



— Version 2024 —

Data Storage 2030



Building a Fully Connected,
Intelligent World



Foreword

Human history is a story of data storage and transmission, from the first oracle bone inscriptions 3,500 years ago, to the advent of papermaking 2,100 years ago, and the emergence of digital storage just more than 60 years ago. We have never stopped in our quest to make knowledge more accessible and data storage more efficient.

In the next decade, the rapid development of 5G/6G, AI, big data, and cloud computing give us the potential to produce yottabytes of data every year, as innovative storage technologies usher in a new era of civilization. The data-centered infrastructure which is efficient, green, and secure will drive us to better understand and explore the smarter world.

Contents

01	Future Data Storage Scenarios	06
----	-------------------------------	----

1.1	Digital Technologies Move Human Society from Informatization to Digitalization	08
1.1.1	Healthcare: Digitalizing Health, Improving Quality of Life	08
1.1.2	Food: Data-driven Food Production for More Bountiful, Inclusive, and "Green" Diets	10
1.1.3	Living Spaces: Whole-House Intelligence Enables Personalized Spaces	11
1.1.4	Transportation: Smart, Low-Carbon Transport Powered by Data Opens up the Mobile Third Space	13
1.1.5	Cities: New Digital Infrastructure Makes Cities More Human and Livable	14
1.1.6	Enterprises: Digital Factory Reshapes Production Models and Enhances Enterprise Resilience	17
1.1.7	Energy: Data Helps Build Energy-Efficient, Low-Carbon Data Centers	20
1.1.8	Digital Trust: Data Security Applications Shape a Trusted Future	22
1.2	The Digital Economy Leads Humankind into the Yottabyte Era	24
1.2.1	As the Data Volume Increases from 175 ZB to 1,003 ZB, We Enter the Yottabyte Era	24
1.2.2	Diverse Data Applications Generate Many Different Types of Data	25
1.2.3	AI Promotes Data Awakening, Heralds a Golden Age of Data, and Unlocks the Value of Multi-Modal Data	25
1.2.4	The Surge of Cloud and Internet Data Necessitates Changes to Data Architecture, Paving the Way for Long-Term Memory Storage	26
1.2.5	70% of Data Generated on the Cloud, Edge, and Devices Is Concentrated, Resulting in Intensive Large-Scale Data Centers	26
1.2.6	By 2030, Endpoints Will Be the Primary Source of New Data. However, the Proportion of Data Generated by Edge and Data Centers Will Also Increase in the Future	27

02	Vision and Key Features of Data Storage by 2030	28
----	---	----

2.1	Advanced Media Application	30
2.1.1	Advanced Media Technology	32
2.1.2	Media Application Innovation	35
2.2	Data-Centric Architecture	38
2.2.1	Decoupling Storage and Compute Resources and Enhancing Compute with Storage	39
2.2.2	Coupling Storage and Compute Resources and Eliminating Repeated Computing Through Queries	40
2.2.3	Cluster Storage	42

2.3 Intrinsic Data Resilience	43
2.3.1 Proactive Data Protection	43
2.3.2 Data Zero Copy	44
2.3.3 Zero-Trust Storage	46
2.4 Intelligent Data Fabric	48
2.4.1 Automatic Data Orchestration	48
2.4.2 Cross-Region Data Collaboration	50
2.4.3 Storage Network	51
2.5 Data Intelligence	52
2.5.1 Service Interface for Content Consumption	53
2.5.2 Data Semantics Extraction	53
2.5.3 Multi-Modal Data Analysis	53
2.5.4 Adaptive Data Modeling	54
2.6 Sustainable Storage	55
2.6.1 Storage System-Level Energy Saving	55
2.6.2 Data Transmission Energy Efficiency Improvement	57
2.6.3 Chip-Level Energy Saving Technologies	58
2.6.4 Green and Intensive Storage Standards	60

4.1 Appendix A: References	62
4.2 Appendix B: Acronyms and Abbreviations	64
4.3 Appendix C: Acknowledgment	66



Future Data Storage Scenarios

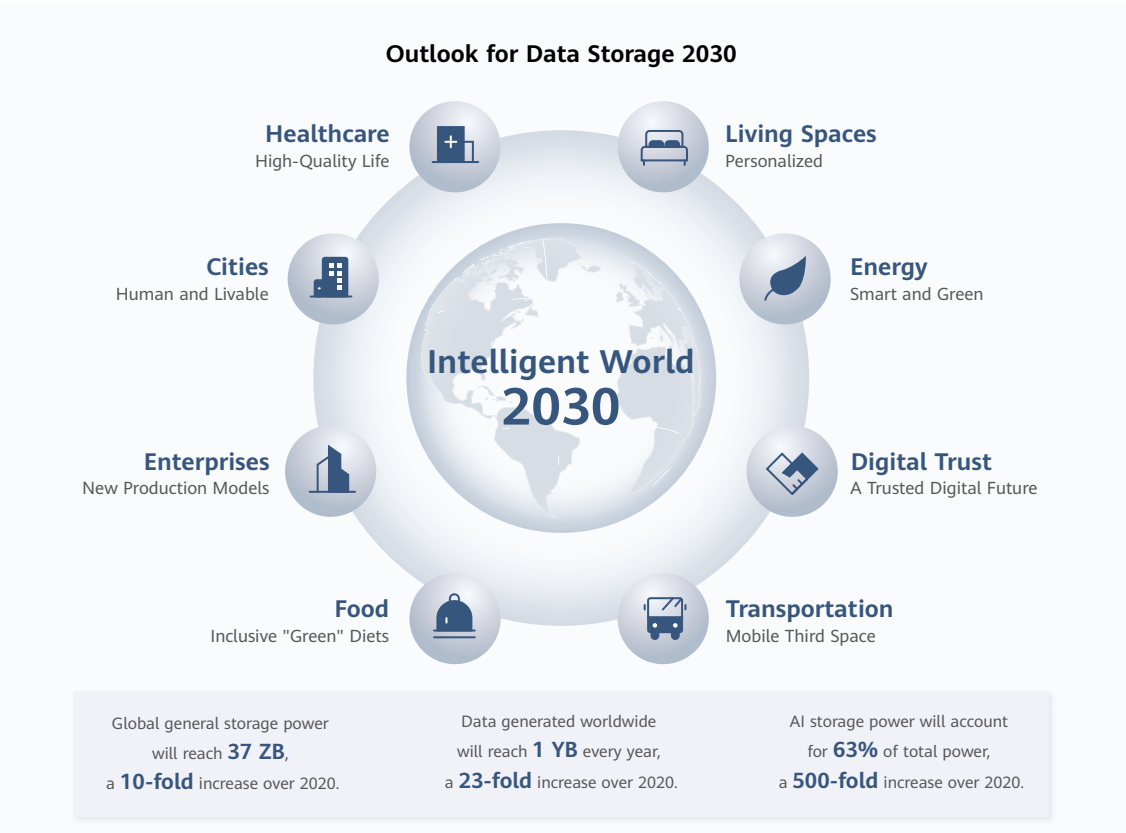


Figure 1-1 Future Data Storage Scenarios



A decade ago, humanity generated just a few zettabytes of data every year, and mobile Internet, cloud computing, and big data were still in their infancy. Today, these technologies are profoundly changing our world, and new technologies such as AI, blockchain, 5G/6G, AR/VR, and the metaverse are driving human society into a new stage of an intelligent world.

By 2030, we will be producing yottabytes^[1] of data every year. Compared to 2020, annual data growth will be twenty-three times higher, general storage power in use will be ten times higher, and AI storage power will have increased by a factor of five hundred^[2]. The digital and physical worlds will be seamlessly converged, allowing people and machines to interact perceptually and emotionally. AI will be ubiquitous and will serve as scientists' microscopes and telescopes, enhancing our understanding of everything from the tiniest quarks to vast cosmological phenomena. Industries which already make extensive use of digital technology will use AI to become more intelligent.

In the next decade, digital technologies will help us move to an intelligent world – a process of the same epochal significance as the age of discovery, the industrial revolution, and the space age.

■ 1.1 Digital Technologies Move Human Society from Informatization to Digitalization

Healthcare: Digitalizing Health, Improving Quality of Life

Over the past decade, the health of humanity as a whole has improved markedly. According to the World Health Organization's (WHO) World Health Statistics 2021, global life expectancy at birth has increased from 66.8 years in 2000 to 73.3 years in 2019. The pace of population aging is accelerating worldwide. Projections indicate that 16.5% of the global population will be 60 years old or over by 2030. This is expected to drive a surge in demand for healthcare services^[3]. According to the WHO's 2019 findings, spending on health is growing faster than the rest of the global economy, accounting for 10% of global gross domestic product (GDP). The WHO also predicts that by 2030, there will be a global shortfall of 5.7 million nurses and 10 million health workers in total. At the same time, we are seeing wide disparities in the global distribution of medical resources – disparities that become especially clear when viewed in terms of population growth.

Looking to the future, new methods of reducing healthcare costs, diversifying healthcare resources and services, and creating new prevention and treatment methods are desperately needed to increase quality of life and make medical treatment more accessible and affordable for all. Many innovative solutions are emerging that may find application within the next ten years. Real-time health monitoring and health data modeling can help weave disease prevention into the fabric of our daily lives. This shifts the paradigm governing our healthcare system from treatment to prevention and covers the following scenarios:

Building a knowledge graph for better health management

The growing popularity of wearables and portable monitoring devices combined with advances in technologies such as the Internet, IoT, and

AI will make personal health data modeling a realistic prospect in the near future^[4]. User-specific knowledge graphs can be built based on data, including health indicators, medical diagnoses, and treatment results. They can compare and analyze these knowledge graphs to formulate personalized health solutions. We can take intervention measures which include guidance on nutrition, exercise, and sleep, as well as mental health support to incrementally improve our lifestyles. For example, a digital health company built a knowledge graph to examine relationships between diet and disease. The company used the knowledge graph to help individuals improve their sleep quality and manage their weight. The health management survey conducted by the company showed that the participants recorded an average 35 minutes more sleep daily, and a total body weight roughly 1.5 kg lighter across the year, which translated into lower probability of disease.

Making infectious disease prediction more accurate

New digital technologies, such as natural language processing, can also broaden the amount of data that can be used for epidemiological management. These technologies allow public health institutions to collect and analyze news articles, reports, and search engine indexes to track major public health events around the world. These institutions can extract valid information from the collected data, build new scientific models, and conduct intelligent analyses of the data so that they can respond to incidents faster and more effectively. For example, a technology company has used natural language processing and machine learning to gather data from hundreds of thousands of public sources, including statements from official public health organizations, digital media outlets, global airline ticketing agencies, as well as livestock health reports and population demographics, to analyze the spread of disease 24 hours a day.

More accurate drug trials shifting treatment from "one-size-fits-all" to "bespoke"

AI can help doctors develop personalized treatment plans by analyzing thousands of pathology reports and treatment plans, and determining which would be most appropriate for each patient. One research institute in Singapore has even created an AI-powered pharmaceutical platform that optimizes medication dosages. The platform can quickly analyze a patient's clinical data, provide the patient with a recommended drug dose or combination regimen based on their specific condition, and revise tumor sizes or biomarkers levels based on available data. In addition, doctors can use the data to determine new courses of treatment for patients.

Achieving safe & precise identification of cancer cells with AI

Precision medicine can help the fight against cancer. During traditional radiation therapy, the radiation typically also kills a large number of healthy cells since the targeted area is quite broad. With the help of AI technology, adaptive radiation therapy (ART) systems can automatically identify changes in lesion positioning and more accurately outline the target areas for radiation treatment. This helps focus the radiation on just the cancer cells and reduces damage to healthy tissue. AI is already enabling accurate identification and automatic contouring of target areas for various types of medical imaging, including CT, ultrasound,

and MRI. A contouring workload that would once have taken hours can now be completed in less than a minute, and the damage caused by radiation therapy to healthy tissue can be reduced by 30%.

By 2030, we will be able to track our own physical indicators in real time with sensitive biosensor technologies and intelligent devices. We will also be able to build health knowledge graphs to manage our health independently, reducing the reliance on doctors.

Driven by ICT technologies, portable medical devices powered by advanced software and hardware, cloud-edge-device computing, and stable networks will be available in grassroots-level hospitals, communities, and households. These devices will collect medical data in real time and upload the data to the cloud for processing. Thanks to the big data knowledge base and AI scheduling, users will be able to access coordinated telemedicine services and track their health from the comfort of home. Building knowledge graphs on the cloud requires the large-scale deployment of storage power.

Huawei predicts that by 2030, the global general storage power capacity will reach 37 ZB, a 10-fold increase over 2020. AI storage capacity will account for 63% of total capacity, a 500-fold increase over 2020.



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

Food is a necessity for all, so the UN has made "Zero Hunger" one of its Sustainable Development Goals (SDGs) for 2030^[5]. Current estimates show that nearly 690 million people are hungry, and if recent trends continue, the number of people affected by hunger would surpass 840 million by 2030. The agriculture workforce is shrinking: According to the International Labor Organization, the proportion of the world population working in agriculture has dropped from 43.699% in 1991 to 26.757% in 2019. Arable land per capita is decreasing: According to World Bank data, arable land per capita has fallen from 0.323 hectares to 0.184 hectares from 1968 to 2018 – a drop of 43%. Overuse of pesticides is causing severe soil pollution: According to statistics, 64% of global agricultural land (approximately 24.5 million square kilometers) is at risk of pesticide pollution, and 31% is at high risk. Simultaneously, the focus of people's diets around the world is shifting to more nutrition and food safety standards. For example, 13,316 food products in China received some kind of green certification in 2018. This number increased to 14,699 in 2019, up 10.4% YoY. This higher demand for green-certified products results in higher requirements on agricultural conditions and technologies.

As we move towards 2030, technology and data are key to empowering agriculture, helping it overcome traditional growth constraints, increasing food production across the board, and bringing "green" food to every table around the world. Digitalized agriculture will cover the following scenarios:

Using accurate data, not experience, to guide cultivation

As the saying goes, there is "a time to plant, and a time to pluck up that which is planted." Decisions on when to sow, fertilize, and use pesticides are highly informed by personal experience. However, this leaves a lot of room for uncertainty, and whether any given year yields a good harvest is still ultimately up to fate.

ICT technologies bolster agriculture with accurate data based on analysis of soil moisture, ambient temperature, crop conditions, terrain, climate models, and pests. Precise controls are provided for optimally paired soil and crops. With maize, for instance, data-powered adaptive sowing can increase crop yield by 300 to 600 kilograms per hectare of land.



Taking a digitalized, factory-like approach to protect agricultural production from environmental conditions

One typical example of industrialized agriculture is indoor, vertical farms that use data to build standardized growth environments without needing to consider geographical constraints. In vertical farms, farmers are able to artificially create ideal environments for their crops by precisely controlling light, temperature, water, and nutrient delivery based on the needs of each crop at every step of the process. Vertical farms don't require pesticides or soil, and reduce agricultural water waste. They are not affected by climate, providing consistent and ideal conditions for fresh produce. These smart agricultural models are globally replicable, as the ICT control system and data model used in one vertical farm can be used anywhere else to achieve almost identical results. Recent pilot programs for vertical farms have found that, if harvested every 16 days, a 7,000

square meters area can yield a staggering 900,000 kilograms of vegetables every year.

By 2030, ICT technology will enable us to connect key agricultural production factors, such as farmland, farm tools, and crops, and collect and utilize data on attributes like climate, soil, and crops. New plant modes like vertical farms and precise data analysis will lay the way for yield-boosting agricultural processes.

Huawei predicts that by 2030, the data generated worldwide will reach 1 YB every year, a 23-fold increase over 2020. With the wider application of data in agriculture, we will build a more resilient and green food system. The total amount of data generated by global agriculture will reach 4 ZB each year, a 23-fold increase over the amount in 2020.

Living Spaces: Whole-House Intelligence Enables Personalized Spaces

As demand for personalized home experiences continues to rise, ICT-enabled smart home technology is gaining popularity. A survey found that about 80% of millennials and 69.2% of baby boomers are interested in smart home technologies^[6]. In the UK, 80% of consumers are now aware of smart home technology and it is second only to mobile payments in consumer awareness of a basket of tech trends. Interoperability has risen as one of the most important buying considerations. Interest in smart living spaces that offer enhanced convenience and safety is also on the rise.

Digitalization and data enable future home experiences, of which scenarios are as follows:

Digital cataloguing and automated delivery for offsite storage

New communities will deliver comprehensive

services to residents, powered by the Internet of Things (IoT), Gigabit fiber networks, and other new advanced infrastructure. Services such as virtual community events and smart pet management will bring residents and their communities more closely together. Groundbreaking new design concepts will also start changing the way our homes look at the household level.

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

Whole-house intelligence that understands usage and creates intuitive experiences

Smart home systems collect data from a wide range of smart home appliances and sensors, over highly-reliable, high-speed networks that reach every corner of your home. They use AI engines to determine what is happening and run appropriate applications. The AI engines, in turn, make informed decisions on how to configure your home appliances, which could be taken independently or in collaboration with other systems, to meet your needs in real time. When implemented properly, smart home systems deliver immersive, personalized, and intelligent experiences that evolve as your usage needs change. The variety of smart home appliances we will see in the coming years is expected to explode. They will work together to intelligently anticipate and meet your needs in different situations. For example, a sleep support solution could easily be created for the bedroom by designing a system that automatically adjusts the softness of your mattress and pillow to suit your body and sleeping habits, and changes

the bedroom lighting to stimulate the production of melatonin – the hormone that helps you fall asleep. Bedroom speakers could play music to relax you, and air conditioners could keep track of temperature, humidity, and oxygen levels.

In 2030, your home may be full of smart appliances that bring a new level of interactivity to your lifestyle and entertainment. The building you live in may be supported by a great variety of smart control systems, and smart functions may be more widely available in your local community. However, none of this will be possible without connections that deliver high bandwidth and extremely low latency.

Huawei predicts that by 2030, 25% of homes will have access to 10 gigabit fiber broadband. The number of global smart home households will soar to 1.8 billion and the annual data volume will be a staggering 23 ZB.



Transportation:

Smart, Low-Carbon Transport Powered by Data Opens up the Mobile Third Space

Travel by private cars is an important part of modern life. In 2020, the vehicle-miles traveled across the US totaled 2.83 trillion miles. In Europe, vehicles travel more than 12,000 kilometers a year on average. Transport systems now face many challenges: Traffic jams are becoming increasingly common and transportation accounts for 26% of global carbon emissions^[7].

All of the key elements (vehicles, traffic lights, pedestrians, etc.) need to be connected using ICT technologies and are infused with big data to support decision-making, so that each phase of a journey can be more intelligent and less carbon-intensive. Digitalized transportation will cover the following scenarios:

Self-driving vehicles in the fast lane

Self-driving vehicles are achieving higher levels of automation, from L2 and L3 to L4 and L5. Buses, taxis, low-speed logistics, and industrial transport (logistics and mining) will be the first commercial applications of autonomous driving.

Low-speed public roads: Self-driving vehicles have delivered positive results in fields such as logistics and distribution, cleaning and disinfection, and patrolling. Unmanned vehicles for logistics and distribution can successfully drive at low speeds on roads with less complicated conditions. This means they can provide safe unmanned delivery services on public roads. Low-speed unmanned vehicles have provided valuable support during the fight against COVID-19, especially in the transportation and distribution of medical supplies, cleaning and disinfection, patrolling, and checking temperatures.

High-speed semi-closed roads: Heavy trucks drivers are expensive, and they frequently breach rules by overloading their vehicles and working overtime. So autonomous driving of heavy trucks would quickly help industries cut costs and work more efficiently, making this a compelling business case. According to a Deloitte report on smart

logistics in China, technologies like unmanned trucks and artificial intelligence will mature in a decade or so, and will be widely used in warehousing, transportation, distribution, and last mile delivery.

Special non-public roads: Autonomous driving demonstrates its commercial value with high levels of safety and efficiency in environments like mines and ports. While working autonomously, many mechanical vehicles, such as mining trucks, excavators, and bulldozers can work together. In the event of a fault or danger, the commander can remotely pilot the vehicle to a safe area from the control center. At the Yangshan Port in Shanghai, 5G-powered L4 smart-driving heavy trucks can drive at speeds of up to 80 km/h, and the distance between vehicles can be shortened to 15 meters. Thanks to the centimeter-level precision of the BeiDou GPS system, the vehicles can come to a stop within 15 seconds of a command, with an error of only 3 centimeters. The use of autonomous vehicles has brought a 10% improvement in vessel loading/unloading times.

