS	teraFLOPS
THz	THz
TIM	Thermal Interface Material
TSN	Time Sensitive Networking
TTD	Time to Detect
TTR	Time to Repair
UB	Unified Bus
URLLC	Ultra-Reliable Low-Latency Communication
UWB	Ultra-Wideband
VLEO	Very Low-Earth Orbit
VPN	Virtual Private Network
VR	Virtual Reality
VST	Video See-Through
W2W	Wafer-to-Wafer
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing
WDMA	Wavelength-Division Multiple Access
Wi-Fi 6	Wireless Fidelity 6
Wi-Fi 7	Wireless Fidelity 7
Wi-Fi 8	Wireless Fidelity 8
WLAN	Wireless Local Area Network
XR	Extended Reality

Data Prediction Methodology

Humankind has always been driven to explore, and bring certainty to this uncertain world.

We believe that collaboration is essential for future exploration. Over the past three years, in addition to our own Huawei experts who have been building their own insights and plans for the future, our GIV@2030 research team has also been communicating extensively with industry scholars, customers, and partners. We have studied data, methodologies, and insight reports from prominent players around the world, including the United Nations, GSMA, and other third-party consulting firms.

The GIV@2030 research team used trend extrapolation and time series forecasting to arrive at their findings about the future.



Definitions of the Metrics

Category	No.	Indicator	Definition	Prediction for 2030
Connectivity	1	Number of global connections	Total number of connected people and things worldwide	200 billion
	2	Number of wireless cellular connections	Total number of people and things connected through wireless cellular technologies worldwide	32.5 billion cellular connections and 80 billion cellular-based passive connections
	3	Penetration rate of gigabit/10 gigabit home broadband	Proportion of gigabit or higher home broadband users to global households, and proportion of 10 gigabit home broadband users to global households	Penetration rate of gigabit or higher home broadband: 60%; penetration rate of 10 gigabit home broadband: 25%
	4	Number of active wireless AI agent users worldwide	Number of network users with AI agents	6 billion
	5	IPv6 address penetration rate	Percentage of connections and applications using IPv6 addresses worldwide	90%
	6	Number of VR/AR/MR users	Total number of VR/AR users worldwide	1 billion
	7	Number of robots per 10,000 workers working together	Number of robots per 10,000 manufacturing workers working together	1000
	8	C-V2X penetration rate	Proportion of C-V2X vehicles to vehicle parc	60%
	9	Monthly wireless cellular network traffic per user	Average monthly cellular network traffic per user (GB), including users using multiple devices	600 GB, 40x more than 2020
	10	Monthly average home network traffic	Average monthly network usage of a family	1.3 TB, 8x more than 2020
	11	Number and penetration rate of 5G industry private networks	Number of 5G industry private networks (including virtual private networks)/Proportion of large- and medium-sized enterprises using 5G industry private networks	Number of 5G industry private networks: 1 million; penetration rate of 5G industry private networks among large- and medium-sized enterprises: 35%
	12	Penetration rate of gigabit/10 gigabit enterprise Wi-Fi	Proportion of 1 Gbps and 10 Gbps bandwidths for WLAN access on enterprise campuses	Gigabit enterprise Wi-Fi penetration rate: 14%; 10 gigabit enterprise Wi-Fi penetration rate: 84%
	13	FTTR broadband penetration rate	Number of global FTTR-H home users Number of global SME FTTR-B users	FTTR-H home broadband penetration rate: 31% FTTR-B broadband penetration rate of SMEs: 41%
	14	Average AN level of carriers	Average AN level of global carriers	L4
	15	Global penetration rate of home cloud storage	Percentage of high-speed cloud storage in home scenarios	35%

