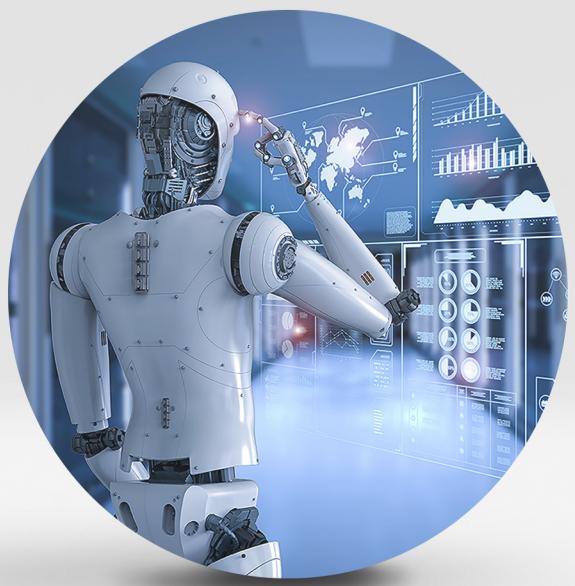




Data Center 2030



Building a Fully Connected,
Intelligent World

Foreword

Zheng Weimin

Science and technology are advancing far faster than we could have ever imagined. People at the cutting edge of this advancement note that the convergence of exponential technologies – those that double in performance over a short period of time – will open up a new horizon for social development over the next decade, redefining every aspect of our work and lives. Artificial intelligence is the latest in this line of world-changing technologies. In China alone, the size of the AI industry is estimated to exceed CNY450 billion by 2025, driving spillover in adjacent industries to the tune of CNY1.6 trillion.

Data, algorithms, and computing power are the three pillars of the AI industry. With recent innovations in AI foundation models like Generative Pre-trained Transformers – the "GPT" in ChatGPT – many say the computing industry is having its "iPhone moment". The driving force behind this progress is computing infrastructure made up of large AI clusters, which are used to train foundation models with thousands and even tens of thousands of neural processing units (NPUs). In essence, computing power has become the core pillar of AI and the metaverse, and it plays an increasingly important role in the digital economy.

There are three types of computing clusters: those for high-performance computing (HPC and supercomputing clusters/centers), AI computing (both clusters and centers), and general-purpose computing clusters (which take the form of data centers that employ cloud and big data technologies). In the past, these three types of computing clusters were mainly siloed. Moving forward, though, the deployment of these clusters will be increasingly integrated. For example, combining HPC and AI can greatly improve the computing efficiency of traditional HPC. Another example is short-range weather forecasts, which combine AI, big data, and scientific computing. Future data centers will be integrated computing clusters that provide diversified computing power to meet the unique needs of all industries along their intelligent transformation journey.

In addition, future data centers need to account for disparate energy structures between different regions, as well as service requirements between different industries in the same region, in order to provide greener computing power and meet the requirements of highly latency-sensitive applications. In China, the interconnection and unified scheduling of computing power is a basic prerequisite for the "Eastern Data, Western Computing" project. I believe this is the inevitable path that computing power will take.

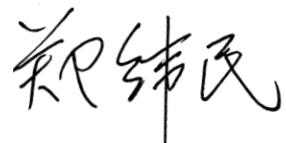
Of course, there are challenges. For current application scenarios, latency caused by insufficient bandwidth is an unavoidable bottleneck for interconnecting computing power. With existing architectures, for example, it can take up to five days to transmit 4 TB of raw data from Beijing to Wuxi – and that's using the fastest networks with no faults throughout the process.

China has put forward the concept of a "computing network", which aims to connect all computing centers across the country to form one big network computer. To make this computing network possible, we have to improve the efficiency of computing power transmission, and realize high-bandwidth, low-latency connections through network integration.

We also have to get major enterprises in the industry to work together. For example, we can shield the differences between heterogenous infrastructures, and use approaches like unified programming frameworks, compiled resource management, and scheduling software to implement interconnected computing power and unified scheduling and management of resources. This will help to steadily advance the "Eastern Data, Western Computing" initiative, and optimize the use of computing resources for the whole of society.

Huawei's *Data Center 2030* comes at just the right time. This report draws on Huawei's over 30 years of innovation in ICT and 10+ years of experience in helping its customers build out digital infrastructure. The report crystalizes the thoughts and outcomes of many rounds of discussion by experts inside and outside of Huawei. It starts with the application scenarios that future data centers will support, and forecasts their development roadmap towards the year 2030. The paper systematically describes the key technologies that will power future data centers, and proposes for the first time that future data centers will be computers that feature high degrees of efficiency in energy, computing, data, transmission, and operation. The report also points out the direction of integrated innovation between computing, storage, and networking, and concludes with reference architecture for next-generation data centers. It's a valuable reference for related industries.

In the next decade, the arrival of an intelligent world underpinned by big data, AI, and the metaverse will continue to gain momentum. As the world moves forward towards an intelligent future, the construction and development of data center infrastructure will become increasingly important. From concepts like the convergence of diversified computing power to the interconnection of computing power, the whole industry needs to continuously explore and innovate. Working together, we are better positioned to contribute to a strong global computing industry and a faster-growing digital economy.



Academician of CAE Member

Foreword

Joe Weinman

Artificial intelligence (AI) is the ability of machines to perform tasks that normally require human intelligence, such as reasoning, learning, decision making, and creativity. AI has the potential to disrupt many aspects of human society, such as economy, education, health, security, and culture.

Illustrating the point, the above paragraph was written by an AI Chatbot, thus demonstrating the blurring of boundaries between human capabilities and emerging digital technologies. But it's not just AI: virtually every technology is seemingly accelerating in price-performance, or, after showing nothing or only incremental progress for decades, suddenly disrupting a market in what evolutionary biologists call "punctuated equilibrium." There are hundreds of examples in the ICT industry, including large language models, diffusion models, blockchain, web3, quantum computing, new communications protocols, neuromorphic chips, soft robots, swarm intelligence, homomorphic encryption, etc. These technologies become even more powerful when combined in new ways, and enjoy accelerated adoption when they exploit existing platforms and ecosystems, such as cloud computing, mobile devices, and the Internet.

There are infinite uses for these emerging technologies, but we can roughly categorize their strategic application into four broad categories that can create competitive advantage, which I call "digital disciplines": information excellence, solution leadership, collective intimacy, and accelerated innovation.

Information excellence is the use of digital technologies to improve processes, resource utilization, and organization structures. For example, algorithms can best plot and revise logistics routes, and an AI-based solution can achieve more accurate radiology diagnostics results more quickly, reliably, and repeatably.

Solution leadership exploits smart, digital, connected products and services, that thereby have near-infinite potential for personalization and extensibility. Examples include smart homes, hospitals, ports, and cities.

Collective intimacy exploits millions or trillions of data points to create highly personalized services. Algorithms can analyze viewing data to suggest content, purchasing data to suggest cross-sells, or diagnostic, genetic, epigenetic, microbiomic, and pharmacological data to create patient-specific therapies.

Accelerated innovation can make innovation faster, cheaper, and better—digitally. For example, AI can accelerate drug discovery by rapidly combing through tens of thousands of scientific papers and experimental results, or by determining protein structure and dynamics.

Ultimately, however, the most advanced algorithms incorporated into the greatest applications using the most comprehensive data sets must have a real-world physical implementation. This is where centralized enterprise and cloud data centers and distributed edge computing and devices—made up of processing, storage, networking, security, and management—must all work together, to run the applications, at scale. They must meet a wide range of goals, including cost, performance, implementation schedules, reliability, security, visibility, manageability, privacy, and sustainability. To meet shifting goals, they must also ensure flexibility and forward and backward compatibility.

And, in today's environment of intense global competition, rapidly evolving technologies, products and services, various economic and natural disruptions, and fickle customers, companies must undergo continuous strategic reevaluation and continuous digital

transformation. Any examination of marketplace dynamics shows that companies can rarely rest on their laurels.

With this in mind, Huawei's *Data Center 2030* report provides an invaluable roadmap to the technologies and solutions of the coming decade, enumerating key issues facing the industry while reviewing—in clear, understandable terms—the palette of options becoming available to the IT manager or CIO, and their associated trade-offs. The report spans networking strategies and distributed compute fabrics; ranges from underwater data centers to those that will soon be in orbit; addresses green energy strategies and dynamic microgrids; highlights choices along the cloud-edge-device continuum; and covers dozens more topics of essential interest to those planning for today—or for the next decade.

While some may view these insights as too detailed, the fact remains that economic success today—whether for a garage start-up or an entire nation, is almost always inseparable from digital capabilities, successfully implemented and executed. The size of the global economy will exceed US\$100 trillion this year , and although the portion of it that is digital varies by country and based on what is included as “digital,” has been estimated to be 10% , 23% , or even 65% , ranging from irrigation sensors in Brazil to mobile payments in Nigeria to fish markets in India to supercomputers in China. To survive and thrive in this new digital economy, and thus in the global economy, any executive would do well to clarify their organizations' strategy and competitive posture, focus on creating digital advantage, and study the *Data Center 2030* report and leverage the insights contained within it to best implement and evolve their strategy, implementation, and roadmap. As Sun Tzu said, “Those who are victorious plan effectively and change decisively.”

Joe Weinman

Digital Strategist
Author, *Cloudonomics* and *Digital Disciplines*
CEO, XFORMA LLC

Foreword

Jerry Kaplan

My first encounter with cloud computing took place at summer camp in the late 1960s. While other “campers” were naturally focused on playing sports, hiking in the woods, and swimming in the cool lake water, I was fascinated by the sudden appearance of a Teletype Model 33 terminal, hooked up through a phone line to a central computer running the Dartmouth Time Sharing System hundreds of miles away. After a brief introduction to the BASIC computer language, I was hooked: I spent most of the summer sitting in a dark basement writing a program to play the card game “blackjack”. But my early career as a teenage programmer came to an abrupt halt when the camp got the bill for the computer and phone time!