Public roads: Robotaxis are an obvious business model for self-driving companies. According to one study, robotaxis could replace 63% of carshare and taxis and 27% of public transport. Autonomous driving technologies will lead to more innovative changes. Cars can become the mobile third space, catering to many different scenarios. This will disrupt the business models of existing industries. Self-driving food trucks may become the standard of the future, and dinner with friends and family may take on a whole new form: After you book a lunch, a self-driving food truck will pick you up and carry you along whatever scenic route you choose. You can enjoy the views while dining and chatting, all within a private space. This model would eliminate the need to visit restaurants and ensure privacy during the meal.

Urban air mobility

In the future, airspace will become an important resource for urban transportation. An efficient air-based urban transportation network will greatly free up roads, reduce travel times, and improve the efficiency of logistics and emergency services.

Air emergency rescue systems: Over the past decade (2010–2020), skyscrapers have sprung up in many cities, creating new safety hazards. Firefighting and emergency medical services in skyscrapers will be a new challenge for cities. Air emergency rescue offers a new solution to these challenges, allowing firefighters and medical personnel to better protect lives and property by quickly reaching higher floors to put out fires or assist people.

Air metro/air taxis: Convenient and efficient transportation is one of the core needs of urban residents. eVTOL will prove to be an effective tool to improve the urban transport experience. Four-seat aircraft of multiple companies are now capable of reaching cruise mileage of about 100 km. Pilot projects have begun for air passenger transport services. In 2019, a Chinese company launched the world's first urban air mobility (UAM) service in Zhejiang, cutting road trips that normally took 40 minutes to a five-minute air hop.

UAM scenarios require a fast and stable space-air-ground integrated network and positioning system, cost-effective and reliable visual sensors and lidar, secure and stable automatic flight algorithms, and an efficient, real-time command and dispatch platform.

Future transport will be a multi-dimensional innovative system. The shift to electric, autonomous, shared, and connected vehicles will create an intelligent, convenient, low-carbon transport experience to reshape the transport experience, create innovative mobility services, enable more efficient sharing of transport resources, alleviate traffic congestion, and reduce the environmental pollution caused by traffic. This is how we will resolve the conflict between the surging demand for transport and the urgent need to decarbonize.

Huawei predicts that by 2030, electric vehicles will account for 82% of all global vehicles sold, and the penetration rate of new self-driving vehicles in China will reach 30%. The vehicle-level storage capacity will exceed 500 PB.

Cities: New Digital Infrastructure Makes Cities More Human and Livable

Rapid advances in new technologies, such as 5G, cloud, AI, blockchain, and intelligent sensing are opening up more possibilities for smart cities, which represent the best places to create new applications for these technologies. In 2020, nearly 1,000 exploratory smart city projects were underway worldwide. In 2020, this spending totaled nearly US\$124 billion, an increase of 18.9% over 2019^[8].

Advancing digital transformation has become one of the key pathways to sustainable development for the world's leading cities. Digitalization and data support the following future scenarios:

Nanosensors track the pulse of the city

Digital cities depend on data, which comes from a wide array of sensors scattered throughout the city. While there are various types of sensors in use currently, one particularly cost-effective and disruptive technology — nanosensors — is expected to drive the next revolution. As such, the MIT Technology Review listed Sensing City as one of the 10 Breakthrough Technologies 2018.

Graphene gas nanosensors are ultra-sensitive to odors. An American university has created a novel type of nano-coating using graphene, and when the coating is applied as a nanofilm on gas

sensors, it delivers a 100-fold increase in molecular response compared to the best available sensors that use carbon-based materials. In the future, these sensors will be able to accurately identify hazardous, toxic, or explosive gases in the air, thereby greatly enhancing cities' capability to detect dangerous substances.

Nanocrack-based acoustic sensors are able to recognize specific frequencies of sound, which perform far better than conventional microphones at separating out sounds in a given frequency range. For example, when these nanosensors are placed on the surface of a violin, they can accurately record every note of a tune and "translate" it so that a connected device can accurately recreate an electronic version of the tune. When this kind of sensor is worn on the wrist, it can accurately monitor a person's heartbeat. Breakthroughs in this technology will greatly enhance acoustic monitoring in urban infrastructure.

All-optical information exchange, 10 gigabit interconnection

The digital transformation of cities poses challenges to massive flows of information, however, 10 gigabit interconnected all-optical cities can unleash tremendous value and growth potential. In April 2021, Shanghai became the world's first all-optical smart city. Owing to its F5G optical network, the city is able to deliver stable connections with latency below 1 millisecond anywhere in the urban area. The deployment of this high-speed optical network has laid a solid foundation for Shanghai's future digital transformation. The future architecture of an all-optical city will consist of four parts:

All-optical network access: All network connections will be optical, including homes, commercial buildings, enterprises, and 5G base stations. The all-optical transmission network will be extended into edge environments like large enterprises, commercial buildings, and 5G base stations. This will enable the digital transformation of many different industries, and support the development

of F5G+X and 5G2B industrial applications.

All-optical anchors: Connections originating in home broadband, enterprise broadband, 5G networks, or data centers will be routed and transmitted through all-optical networks; optical networks will support multiple different fixed access technologies and provide one-hop connections to the cloud.

All-optical switching: One-hop access to services through urban optical networks. All-optical cross-connect technologies are used to build multi-layer optical networks that support one-hop access to services; high-speed inter-cloud transmission; and high synergy between cloud and optical networks.

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

Intelligent hubs cut out the human factor from urban management

With the breaking down of data barriers, AI evolves from partial intelligence to all-scenario intelligence, creating a new public governance subject. Future cities need a powerful smart central platform that aggregates massive data from all corners of the city, on the other hand, the platform transforms data into an advanced city governance capability, benefiting thousands of industries and greatly improving city governance efficiency and user service experience. Early-stage explorations by Toyota: In Toyota's plan for the city of the future, each house, building, and vehicle will be equipped with sensors. Data from these sensors will then be aggregated into a city's intelligent hub. Then AI will be used to analyze people's surroundings and then guarantee the safety of pedestrians and drivers by keeping them separated.

Proactive, precise provision of government services



Machine recognition technology makes contactless services possible. Today, in most of China's developed provinces, citizens do not need to go to government offices to access government services. They can now access them directly through their smartphones. Over the next decade, the digitalization of government services will be taken to the next level.

In the future, as governments aggregate more massive data and their AI technologies mature, they will be able to deliver government services in a more precise and proactive way; manage their municipalities more efficiently; and improve their service experience. Let's look at smart care for the elderly as an example. Communities in Shanghai have installed smart water meters for elderly people who live alone and agree to the installation. If the total water used within a 12-hour period falls below 0.01 cubic meters, the meter will send an alarm to the central network and community workers. These workers will then visit the elderly person in question to check whether everything is normal. Such attentive care

demonstrates a real human touch and concern for the elderly.

ICT technologies like 5G, optical networks, AI, cloud, blockchain, and intelligent sensing will all be rolled out rapidly over the next decade. Cities will soon welcome a period of 10 gigabit connectivity, with 10 gigabit wireless services becoming available to organizations, homes, and individuals.

The application of these ICT technologies in cities will significantly boost our ability to make use of limited resources, to manage our cities efficiently, and to give citizens a positive experience. ICT technologies will help cities achieve their sustainable development goals, and make our cities more human and livable.

Huawei predicts that by 2030, cities contribute to 96% of the data, and 42% of the data emanates from monitoring, scheduling, and management of infrastructure linked to cities.

Enterprises:

Digital Factory Reshapes Production Models and Enhances Enterprise Resilience

Over the next decade, population ageing will lead to a huge worldwide labor shortage. According to a report published by the UN, the global population aged 65 and over is projected to exceed 12% of the total population by 2030, while the global population aged under 25 will decrease from 41% in 2020 to 39% in 2030. By 2030, we can expect a deficit of 85.2 million workers around the world. Take manufacturing for example. By 2030, this sector is estimated to face a global labor shortage of 7.9 million workers, leading to an unrealized output of US\$607.14 billion^[9].

Consumer demand is set to become much more diverse, which will profoundly change production models, forcing businesses to innovate. For example, as the "singles economy" gains traction, companies can rapidly adjust their products by targeting solo dining, small home appliances, and mini karaoke booths. In addition, companies need to take the initiative and stimulate demand through emotional appeal, and rapidly produce combinatorial designs for product appearance, images, and implications. For example, they can customize limited-edition products or launch co-branded products within the shortest time possible.

What's more, black swan events pose new challenges to production continuity. Due to the pandemic, it is estimated that global GDP suffered

US\$3.94 trillion in lost economic output in 2020. The top risk to enterprise growth is supply chain disruptions. Therefore, determining how to use and protect data, reshape production modes, and enhance supply chain resilience are now vital challenges that companies must take very seriously.

Collaborative robots

More and more enterprises are subject to labor shortages, which requires new forms of productivity come in. Collaborative robots are a type of industrial robot. They were initially designed to meet the customized and flexible manufacturing requirements of small- and medium-sized enterprises. To combat labor shortages, they are now an essential complement. Collaborative robots are suitable for jobs that people are unwilling to do, such as highly repetitive work like sorting and packaging. Collaborative robots have several unique advantages:

Safer: Collaborative robots are compact and intelligent, and their sophisticated sensors enable them to stop in an instant. They can work closely together with human workers on the production line to get the work done.

Faster and more flexible deployment: Collaborative robots feature user-friendly programming, such



as programming by demonstration, natural language processing, and visual guidance. They can be placed in new positions at any time, and programming and commissioning can be completed rapidly, so they can start working very quickly.

Lower total cost of ownership (TCO) and shorter payback period: The price and annual maintenance cost of collaborative robots are significantly lower than those of traditional industrial robots. According to China's Forward Industry Research Institute, the average selling price of collaborative robots has halved over the past several years.

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

Autonomous mobile robot

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

awareness. They feature dynamic route planning, flexible obstacle avoidance, and global positioning. The AMRs used in industrial manufacturing and logistics are mainly powered by the simultaneous localization and mapping (SLAM) technology, laser navigation, visual navigation, and satellite positioning to enable autonomous navigation. AMRs make automated and unmanned logistics possible. This includes unmanned operations for sorting, transporting, and storing goods.

Digital simulation and flexible manufacturing

To respond to changing market conditions and set themselves apart in the face of fierce competition, companies must take the initiative and embrace new production models. That's why an increasing number of companies are looking to concepts like flexible manufacturing. Flexible manufacturing relies on ICT technologies. Simulation, modeling, VR, and other ICT technologies can be used to simulate the entire new manufacturing process. This will reduce the cost of new product development and design, and support more accurate adjustment costs and capacity planning. In addition, the intelligent task scheduling system schedules the production tasks and allocates production materials and tools based on known features such as the factory's production capacity, order complexity, and delivery deadlines. Flexible manufacturing uses ICT technologies such as visual programming, natural language interaction, and action capture to help factories reprogram



and define equipment quickly and easily. This will help companies promptly meet consumer demand for flexible manufacturing. Flexible logistics uses ICT technology to effectively manage warehousing and logistics, which prevents omissions and other errors in the shipment process. Take furniture producers as an example. With large-scale customization, every board, decorative strip, and handle may need its own identification code or radio frequency identification (RFID) tag to facilitate automated packing and loading, and to support traceability throughout the whole transportation and distribution process. This is the intelligent customized production that centered on consumers.

Resilient and intelligent supply chains that help enterprises respond to crises

More and more companies regard building a resilient and intelligent supply chain as one of their most important strategic priorities. Supply chain visualization uses ICT technology to collect, transmit, store, and analyze upstream and downstream orders, logistics, inventories, and other related information on the supply chain, and graphically display the information. Such visualization can effectively improve the transparency and controllability of the whole supply chain and thus greatly reduce supply chain risks. Supply chain visualization supports the tracking of materials and equipment in upstream activities. Logistics information is displayed in real time, including information on packing, goods logged in, goods logged out, and inspections; goods can even be traced throughout the production process. This enables companies to detect and rapidly respond to any logistics emergency by promptly adjusting logistics routes to ensure the timely and safe delivery of goods.

A remote monitoring system monitors the environment in warehouses in real time. This system uses various sensors to graphically display operations and maintenance (O&M) information such as temperature, humidity, dust, and smoke. This allows the timely detection of any signs of fire or water leakage, which enables prompt



intervention and prevents material losses. Goods can be tracked in real time as they are logged in to and out of warehouses. With the movement of goods, IoT, RFID, and QR code technologies are used to automatically identify and register goods, and the warehousing status data of goods can be accessed remotely in real time.

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

In the future, product design, process design, equipment functions, logistics, and distribution will all be reshaped to become more flexible and serve new people-centric production models. Powered by digitalization, supply chains will be visualized and expand into supply networks. This will enable companies to become more resilient than ever and more capable of responding to volatile markets.

Huawei predicts that by 2030, the digital transformation of enterprises will further promote the application of data services in enterprises. In addition, data services are forecast to account for 87% of enterprises' application expenditures, while AI computing will account for 7% of a company's total IT investment.



Energy: Data Helps Build Energy-Efficient, Low-Carbon Data Centers

Climate change is a global challenge, and many countries have come together to tackle it. At the UN Climate Conference (COP 21) in 2015, parties to the Paris Agreement agreed to intensify efforts to limit global warming to well below 2°C, preferably to 1.5°C, compared to pre-industrial levels, and set the goal of reaching net zero CO2 emissions globally around 2050. In other words, by the midpoint of this century, the CO2 emitted by human activities needs to be matched by the CO2 deliberately taken out of the atmosphere^[10]. At the 75th UN General Assembly in September 2020, China pledged to peak its carbon emissions by 2030 and achieve carbon neutrality by 2060. Concerted efforts are needed to combat climate change and drive the transformation of the global energy in three areas: energy supply, consumption, and carbon fixation.

With the increasing complexity of energy networks and the increasing digitalization of the energy sector, ICT technologies have become an important part of decarbonization solutions. The key questions for global warming now are: How can we further increase the share of renewables in the energy mix? How can we adapt to the new energy mix? And how can we fully harness the power of ICT technologies? Smarter green energy drives sustainable economic development and

supports the following scenarios:

Offshore wind, a promising energy source for the future

In 2020, the worldwide energy installed capacity from renewable sources increased by 280 gigawatts (GW) or 45%. Of this, 114 GW was contributed by wind, an increase of more than 90%. Some European countries are actively exploring offshore power generation. In 2020, the installed offshore wind capacity in the UK and Germany exceeded 18 GW, accounting for 51% of the world's offshore wind capacity. Offshore wind energy still provides only 0.3% of the electricity globally, and there is huge room for expansion.

Wind conditions at sea are better than on land, with wind speeds typically 25% higher than on coastal land and less turbulence, resulting in a dominant and stable wind direction. The capacity of offshore wind turbines can be 3 to 4 times greater than that of inland wind turbines. There are fewer calm periods at sea, so offshore wind turbines can generate power for 3,000 hours a year, which makes for more efficient use of generator capacity. The technology advances have led to a significant reduction in the cost of offshore wind installations, and the offshore

power generation cost in 2040 is expected to be 60% lower than in 2019. The Global Wind Energy Council (GWEC) forecasts that global offshore wind capacity will increase from 29.1 GW today to 234 GW by 2030. Annual installations of offshore wind capacity are expected to grow at 31.5% per year over the next five years. This is a boom time for offshore wind power.

Floating PV, the latest trend in solar PV

According to the Snapshot of Global PV Markets 2021 by the IEA, the total installed capacity of photovoltaics at the end of 2020 was 760.4 GW. In 2020, solar PV accounted for approximately 42% of the total power generation from all new renewable energy sources. Large inland PV power plants are the most common form of PV installation, but there are a number of problems associated with inland solar farms: land acquisition, high costs, and poor performance under high temperatures. Floating PV (FPV) is a new direction for solar PV. Compared with land-based PV (LBPV) systems, installation of FPV systems on water saves land for agricultural use. The lack of obstacles on the surface of the water means less shading loss and less dust. In addition, the natural cooling potential of the body of water may enhance PV performance. In 2020, a research team from Utrecht University in the Netherlands simulated an FPV system on the North Sea. They found that the apparent temperature at sea was much lower than on land. The apparent temperature difference was nearly twice that, at 9.36°C. This study found that an FPV system performs 12.96% better on average on an annual basis than an LBPV system.

As the technologies mature, rapid growth is anticipated in FPV. On July 14, 2021, Singapore's Sembcorp Industries unveiled a floating solar farm deployed on the Temgeh Reservoir. With 122,000 solar panels spanning across 45 hectares (equivalent to about 45 football fields), the 60 megawatt-peak (MWp) solar farm is one of the world's largest inland floating PV systems. According to Rethink Energy, the global FPV market capacity will exceed 60 GW by 2030.

Globally, the estimated potential capacity is 400 GW, meaning that FPV could double the current global installed capacity of solar PV. The floating solar market is set to accelerate as the technologies mature, opening up new opportunities for scaling up global renewables.

Low-carbon data centers and sites

According to the IEA, since 2010, the number of Internet users has doubled, global Internet traffic has increased by 12 times, and the electricity consumed by data centers and transmission networks has increased significantly. The global electricity demand from data centers was about 200 terawatt-hours (TWh), accounting for about 0.8% of global electricity demand. Data networks consumed approximately 250 TWh in 2019, accounting for approximately 1% of global electricity consumption, with mobile networks making up two-thirds of this figure. Data center power consumption in China alone is expected to exceed 400 billion kWh in 2030, accounting for 3.7% of the country's total power consumption. If data center power usage effectiveness (PUE) improves by 0.1, the result will be 25 billion kWh of power saved and 10 million fewer tons of carbon emissions. If all data centers use green power, carbon emissions will be reduced by 320 million tons each year. Green power and PUE optimization are key measures for low-carbon data centers.



In addition to applying renewable energy and free cooling, AI is another effective way to make data centers more efficient and save energy. Sensors in data centers collect data such as temperature, power levels, pump speed, power consumption rate, and settings, which are analyzed using AI. Then, the data center operations and control thresholds are adjusted accordingly, reducing costs and increasing efficiency. AI is used in data center cooling to reduce the energy used for cooling by 40%. According to Datacenter Dynamics, the Boden Type Data Center (BTDC), an experimental data center built in Sweden with funding from the EU's Horizon 2020 programme, has achieved a PUE level of 1.01 by using AI algorithms to achieve synergy between the cooling system and computing loads, server fans, and temperatures, in addition to environmental cooling. As AI technology matures, with green electricity and free cooling, data centers and communication

networks will be more efficient and reach zero-carbon goals eventually.

By 2030, the global carbon emissions need to be reduced by half. For production sites, new energy such as wind and PV can fuel new deployment modes, while on the consumer side, electrification can help achieve electricity substitution goals. ICT not only helps itself but also other industries to reduce carbon missions.

Huawei predicts that by 2030, the power consumption of data centers in China will account for 3.7% of the total China's power consumption (incl. 25%–32% consumed by storage systems). Green energy will play an important role in slashing emissions.

Digital Trust: Data Security Applications Shape a Trusted Future

Driven by digital transformation, interactions between organizations, between organizations and customers, and within organizations, are migrating to the digital world ever more quickly. Valuable digital assets are generated during these processes. Digital trust is a complex system that covers a range of areas, including privacy, security, identity, transparency, data integrity and governance, and compliance^[11]. New technologies such as blockchain, privacy-enhancing technology, and AI and new rules will help shape a trusted digital future.

Smart contracts on the blockchain

Digital assets bring unprecedented quick access and convenience to organizations and individuals with potential risks of theft and misappropriation. Blockchain-based smart contracts contain terms expressed in a digital form on a blockchain, and the recording and processing of these terms are completed on the blockchain. Blockchain technology allows information to be recorded and

distributed, which ensures that the entire process, from contract storage and access to performance, is transparent, traceable, and tamperproofing. Smart contracts have huge market potential in logistics, e-commerce, finance, insurance, and other sectors. According to Capgemini Consulting, smart contracts may help US consumers save US\$480 to US\$960 per mortgage loan, and enable banks to cut costs in the range of US\$3 billion to US\$11 billion annually by lowering operational costs in the US and European markets. Consumers in the US and EU could save US\$45 to US\$90 per year on their motor insurance premiums, and insurers would reduce the cost of settling claims by US\$21 billion a year globally.

New mechanisms for collecting personal information online

More and more laws and regulations concerning the over-collection of data have been passed in recent years. In the context of big data, a fair digital strategy would contain optimized

mechanisms that balance the privacy of individuals and the interests of data users creating value with consumer data. The level of control data subjects has over their own personal information will be further enhanced while preserving the conventional approach of obtaining informed consent. In 2021, China promulgated its first Personal Information Protection Law. This law emphasizes multiple basic principles for protecting personal information, including openness, transparency, knowledge of purpose, and minimization. In the future, regulatory frameworks will be further refined so that users will have more knowledge and control over the ways in which their data is collected and used, and the associated risks.