Category	No.	Indicator	Definition	Prediction for 2030
Connectivity	16	Penetration rate of home cloud computer services	Percentage of cloud computers in home scenarios	17%
	17	Home security service penetration rate	1. Percentage of households with cameras installed 2. Percentage of households with 3D optical sensing terminals installed	1. 2030: China: 24%; global: 15% 2. 2030: 3D optical sensing (elderly care) penetration rate: 8%
	18	Number of global fiber broadband users	Number of global fiber broadband users	1.6 billion
	19	OTN coverage rate of transmission networks in large enterprises, public institutions, and government agencies	OTN coverage rate of transmission networks in key universities and scientific research institutions, large hospitals, and large industrial enterprises, as well as development zones and industrial parks at the county level or above.	100%
	20	Density of OTN integrated access nodes per 10,000 people	Number of OTN integrated access nodes per 10,000 people	Number of integrated access nodes per 10,000 people: 2 by 2025 and 4 by 2030
	21	Density of 100G OTN integrated access nodes per 10,000 people	Number of 100G OTN integrated access nodes per 10,000 people	1
	22	Reliability of network connections for distributed AI foundation model training	Reliability of DCI connections required for distributed training between different data centers	99.9999%
Data Storage	23	Proportion of unstructured data used for decision-making in production	The proportion of unstructured data (text, images, videos, and audio) used for decision-making in production	80%
	24	Proportion of enterprises deploying multi-layer ransomware protection systems covering the storage systems	The proportion of enterprises that deploy multi-layer ransomware protection systems	> 80%
	25	Proportion of hot data stored on SSDs	The proportion of hot data stored on SSDs after usage	100%
	26	Proportion of enterprises accessing active archived data at least once a day	The proportion of enterprises that access active archived data at least once a day	> 60%
	27	Proportion of the cloud and Internet enterprises using diskless reference architectures	The proportion of the cloud and Internet enterprises that adopt diskless reference architectures for IT systems	> 80%
	28	Proportion of endpoint, edge, and core data center data processed by AI in real time	The proportion of endpoint, edge, and core data center data that is processed by AI in real time	> 75%, > 80%, > 90%
Cloud	29	Cloud resource utilization enhanced by flexible compute	Cloud resource utilization enhanced by flexible compute	70%
	30	Peer-to-peer architecture penetration in cloud data centers	Proportion of cloud data centers that evolve from primary/secondary architecture to peer-to-peer architecture	60%
	31	Green power penetration in cloud data centers	Proportion of cloud data centers that use new technologies such as liquid cooling, thermal storage, waste heat recovery, and energy storage, as well as renewable energy such as solar, wind, and nuclear energy	100%

Category	No.	Indicator	Definition	Prediction for 2030
Cloud	32	AI devices facilitating device-cloud synergy	Number of devices that offer advanced intelligent assistance powered by the robust cloud computing power and electric power, such as intelligent connected vehicles	3 billion
	33	"Cloud for devices" compute capacity	Total compute capacity offered by cloud data centers to devices	40 ZFLOPS
	34	Contribution of AI to the GDP from creative and knowledge-intensive industries	Contribution of AI to the GDP from creative and knowledge-intensive industries	75%
	35	AI penetration in scientific computing tasks	AI penetration in scientific computing tasks	50%
	36	Discriminative AI penetration in all AI use cases	Discriminative AI penetration in all AI use cases	50%
	37	Discriminative AI penetration in all enterprise use cases	Discriminative AI penetration in all enterprise use cases	70%
	38	Media content AI-generated or generated with the support of AI	Media content generated by AI or with the support of AI	70%
	39	Jobs impacted by AI agents	Jobs in the world impacted by AI agents	67%
	40	Work hours replaced by AI agents	Work hours that AI agents may replace in total economic working hours	30%
	41	People who need to be reskilled for jobs due to AI	People who need to be reskilled for jobs under the impact of AI	> 40 million
	42	Employees who will have their own intelligent assistants	Employees who will have their own intelligent assistants	1.5 billion
	43	XR device shipment	XR devices shipped globally	60 million
	44	People who venture into the realm of spatial computing, a fusion of virtual and real worlds	People across the globe who venture into the realm of spatial computing, a fusion of virtual and real worlds	500 million
	45	Additional hours spent on spatial intelligence	Average additional hours of each person spent on spatial intelligence everyday	5 hours/day
	46	Areas covered by city-level 3D reconstruction solutions	Areas covered by city-level 3D reconstruction solutions with AI+CG	10,000 square kilometers
	47	Compute demand from data preprocessing and training of VLA models	Compute demand from data preprocessing and training of VLA models that power spatial intelligence	100 ZFLOPS
	48	General-purpose service robot shipment	Shipment volume of general-purpose service robots powered by embodied AI	30 to 50 million units
	49	Enterprise software built or rebuilt with AI	Enterprise software built or rebuilt with AI	80%
	50	Global digital content producers assisted by AI + CG	Global digital content producers assisted by AI + CG	100 million
	51	Interactive media content created with AI + CG	Length of interactive media content created with AI + CG	1 trillion hours