My second encounter with cloud computing came a decade or so later when I was a graduate student in Computer Science at the University of Pennsylvania. Though my Apple II personal computer was adequate for word processing and spreadsheets, it didn’t have the processing power required for advanced Artificial Intelligence programs like the natural language query system I was developing for my PhD thesis. Instead, my program had to run on the University’s sole mainframe, where it occupied virtually all of the CPU time and memory. Soon, the University IT staff started getting complaints from other users that their programs were running too slowly (or not at all), and I was banished to working only at night, when the machine was otherwise idle!

Since that time, the cost of computing and data storage has dropped by more than a factor of one million. (My watch now has more power and memory than was available to the entire University back then.) Given this astonishing progress, you might expect that there would be more than enough computing power to meet all customer demand many times over for the foreseeable future, and at very low cost. And yet, the emergence of a new wave of Artificial Intelligence is once again placing demands on computing infrastructure that suppliers are

struggling to meet. Training a single Generative AI program (GPT-4) was estimated to take 90-100 days on twenty-five thousand GPUs at a cost of over \$100 Million. And that's just one of many such programs currently under development around the world!

Wallace Simpson, the colorful English Duchess of Windsor famously said in 1936 that "You can never be too rich or too thin". To that, I could add that you can never have enough computing resources. No matter how much data center efficiency improves, no matter how many GPUs are built or cloud computing facilities are constructed, our increasingly data-driven future will ensure that demand for computational power will continue to outstrip the available supply.

Huawei, a world leader in the data center industry, is well positioned to take advantage of this ongoing trend. The company's enduring commitment to technological innovation, increasing capacity, and cost reduction, serves as a shining example of how a single company can serve society by developing reliable, plug-and-play infrastructure to power our compute-hungry future.

This report, written in consultation with customers, research centers, and other stakeholders, provides a much-needed roadmap to ensure that progress in data center computing continues at a rapid pace for many years to come. Whether you are in business, research, entertainment, manufacturing, or any other data-intensive industry I encourage you to take full advantage of this opportunity to learn how Huawei's vision for data centers can provide your organization with the means to compete in a world where big data and Artificial Intelligence are the keys to success.

A handwritten signature in black ink, appearing to read "Jerry Yang".

Silicon Valley entrepreneur, author, AI expert

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. In addition, general-purpose computing power will increase 10-fold, and AI computing power will grow by a factor of 500 relative to 2020. 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.

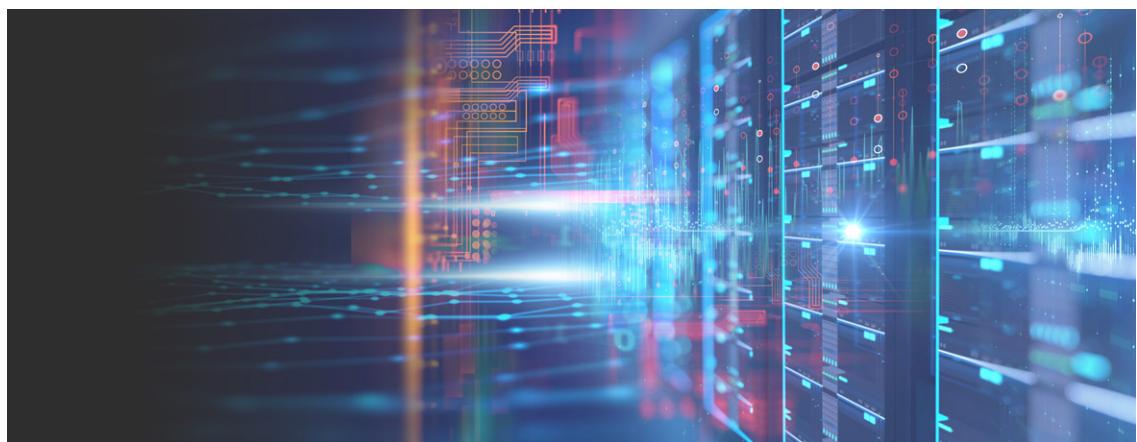
People tend to overestimate more immediate, short-term changes, but underestimate those that are coming in the next ten years. It's because these are harder to see. *Data Center 2030* seeks to demystify what's coming. Making bold assumptions and accurate predictions can often be dialectically opposed. But it's in exploring the overlap that we can create a better future. Looking forward, we will still have 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 of the Board
Chairman of the ICT Infrastructure Managing Board
President of the Enterprise BG
Huawei

CONTENTS

| | | |
|--|---|-----------|
| 01 | Industry Trends | 16 |
| 02 | Future Scenarios and Innovation Directions | 24 |
| AI for All: Creating new productivity | 25 | |
| The fourth paradigm: Exploring the unknown with data-intensive computing | 27 | |
| Spatial Internet: Supporting virtual-physical interaction | 28 | |
| Industrial digital twins: Promoting intelligent upgrade | 29 | |
| Inclusive cloud native: Bridging the enterprise digital divide | 30 | |
| Multi-flow synergy: Improving energy efficiency | 31 | |
| Software and hardware synergy: Improving computing efficiency | 32 | |
| Lossless networks: Improving transmission efficiency | 33 | |
| Socialized data collaboration: Improving data efficiency | 34 | |
| Human-machine collaboration: Improving operation efficiency | 36 | |



03

Our Vision for Future Data Centers and Their Key Technical Features 38

| | |
|---|-----------|
| Vision | 39 |
| Key technical features | 40 |
| 1. Diversity and ubiquity | 40 |
| (1) Big clusters | 40 |
| (2) Lightweight edges | 41 |
| (3) New patterns | 45 |
| 2. Security and intelligence | 49 |
| (1) High security | 49 |
| (2) High reliability | 54 |
| (3) High intelligence | 56 |
| 3. Zero carbon and energy conservation | 58 |
| (1) Green power supply | 58 |
| (2) New energy storage | 60 |
| (3) Liquid cooling | 63 |
| 4. Flexible resources | 66 |
| (1) Disaggregated pooling | 66 |
| (2) Flexible computing | 71 |
| (3) Cross-region and cloud-edge synergy | 76 |
| 5. Peer-to-peer interconnection | 80 |
| (1) Hyper-convergence | 80 |
| (2) High performance | 81 |
| (3) Intrinsic optical capabilities | 84 |
| 6. SysMoore | 92 |
| (1) New computing power | 92 |
| (2) New storage | 94 |

04

Reference Architecture for New Data Centers 104

| | |
|---|-----|
| New data center infrastructure: Driving inclusive green growth with innovative power supply and cooling | 106 |
| New computing infrastructure: Building a data-centric, diverse computing system | 108 |
| New resource scheduling: Implementing application-centric, flexible scheduling | 108 |
| New data management: Realizing instant visualization for data flow systems | 110 |
| New collaboration service: An open architecture to connect democratized computing power | 111 |
| New intelligent management: Enabling AI-driven, automatic data center O&M | 112 |

05

Development and Call to Action 116

| | |
|---|-----|
| Appendix 1: Indicator system of key prediction data | 119 |
| Appendix 2: Abbreviations and acronyms | 121 |





Industry Trends





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. Relative to 2020, general-purpose computing power will increase by a factor of 10 and AI computing power will increase by a factor of 500. 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.

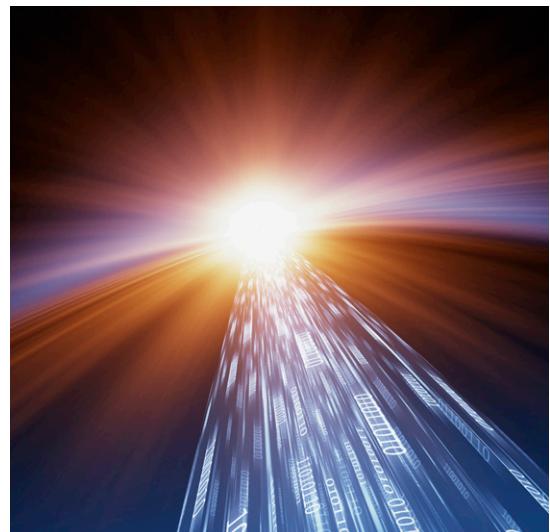
■ Data centers are shifting from consuming a lot of power to prioritizing green development

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.

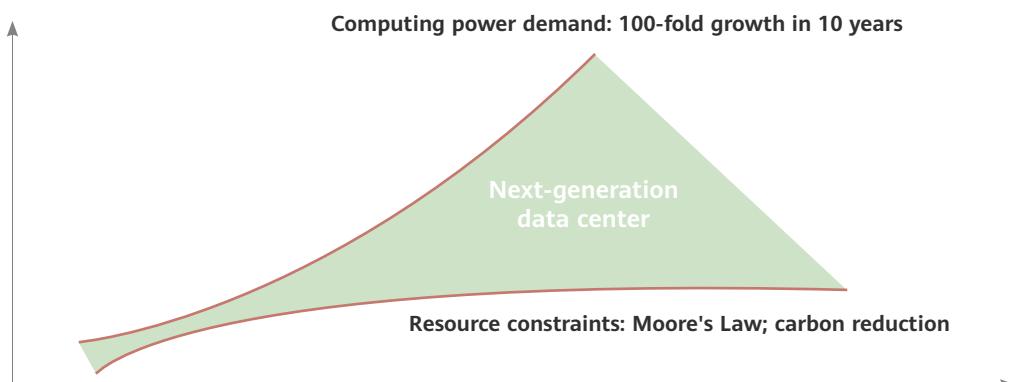


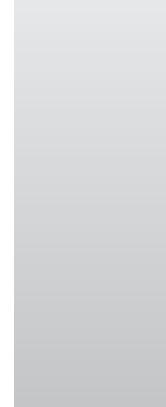
Figure 1-1 Challenges of computing power demand and resource constraints





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.