The General Data Protection Regulation (GDPR) is currently the most stringent privacy and data security law in the world. It was drafted by the EU and took effect on May 25, 2018. In 2020, the US published its Federal Data Strategy 2020 Action Plan, which includes the goals of protecting data integrity, conveying data authenticity, and ensuring data storage security. On May 27, 2020, Japan passed the Act on Improving Transparency and Fairness of Digital Platforms, which was designed to regulate specific digital platforms and enforce obligations to the public on those platforms. These regulations represent a global trend towards antitrust action in the data domain. As antitrust laws are further modernized and adopted, data

users and third-party companies will be granted more data rights against industry giants. This will help develop a healthy digital trust ecosystem, and prevent large platforms from committing digital security violations or engaging in other behaviors that compromise fair competition, such as illegally obtaining, abusing, and trading personal data.

Blockchain, AI, and other technologies will be the foundation of a digital, trustworthy, world, as they provide better personal privacy and asset protection, can accurately highlight disinformation, and mitigate fraud or data theft risks. Further, privacy-enhancing computation ensure data shared among multiple parties is encrypted without risk of private information leakage.

Huawei believes that by 2030 half of all computing environments will use privacy-enhancing computation, and 85% of enterprises will use the blockchain. While privacy-enhancing computation, blockchain, and IPFS technologies all offer clear advantages in security, it will cause a huge increase in encrypted and distributed ledger data, generating 17 ZB of new data each year. Additionally, over 80% of enterprises are expected to deploy multi-layer ransomware protection systems which cover the storage systems.



1.2 The Digital Economy Leads Humankind into the Yottabyte Era

By 2030, the digital economy will account for 60% of the global economy, and data will become the basis for industry digitalization.

Today, technologies and industries are transforming, and the digital economy is thriving, changing the way people live and work and influencing economic and social development, global governance, and civilization. According to the 2020 Global Digital Economy report by China Academy of Information and Communications Technology, the global digital economy was worth US\$31.8 trillion in 2019, accounting for about 36% of global GDP. The digital economy has maintained rapid growth, developing significantly along the way. The added value of the digital economy has reached CNY35.8 trillion, accounting for 36.2% of GDP and contributing 67.7% to GDP growth.

By 2030, the global digital economy will account for 60% of the global economy. The digitalization of conventional industries is speeding up. By

2030, the output value of the digital industry will reach 9%, a catalyst for economic growth. The digitalization of conventional industries requires digital tools to become more Internet-based, intelligent, and automated, to expand customer scope, to reduce costs, and to improve efficiency.

By 2030, 45% of industries will be digitalized, enabling us to better understand the world and promote AI and smart manufacturing.

Technology allows us to better observe, monitor, track, and process human, social, and earth activities, creating data that enables us to understand and describe the world more accurately than ever. Data and machine learning technologies are driving the development of AI, which will allow for increased automation of services, processes, and communication. By providing customized products based on customer preferences, AI will take efficiency and productivity to a new level.

As the Data Volume Increases from 175 ZB to 1,003 ZB, We Enter the Yottabyte Era

According to IDC and Huawei GIV team, the amount of data generated globally every year increases rapidly with the development of digitalization, from 2 ZB in 2020 to 175 ZB in 2025. In 2030, this figure is expected to reach 1,003 ZB, marking the start of the yottabyte era (1 yottabyte = 1000 zettabytes).

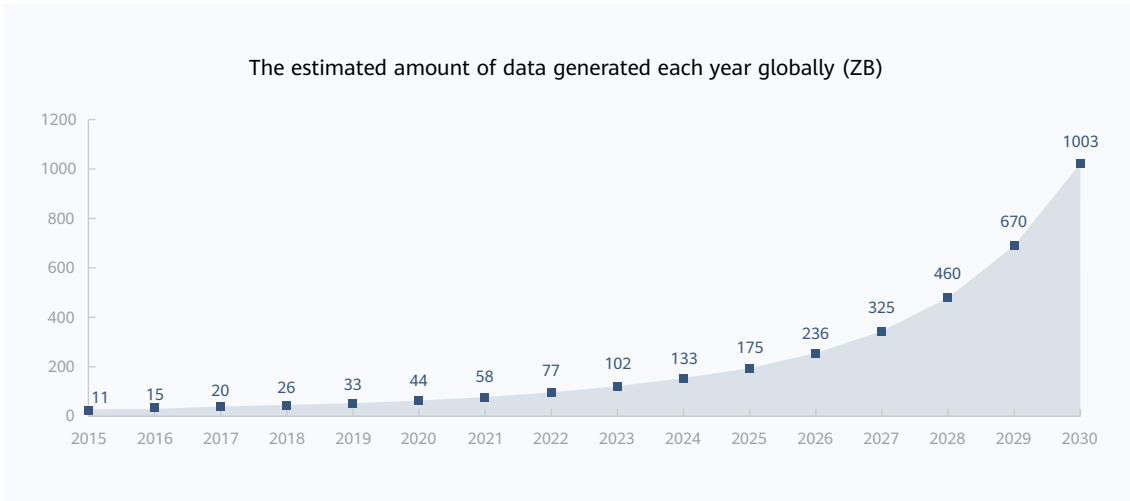


Figure 1-2 Prediction of the total amount of new global data generated each year



Diverse Data Applications Generate Many Different Types of Data

As the wave of digitalization sweeps across industries, data applications are becoming increasingly diverse. In addition to conventional database applications, new applications such as distributed databases, big data, and high-performance computing (HPC) are emerging. On average, an enterprise now has more than 100 types of data applications.

Evolving digital and mobile technologies have significantly changed how enterprises interact with customers. Internet applications, such as mobile apps, have become the most effective platform for driving customer purchases, and they have led to rapid service growth. The resulting surge in the volume of structured data has made core system workloads less predictable and more volatile. To cope with this, enterprises need to have core systems with flexible resource scalability, so that they can quickly expand resources during peak times and release idle resources during off-peak times to avoid waste. In addition, the multi-read and multi-write capability is becoming a mainstay feature in core systems as it ensures high system reliability.

Unstructured data is becoming a key enterprise

asset because it contains a lot of valuable information and comes in different formats, such as text, images, videos, and audio files. By 2030, 1 yottabyte (YB) of data is expected to be generated globally each year, and more than 80% of it will be unstructured. Unstructured data is widely used in enterprises. 56% of enterprises use AI for at least one business function to analyze and process unstructured data in many different scenarios. The increasingly improved enterprise data governance capabilities enable data-driven service growth. Enterprises have also started to use unstructured data in systems that make decisions about production. Examples include online real-time credit approvals in the finance industry and pathological analysis in healthcare. By 2030, it is estimated that 80% of unstructured data will be used to support production-related decisions.

AI Promotes Data Awakening, Heralds a Golden Age of Data, and Unlocks the Value of Multi-Modal Data

Large AI models are having an unprecedented impact on our daily lives, propelling us towards a more intelligent world and ushering in an era of data awakening. Data, as one of the three elements of AI, determines how far AI can go,

so the different types of data and how they are stored and accessed are very important.

First, the amount of hot data has skyrocketed. Statistics show that in China the new data stored in 2023 represented only 2.9% of the total data generated that year. A massive amount of data was discarded at the source and not stored. As AI continues to evolve, the volume of hot data and its importance are also on the rise. More and more data is being stored, and this can be used as a real-time and valuable input for AI. It is estimated that by 2030, all hot data will be stored on SSDs.

Second, the value of warm and cold data is being reexamined, and these types of data are gradually becoming hotter. Warm and cold data refers to information that is not frequently accessed, such as backup and archived data, so it has traditionally been seen as being less valuable. However, since AI requires extensive data for training, warm and cold data is becoming more important. By incorporating this data into the training process, we can enhance the accuracy and generalization of large AI models while also unlocking the value of previously overlooked data. Additionally, warm and cold data that needs quick access is termed active archived data. It is projected that by 2030, more than 60% of enterprises will need to be able to access active archived data at least once a day.

The Surge of Cloud and Internet Data Necessitates Changes to Data Architecture, Paving the Way for Long-Term Memory Storage

In recent years, the rapid development of cloud computing and Internet technologies has had a significant impact on various industries. The demand for data storage in the cloud and Internet fields is growing at an unprecedented rate. Statistics show that about two-thirds of enterprise-level SSDs are delivered to cloud and Internet vendors.

To cope with the explosive rise in the volume of data and rapid changes in service requirements, cloud and Internet vendors, such as Google

Cloud, are promoting the use of diskless reference architectures. A diskless reference architecture allows local disks of servers to be remotely deployed to form a brand-new architecture consisting of diskless servers and remote storage pools. It decouples compute resources from storage resources and enables flexible resource sharing, which greatly improves resource utilization and scalability while simplifying O&M and reducing energy consumption. With its flexibility and efficient storage resource management capabilities, it is expected to provide strong support for the sustainable development of the cloud and Internet industries and become the mainstream architecture. Estimates suggest that by 2030, more than 80% of cloud and Internet enterprises will use the diskless reference architecture.

70% of Data Generated on the Cloud, Edge, and Devices Is Concentrated, Resulting in Intensive Large-Scale Data Centers

The differences between data generated at the edge, endpoints, and core data centers are as follows:

Endpoints refer to devices on the edge of the network, including PCs, mobile phones, industrial sensors, automobiles, and wearables. By 2030, more than 75% of endpoint data will be processed by AI in real time.

Edge refers to servers and devices that process enterprise-level loads. Instead of being located in core data centers, these servers and devices are placed in server rooms, workplaces, or wireless base stations of branch offices to facilitate data processing with reduced network latency. By 2030, more than 80% of edge data will be processed by AI in real time.

Core data centers refer to large-scale data centers, including enterprise data centers, IDCs, and the cloud data centers of public cloud vendors. By 2030, more than 90% of data at core data centers will be processed by AI in real time.

By 2030, Endpoints Will Be the Primary Source of New Data. However, the Proportion of Data Generated by Edge and Data Centers Will Also Increase in the Future

As the number of endpoints continues to increase, endpoints will remain the main source of new data by 2030. It is predicted that the new data they generate will increase by 14 times, accounting for 52% of the total new data by 2030. This is due to the dramatic growth of smart vehicles, wearable devices, and industrial IoT.

There will also be a significant increase in edge devices by 2030. 5G MEC, CDN, remote and branch offices (ROBO), and high-tech media processor applications will become universal and home digital processing centers will begin to scale. In the future, every family will have a digital processing center that connects all home digital or intelligent endpoints, such as mobile phones,

wearable devices, and smart home appliances like refrigerators, to store and process data and assist with everyday routines. By 2030, the data generated at the edge will have increased by 22 times, accounting for 21% of the total data generated.

The cloud is a key node for data aggregation, processing, backup, replication, and transfer. As each operation generates new data, data operations in the data center have an amplification effect, which is enhanced as more data is aggregated in the data center in the future. By 2030, the data generated by data centers will have increased by 18 times, accounting for 27% of the total new data.

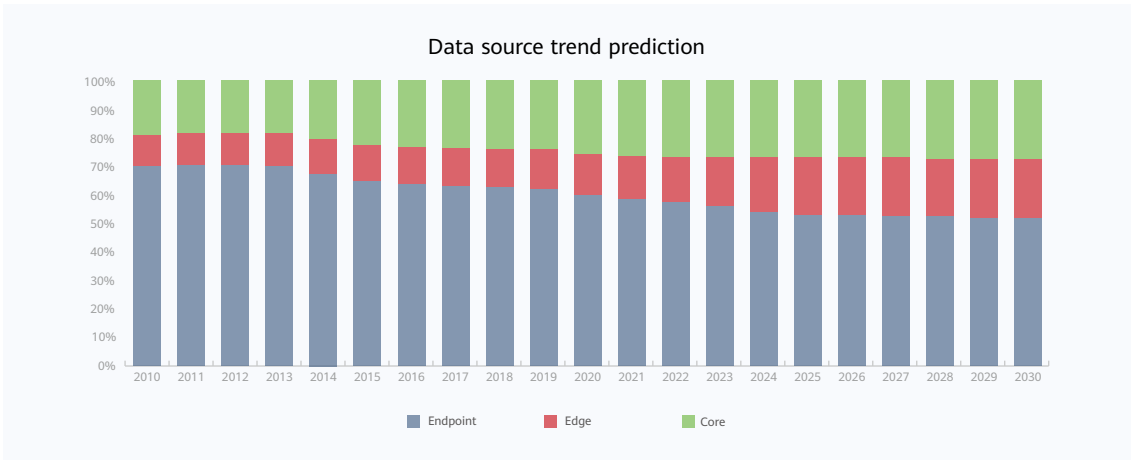


Figure 1-3 Data source trend prediction

A large proportion of data generated on endpoints will be stored in data centers through application systems and backup systems by 2030. The optimization of network and bandwidth make it more convenient and secure to store data, such as web disks, photos, account information, and applications, in data centers. Take the application account as an example. You can use one account to log in to the system via different clients and access the same view and services based on the account and its status stored in the data center.

About 65% of stored data will be stored in data

centers by 2030. Data will be periodically backed up to data centers rather than being stored in endpoints. With more applications requiring real-time processing and low latency, data storage scenarios at the edge will become diversified. This includes intelligent driving training endpoints, real-time edge stream processing, 5GMEC, and VR/AR edge centers. The proportion of data processed at the edge will reach 10% by 2030.

Scattered data will be concentrated to data centers, allowing us to more easily mine data value and lay a solid foundation for digitalization.



Vision and Key Features of Data Storage by 2030

Over the next 10 years, the compound annual growth rate (CAGR) of data is expected to reach close to 40%, and data types are diversified. A single storage media cannot meet diversified data storage requirements. Diversified storage media are required to cope with challenges such as high storage costs, high power consumption, and poor durability. Diversified mass data promotes the development of diversified advanced media and media applications. Intelligent data reduction and joint data coding technologies will increase storage capacity density by several times.

The fast-growing data volume is contradicted by the slow-growing data processing capability, and the data storage capacity and data development are severely unbalanced. The classic CPU-centric architecture cannot meet the requirements of mass data storage and processing. Therefore, the entire architecture needs to be reconstructed with data as the core^[12]. The new architecture supports storage-compute decoupling at the macro level, and computing in-memory at the micro level. The high-throughput, ultra-low latency, and high-scalability interconnection bus break the resource boundary and form a resource pool of processors, memory, and storage, supplementing computing through storage and improving data processing efficiency by several times.

The mounting data transfer requirements and increasingly severe data gravity have formed a fundamental contradiction limiting the value of data. The intelligent data fabric supports cross-region intelligent and efficient data flow, breaks space constraints, achieves what you see is what you get (WYSIWYG), and improves data flow efficiency by one hundredfold. Intrinsic data resilience separates



data use rights, management rights, and ownership rights, promoting trusted data flow^[13]. A secure and reliable data application environment must be built through proactive defense to ensure data privacy and improve trusted data transfer efficiency by thousands of times.

Complex storage systems cannot meet the intelligent data service requirements of emerging multi-cloud applications. Therefore, data service logic needs to be decoupled from data intelligence. Future data storage will have new data awareness and understanding capabilities, supporting the rapid growth of data services in thousands of industries.

The continuous storage energy consumption increase still lags behind the global low-carbon development targets, placing new requirements on the green and low-carbon capabilities of storage. New energy-saving materials, transmission of data with optical signals instead of electronic signals, and dynamic energy-saving technologies promote chip energy saving. New liquid cooling heat dissipation and intelligent system control technologies decrease energy consumption across the entire system. System-level multi-dimensional and intelligent resource control technologies will reduce emissions throughout the data lifecycle, improving energy consumption efficiency by several times in the future and supporting the sustainable development of the future data industry.

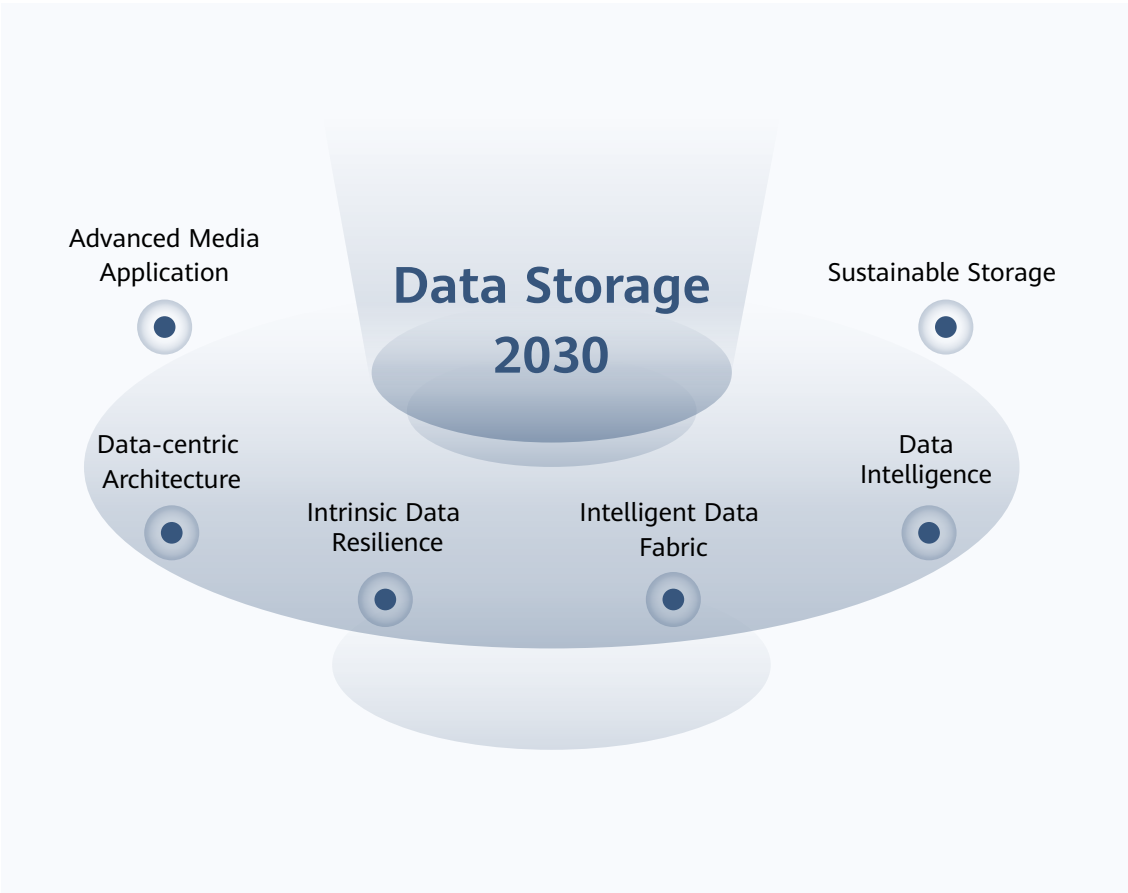


Figure 2-1 Six key features of data storage by 2030

To sum up, future storage will have the following six key features: advanced media application, data-centric architecture, intrinsic data resilience, intelligent data fabric, Data Intelligence, and sustainable storage.

2.1 Advanced Media Application

The evolution of large AI models to being multi-modal is gradually awakening data. More and more video and image data will be saved for training. It is estimated that from 2030, 1 YB of data will be generated each year. The volume of data used for large AI model training is expected to increase more than one thousandfold to 400 EB. Nearly 50 ZB of valuable data will need to be stored every year, a 23-fold increase compared with 2020. Storage media must provide large capacity with cost-effectiveness and low energy consumption, featuring high reliability, high scalability, durability and high security. In addition,

storage devices must have data computing and analysis capabilities to obtain data faster.

Different media have their own advantages and disadvantages. Therefore, multiple media need to be combined to cope with challenges. The evolution trend of different media suggests that the media capacity density will increase by 10 times by 2030. However, compared with the 23-fold increase of data volume, the growth of media capacity density still lags far behind. Media application innovation is required to fill the gap.