Category	No.	Indicator	Definition	Prediction for 2030
Cloud	52	Software development efficiency increase with convergence pipelines	Efficiency increase of different software development after large-scale application of convergence pipelines	10-100 times
	53	Man-machine speech recognition accuracy	Man-machine speech recognition accuracy enabled with natural language processing (NLP)	> 98%
	54	Market space of smart contract platforms	Market space of smart contract platforms that execute and manage an array of functions and services	USD15–25 trillion
	55	Web 3.0 zero-knowledge proofs	Zero-knowledge proofs executed by Web 3.0 applications	> 90 billion
	56	Smooth cloud migration efficiency increase for applications	Efficiency increase due to fully automated and unattended cloud migration for applications	10 times
	57	Cloud governance compliance risks eliminated in the early stages	Cloud governance compliance risks eliminated in the early stages using AI and other technologies	90%
	58	Faults that can be detected within 1 minute, demarcated within 5, and recovered within 10	Faults that can be detected within 1 minute, demarcated within 5, and recovered within 10	80%
	59	Cloud resource expenses saved with FinOps	Cloud resource expenses saved for global users every year with FinOps	> USD200 billion
	60	Increase in intensity and complexity of cyber attacks	Increase in intensity and complexity of cyber attacks across the globe	10 times
	61	Companies that will implement zero trust policies with cloud security as the core	Global companies that will implement zero trust policies with cloud security as the core	> 95%
	62	Organizations that will adopt AI-powered cybersecurity products	Global organizations that will adopt AI-powered cybersecurity products	80%
	63	Companies that have their own security models	Global companies that have their own security models	50%
	64	Companies that adopt an AI-powered cloud native SecOps model	Global companies that adopt an AI-powered cloud native SecOps model	70%
Energy	65	Proportion of renewables in global electricity generation	The proportion of electricity from renewables in the total electricity generated worldwide	65%
	66	Global total PV installations	The total capacity of PV plants installed worldwide	6,000 GW
	67	Proportion of digital infrastructure using green energy	The proportion of digital infrastructure that is powered by green energy	> 80%
	68	Proportion of PV plants using AI technologies	The proportion of PV plants that deploy AI technologies	90%
	69	Number of ultra-fast charging vehicle models	The number of vehicles that support ultra-fast charging	> 60%
	70	Annual ESS capacity addition	The total capacity of energy storage systems added each year worldwide	140 GW
	71	Annual energy charged to EVs	The total amount of energy charged to electric vehicles each year	1.1 trillion kWh

Category	No.	Indicator	Definition	Prediction for 2030
Vehicle	72	Autonomous vehicles as a proportion of new vehicles sold	Autonomous vehicles as a proportion of new vehicles sold annually in China	L2: 90%; L3 and higher: 30%
	73	Electric vehicles as a porportion of new vehicles sold	Electric vehicles as a porportion of new vehicles sold annually in China	82%
	74	In-vehicle computing power	Computing capabilities within a vehicle	Over 5000 TOPS
	75	General obstacle detection accuracy	Accuracy of detecting general obstacles during intelligent driving	Over 99%
	76	Safety of intelligent driving surpassing human driving	Level of safety of intelligent driving compared to human driving (measured by accidents)	10 times safer than human driving
	77	Miles per intervention	How far a vehicle can travel before human intervention is required during the intelligent driving process	621 miles (equivalent to 1000 km)
	78	Daily driving distance simulated in system	Daily virtual simulation capability of the model for simulating interactions between traffic participants constructed in the simulation system	Over 6.21 billion miles (equivalent to 10 billion km)
	79	Daily driving distance achieved by cloud-based parallel simulation	Intelligent driving testing and verification capability built in the cloud	Over 6.21 million miles (equivalent to 10 million km)
	80	Improvement of E2E response delay	Comparison of the end-to-end response delay capability of intelligent driving and human driving	Two times improved
Electric	81	Proportion of AI computing power	Proportion of AI computing power to final power in the power system	90%
	82	Edge intelligence adoption	Intelligent edge deployment rate in the power system	>75%
	83	Proportion of cloud-based applications of electric power companies	Degree of cloud-based resources, platforms, and applications in power systems	>75%
	84	Coverage of next-generation HPLC technology	Application ratio of next-generation HPLC technology in power systems	100%
ICT Services	85	O&M agent coverage	Proportion of carrier ICT O&M scenarios with agents	45%
	86	O&M copilot coverage	Proportion of carrier ICT O&M scenarios with copilots	100%
	87	Digital twin penetration rate	Proportion of carriers that have deployed digital twin systems	30%
Data Center	88	General-purpose computing power of a cluster	Effective computing power of a single cluster with software and hardware tuning	>70EFplops
	89	AI computing power of a cluster	Effective computing power of a single AI cluster	750 EFLOPS
	90	Cluster storage capacity	Effective storage capacity of a single cluster	Exabytes
	91	Percentage of data collaboratively processed by clouds and edges	Ratio of data requiring edge and data center processing to the total data volume	0.8