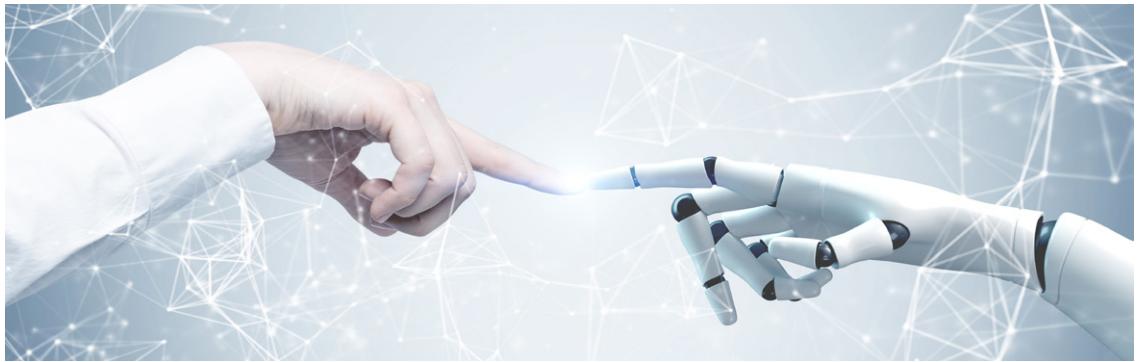
■ AI for All: Creating new productivity

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, from 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.

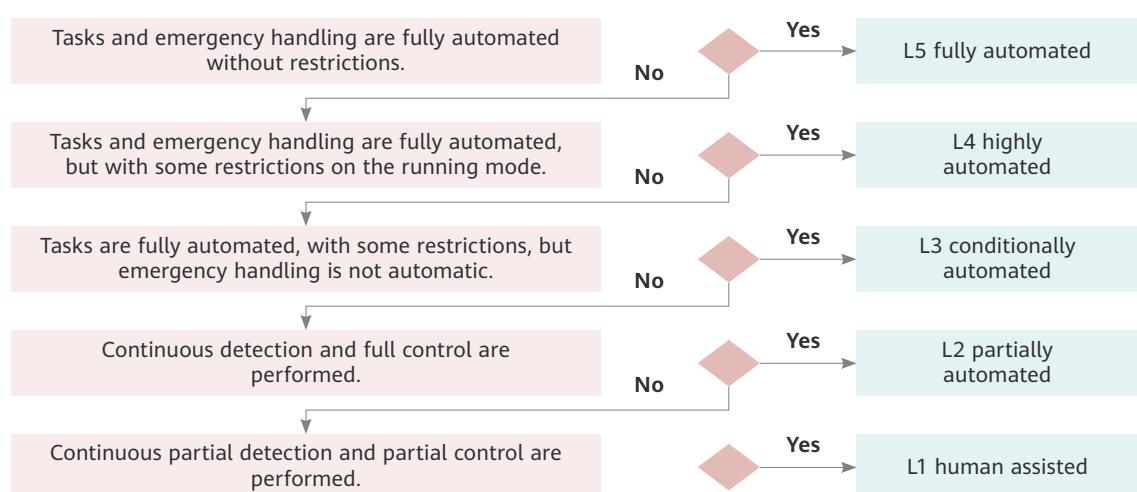


Figure 2-1 Five levels of data center automation





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.

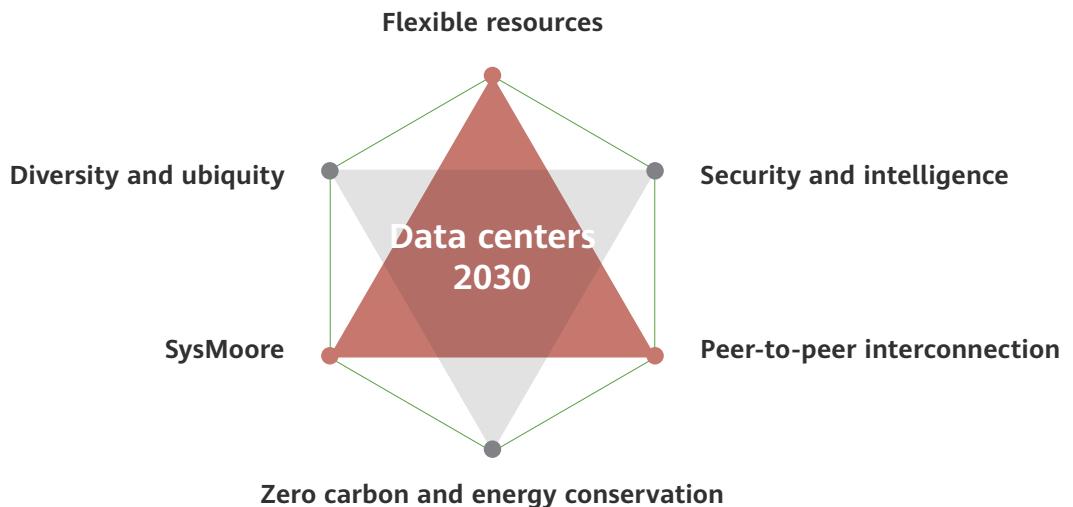


Figure 3-1 Key technical features of data centers 2030



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 XPU 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.

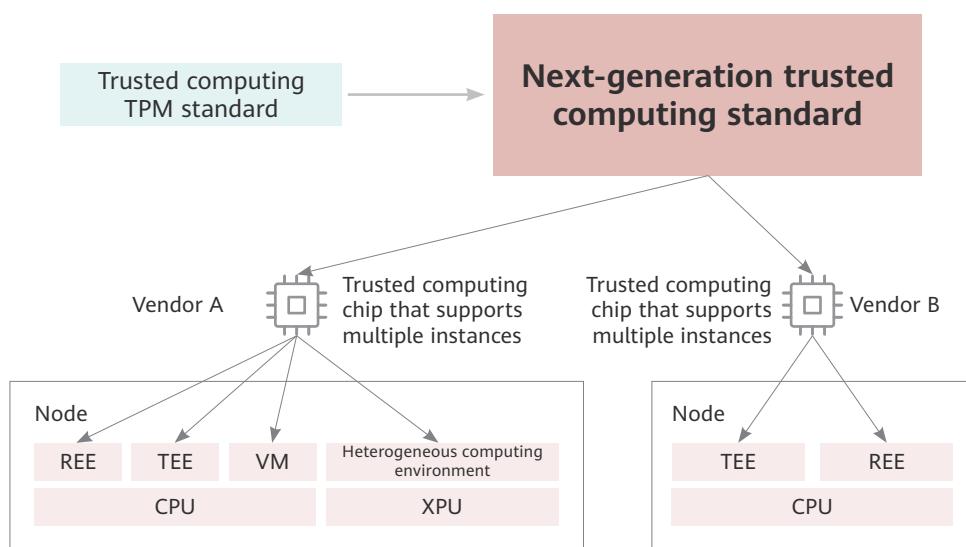


Figure 3-2 Trusted computing in the new computing paradigm

- **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 cooling 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.

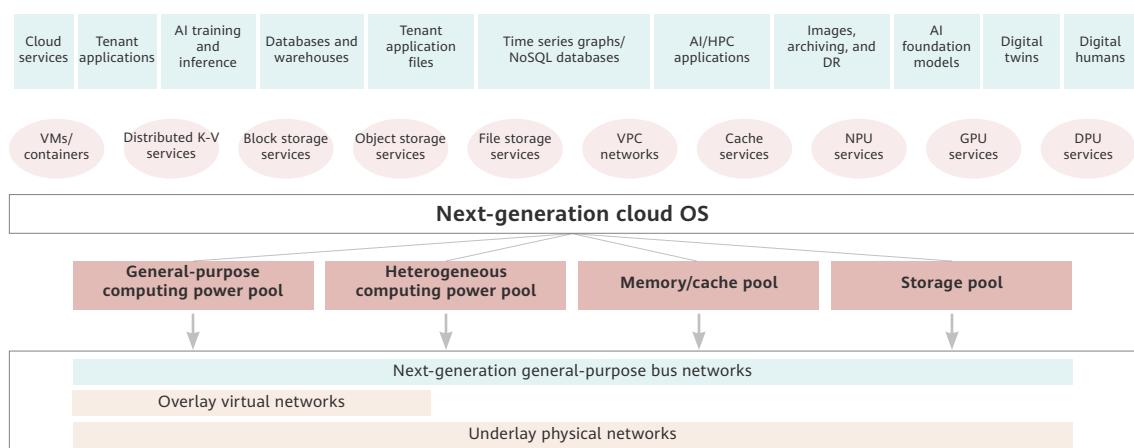


Figure 3-3 Disaggregated pooling in cloud data centers

- **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

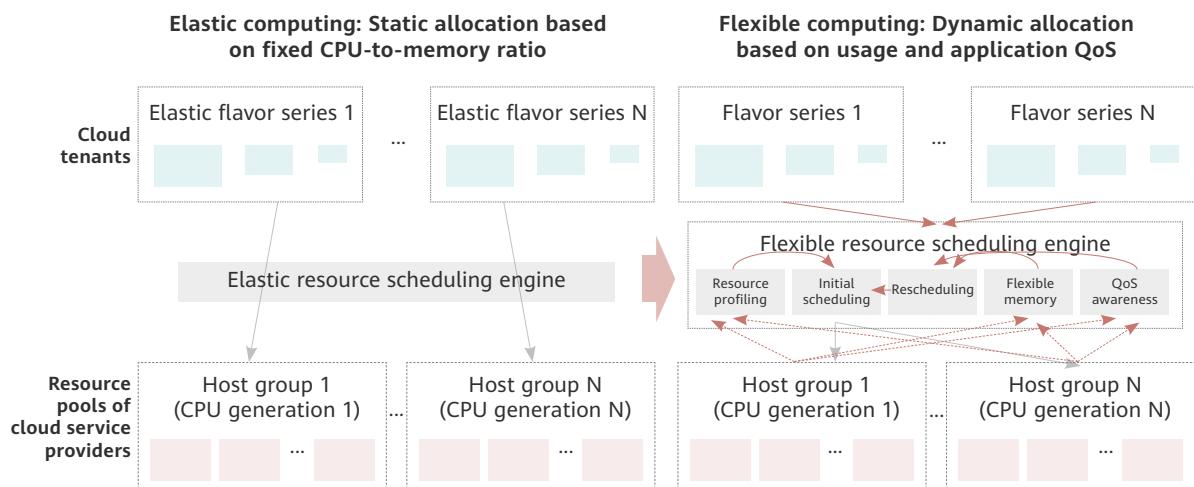


Figure 3-4 Comparison of elastic computing and flexible computing

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), PCIe, CXL, 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.