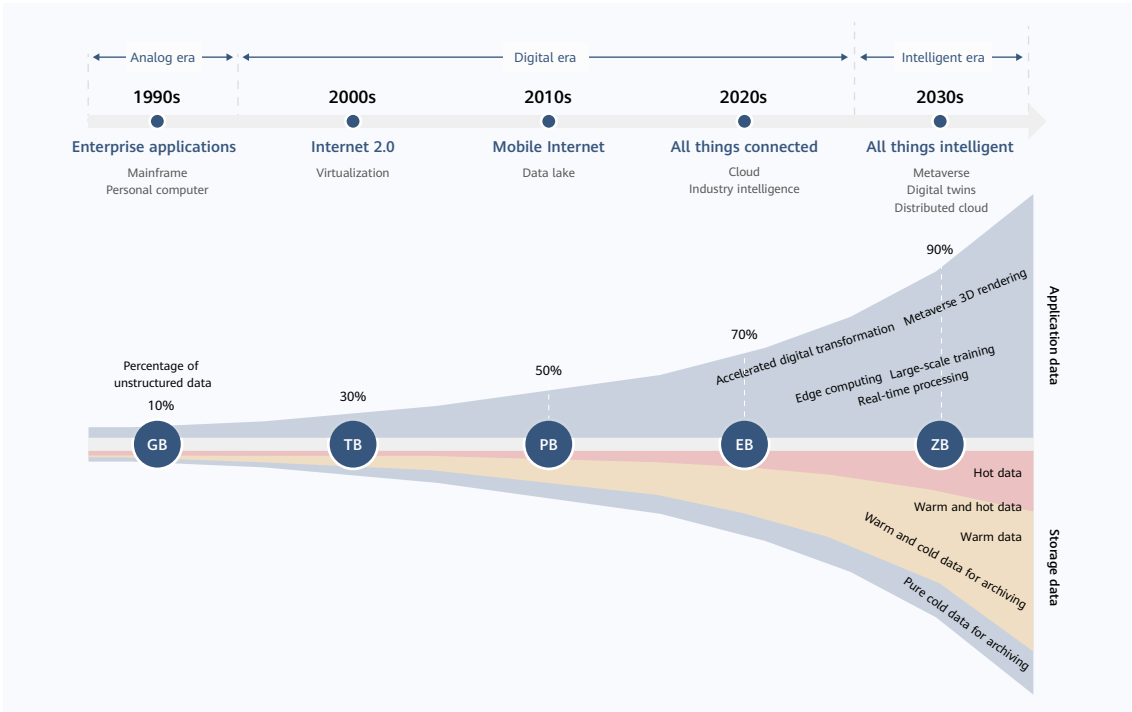


Figure 2-2 Data volume growth trend

Data can be classified into hot, warm, and cold data based on the access frequency. Different data is stored on different storage media.

Hot data: accounts for around 30% of total data. Of this, real-time data processing of AIoT, edge computing, robotics, and autonomous driving requires data access in nanoseconds. Such data is considered extremely hot, accounts for 1.5% of the total storage capacity, and requires memory media with very high performance. Online transaction services such as banking and e-commerce, and industrial manufacturing services such as EDA also require frequent real-time data access. Such data is categorized as common hot data, will increase by more than 35 times, and requires high performance flash storage media.

Warm data: Data-intensive services, such as HPDA, need to analyze a large amount of data but do not have high requirements for access frequency and real-time performance. Such data accounts for 60% of the total data and is expected to increase by more than 25 times by 2030. Apart from the need for large capacity media, it is also cost-sensitive with high sensitivity to power consumption, and requires cost-effective storage media.

Cold data: Historical documents, national archives, and other data that needs to be stored for a long time as required by laws and regulations are seldom accessed and therefore categorized as cold data. Such data accounts for around 10% of the total and is expected to increase by nearly 20 times by 2030. The long-term storage of such data requires high storage reliability and long-life storage media.

To train large AI models, increasing amounts of cold data will be activated to become warm data. Traditionally, data is stored in 3 tiers—hot, warm, and cold—in a ratio of 2:3:5. Now, cold and warm data will be merged into a combined tier. Data will be stored in two tiers—a hot tier and a warm/cold-combined tier—in a ratio of 3:7. The proportion of warm data as a percentage of all data will exceed 60% and sit somewhere close to 70%. The activated value of cold data will prompt data awakening.

Advanced Media Technology

Diversified data drives diversified storage media, requiring greater competitiveness in different application fields. Memory media for storing extremely hot data will be mainly DRAM supplemented by SCM, and memory tiering will become a new form. All hot data media will be NAND Flash, and Flash technology will evolve towards high density and low latency. In warm and cold data storage media, magnetic tapes are expected to continue to evolve towards high density and high concurrency, while optical disks will move towards larger capacity, higher concurrency, and longer service life.

1.Hot Data Media Technology

Memory plays a very important role in the computer system architecture for caching programs and data. With the development of data-intensive applications, the amount of data to be processed increases from GB-level to TB-level, driving memory media to provide larger capacity, lower power consumption, and higher concurrency.

(1)The memory architecture will become multi-layered.

DRAM currently dominates memory media. Due to the limited space for improving the capacity density of processes for chips below 20 nm, the 10 nm-chip process will keep developing for 10 years. With the increasing requirements of big datasets for large memory, the development of new media technologies such as SCM promotes the multi-layer memory architecture and gradually complements DRAM.

(2)SCM will continue to explore new scenarios.

SCMs that are based on new materials and structures have a performance comparable to DRAM and have the novel feature of data persistence. Computing in-memory (CIM) implemented by SCM has been used to supplement DRAM in certain fields, achieving a

good acceleration effect. In the future, the new ecosystem centering on SCM will be enriched. Various SCM media with persistence capabilities allow fast, high-performance access to hot data. As for processors in existing storage systems, a large amount of time is typically consumed in I/O waiting. Innovative in-memory persistent storage subsystems are likely to resolve this problem in the future.

(3)Continuous evolution of NAND Flash in the 3D stacking accelerates the replacement of HDDs.

Compared with HDDs, SSDs have obvious advantages in terms of performance, power consumption, and capacity. In the consumer market, HDDs have largely been replaced by SSDs, while in the enterprise market, the replacement of HDDs by SSDs is expected to accelerate.

SSD development involves increasing the number of stacking layers, thereby increasing the storage capacity per unit silicon wafer area and reducing the cost per unit storage space. However, as the number of stacking layers increases, the depth-width ratio (the ratio of the hole depth to the hole diameter) of stacking memory holes increases, which brings greater challenges to the etching and deposition processes and limits the continuous increase of the number of stacking layers. To further improve the storage density and increase the effective area percentage of NAND arrays, the three-dimensional (3D) architecture of stacking peripheral CMOS circuits and NAND arrays will become the mainstream in the future.

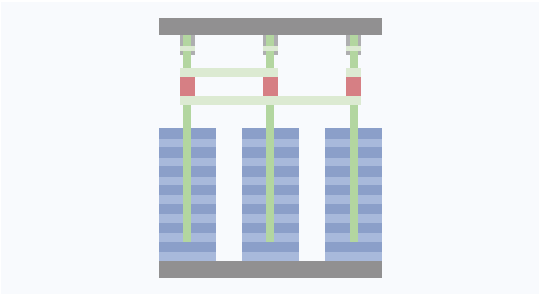


Figure 2-3 Working principle of the 3D NAND

Through stacking and 3D architectures, the capacity density per chip area is expected to increase by 10 times in 2030 compared with 2021. However, due to factors such as technology complexity and process yield rate, the cost of SSD in 2030 will not see a 10-fold drop in costs compared with 2021. Furthermore, due to impact of processes, interference of internal cabling, and an increase in density, the bottom-layer bit error rate of SSDs may further deteriorate, which poses a new challenge to an error correction algorithm with a low bit error rate, a low latency, and a high throughput.

2.Warm Data Media Technology

The evolution trend of SSDs and HDDs suggests that HDDs will still be cost-effective in 2030. As a result, HDDs will remain the dominant media in warm data storage scenarios requiring cost-effectiveness^[14].

HDD technology improvements center around density improvement. Because the magnetic recording of HDDs can only be attached to the surface of the substrate, the density can only be improved by increasing the number of disks and enhancing the magnetic density. Restricted by the HDD form and superparamagnetism, the capacity density of HDDs is close to the limit. Therefore, short-term HDD density improvement will evolve towards breaking through form and superparamagnetism restrictions, such as ultra-thick HDDs and energy-assisted magnetic

recording (EAMR) (HAMR and MAMR)^[15]. Long-term technology evolution includes breakthroughs in magnetic recording technologies and materials, such as skyrmion and magneto-optical and magneto-electrical combination technology and materials.

3.Cold Data Storage Media Technology

By 2030, cold data storage media will still be mainly magnetic tapes and optical disks. Featuring high reliability, long service life, and low requirements on storage environments, optical disks are more suitable for ultra long-term storage of cold data. Magnetic tapes are mainly used for medium- and long-term storage of cold data. In the data-driven intelligent era, data becomes hotter, resulting in two new requirements for cold data storage media: low cost and fast read speed.

(1)Magnetic Media Technology

Magnetic tape (tape for short) recording implements data storage through moving tapes. Tape capacity is usually expanded via space folding. For example, the media recording area of LTO-9 is 100 times that of HDDs in the same period. Currently, the capacity density of tapes is about 1/100 of that of HDDs. In the future, the capacity of tapes is expected to exceed 100 times that of HDDs by leveraging magnetic domain miniature, high-precision servo control, and ultra-low bit error rate (BER) magnetic channel coding technologies.

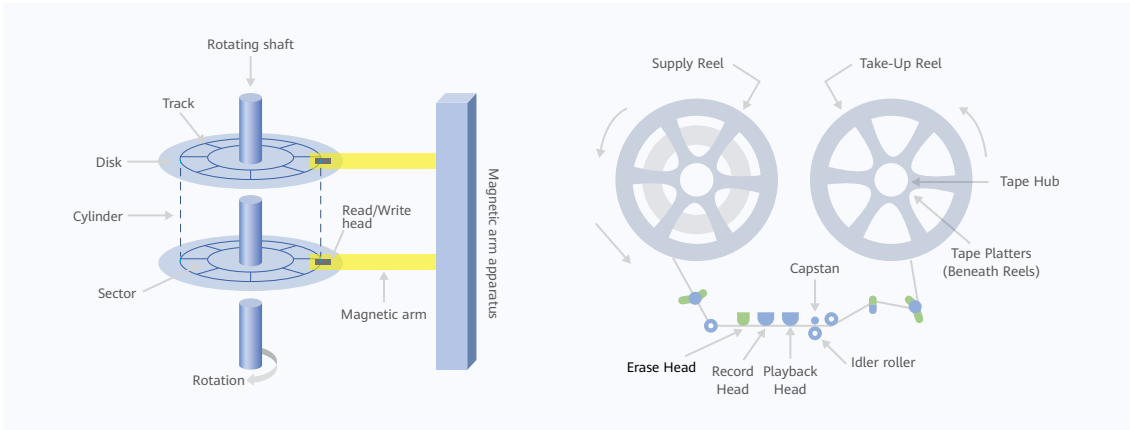


Figure 2-4 Working principle of magnetic storage (disk and tape)

Tape-based linear motion enables more heads to concurrently read and write data. Currently, the concurrent bandwidth of 32 heads in LTO-9 is more than twice that of HDDs. In the future, the bandwidth will be 10 times that of HDDs. The working principle of tapes suggests that they have innate advantages in sequential read/write. However, during random read/write, the head positioning time increases along with the capacity, affecting real-time data access^[16]. In the future, data access will be faster for devices with high bandwidth, and data layout and scheduling algorithms can further improve real-time data access performance. The material suggests that their service life is significantly affected by the environmental temperature. When the temperature ranges from 35°C to 40°C, the service life of tapes decreases dramatically, increasing the risk of data loss. In the future, new magnetic materials, manufacturing processes, and efficient environment control technologies need to be further explored to prolong the service life of tapes.

(2)Optical Storage Media Technology

In the future, optical storage media will have larger capacity and cost less. Currently, Blu-ray storage is the mainstream optical storage media. Blu-ray was initially used in the consumer field, but its capacity is only 500 GB/disk, and the throughput of a single optical head is only over 40 MB/s^[17]. In the future, optical storage will make breakthroughs in technologies such as super-resolution, multi-level, multi-dimensional, mirror ultra-multi-layer, and body material to increase the capacity of optical storage to 300 to 700 TB/disk and the throughput to 100 MB/s. In 20 years,

the capacity of a single optical disk is expected to reach 100 TB.

Due to the requirement for long service life in cold data storage, another major future challenge for optical storage is how to ensure that data in optical storage media can be securely and accurately read after thousands of years^[18].

Super-resolution optical storage technology: Optical storage records information by using a laser beam to physically and chemically change the recording material. Reducing the wavelength and increasing the numerical aperture can reduce the size of the laser spot and improve the recording density of optical storage. However, the wavelength and aperture are limited by the diffraction limit. In the future, the diffraction limit is expected to be exceeded through the use of multi-beam superposition interference. This will help further improve the recording density and increase the capacity of optical disks.

Multi-dimensional/multi-level optical storage technology: Unlike single-dimensional optical storage that can record only a single bit, multi-dimensional optical storage can record multi-bit information. The technology that is currently under research is five-dimensional optical storage. Five-dimensional optical recording stores data in five different dimensions: three spatial dimensions of storage media, polarization dimension, and intensity dimension. In the future, five-dimensional optical storage is expected to resolve the spatial interference of optical signals and develop towards six or more dimensions, further improving the capacity density of optical disks.

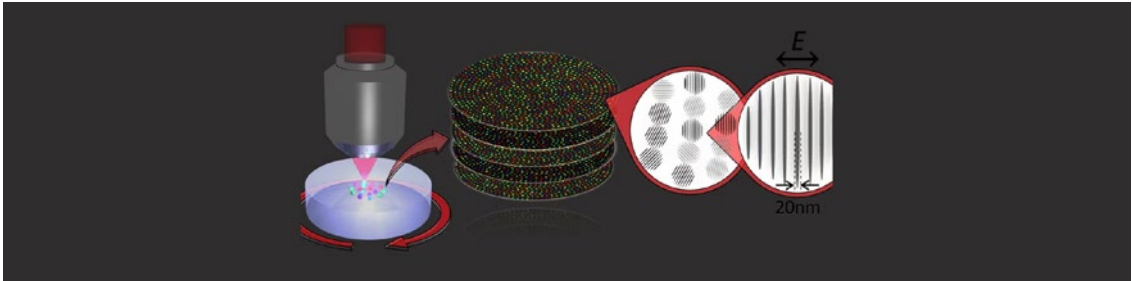


Figure 2-5 Working principle of optical storage

Multi-layer/Body material optical storage technology: The storage density of optical disks can be improved by superimposing the number of disk layers. For example, Blu-ray storage is commercially available with six layers. In the future, the industry is expected to solve the issue of inter-layer optical interference, allow dozens and even hundreds of layers to be achieved. Holographic optical recording uses phase change body materials to record information at different layers and angles inside the storage media. By combining multi-layer and body material recording technologies, optical storage can reach even higher densities resulting in a storage capacity of more than 100 TB/disk.

Servo drive technology: An optical disk drive includes a laser and an optoelectronic modulation device. Currently, the femtosecond laser and optoelectronic modulation device used in multi-dimensional optical storage are costly. With the development of the femtosecond laser industry, further breakthroughs in high-frequency and high-voltage optical circuit technology are expected to be made in the future to reduce the cost of gemstone-level crystals so that they can be put into large-scale commercial use in the optical storage industry. Limited by the write principle of optical storage, the read/write bandwidth of a single laser is only dozens of megabytes per second. In the future, the high-precision servo control technology is expected to implement parallel reads/writes of multiple optical channels and improve the throughput.

Media Application Innovation

1. Media Process Technology

The semiconductor manufacturing process and physical limits of media structure mean that the integration of media such as SSDs and DRAMs cannot be continuously improved. In the future, wafer-level innovation, chiplet-level innovation, and interface and protocol innovation can further improve media density and service life, reduce media power consumption, and enhance media reliability.

Wafer-level innovation: The die-on-board (DOB) technology can integrate storage chips into circuit boards to provide higher density and better performance. The Wafer-Scale technology directly uses wafers of multiple NAND dies without cutting or packaging the wafers, achieving higher density, faster speed, and higher reliability. The Wafer-Scale technology is currently immature. Problems such as manufacturing of ultra-large chips, function management and monitoring of chips, cross-chip connection, chip heat dissipation, and reliability management need to be solved. In the future, advanced process technologies, innovative chip design methods, and intelligent test methods are expected to be used to achieve higher

capacity and better durability while maintaining the advantages of high density and low power consumption.

Chiplet-level innovation: Chiplets can integrate different functional modules into separately packaged chips to achieve better flexibility and scalability, higher performance, and higher power efficiency. Currently, the Chiplet technology still faces many technical challenges, such as inter-chip communication and synchronization, cache consistency, and transmission rate matching. In the future, technologies such as intelligent control algorithms, efficient chip cache consistency protocols, encapsulation processors inside storage media, heterogeneous processors, and accelerators are expected to encapsulate computing chips and media chips together to build a Chiplet media that integrates storage and computing, achieving high performance, low power consumption, and easy scalability.

Interface and protocol innovation: As media become diversified, data transmission between multiple media interfaces has a large protocol conversion overhead, which can be greatly improved in terms of performance, security, and

universality. Zone Namespace (ZNS) is a high-speed storage protocol used for flash storage devices. It supports efficient space management based on smaller data blocks, alleviates the performance imbalance of SSDs, and improves the performance of garbage collection and data migration of SSDs. Currently, issues such as compatibility and application migration need to be resolved. Plogs are used for persistent data storage management. They can transmit and

process data between different storage systems across multiple storage media. The Plog protocol uses the automatic retransmission and self-healing mechanisms to ensure data consistency, reliability, and integrity, improving data transmission and access efficiency. In the future, with the continuous development of diversified media technologies, new high-performance interfaces and protocols need to be defined to further improve compatibility and data access efficiency.

2.New Data Coding

Data coding technologies include compress encoding for reducing data volume (Sayood, 2017), error correction coding for preventing data errors, and erasure coding for preventing data loss (Peterson & Weldon, 1972). They are the core technologies that achieve large storage space and long storage period. In the future, storage systems that integrate diversified media for mass data call for breakthroughs in data coding technologies through intelligent data compression, joint coding, and intelligent data classification to improve effective storage capacity, save energy, and ensure long-term reliability.

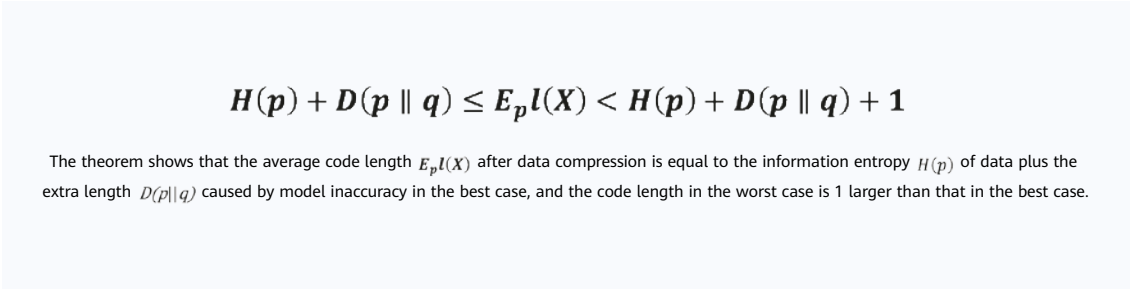


Figure 2-6 Lossless data compression theory

Intelligent data compression: Data compression is the process of using short-bit data to represent information based on a specific coding mechanism. In data storage, lossy compression coding and lossless compression coding coexist. The current lossy coding cannot break the classical rate-distortion theory. In the future, semantics extraction and compression technologies need to be explored, rate-distortion functions need to be extended, and a new theoretical system needs to be established to make technical breakthroughs in lossy compression. Mainstream lossless compression methods in the industry use LZ and entropy coding as the core, and the effect of compressing unstructured data is poor. The compression method based on statistics

and dynamic prediction models can effectively improve the reduction ratio of unstructured data. However, the models depend on data and expert experience and develop slowly. AI-based prediction models can surpass expert-designed predictors through automatic extraction of data features and self-learning of models. The existing AI-based compression algorithms face the problems of poor generalization capability and high computing power consumption. In the future, transfer learning, meta-learning, and large model technologies are expected to improve the model generalization capability and algorithm efficiency, improving the reduction rate by several times in the storage system.

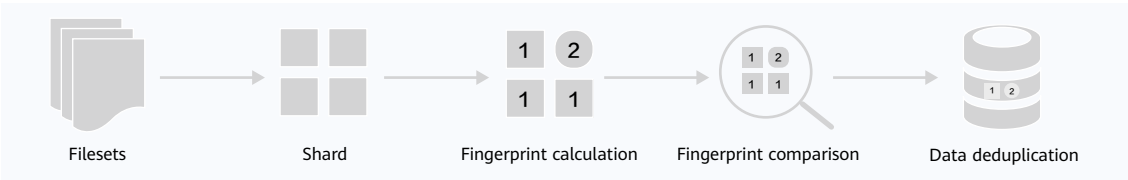


Figure 2-7 Basic principles of data deduplication

Data deduplication: The deduplication technology deletes duplicate data blocks by identifying the content at the data block level. With the emergence of processor technology and new storage media, the deduplication technology is gradually shifting from offline to online processing. The data deduplication granularity is shrinking and changing from file-level deduplication to recent byte-level similarity-based deduplication, creating challenges for the computing power

and I/O throughput of the system. In terms of deduplication of diversified mass data, the deduplication ratio in high-dimensional data scenarios is several orders of magnitude lower than that in structured data scenarios. With the development of semantics deduplication technologies in the future, the storage efficiency of unstructured data is expected to be fundamentally improved.