Category	No.	Indicator	Definition	Prediction for 2030
Data Center	92	Digital access rate of enterprises' production devices	Ratio of enterprises' production devices that can be accessed through edge data centers to enterprises' total production devices, after being IoT-enabled and intelligentized	0.8
	93	Data security investment ratio	Proportion of data security investment to the total data center investment	0.2
	94	System-level availability	System-level availability = System's annual MTBF/(System's annual MTBF + MTTR)(MTBF is short for mean time between failures, and MTTR is short for mean time to repair)	0.99999
	95	Disaster recovery (DR) coverage of important data	Percentage of important data and associated application systems for which DR is available	1
	96	Automation level	L1: human assisted; L2: partially automated; L3: conditionally automated; L4: highly automated; L5: fully automated. (Automation includes automatic predictive troubleshooting and analysis, automatic emergency handling, and AI-powered energy efficiency management. L4 indicates operations approaching truly unmanned, and L5 indicates operations without any human involvement.)	L4
	97	Power usage effectiveness (PUE)	Total data center power consumption/IT equipment power consumption	1.0x
	98	Renewable energy factor (REF)	Renewable energy consumption/Total data center power consumption	0.8
	99	Water usage effectiveness (WUE)	Water consumption/IT equipment power consumption	0.5 L/kWh
	100	DC-level resource pooling rate	Ratio of compute, storage, and network resources available for global scheduling to all resources in a single DC	0.8
	101	Proportion of new cloudnative applications	Ratio of new cloud-native applications to all new applications	0.9
	102	Resource allocation granularity	Granularity of compute, storage, and network resource allocation, scheduling, and billing	Functionlevel
	103	Penetration rate of hyper-converged interconnection bus technologies	Penetration rate of unified hyper-converged interconnection bus technologies	0.6
	104	Penetration rate of hyper-converged Ethernet networks	Ratio of converged networks of general-purpose computing, high-performance computing, and storage to all networks of data centers	0.8
	105	Penetration rate of optical + computing collaboration	Ratio of cluster computing power that supports computing power collaboration using spine layer on the all-optical direct connection AI parameter plane to the total computing power of the cluster	0.5
	106	Penetration rate of optical + storage	Ratio of data transmitted in cross-WAN highspeed mode using all-optical direct connection SSD to the total transmitted data collaboration	0.5

Category	No.	Indicator	Definition	Prediction for 2030
Data Center	107	Percentage of all-flash storage	Ratio of all-flash storage capacity to the total capacity of data centers	0.8
	108	Penetration rate of RDMA storage networks	Percentage of RDMA-based storage networks	0.8
	109	Percentage of data processed near memory or in memory	Ratio of the amount of data processed near memory or in memory to the total amount of data processed	0.3
Campus	110	Intelligent platform utilization	Percentage of campuses in China that use intelligent platforms	0.3
	111	Light field holographic rendering computing power	A single user's demand for light field holographic rendering computing power in digital twins	10 teraFLOPS
	112	AI penetration rate	AI penetration rate of new campus services	1
	113	Optically-sensed gas density precision	Gas density sensing precision of the HCS system that combines communications with IoT sensors	< 10 ppm
	114	Optically-sensed illuminance precision	Illuminance sensing precision of the HCS system that combines communications with IoT sensors	< 0.1 lux
	115	Optically-sensed humidity precision	Humidity sensing precision of the HCS system that combines communications with IoT sensors	< 2%RH
	116	Optically-sensed temperature precision	Temperature sensing precision of the HCS system that combines communications with IoT sensors	<0.2° C
	117	Human body imaging precision of radar	Precision of detecting human body position and contour based on multi-radar collaboration	Centimeterlevel imaging
	118	Number of XR users	User quantity of XR, which is an important means to support immersive interaction	1 billion
	119	TSN jitters and latency	Necessary jitters and latency from sensing to collaboration of human-machine interaction supported by asynchronous TSN	10 μs–100 μs
	120	Miniature holographic bandwidth	Bandwidth capacity required to display a 10-inch miniature hologram	12.6 Gbit/s
	121	Simulation computing power	Single-user computing power required to shorten network latency through computing to simulate real user behaviors	1 TeraFLOPS
	122	Area of the map drawn in real time	The capability of the data foundation service to draw a spatial map in real time	1 million square meters
	123	Agent density of real-time interaction	Number of agents that support real-time interaction in the same physical space	10000
	124	Wireless access bandwidth on campuses	Peak rate of an AP and cell capacity of the mobile network	100 Gbit/s
	125	Deterministic latency	Wireless air interface latency of an AP	1 ms with reliability of 99.9999%

Category	No.	Indicator	Definition	Prediction for 2030
Campus	126	Multi-modal wireless coverage	The medium coverage distance of multi-modal Wi-Fi with microwatt-level power consumption	2–10 km
	127	10GE Wi-Fi penetration rate	Global utilization of 10GE Wi-Fi on enterprise campuses	0.4
	128	Penetration rate of 5G industry private network	Global utilization of 5G industry private network on enterprise campuses	0.35
	129	Proportion of investment in data security	Proportion of investment in data security	0.2
	130	Campus PV LCOE	LCOE of PV power generation on campuses	US$0.01 per kWh

HUAWEI TECHNOLOGIES CO., LTD.
Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P. R. China
Tel: +86-755-28780808
www.huawei.com