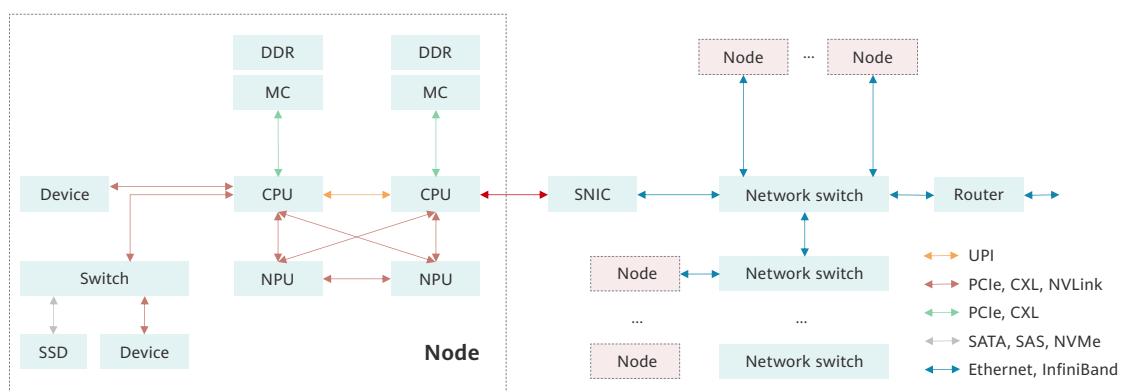


Figure 3-5 Multiple interconnection protocols in computing systems

- **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

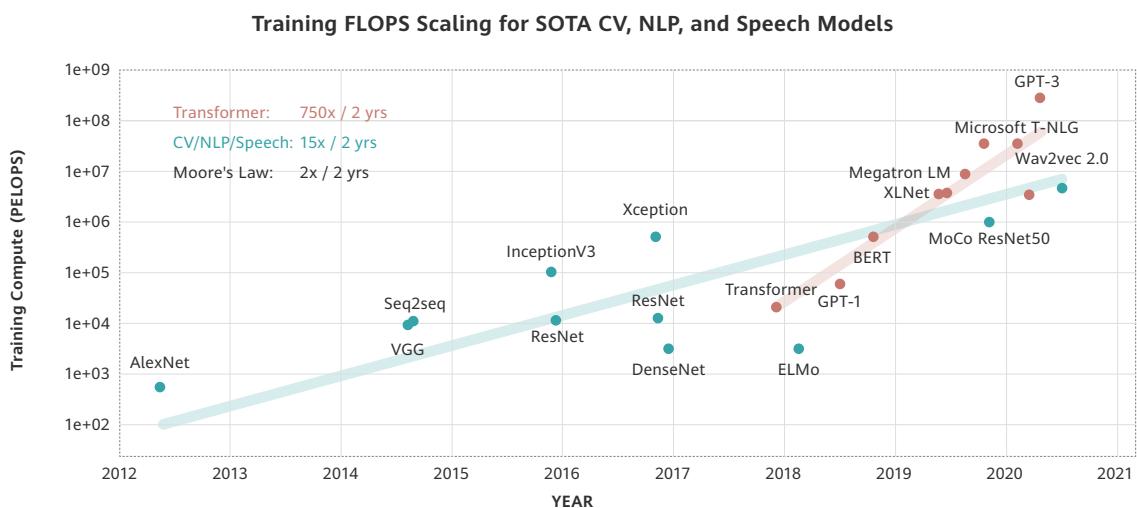


Figure 3-6 Different models' computing power demands

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.

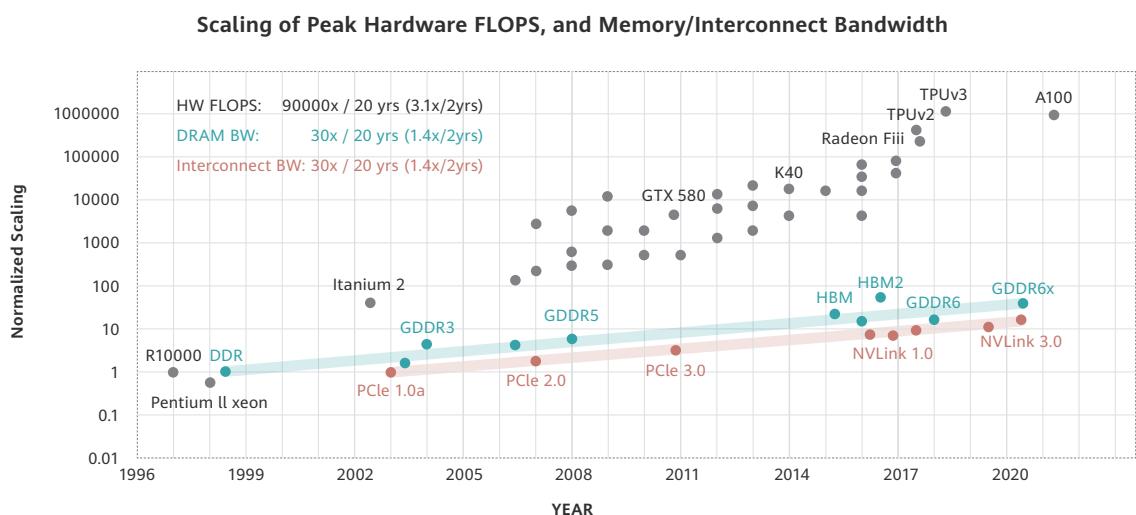


Figure 3-7 Computing power, interconnection bandwidth, and memory bandwidth

- **Lossless data center networks**

ChatGPT is driving the emergence 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- μ s-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 $16 \times 100G$ 1.6TSR solution was initiated in the IEEE 802.3dJ project. Some companies have expressed their expectations toward the $8 \times 200G$ 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.

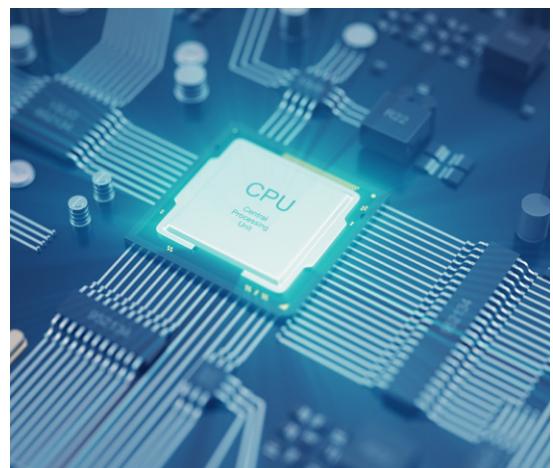


- **Co-packaged on-board optics**

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 ($< 5 \text{ Pj/bit}$). 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- μ s 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

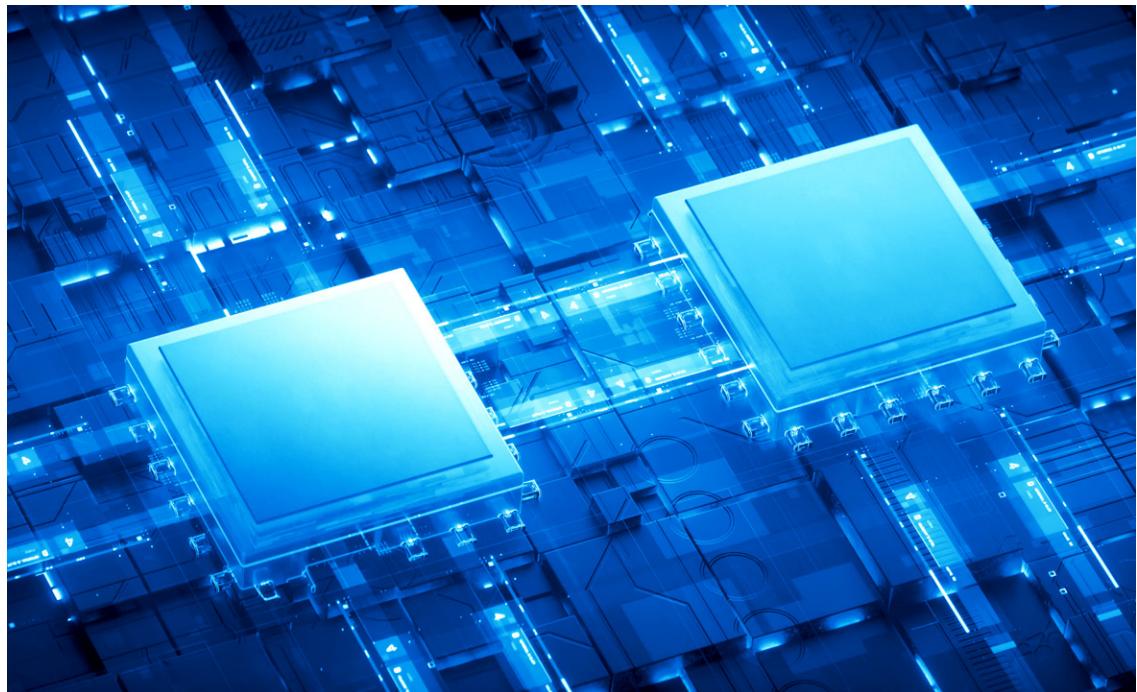
(1) New computing power

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.