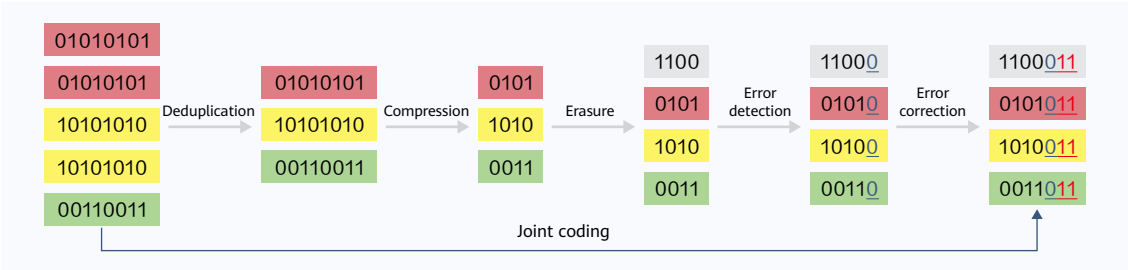


Figure 2-8 Data joint coding

Data joint coding: Shannon's separation theory (Shannon, 1948) proves that separate design of source coding and channel coding can achieve optimal system performance when the code length tends to be infinite. If the code length is limited, combining source coding and channel coding may achieve gains (Jiang & Bruck, 2008). In the future, joint coding can be designed to achieve higher-density storage, simplify the system, and reduce power consumption.

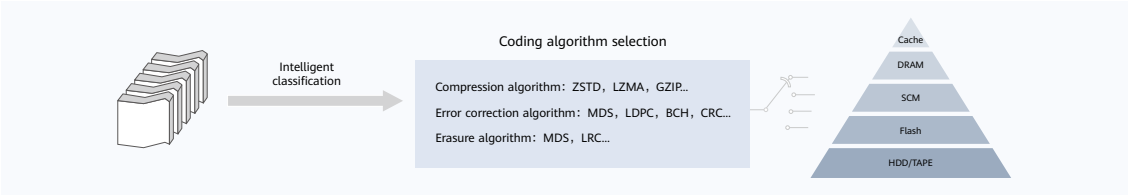


Figure 2-9 Intelligent classification

Intelligent data tiering and classification: Storage is a system with diversified and hierarchical media. The reliability, latency, bandwidth, and cost of different storage media vary greatly and, therefore, a matching data coding algorithm (Kim, Gupta, Urgaonkar, Berman, & Sivasubramaniam, 2011) (compression, error correction, or erasure coding) must be selected. In the future, breakthroughs in intelligent data classification technologies are required to optimally match different data coding technologies with different media, improve data density and reliability, and reduce latency.

2.2 Data-Centric Architecture

Driven by new data-intensive applications such as big data, AI, HPC, and IoT, the data volume is increasing explosively with a compound annual growth rate (CAGR) of nearly 40%. Hot data will account for more than 30% of the total data volume. With the slowdown of Moore's Law and Dennard Scaling, the annual growth of CPU performance has reduced to 3.5%. The fast-growing data volume and the slow-growing data processing capability present a challenge to the data industry, as storage capacity and data development are severely unbalanced.

In a typical CPU-centric data center architecture, uneven distribution of services in space and time lead to low utilization of local storage resources. The idle rate of local memory and storage exceeds 50%^[19]. In addition, data movement and repeated data format conversion consume a large amount of CPU resources, resulting in a low data processing efficiency.

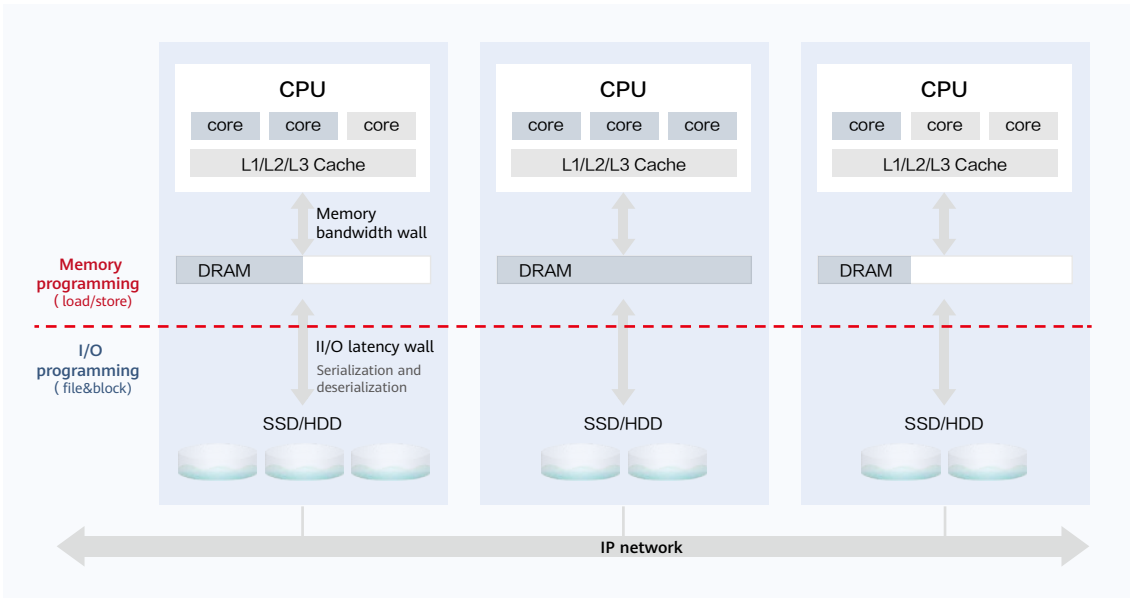


Figure 2-10 CPU-centric architecture

Storage and computing resources cannot be efficiently used in the existing data center architecture. To improve data processing efficiency and storage resource utilization, the future data center architecture needs to shift from CPU-centric to data-centric, including:

1.Storage-compute decoupling at the macro level: Computing and storage resources are independently deployed and interconnected through high-throughput^[20] data buses for unified memory semantics access, implementing decoupling and flexible scheduling of computing and storage resources and maximizing resource utilization.

2.Computing in-memory at the micro level: Data is processed nearby with data as the core, reducing unnecessary data movement. Dedicated data processing computing power is deployed at the edge of data generation, on the data flow networks, and in the data storage system. The convergence of network, storage and computing improves the data processing efficiency.

3.Highly scalable cluster storage: The scale-out and scale-up capabilities are increased by a factor of several dozens. Cluster storage can be scaled out from dozens of controllers to hundreds of controllers and EB-level capacity is available. Hundreds of xPUs can be scaled up to thousands of xPUs, achieving acceleration via near-storage data processing.

Decoupling Storage and Compute Resources and Enhancing Compute with Storage

Storage-compute decoupling will no longer be limited to the decoupling of CPUs from SSDs and HDDs but will instead completely break the boundaries of various storage and computing hardware resources and builds them into independent hardware resource pools (such as CPU pools, DPU pools, memory pools, and flash storage pools) for elastic expansion and flexible sharing of hardware. The storage-compute decoupling architecture has three features: storage resource pooling, memory-based storage that supports global memory semantic access, and NPUDirect Storage.

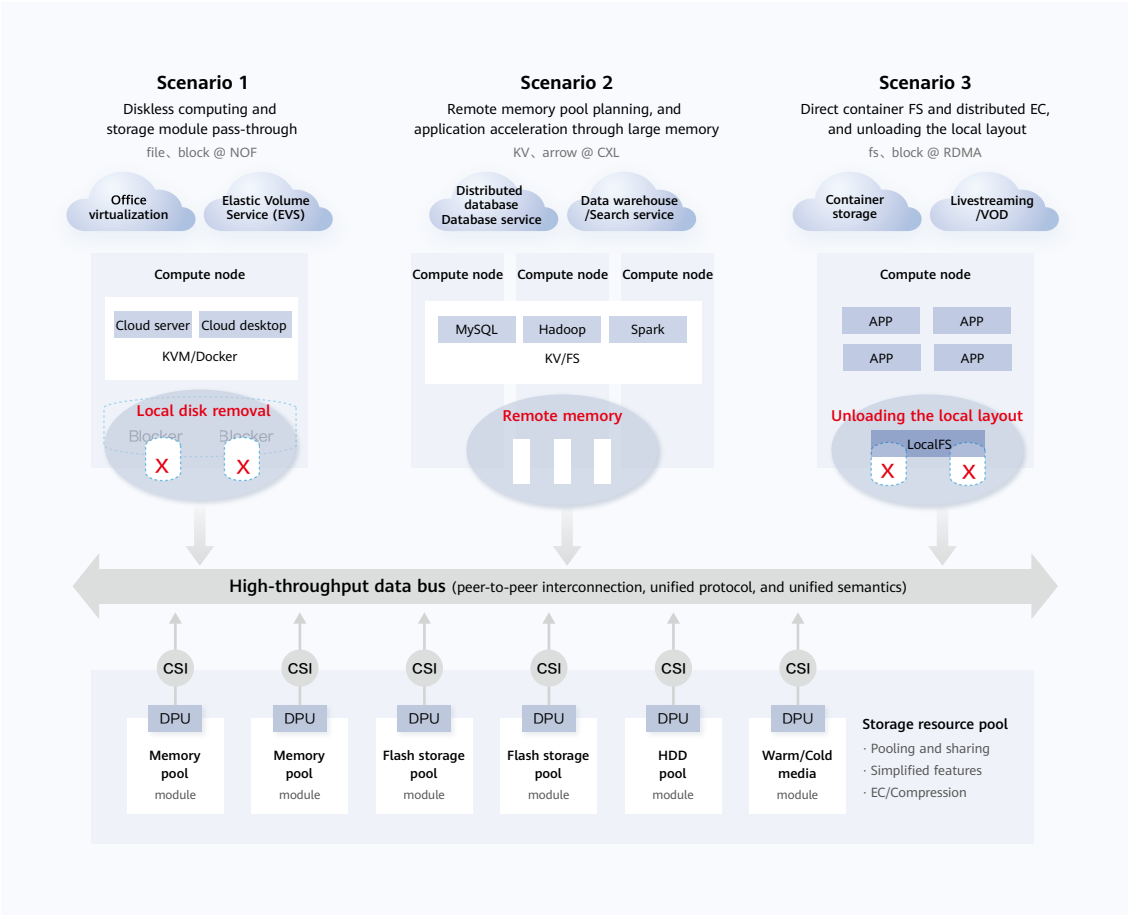


Figure 2-11 Storage-compute decoupling architecture

1.Storage Resource Pooling

The new storage-compute decoupling architecture remotely deploys local disks of servers to form diskless servers and remote storage pools. In addition, the remote memory pool is used to expand local memory, implementing true storage and compute decoupling and greatly improving storage resource utilization. In service scenarios, virtual disks with different performance and capacity and pooled memory space can be

selected based on application requirements. *First*, storage resource pooling prevents space waste caused by local storage space over-configuration. *Second*, resource pooling prevents data from flowing across buses and devices, reducing data movement, improving performance, and reducing power consumption. Finally, if a server is faulty or replaced, data migration is not required.

The NVMe over RDMA network technology implements SSD pooling and provides consistent

local access performance for remote access to SSDs. In the future, new memory networks (such as CXL and Unified Bus), intelligent tiering of memory media, and unified addressing technologies are expected to be used to implement memory pooling, expand the memory capacity by 10 times, and reduce the cost of large memories for applications.

2. Memory-based Storage that Supports Global Memory Semantic Access

Traditional applications access data via file, object, and block interfaces. The I/O stack protocols are complex, and the application I/O overhead exceeds 30%. Memory semantics and memory data format are used for data access, achieving zero I/O stack overhead, zero format conversion overhead, and zero data flow overhead. Currently, memory semantics data access still faces challenges from the application data access interface ecosystem and memory semantics network standardization. In the future, a unified memory semantics standard protocol is expected to be formed for memory semantics data interworking and further improvement of data access efficiency.

3. NPUDirect Storage

The traditional interconnection bus is CPU-centric. The CPU becomes the system bottleneck, and the system cannot be expanded on a large scale. The protocol types are different from each other, and the protocol conversion is repeated, reducing the system efficiency. Different devices have different communication semantics, and the data format conversion is repeated, causing extra overheads.

The high-throughput data bus needs to be defined to support peer-to-peer communication among devices, eliminate protocol conversion, and simplify data access. The high-throughput data bus has the following features:

- Peer-to-peer connection: The CPU-centric structure is broken. CPUs, DPUs, and storage devices are interconnected in peer-to-peer mode. Data access does not pass through CPUs. Heterogeneous and diversified data processing devices directly access data in peer-to-peer mode, improving data migration efficiency.
- Unified protocol: Different communication requirements are abstracted in devices, cabinets, and data centers. Unified basic protocol functions are formulated and unified access protocols are used between processors and storage devices and among different storage devices.
- Unified semantics: Different access requirements are abstracted into unified access semantics, and data can be shared and accessed across systems and devices of different types.
- High throughput: The bandwidth of a single SSD will evolve to 25 GB/s, and that of the memory will support 100 GB/s, with a latency of less than 50 ns. New buses are used to interconnect SSDs, memories, and processors, as well as extend to inter-rack interconnection. In addition, new buses need to meet the requirements of high bandwidth for large-block data transmission and low latency for small-block data transmission. In the future, buses need to support TB/s-level bandwidth and 10 ns-level latency.

Coupling Storage and Compute Resources and Eliminating Repeated Computing Through Queries

In the data-centric processing paradigm, data processing has shifted from general computing to professional data processing, and from data migration to processors to near-data computing power deployment. Data is processed with the most appropriate computing power near data, and data is processed nearby at the edge of data generation, during data movement, and in the data storage system. As a data carrier, data storage not only provides data access services, but also near-data processing acceleration services. There are three main methods for near-data processing: diversified storage and computing convergence, convergence of data storage and networks, and convergence of data processing and networks.

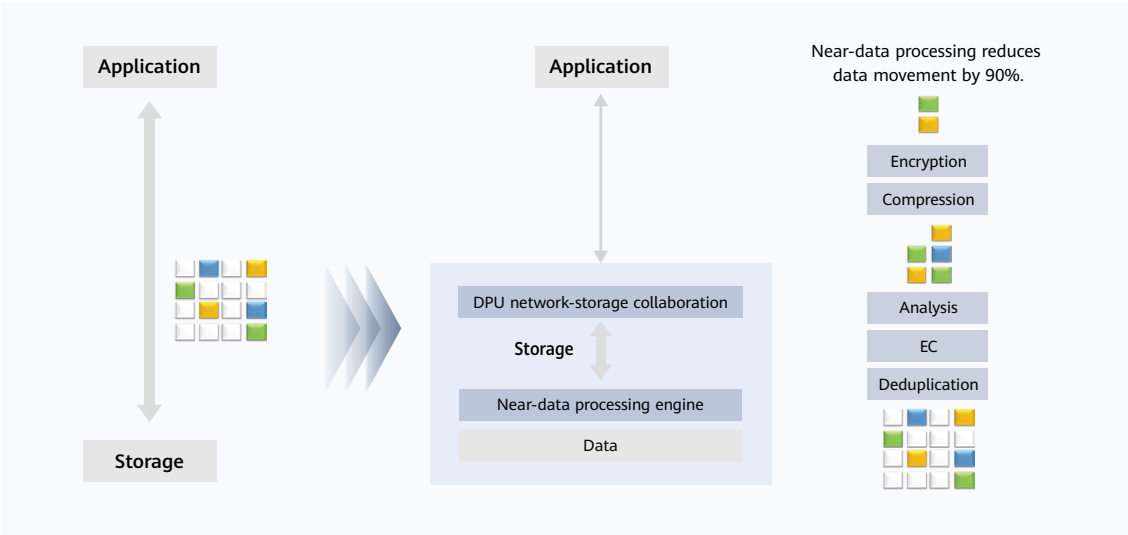


Figure 2-12 Working principle of storage and compute convergence

1.Diversified Storage and Computing Convergence

Storage and computing convergence is a technology used to offload host processing to memories to reduce data migration and network latency, overcome bandwidth bottlenecks, and improve data processing efficiency. Storage and computing convergence includes storage and computing integration (SCI) and computing in-memory (CIM)^[21].

SCI integrates the instruction computing unit and operator unit on the storage component for data preprocessing, for example, adding a solidified data preprocessing unit (such as a compression engine or an encoding engine) to an SSD or a memory to accelerate data processing, or, integrating a large-capacity memory into the processor to reduce data access and, in turn, improve data processing efficiency. In the future, how to define efficient forward-compatible instruction sets and new operator abstraction in the aforementioned scenario is still a challenge that needs to be resolved but it is expected that implementation of efficient data processing in common scenarios through common instruction set research and customized operators will be achieved.

CIM uses a non-Von Neumann architecture and integrates storage units and computing logic to break the boundary between computing and storage and transfer a small amount of data during data processing. Compared with the traditional Von Neumann architecture, CIM improves energy efficiency by more than 10 times. Due to the limitations of current storage media, there are still great challenges to overcome in digital-to-analog conversion efficiency, computing precision, and scale. In the future, breakthroughs are expected to be made by improving storage media and discovering new media materials.

2.Collaboration of Data Storage and Networks

Convergence of data storage and networks senses the storage semantics to offload data storage services and schedule data flows, improving data access performance and accelerating data application services. Currently, it can be potentially applied in offloading storage access protocol (file protocol, object protocol, KV, etc.), accelerating storage I/Os (data passthrough and I/O zero copy), and offloading data layout (such as index). Storage services can be flexibly offloaded to intelligent network interface cards (NICs). However, this faces challenges in terms of programming friendliness and operational efficiency. In the future, efficient storage operators

are expected to be defined to achieve greater flexibility and higher performance.

3.Collaboration of Data Processing and Networks

By collaborating with the network, data processing in hosts by general-purpose processors is offloaded to dedicated data processors, such as security data processing (such as SHA256 and lattice-based cryptography), data compression (ZSTD, LZ, and CDC), data protection (EC), and data analysis (Scan, Filter, and Merge). Dedicated data processors represented by DPUs feature lower costs, lower power consumption, plug-and-play, and swap-and-play. They accelerate data processing during data flow, release the computing power of general processors, and improve the performance of big data, HPC, and databases by multiple times.

Cluster Storage

In a large-scale AI computing center, neither a single storage node nor scalability to any fewer than 100 nodes can meet the compute cluster's requirements for hundreds of PBs of capacity, retrieval from hundreds of billions of files, and hundreds of TB/s of bandwidth. By 2030, we expect that a storage cluster will be scalable to thousands of nodes. In addition, more and more data reads and writes are being offloaded to xPUs to achieve near-data processing and improve efficiency. Storage can support an increasing number of xPUs. In the future, millions of xPUs will be able to work concurrently and will be elastically scalable. The cluster storage capacity is expected to increase one hundredfold to hundreds of PBs, and this will meet the demands for both high performance and large capacity.