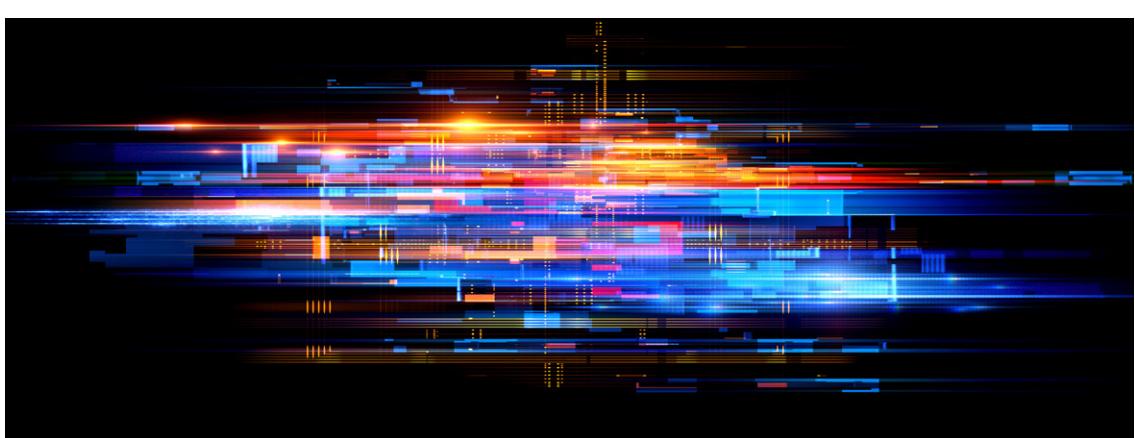
- **Large-scale application of non-silicon-based computing**

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.

(2) 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 1a process technology, 1a nm DRAM currently boasts the industry's highest bit density of 0.315 Gb/mm². 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 4F² 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:

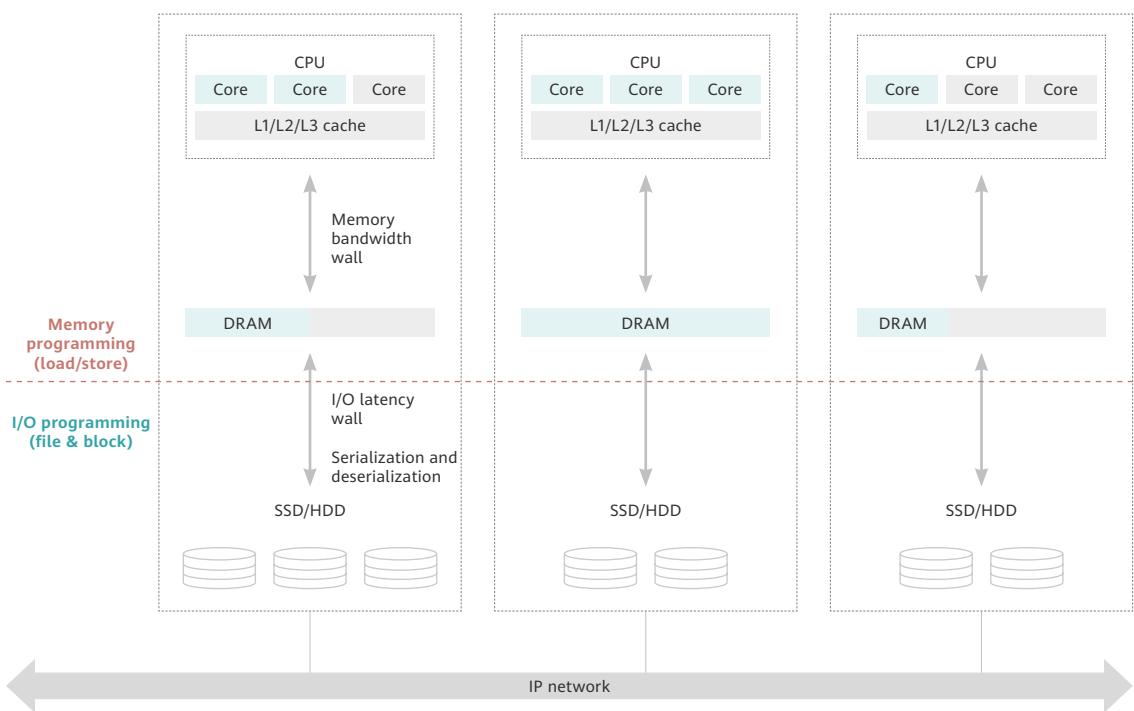


Figure 3-8 Traditional CPU-centric architecture

- (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

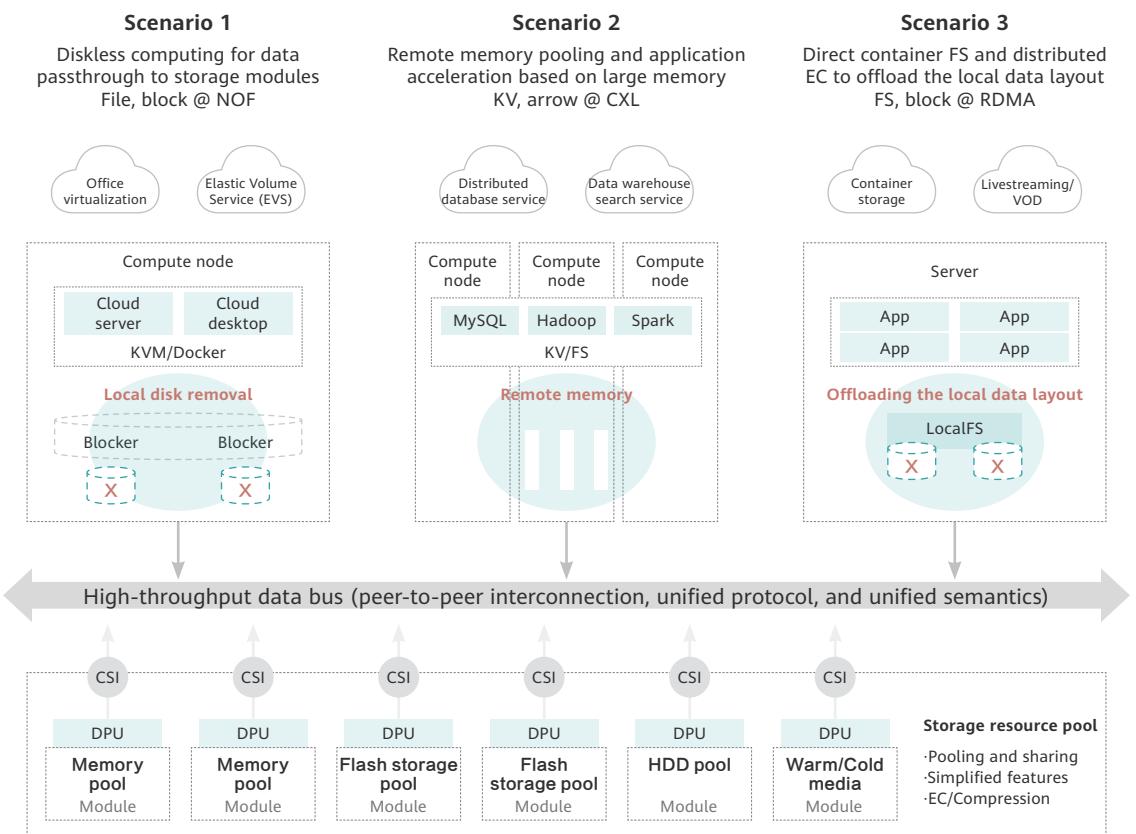


Figure 3-9 Decoupled storage-compute architecture

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

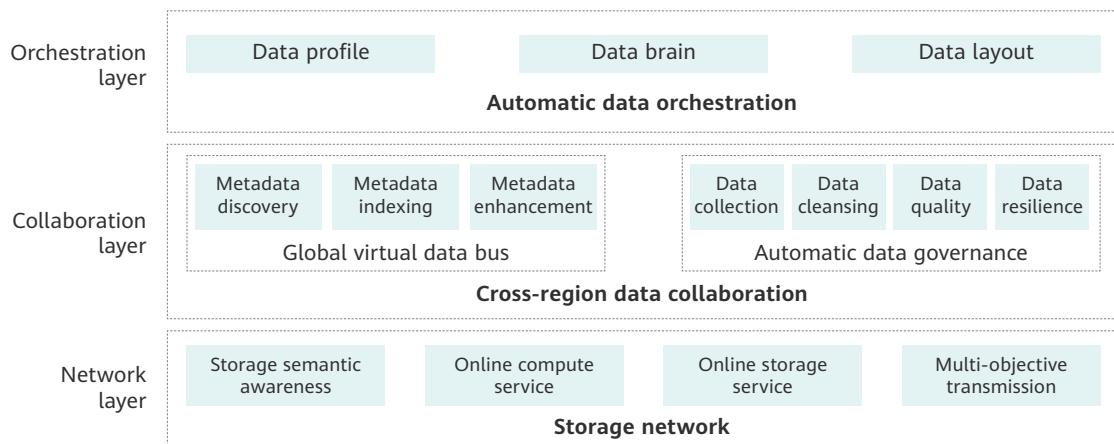


Figure 3-10 Intelligent data fabric framework

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.