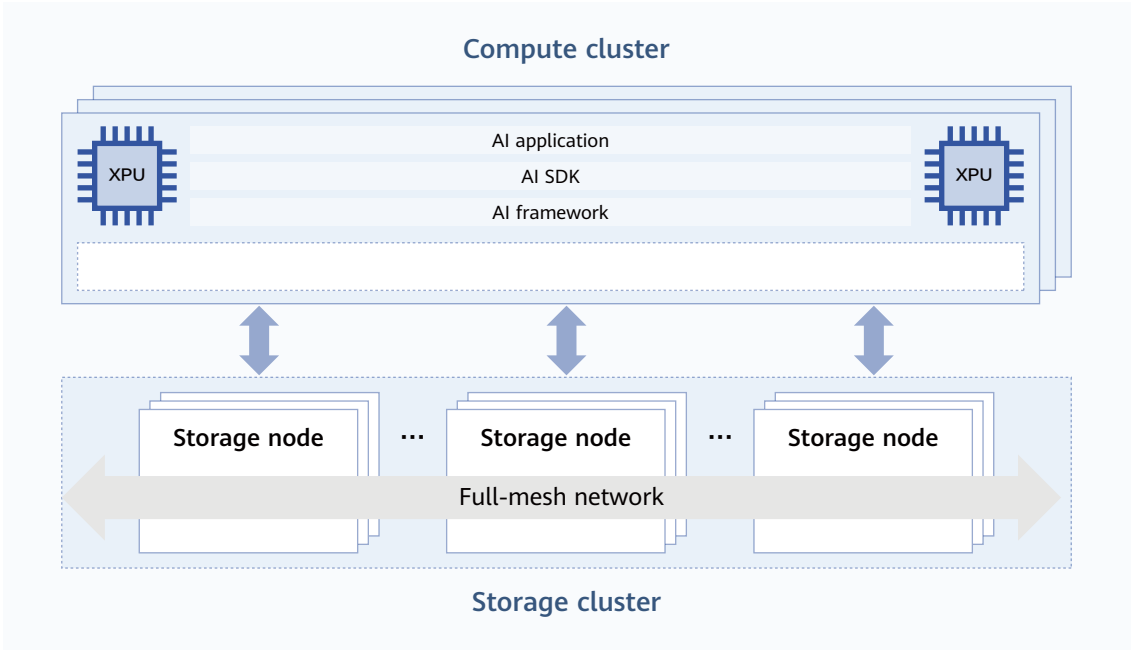


Figure 2-13 Cluster storage

2.3 Intrinsic Data Resilience

As a new production factor, data is becoming increasingly valuable, increasing the attack surface and attack intensity. The current border-based passive defense system cannot meet future data security requirements^[22], and data privacy protection requirements are mounting during data value release. Privacy computing centering on "usable but invisible" data converts and releases data value while fully protecting data and privacy. Data transfer is necessary for releasing data value. The replicability, sharing, and unlimited supply of

data means that protecting data property rights, use permissions, and control rights during data transfer is the primary issue the current data infrastructure must resolve.

In the future, data infrastructure will feature **intrinsic data resilience**. As such, we need to make continuous breakthroughs in technologies such as proactive data protection, zero data copy, and zero-trust storage.

Proactive Data Protection

Research on data security attack and defense situation shows that the current passive defense security system cannot effectively defend against virus attacks such as ransomware. A proactive data protection security system needs to be built from multiple technical directions, such as data security situational awareness, data timeline travel, native anti-tampering, and multi-dimensional linkage response.

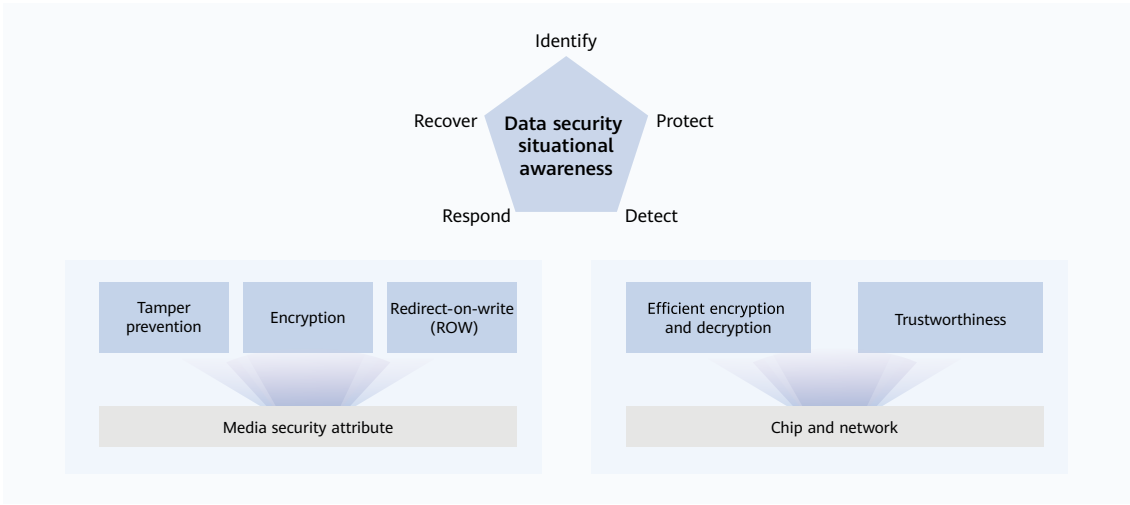


Figure 2-14 Proactive data protection

Data security situational awareness: The data security situational awareness technology collects the data access behavior, data information entropy, internal data correlation, and data distribution within a certain period of time, and dynamically measures and evaluates data security risks and threats based on the big data analysis technology to support subsequent self-defense decision-making and actions. Currently,

the industry is facing challenges such as low accuracy and efficiency of threat detection and situational awareness capability and an insufficient dynamic threat evaluation capability. Detection accuracy and performance are expected to gradually improve in the future. The ability to detect unknown data threats through research on the sampling theory of mass data, converged processing of heterogeneous data, and activity

identification based on incomplete information are also expected to be enhanced.

Data timeline travel: If data is damaged due to internal and external attacks, the data infrastructure must be able to quickly restore the damaged data to any historical point in time to achieve zero data loss. In addition, to implement attack source tracing, the data infrastructure must have the finest-grained data replay capability to support the adjustment and optimization of data security policies. The industry is currently facing the challenges of quickly and accurately locating the time point of damaged data and automatically tracing behavior. In the future, technologies such as I/O-level data recovery and root cause analysis based on cause-effect reasoning are expected to be used to retrace the data timeline.

Native anti-tampering: Currently, the data anti-tampering capability is mainly implemented by the system-level data access control technology. However, due to the large attack surface of the system, it is difficult to effectively ensure data

anti-tampering. In the future, the system-level data access control technology and the physical anti-tampering attribute of media are expected to be combined to implement the physical anti-tampering capability.

Multi-dimensional linkage response technology: The multi-dimensional linkage response technology requires cross-device collaboration among network devices, security devices, endpoint detection and response (EDR) devices, and storage devices to implement multi-dimensional closed-loop threat processing and prevent threats from spreading. The main challenges in the industry currently lie in autonomous decision-making and response technologies, that is, how to develop intelligent response policies to provide customers with convenient and effective alternative solutions. Future technical breakthroughs in areas such as AI security analysis, causal analysis, and inference are expected to make autonomous decision-making and response more intelligent, thus achieving real quick and accurate response.

Data Zero Copy

The value release process of data elements is divided into three phases. The first phase focuses on supporting business system operation and promoting digital transformation and intelligent business decision-making. The second phase allows external enablement of data flows to aggregate and integrate high-quality data from different sources in new businesses and scenarios, achieving win-win and multi-win results. The efficiency between data sharing and data access control during this phase needs to be improved. Technologies such as cryptography-based access control, data self-protection, efficient and transparent audit, and efficient network encrypted transmission can be used to implement efficient data flow and use while ensuring data sovereignty and security. The third phase is borderless zero copy, which eliminates data silos to the maximum extent. The zero data copy access technology is used to break data boundaries and implement data sharing.

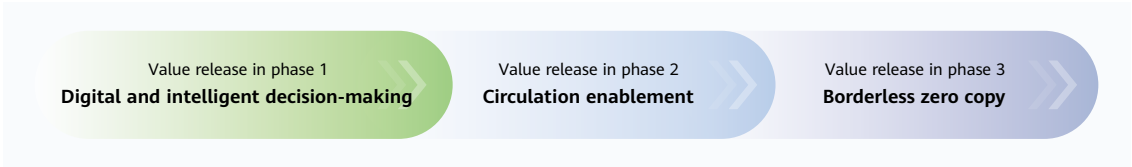
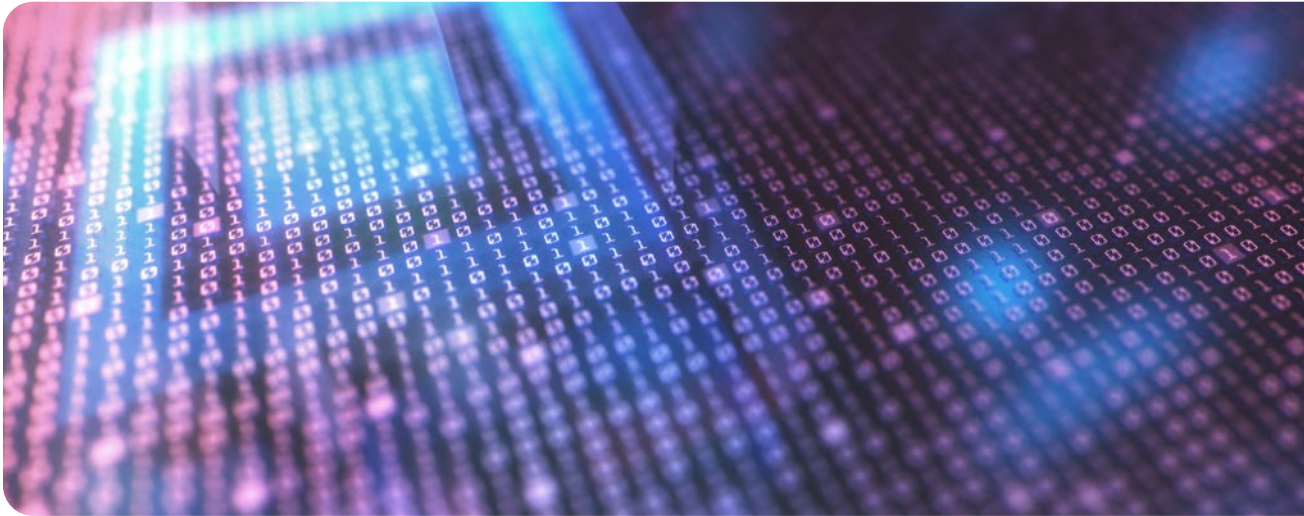


Figure 2-15 Data value release model



Cryptography-based access control:

Cryptography is used to ensure data confidentiality and thus users who do not comply with access control policies cannot decrypt data. Typical attribute-based encryption (ABE) solutions support complete access control policies of any logic which, compared with traditional one-to-one public key encryption, offer a one-to-many setup to slash communication overheads and encryption/decryption calculation overheads for key nodes. Future technologies for encryption must be able to implement policy judgment and randomized processing on a ciphertext that is leaving a trusted domain, so as to prevent any data that *does not* comply with predetermined access control policies from leaving the trusted domain.

Self-protecting data: Data security has evolved from system-centric management and control to data-centric full-lifecycle security protection, with cryptographic computing used for data privacy. Self-protecting data refers to a type of technologies that make data "usable but invisible". However, there are problems of associated information privacy leakage, while the data use scope, use mode, validity period, and access permission are difficult to restrict. In the future, the "data capsule" solution is expected to encapsulate

access policies, use control policies, and encrypted data to ensure that data owners can control data and implement secure data transfer.

Efficient and transparent auditing: The mainstream technology for trusted data auditing uses the blockchain. However, the blockchain technology has problems such as high overheads, low consensus algorithm efficiency, and data storage redundancy. Efficient and transparent auditing must be used to build a certificate anti-tampering auditing solution, to implement more efficient trusted data storage and meet the read/write latency requirement in the actual production process.

Zero data copy access: Due to the differences between application data models, most applications are siloed with independent data copies. In the future, application data models are expected to be deployed at the data storage layer and automatically generated based on the same data to eliminate data silos. In addition, technologies such as fine-grained access control and trusted network transmission based on chip certification are used to implement efficient data access across trusted domains.

Zero-Trust Storage

Zero-trust storage is an extension of the zero-trust model, aiming to solve storage security problems such as data leakage, integrity damage, and data availability damage. In zero-trust storage, all data read and write are considered unverified. Based on the minimum authorization principle, access subjects, data, and data operation implement minimum-granularity data access control through continuous verification and dynamic authorization. To achieve zero-trust storage, we need to make breakthroughs in data storage and usage environment security and full-path data security encryption.

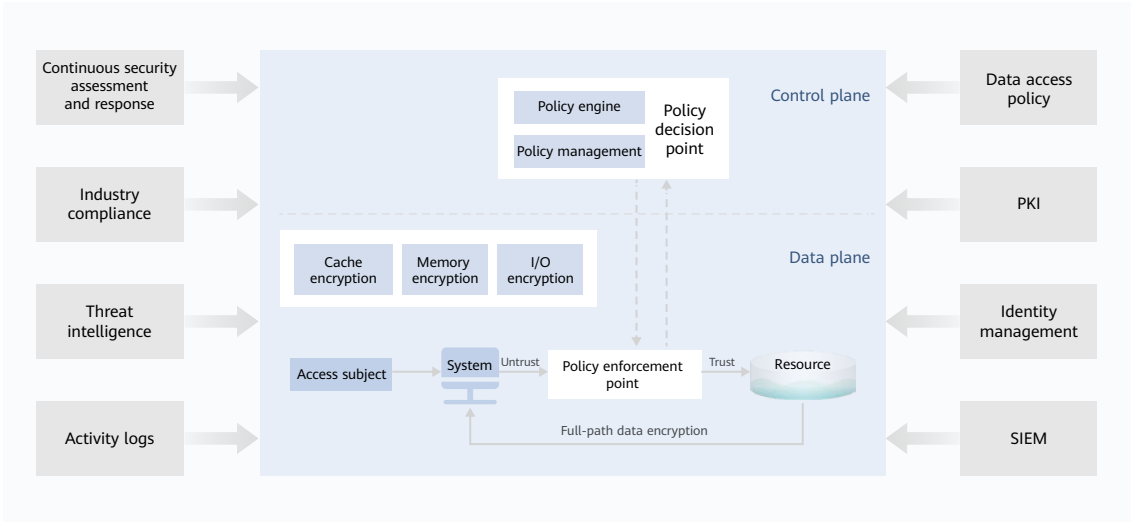


Figure 2-16 Zero-trust storage

Mandatory data access control: Based on the minimum authorization principle, fine-grained data access control uses the mapping among data access subject characteristics, data attributes, and fine-grained data processing to ensure that the minimum-granularity dataset can be accessed and used only by subjects under specific conditions. In the future, the design of access control policies will become increasingly complex due to the huge number of authorized entities and data, complexity of data processing, and uncontrollable conditions, and an improper access control policy can cause major security risks. To cope with this challenge, technologies such as formal verification, automatic policy generation, and compliance auditing will ensure policy consistency and correctness, and solve issues such as those involving large-

scale formal verification performance, automatic policy generation mechanism, and complex rule matching.

Full-path data encryption: In the current border-based data security system, the full-path data is not safe, which may cause data leakage. Therefore, we need to consider encrypting the full-path data processing from memory, storage I/O, network I/O, and cache, and sharing native data security capabilities through unified key management.

Privacy computing: To ensure data privacy and security during computing, secure data computing technologies have emerged, including Federated Learning for AI^[23], hardware-level Trusted Execution Environment, secure multi-party

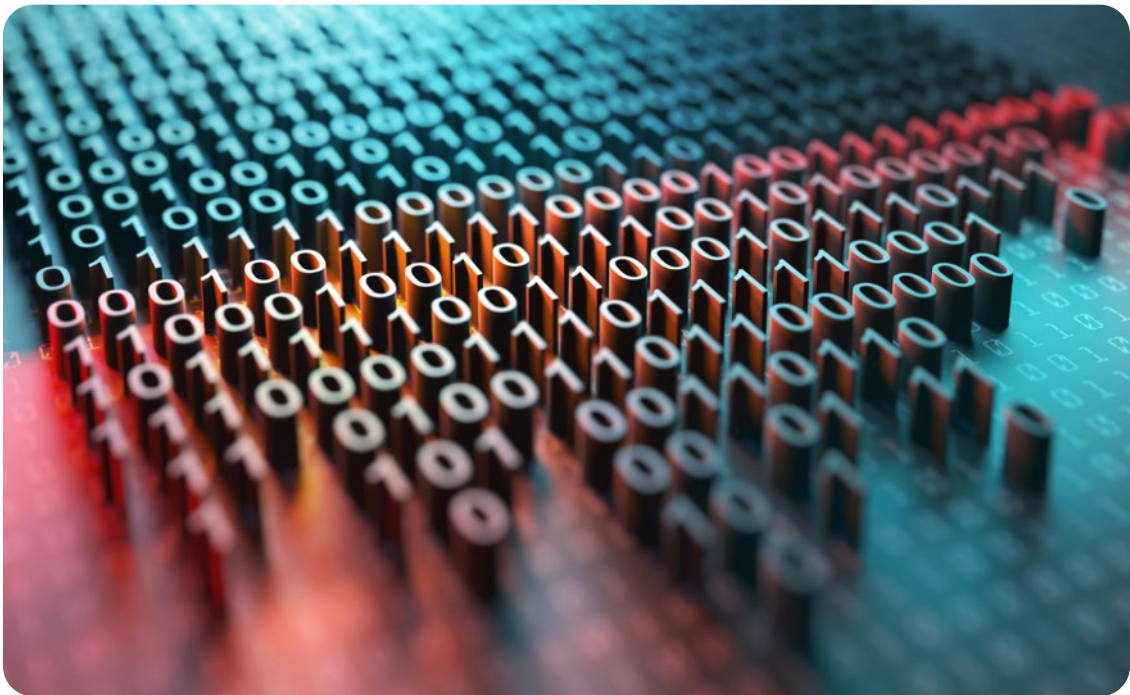
computation using cryptographic algorithms, and zero knowledge proof.

1.Trusted execution environment (TEE): The main challenge of implementing the hardware isolation technology for sensitive data processing is that the completeness of the hardware security isolation mechanism cannot be proved by data, which may cause security vulnerabilities. However, compared with the cryptography technology, TEE has little impact on performance. In the future, TEE-based privacy computing will become a common industry requirement. It is estimated that this technology will be used in more than 50% of data processing scenarios by 2030.

2.Cryptography-based homomorphic encryption and secure multi-party computing technology, whose security can be proved mathematically, have become the industry's most ideal privacy computing technology. However, the main challenge is that its performance is more than 10,000 times lower than that of general

computing and needs to be greatly improved to meet application requirements. With the maturity of approximate computing, homomorphic encryption and secure multi-party computing are applied in specific fields such as health data sharing. In the future, the hardware acceleration-based homomorphic encryption and secure multi-party computing technology will be widely put into commercial use in high-security application scenarios in industries such as finance and healthcare.

3.Multi-party computing is based on secret sharing among multiple parties. If cryptographic methods such as zero-knowledge proof are used to implement multi-party computing, the performance overhead is high. With its mathematically proven security, using TEE to implement secret sharing among multiple parties greatly improves multi-party computing performance, showing potential for widespread application in the future.



2.4 Intelligent Data Fabric

The continuous development of digital technologies has led to a large number of requirements for cross-domain data flow, posing higher requirements on data availability and quality. However, geographical barriers and difficulties in data governance restrict the free flow of data and finally result in data gravity. Data fabric is to coordinate distributed data sources automatically and dynamically, provide integrated and reliable data across data platforms, and support the use of a wide range of different applications^[24]. Based on technologies such as AI and knowledge graph, intelligent data fabric continuously identifies and coordinates

correlations between available data points to help customers mine value. For networks that transmit data among the edge, data center, and cloud, intelligent data fabric enables continuous analytics on already-existing, discoverable, and inferable metadata assets to integrate cross-platform data and provide efficient data flow and processing for applications. To better implement intelligent data fabric, continuous breakthroughs are needed in technical directions such as cross-domain data collaboration, automatic data orchestration, and efficient and fast storage networks to eliminate the data gravity problem.

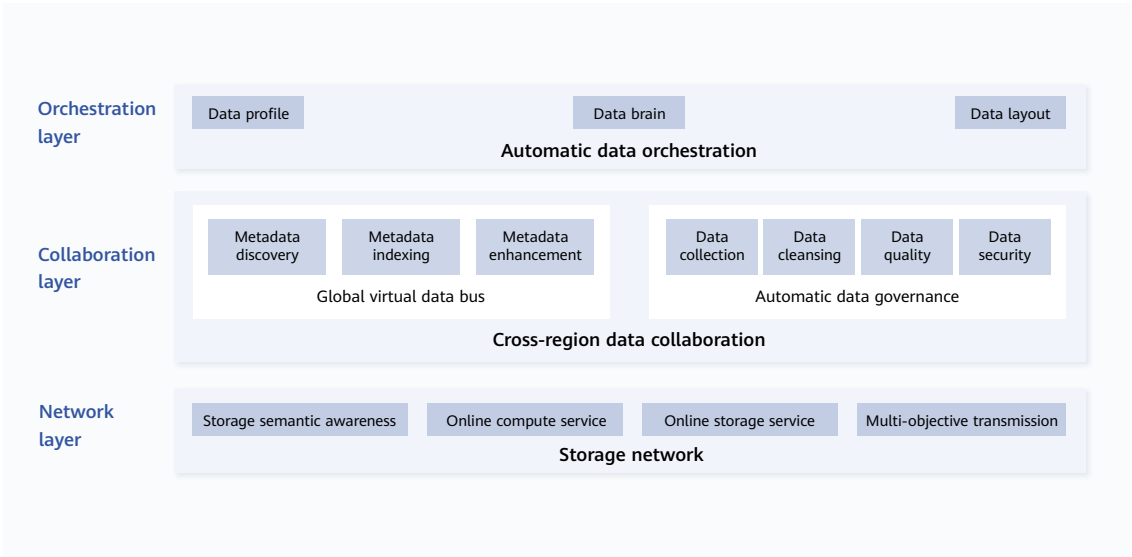


Figure 2-17 Intelligent data fabric framework

Automatic Data Orchestration

The present data content is inadequate at detecting network status, and application intents are unable to transmit to networks effectively. As a result, this unequal distribution of data and networks causes access delays and low network utilization. To address this issue, it is necessary to develop data profiles and data brains that achieve optimal data placement without compromising on service performance.