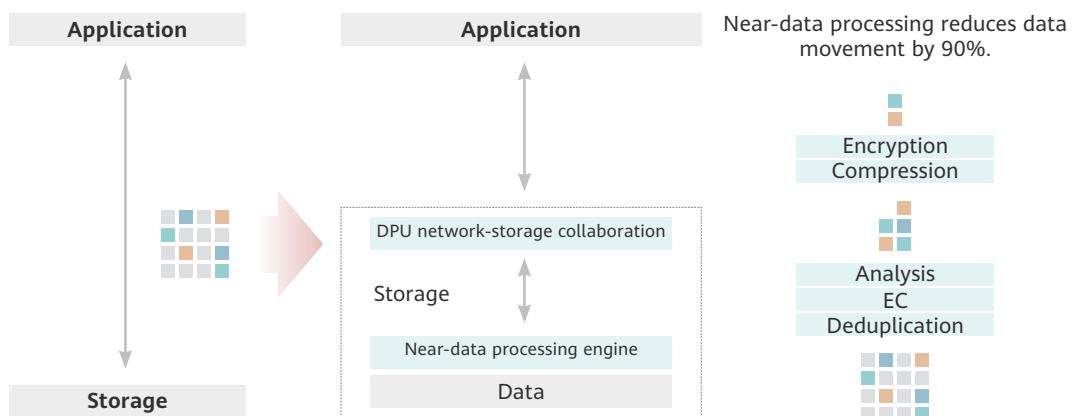
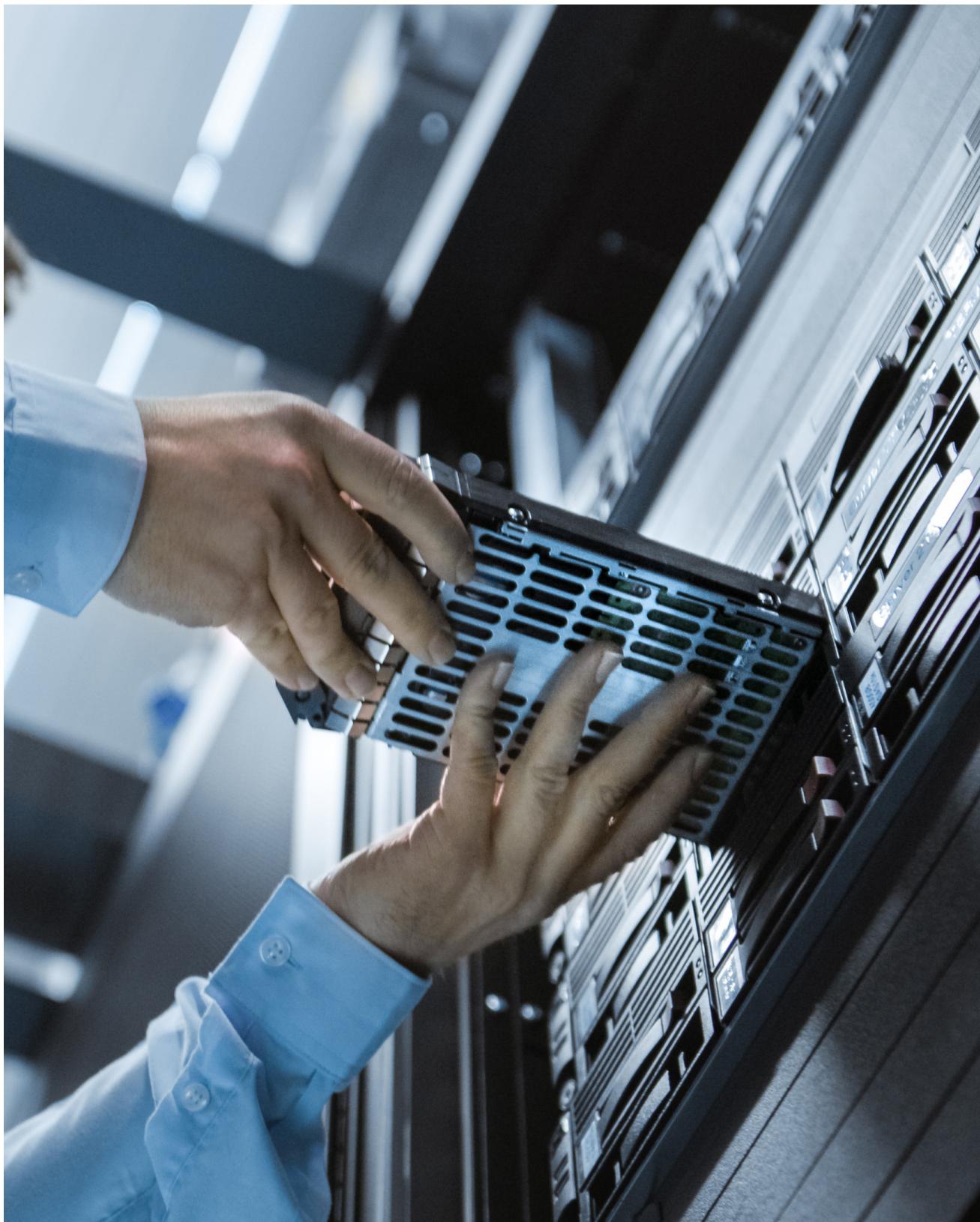


Figure 3-11 Application-oriented data acceleration

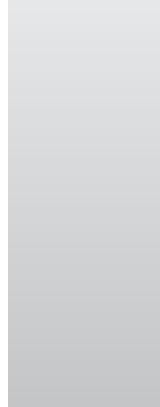






Reference Architecture for New Data Centers

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.

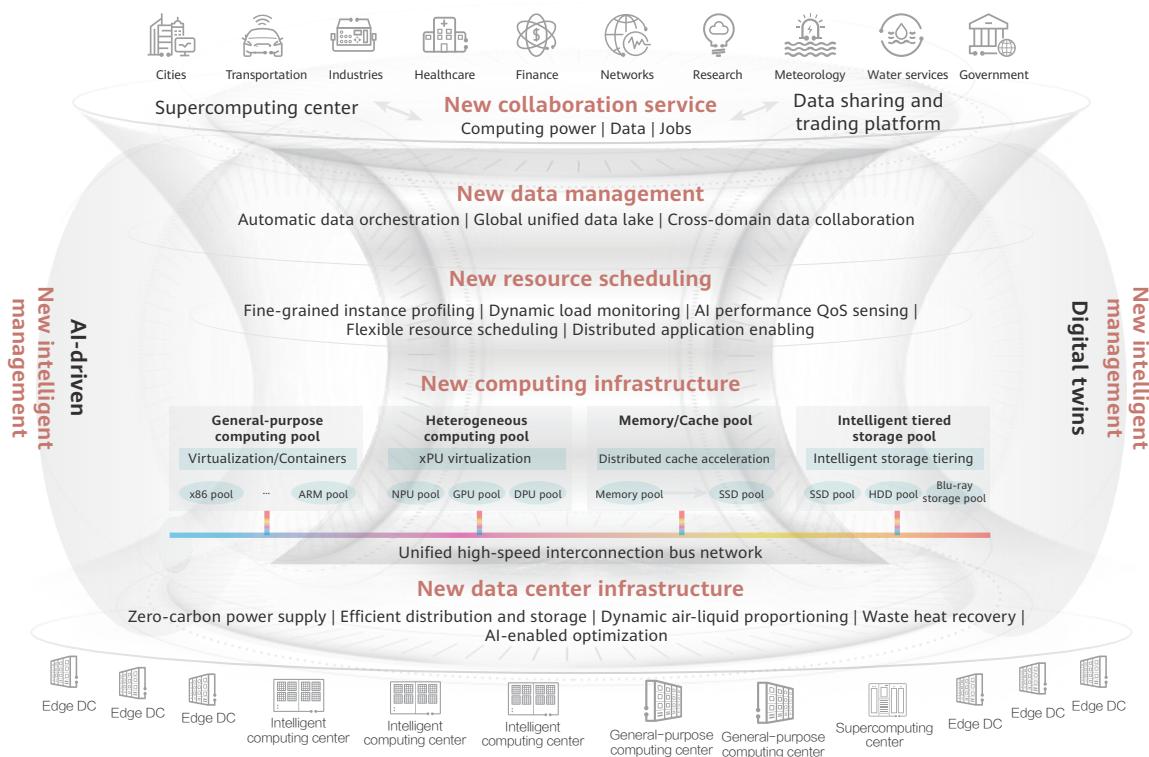


Figure 4-1 A reference architecture with six key features for new data centers

■ 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.

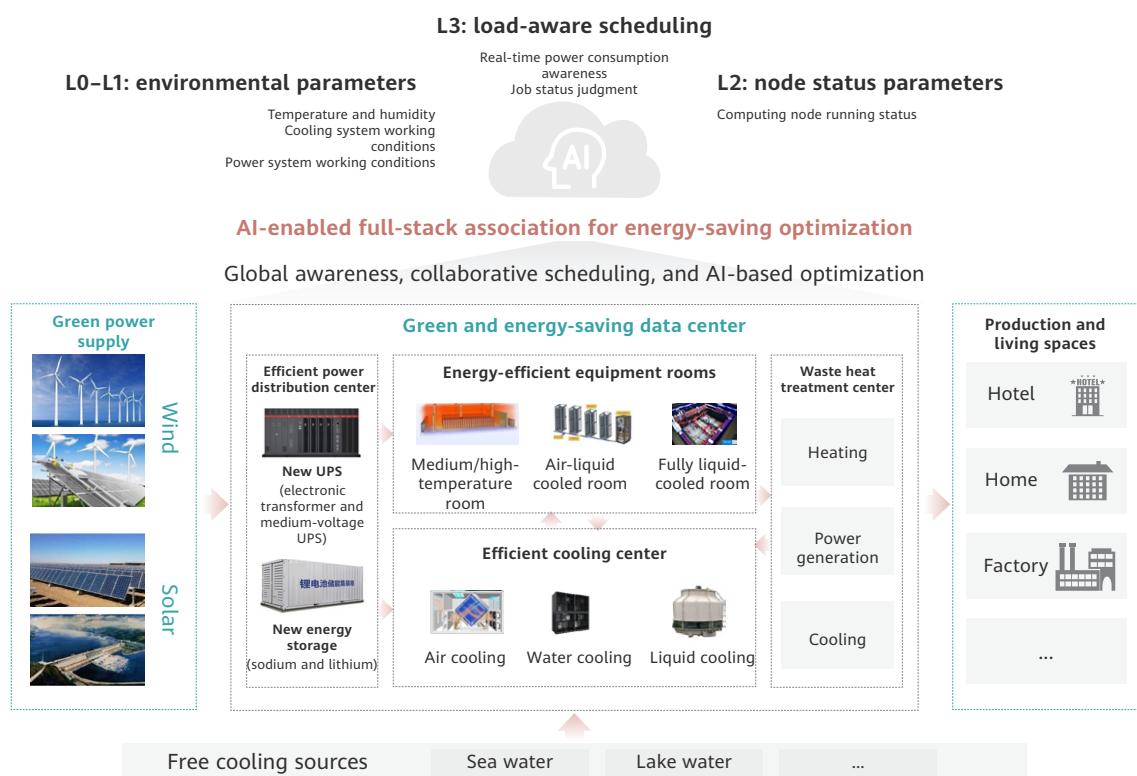


Figure 4-2 New reference architecture for data centers with green power supply and dynamic cooling

■ 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.

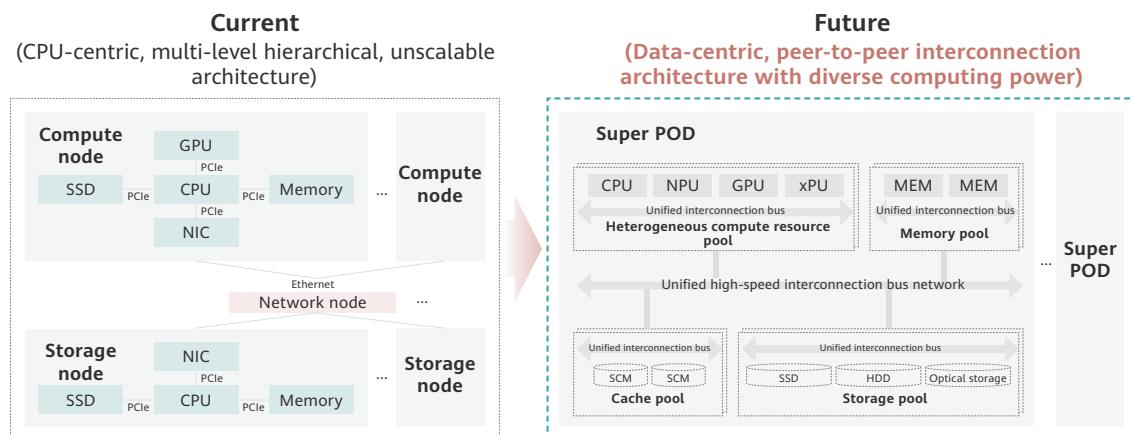


Figure 4-3 An architecture of the data-centric diverse computing system

■ 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.

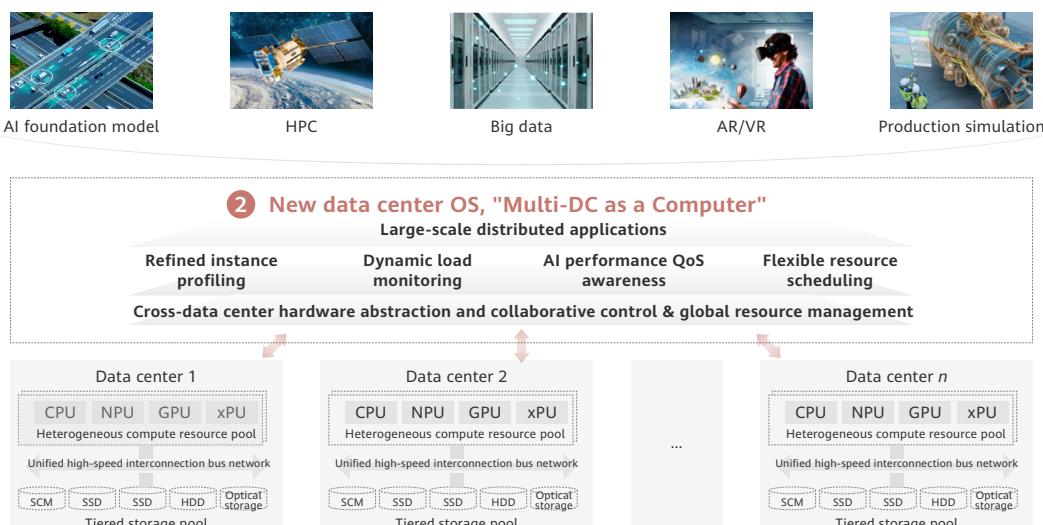


Figure 4-4 An architecture of the next-generation data center resource scheduling system

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.

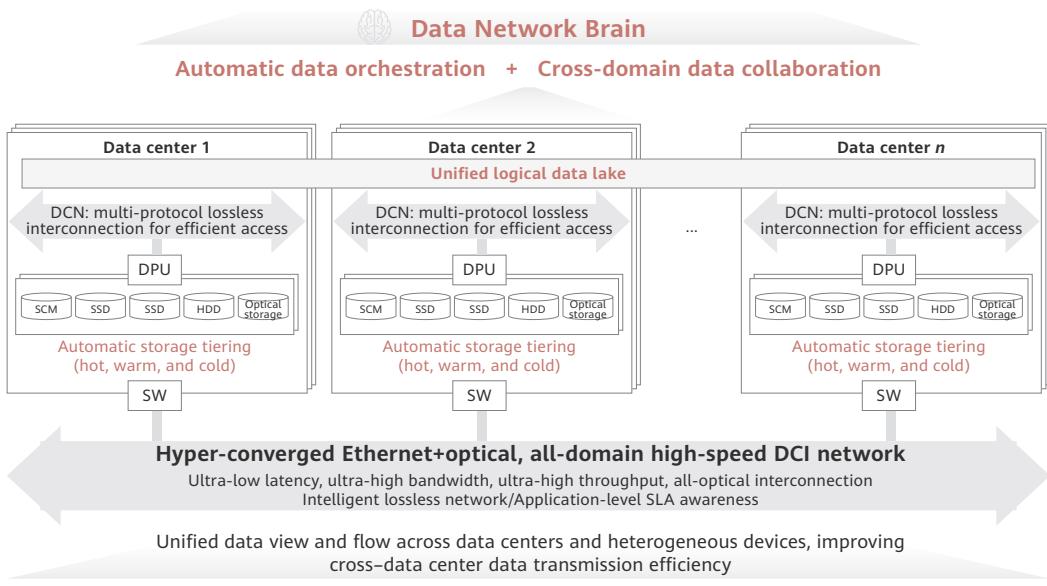


Figure 4-5 An architecture of the next-generation global data management system

■ 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.

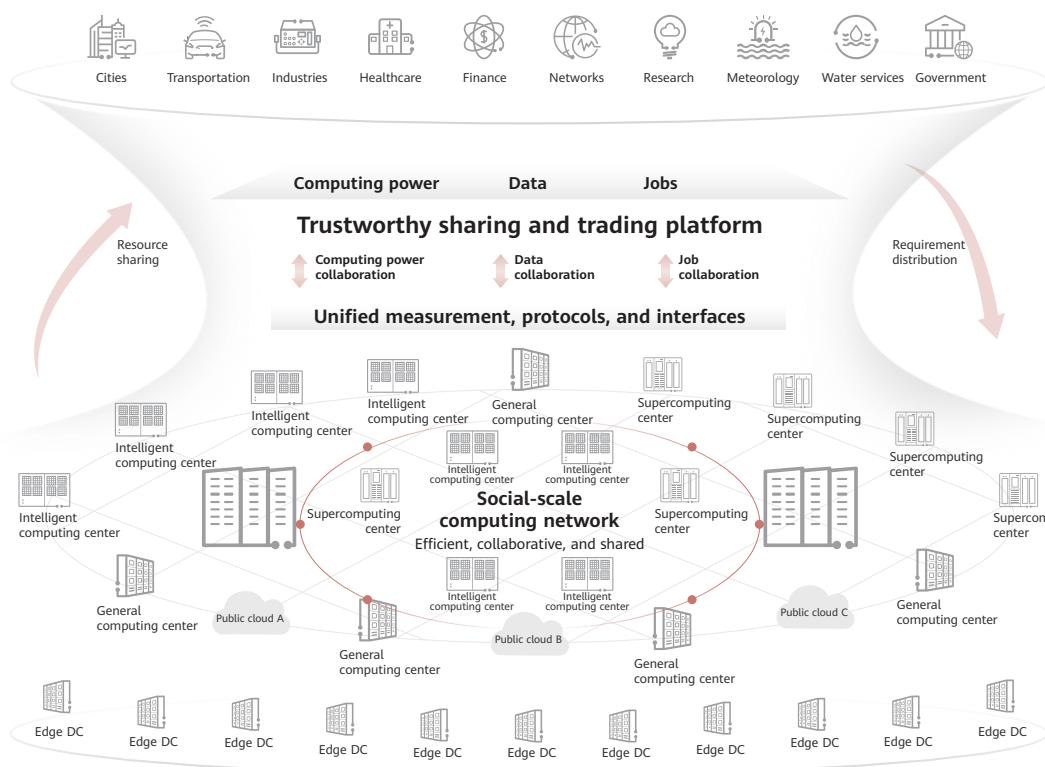


Figure 4-6 An open, shared data center collaboration architecture based on unified standards