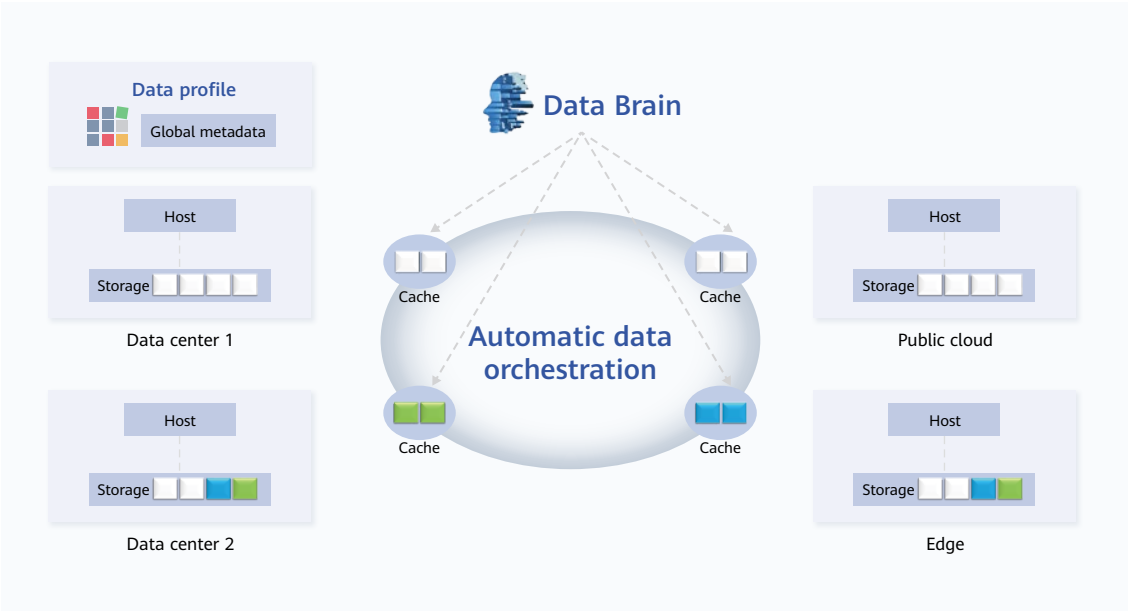


Figure 2-18 Automatic data orchestration framework

Data profiling: This technology senses application characteristics via the storage network status, spatiotemporal information of data blocks, and application labels. The current service awareness lacks granularity and precision. However, there is potential for the use of advanced technologies like deep graph neural networks and cause learning to create multi-dimensional data profiles which take into account aspects such as data gravity, data volume, data activity, network bandwidth, and latency. These technologies help achieve accurate service awareness.

Data brain: Current data orchestration is plagued by scattered data, dimension explosion, a lack of standardization, high technical requirements on developers, and unawareness of customer applications, posing high requirements on data flow management and control in multi-cloud scenarios. The universal data orchestration platform cannot meet the requirements of multiple parties. New technologies such as intent-driven API, machine learning, and big

data analysis can help generate the optimal data layout policy based on industry applications, and provide powerful security management and audit capabilities that realize full-process automatic data orchestration.

Data layout: This technology places data to the optimal location based on service policies, so that users can access nearby data by content name and obtain the best experience at the minimum cost. For example, cross-region backup mode can help store cold data in a region with lower operation costs. The current data layout has problems such as poor data sharing between services, long tail data access, low hit ratio of data cache, and high network bandwidth usage. Breakthroughs in separating service logic from data logic, data network coding, and data prefetch and eviction algorithms are expected to implement adaptive data cache and nearby read/write cache acceleration. This cost-efficient and application-unaware measure will improve data access while simplifying data retrieval and utilization.

Cross-Region Data Collaboration

Typically, enterprises store data in multi-region data centers or using multiple heterogeneous cloud providers to provide unified compute-storage services for lower costs and better infrastructure capabilities. The distribution of assets, software, and applications across multiple data centers or clouds necessitates cross-region data collaboration and integration. Such trends can be divided into the following two directions:

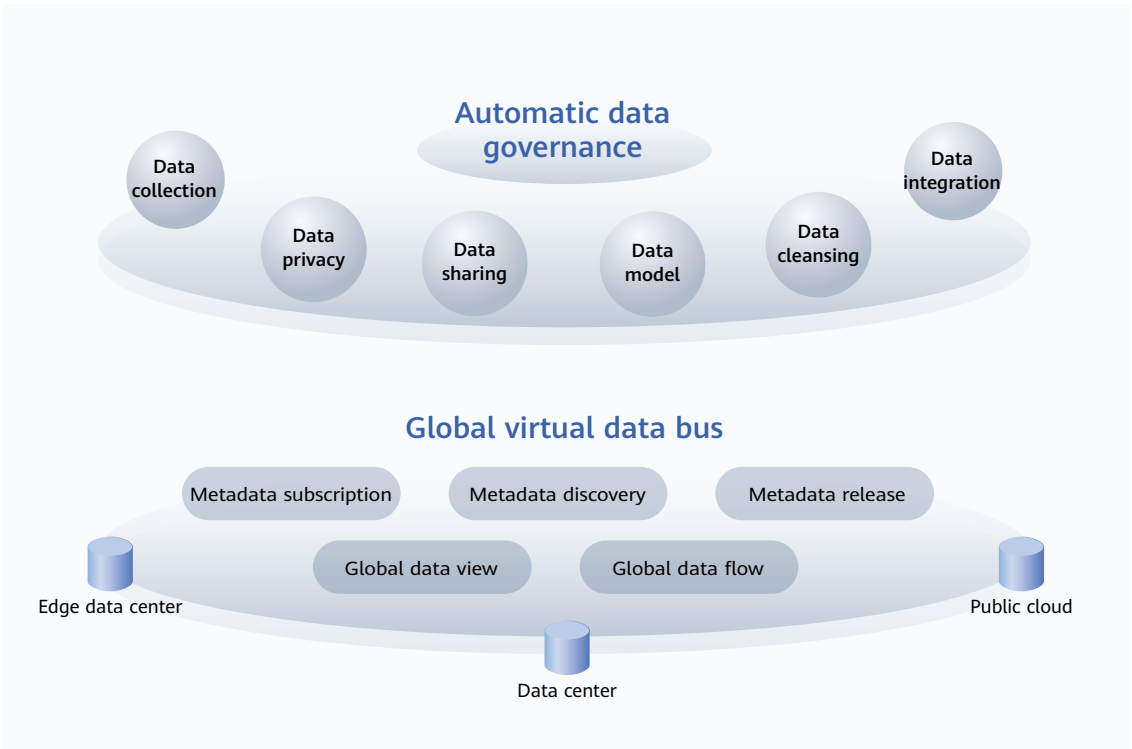


Figure 2-19 Cross-region data collaboration framework

Global virtual data bus: In public clouds and enterprise data centers, data is managed in different regions, and thus a large number of data silos exist. In the future, the release, discovery, and subscription of metadata can help realize efficient on-demand interconnection and build a global virtual data bus. Such virtual data bus must have a unified data namespace and transparent data flow capability to provide cross-cloud global data space and secure, efficient, and easy-to-use data networks.

Automatic data governance: To enhance the efficiency of the process from data collection to

data value mining, data from different sources and of different types requires interconnection and efficient collaboration. This can be implemented through unified and standardized data models and systems that provide automatic processing capabilities and integrate basic functions such as data collection, cleaning, integration, quality improvement, and security assurance. The current data governance technology is not mature. Breakthroughs need to be made in heterogeneous data integration, data lineage management, and data classification and grading to develop unified, efficient, and intelligent data cube services to effectively improve data quality and availability.

Storage Network

As the amount of data generated continues to increase, data silos will become even more common. It will hence be crucial to facilitate cross-region data flow to ensure better data management. However, the long network latency and low system efficiency of data access severely hinder the development of data applications. Therefore, an efficient and fast storage network needs to be constructed to implement data access that is insensible to applications and regions. Based on current storage trends, future storage networks must provide the following capabilities:

Storage semantic awareness: Traditional networks are aware of only network semantics, such as IP addresses and TCP/UDP port numbers, and treat all network packets in the same way. Future intelligent data network can further sense storage semantics, for example, identify and prioritize packets based on storage semantics to implement policy-based forwarding, identify association between packets to schedule co-flow, and route to fit storage I/O semantics. These features will enable differentiated processing of storage packets and optimized utilization of limited network resources, which will facilitate frequent data exchange among different nodes.

Online compute service: Intelligent data networks will expand network computing capabilities beyond solely packet forwarding and routing capability common in traditional networks. The abstract operator is used to design a Turing-complete instruction set to develop an efficient data processing engine that can be carried by a network forwarding device or device-side NIC. A network forwarding device implements associated-channel processing of data, and performs computing processing such as data encryption, decryption, compression, redundancy removal, and verification during data migration, thereby realizing real-time parallel processing of computing and transmission. A device-side NIC implements near-data computing to save data migration bandwidth and deliver a low-latency service. In addition, it

also carries the protocol conversion of interfaces such as Smart Data Accelerator Interface (SDXI)/ NVMe to provide hardware-based data flow capabilities.

Online storage service: While current networks generally transfer data packets, future networks will use the forwarding and processing capabilities of a large number of data packets. This provides diversified associated-channel storage services, such as distributed locks, metadata cache, and transaction concurrency control, to achieve sub-RTT service response time and greatly improve data access efficiency.

Multi-objective transmission: In terms of network control protocols, traditional TCP/IP networks are designed for network survivability and deliver either high throughput or low latency, but this is not a compromise with new-gen technologies such as RDMA over WAN, F6G, and all-optical networks. In terms of network routing protocols, traditional networks are designed for a single objective, whereas modern storage networks handle both real-time database query requests (low latency) and large file transfer requests (high throughput), with a host of multiple objectives, such as the shortest path, maximum network utilization, and load balancing achieved. Therefore, future networks are expected to run on multi-objective protocols to meet diversified data services.



2.5 Data Intelligence

Digital infrastructure in 2030 will be closely related to our everyday lives. Consider how digital twins, metaverse, and ChatGPT are made possible on today's networks, but hindered by limitations in intelligent data processing of emerging multi-cloud applications. Future systems must decouple data logic from data intelligence. The data infrastructure faces three challenges: (1) Scattered data causes silos and impact sharing. (2) Data mining consumes huge resources due to multiple modeling, training, and reasoning processes, which is unsustainable. (3) Complex data management of mass applications affects the efficiency of data pre-processing, severely restricting the development of applications.

Data Intelligence will have benefits in data awareness and understanding and new data services, and support the projected hundred-fold growth in data services in numerous industries. Digital storage is developing towards the ubiquitous, diversified, and cognitive storage trends.

Ubiquitous: Future storage will be miniaturized, portable, green, and intelligent and offer advantages in power consumption, density, and processing. It will be available in new forms (computational, brain-like, and biological DNA storage), and the portable features will enable

faster large-scale commercial availability. In the short term, portable storage will improve data transfer speeds on the device side, edge, data center, or cloud. In the medium and long term, its building-block design will form reliable, secure, and O&M-free storage that can implement real-time data sharing, interaction, and processing.

Diversified: Traditional applications are mainly graphics and images in data format. However, the emerging applications such as brain computer interfaces, bionics, and AI will drive the diversification of data formats and create new data paradigms like vectors, tensors, and retrieval-augmented generation (RAG). and long-term memory storage. From the perspective of data semantics, autonomous driving, drones, and robots will generate a large amount of data with composite semantics.

Cognitive storage: Currently, storage devices only provide the data storage function with multiple access layers, failing to offer ultimate application experience. This will be improved in future, smarter storage that is characterized with cognitive capabilities. It offers advantages in automatic processing and analysis, adaptive modeling, domain knowledge acquiring, and optimized data processing capabilities through learning^[25].

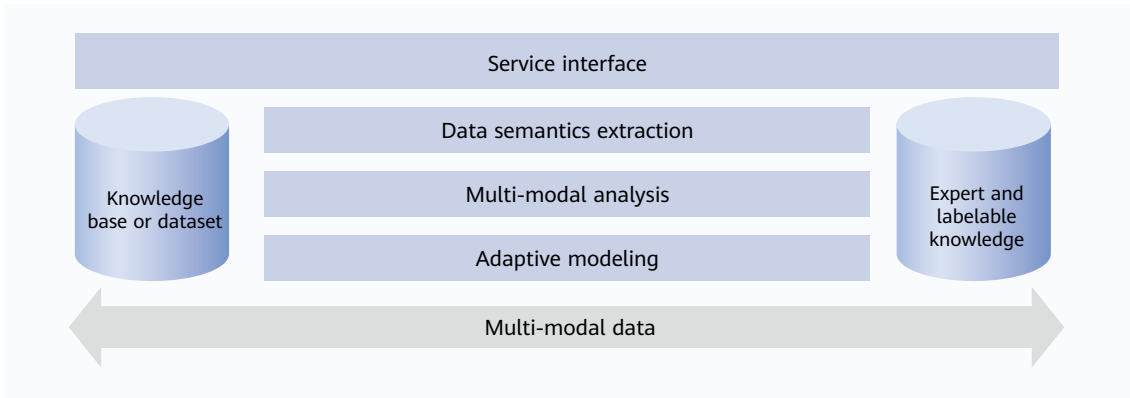


Figure 2-20 Framework of data intelligence

The future of data intelligence technology will follow four major trends:

Service Interface for Content Consumption

Current levels of storage devices provide basic data interfaces such as block, file, and object interfaces, which can further connect to Table format (databases), DataSet vector (training and inference), and asset interfaces (transaction applications). Next-gen data services and APIs, however, will go provide more advanced functionality, better performance, and secure access to data, and support advanced applications to leverage the higher power of data.

Such interfaces are aimed at helping engineering workloads. By simplifying pagination, data streaming, and event-driven architecture, the advances in custom queries, filtering, and programming provide features tailored to specific cases.

In most cases these interfaces will be integrated with NLP technologies to provide services like ChatGPT, in which users can interact with AI to ask questions and receive relevant answers.

For business decision-makers, the data interfaces can use predictive analytics to provide insights and predictions based on historical data. This helps users identify patterns and trends in data that may not be visible through simple analysis. Instead, advanced visualization tech can help the users obtain a more comprehensive understanding of system data.

In summary, future storage will evolve from simple data access to content consumption.

Data Semantics Extraction

Data semantics applies smart AI techniques to extract target information related to service objectives from data sets. In doing so, the original data can be compressed and the system efficiency can be improved.

Current semantic extraction is developed on natural language, knowledge graphs, and deep neural networks, but is limited

by interpretability, accuracy (inference from raw data), and scale of the deep neural network theory. Current techniques run poor generalization and limited deployment, requiring multiple trainings. High availability is based on intact semantics independent of software/hardware and cross-platforms differences, whereas portability requires a semantic scheme for complete data definition and description to drive the standardization and industrialization of data services. In the future, the NLP and pre-trained model technologies will catalyze new breakthroughs in semantics extraction and lossless semantic inference.

Multi-Modal Data Analysis

There is a growing trend of integrated multi-modal, supported with optimized sensory tech that streamlines the collection and processing of structured and unstructured data ^[26]. Consider the workloads involved in autonomous vehicles. Such systems must simultaneously aggregate and process multiple data sources from roads, traffic, in-vehicle sensor, and surrounding environments, realizing situational awareness from which informed decisions can be made.

At the same time, the multi-modal data processing must standardize and process data from different sources so that the data can be exchanged and shared between different applications. Data convergence in the future may have the following modes:

1. Multi-modal data convergence: Data of multiple types such as voice, image, and sensor data are converged and analyzed to obtain enough reliable information, solving the main problems common in a single data source.

2. Multi-layer convergence: Converged data from different layers is, such as bottom-layer sensory data and top-layer semantic information, converged and analyzed to

improve accuracy and depth of data analysis.

3.Multi-source data convergence: Data from multiple sources, such as social media, IoT, and enterprise internal systems, is converged and analyzed to improve data integrity and scope and to discover associations and relationships.

Current multi-modal data convergence and analysis is based on rule, feature, and semantic convergence algorithms, in coordination with machine and deep learning, computer vision, natural language processing,

and sensor technologies. In the future, the multi-modal data convergence analysis will solve the problem of strong dependence on data homogeneous distribution and closed domains. Through spatial transformation, self-supervised learning technology, and AI-Generated Content (AIGC), the multi-modal data convergence analysis can implement cross-modal learning and automatically learn the semantic alignment relationship between modal to improve the precision of modal convergence.

Adaptive Data Modeling

Data Adaptive Modeling is an approach that identifies and learns patterns and structures from data as it is collected, and generates corresponding prediction models. Current models are hindered by inconsistent sampling sets and data drifts caused by differences in the application environment and training models. Data drifts mean that the old models cannot adapt to new environment and need to be retrained. In addition, the data adaptive model needs to quickly adapt to new environments and scenarios, for quick response and efficient prediction.

Currently, adaptive modeling is developed on neural networks and machine learning technologies, requiring dedicated network structures and features. In the future, breakthroughs in incremental and transfer learning, domain adaptation methods, along with Generative Adversarial Network (GAN) must enable adaptive modeling to suit a wider scope in complex and changeable deployments ^[27].

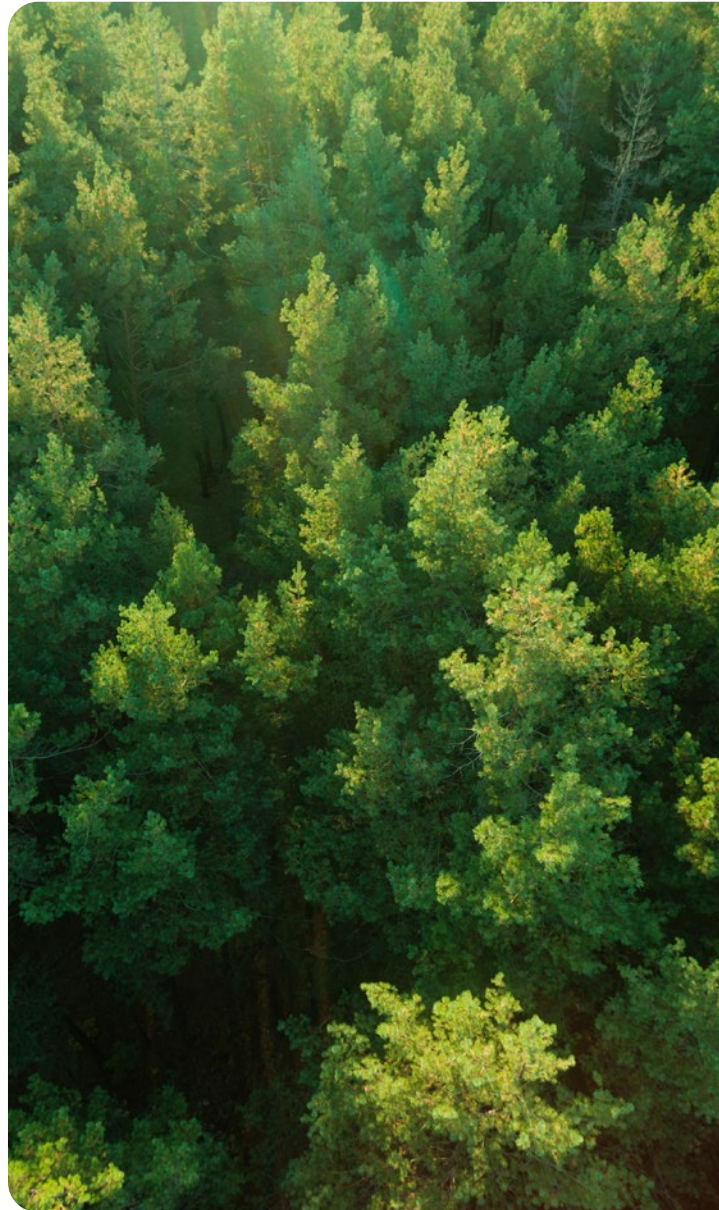


2.6 Sustainable Storage

It is estimated that by 2030, 4% to 6% of the global annual electricity output will be required to read global data once a month. The carbon dioxide generated from this must be absorbed by global trees in seven days. Therefore, in order to build a sustainable data infrastructure, we must reduce the energy consumption per bit of data read/write.