■ 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

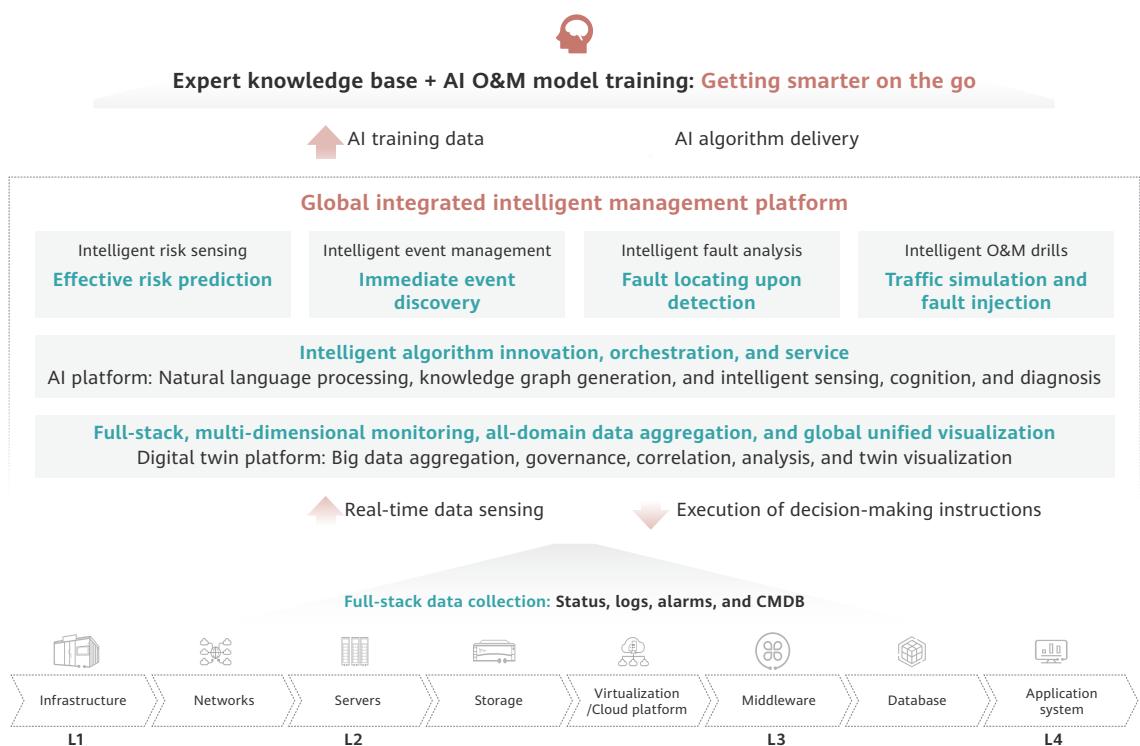
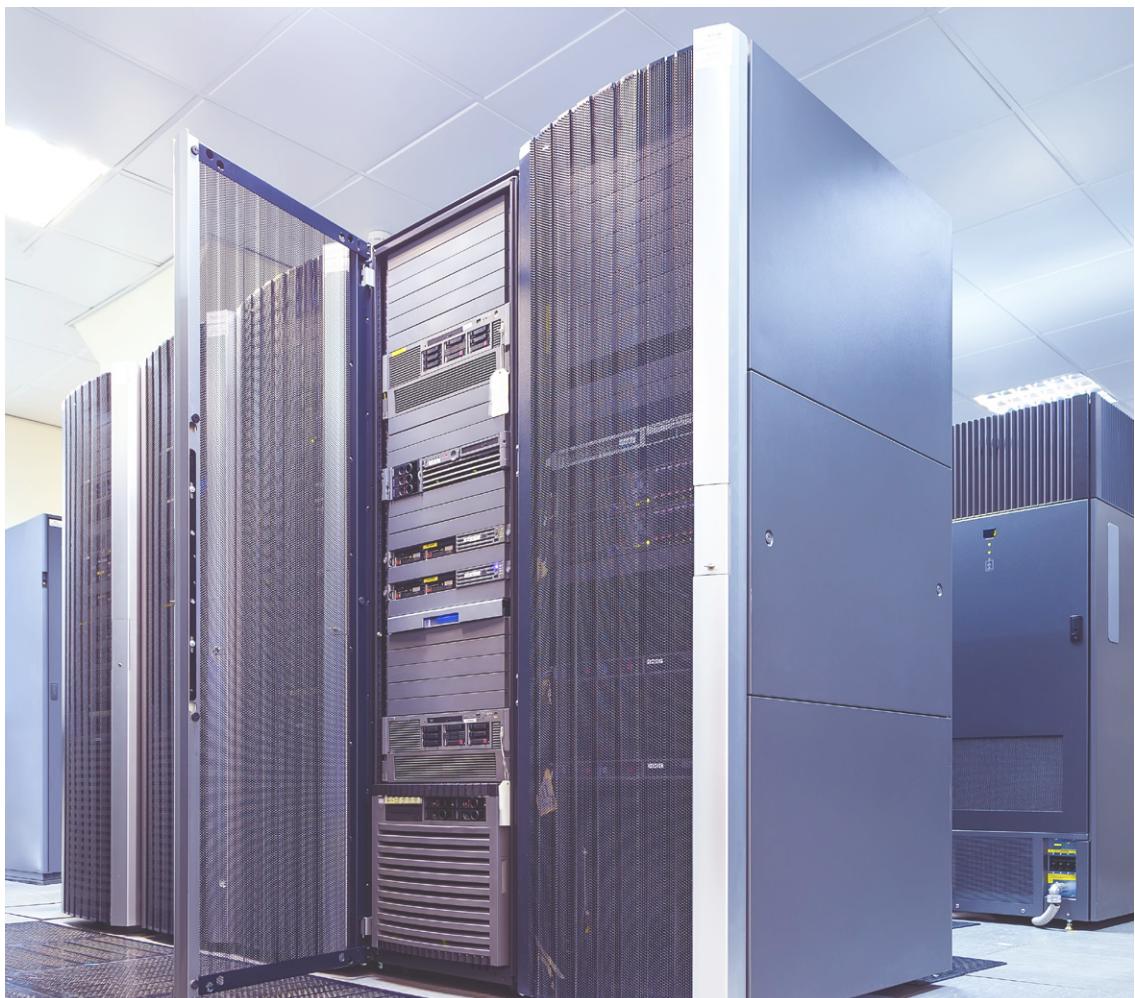


Figure 4-7 Reference architecture of the global integrated intelligent O&M system for new data centers

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).

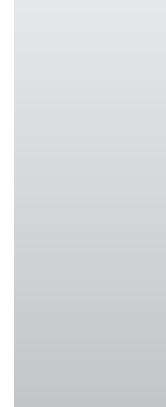






Development and Call to Action

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.

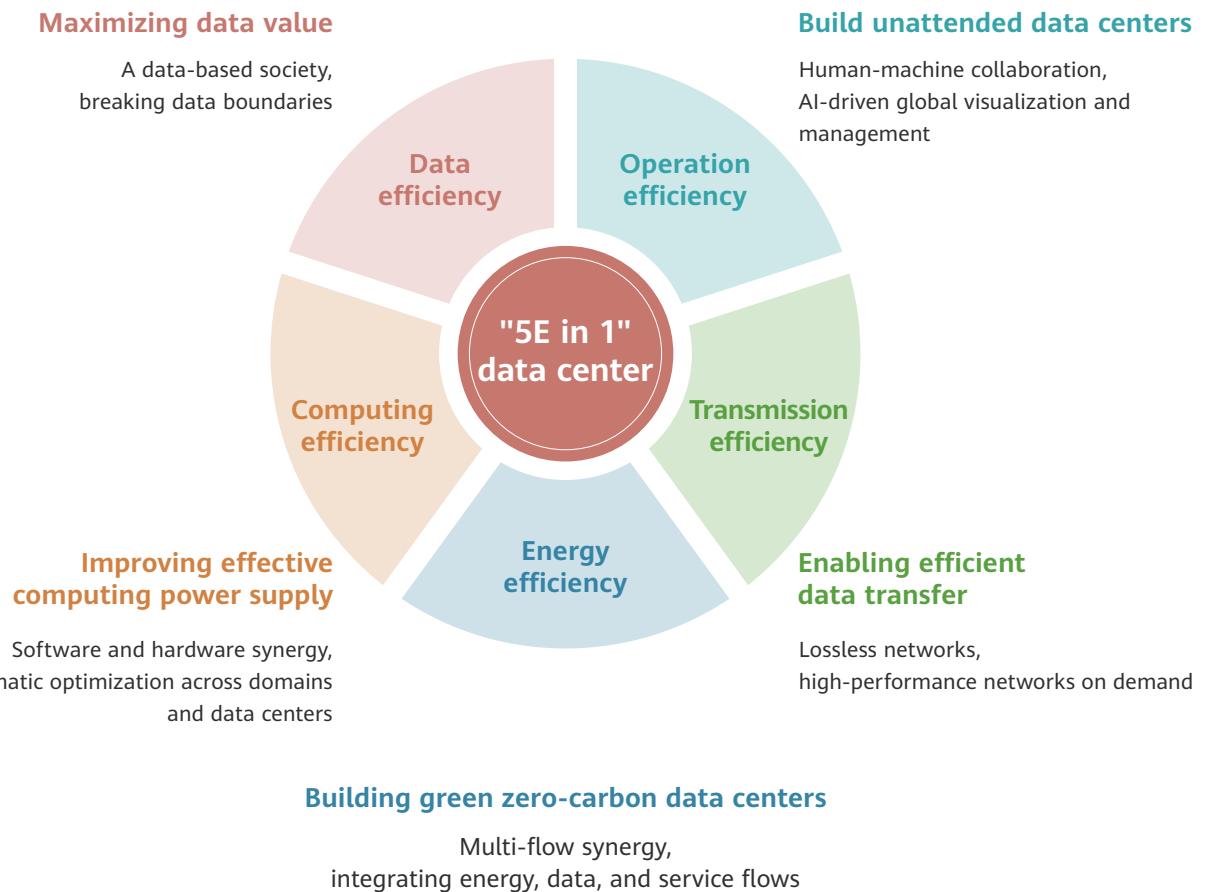


Figure 5-1 5E-in-1 data center

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 | >70 EFlops |
| | AI computing power of a cluster | Effective computing power of a single AI cluster | 100 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 TECHNOLOGIES CO., LTD.

Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P.R. China
Tel: +86-755-28780808
www.huawei.com

Copyright©2023 Huawei Technologies Co., Ltd. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademark Notice

 , HUAWEI , and  are trademarks or registered trademarks of Huawei Technologies Co., Ltd.

Other trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer

Although Huawei endeavors to provide accurate materials and information throughout this document, Huawei does not guarantee the accuracy, completeness, adequacy, or reliability of such materials and information, including but not limited to text, images, data, opinions, suggestions, and analyses. Huawei expressly disclaims any liability for errors or omissions within such materials or information, and gives no guarantees, either explicit or implied, including but not limited to warranties of title and non-infringement of third-party rights. The information in this document is for reference only and does not constitute any offer or commitment. Huawei is not responsible for any actions you take based on the information or content within this document.