Based on the classic Von Neumann architecture, the energy consumption of data transmission between storage and computing units accounts for 60% to 90% of the total energy consumption of the IT system. The energy consumption problem of data-intensive applications is particularly prominent. However, data-centric architecture will solve the problem of high power consumption during data transmission.

In the future, technologies such as low-power-consumption media and transmission of data with optical signals instead of electronic signals will reduce energy consumption. Energy-saving technologies involving storage systems, entire devices, and environment-related solutions will further reduce carbon dioxide and improve energy consumption efficiency. They reduce energy consumption in terms of chips, media, and networks, achieving optimal energy efficiency per bit and minimum carbon emission.



Storage System-Level Energy Saving

These technologies detect the running status of computing, storage, and network devices to identify hot and cold data characteristics and service load rules used to construct a system optimization model. The model can be used to adjust software and hardware working status parameters to achieve optimal energy consumption of the entire system. The system-level energy saving technologies are as follows:

1. Intelligent power consumption optimization for hardware

Historical data analysis based on big data and AI reveals key factors that affect energy consumption, so as to obtain PUE prediction and energy saving benefit models. Optimization algorithms are employed to obtain optimization parameter groups and predict the optimization policy and total energy consumption of devices (such as CPUs, disks, networks, fans, and cooling pumps), to slash energy consumption of the entire system. Current solutions produce insufficient model generalization and samples, and non-real-

time performance, requiring much manual intervention and energy. In addition, AI models have poor explainability, leading to high operation security risks^[28]. Future technologies such as modularization of models will be built on expert experience, probabilistic modeling will use fewer samples, online training/ inference will be more efficient, and domain adaptation will reduce manual intervention workloads, which in turn improve model explainability and slash energy consumption.

2.Energy saving in data tiering

One issue is to reduce electric energy consumed by devices such as servers, storage devices, and network devices in non-working hours. Hot and cold data tiering stores data in magnetic-optical-electric hybrid media based on the data usage frequency, to effectively reduce energy consumption and balance performance and costs. Current data tiering policies and capacity planning are based on manual experience, wasting a lot of resources, and issues such as those related to I/O access modeling, data layout, and prefetch must be resolved. There is a need for AI-based refined models that can ensure performance and minimize energy consumption caused by data access.

3.Heat dissipation technology of storage devices

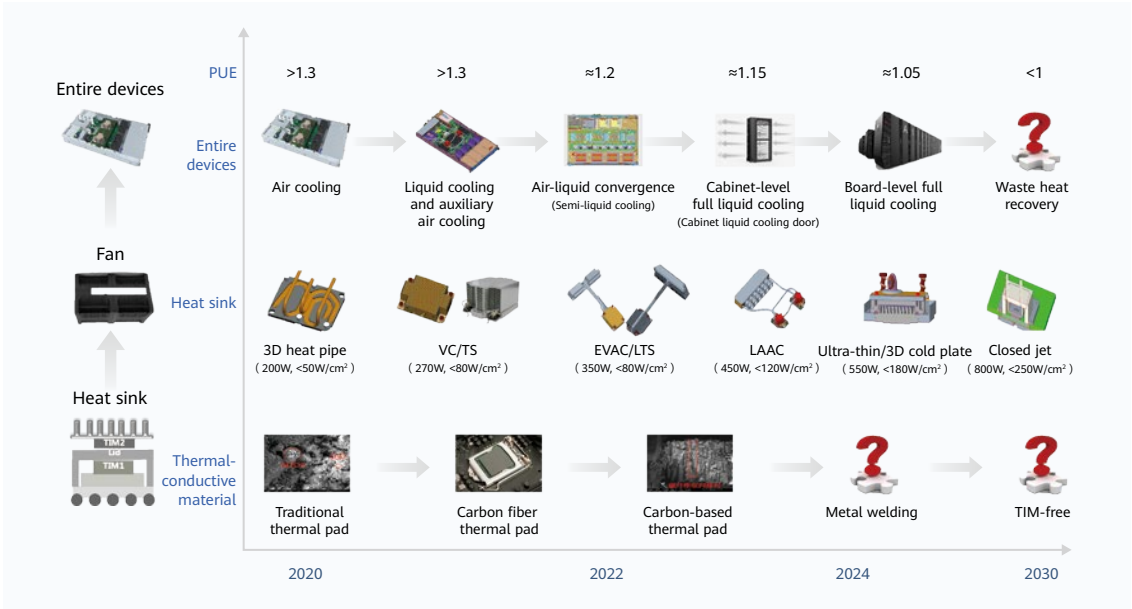


Figure 2-21 Heat dissipation technology of storage devices

Typically, cooling systems account for 30% to 60% of the total power consumption in a data center, making it a key area to slash carbon emissions, specifically by improving the entire-device heat dissipation system and reducing its power consumption^[29]. Consider the air-cooled heat dissipation technology. It runs 5–10 COP but with huge noise generation, which has become an insufficient design for most environments. Future breakthroughs in technologies such as heat conduction-free materials on chips, closed

jet impingement cooling designs, new cooling technologies, and entire-device waste heat recovery will improve heat dissipation efficiency and reduce carbon emissions across all hardware layers. However, issues must be addressed such as those related to the zero-thermal resistance welding, non-aqueous working substance and liquid-cooling material with high specific heat capacity and corrosion resistance, and electric machine conversion efficiency.

4.Resource awareness coordinated scheduling technology

Clean energy (e.g. solar, wind, water, geothermal, biological, and nuclear) does not produce any greenhouse gases. To reduce carbon emissions, large data centers are intensively deployed in Western China which is home to rich clean energy while small-scale ultra-low-latency edge data centers are deployed in Eastern China to support local services and reduce data migration. Data placement policies and cross-DC scheduling engines are needed to dynamically detect the location, status, availability, and heterogeneity of computing, network, and storage resources as well as regional resource pricing and carbon emission standards in real time, to achieve cross-DC data allocation. This achievement combines with intelligent data collaboration to build a global unified framework. This facilitates data extraction, analysis, and aggregation across DCs to achieve optimal computing, data

Data Transmission Energy Efficiency Improvement

Current network communication devices account for about 15% of the total energy consumed in a data center. Driven by new applications (AI and big data analysis), data centers will require higher transmission bandwidth, leading to higher energy consumption. As the transmission speed reaches 400G and even 800G, power consumption will become a bottleneck for improvement of network bandwidth. It is estimated that the electricity expenditure will account for about 95% of the annual operating expense of data centers in 2030, and networks will account for 20% of the total power consumption of data centers. Therefore, optimizing the energy efficiency of communications networks is a priority.

Mainstream data center network solutions use the optical-electrical-optical conversion process and

electrical signal processing, which are the common areas of excess power consumption, making them the obvious starting points for energy savings. One solution is optical switching, which directly maps optical signals to outputs, requiring no optical-to-electrical conversion, to provide 10 TB-scale bandwidth, ns-scale latency, and TB-scale performance per watt. Current optical switching is based on the time switching technology, with the optical path switching latency as high as dozens of milliseconds. The optical-electrical hybrid technology can be used to build a high-throughput network, and breakthroughs will see nanosecond-level optical switching technology and high-speed switching algorithm, to achieve an all-optical data center network with low power consumption.



Chip-Level Energy Saving Technologies

Chips account for most of the energy consumption in current storage systems, making it critical to reduce the energy consumption of chips. As chip components are increasingly integrated, heat dissipation per unit volume is increasing. However, the limited heat dissipation speed of materials restricts chip performance. How to increase the chip computing power and control the chip energy consumption becomes a big challenge. Technologies such as heterogeneous and diversified computing power integration and on-chip dynamic intelligent energy efficiency management can realize both high computing power and low power consumption.

Chip-level energy saving technologies have the following research directions:

1.Low power-consumption raw material

The chip integration density is expected to keep increasing with progress made as follows: emerging chip materials such as cold source structures, oxide materials, and carbon-based nanomaterials; packaging technologies such as

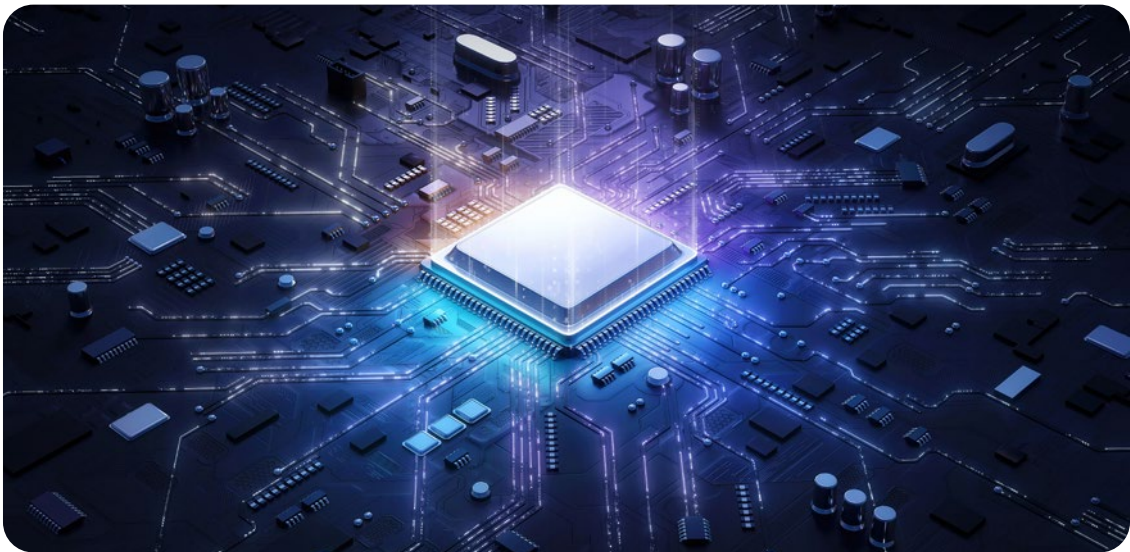
3D packaging and wafer level chip packaging; and low-power consumption technologies such as complementary field-effect transistors (CFETs).

2.High-density and low-power consumption processes

As chip components are becoming smaller and smaller, the energy consumption decreases accordingly. However, this classic physical law is no longer applicable on a nanometer/angstrom scale. In the future, DTCO\STCO technologies are expected to find an optimal chip design and lithography process, to ensure future solutions can integrate 100 billion transistors.

3.Chip energy consumption management

Technologies for on-chip energy consumption management can reduce energy consumption by controlling the chip voltage and clock frequency. Currently, the chip voltage and clock frequency are controlled by chips and maximized according to module requirements, leading to huge energy waste. The on-chip energy consumption





management technology can control the voltage and the clock frequency of the core sub-modules at the core level, to ensure proportional changes in energy consumption and computing power. Future-oriented AI and sensor technologies can be used to implement power prediction, power capping, and component power consumption control, achieving the optimal energy efficiency ratio at the component level.

4. Digital processing specialization

As Moore's Law slows down, the performance improvement of a single CPU is a bottleneck, while annual growth rate of computing power is less than 50%, and the gap between supply and demand is widening. As Dennard Scaling comes to an end, using multiple cores to improve computing power greatly increases energy consumption. The conventional general-purpose processor architecture cannot meet the development needs of diversified applications, requiring specific,

tailored architecture designs to meet different computing power needs and achieve low system power consumption^[30].

The current field-specific architecture provides diversified computing power through efficient parallel forms, hierarchical memory structures, hybrid precision, and field-specific programming languages. Due to differences in system architectures, instruction sets, and programming models, diversified computing power faces challenges such as difficult cross-platform program running and high programming complexity, fueling the need for technical breakthroughs in unified instruction sets, heterogeneous resource abstraction, efficient resource scheduling, and heterogeneous programming models. This will be the foundation for large-scale multi-system heterogeneous software platforms that integrate compilers, programming languages, acceleration libraries, and development tools.

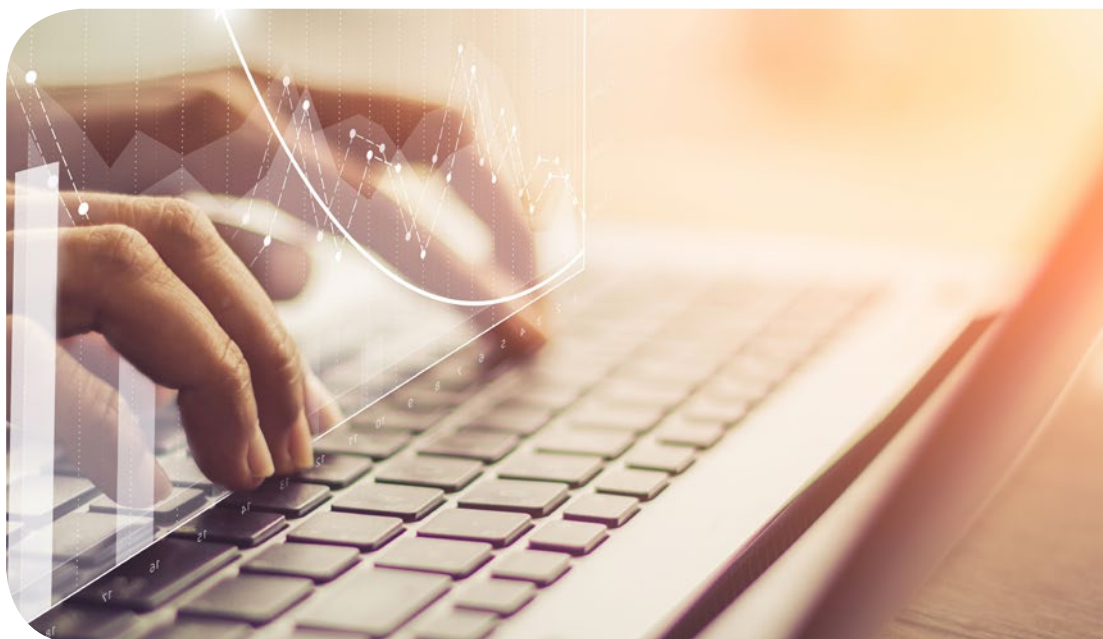
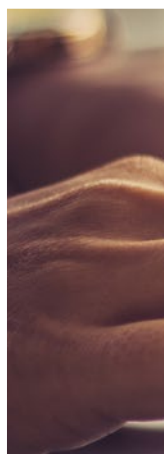
Green and Intensive Storage Standards

Data centers in China consumed about 270 billion kWh of electricity in 2022, representing an increase of 25% compared with 2021 and accounting for 3.1% of the total power consumption of the whole society. It is estimated that the data center energy consumption will double by 2030, resulting in high carbon emissions and heavy pollution. The storage industry urgently needs to gradually improve green and intensive storage regulations and standards in line with the national "carbon peak and carbon neutrality" strategy.

Current green storage standards for data centers cover energy efficiency simulation models,

energy saving technologies, LCA carbon emission evaluation, carbon emission reduction and low carbon emissions, and recycling, but to date there is no unified standards for the storage industry. This needs to be addressed in the future to cover key indicators such as the carbon footprint throughout the data lifecycle, chip control interface, data transmission power consumption, and energy efficiency, carbon emission intensity, and renewable energy utilization of storage devices. Such achievements in energy consumption will create the benchmark from which a comprehensive evaluation system will be formed for a green, low-carbon storage industry.





Data Storage 2030 Initiative

Data storage is the cornerstone of digital infrastructure that supports the globalization of the digital economy. The industry will experience YB-scales of data by 2030, requiring collaboration to make breakthroughs to build better data storage. We hope that industry collaboration and innovation focus on the following:

1. Diversified media, innovative applications, and improved capacity density and per-bit energy efficiency.
2. Breakthroughs in system architecture that expand beyond the traditional von Neumann architecture, promotion of the construction of the data-centric system architecture, high-throughput P2P interconnection bus, and unified standards and protocols, as well as new-gen infrastructure.

3. Storage power development that boosts computing in storage, establishes a storage power indicator-based model based on the whole process of data processing, and improves efficiency.

4. Zero-trust storage systems that separate property, use, and control rights, establish unified standards for data gravity indexes, and streamline mobility of trusted data flows.

5. Green and intensive standard systems that feature optimal energy efficiency per bit and carbon emission must be built on sustainable IT. This will contribute to the shift from environmental-centered energy efficiency to IT-oriented and sustainable energy savings.

Let's work toward the new era of digital infrastructure together.

Appendix A: References

- [1] Seagate and IDC, Data Age 2025, May 2020
- [2] Gartner, "Forecast: Hard-Disk Drives, Worldwide, 2020-2026", 2022. <https://www.gartner.com/document/4014430>
- [3] World Health Organization. World health statistics 2021: monitoring health for the SDGs, sustainable development goals. 2021. <https://apps.who.int/iris/handle/10665/342703>
- [4] Deloitte China, Digital health whitepaper, 2021
- [5] 2030 Sustainable Development Goals in China, SDG China, <http://sdgcn.org/sdg2.html>
- [6] realtor.com, Smart Home Technologies Reshape Real Estate Preferences in 2020, <https://www.realtor.com/research/smart-home-tech-2020/>
- [7] World Economic Forum, Raising Ambitions: A new roadmap for the automotive circular economy, 2022, https://www3.weforum.org/docs/WEF_Raising_Ambitions_2020.pdf
- [8] IDC, Worldwide Smart Cities Spending Guide, 2021
- [9] Korn Ferry, Future of Work---The Global Talent Crunch, 2018, <https://www.kornferry.com/content/dam/kornferry/docs/pdfs/KF-Future-of-Work-Talent-CrunchReport.pdf>
- [10] United Nations Environment Programme, Emissions Gap Report 2020, 2020, <https://www.unep.org/emissions-gap-report-2020>
- [11] Abbosh O., Bissell K., Reinventing the Internet to Secure the Digital Economy, 2019, https://www.accenture.com/_acnmedia/thought-leadership-assets/pdf/accenture-securing-the-digital-economy-reinventing-the-internet-for-trust.pdf
- [12] Hennessy, John L. and Patterson, David A., Computer Architecture, Fifth Edition: A Quantitative Approach, Morgan Kaufmann Publishers Inc., 2011
- [13] China Academy of Information and Communications Technology (CAICT), Data Elements White Paper, 2022
- [14] Gartner, HDD and SSD market forecast, 2021
- [15] Yang S, Zhang J. Current Progress of Magnetoresistance Sensors. Chemosensors, 2021
- [16] Takeshi H., Hitoshi N. A study on high-density recording with particulate tape media for data storage systems, Synthesiology, 2017

- [17]SONY, Optical disc archive generation 2 white paper, 2016
- [18]Yuan X., Zhao M., Guo X., Li Y., Gan Z. and Ruan H., Optical tape for high capacity three-dimensional optical data storage, Chinese Optics Letters, 2020
- [19]Shu Jiwu, Outlook for a new storage-compute decoupling architecture technology, Communications of the China Computer Federation, 2022
- [20]Fan Dongrui, Ye Xiaochun, Bao Yungang, Sunninghui, The road to self-developed high-throughput computer of China, HPC Development Strategy of China, 2019
- [21]Conte T. M., DeBenedictis E. P., Gargini P. A. and Track E., Rebooting Computing: The Road Ahead, IEEE Computer Society Press, 2017
- [22]Wu Jiangxing, Cyberspace Endogenous Security Development Paradigm, China Science: Information science, 2022
- [23]Yin X. F., Zhu Y. M., Hu J. K., A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions, ACM Computing Surveys, 2022
- [24]Gupta A., Key Pillars of a Comprehensive Data Fabric, Gartner, 2021
- [25]Microsoft Azure, Knowledge store in Azure Cognitive Search, 2023, <https://learn.microsoft.com/zh-cn/azure/search/knowledge-store-concept-intro?tabs=portal>
- [26]Baltrušaitis T., Ahuja C., Morency L. P., Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019
- [27]Wilson G., Cook D. J., A Survey of Unsupervised Deep Domain Adaptation, Association for Computing Machinery, 2020
- [28]Yu Y., Wu C., Zhao T., OPU: An FPGA-based overlay processor for convolutional neural networks, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2020
- [29]State Information Center, Research Report on Green and High-Quality Development of Data Centers, 2022
- [30]Hennessy J. L., Patterson D. A., A New Golden Age for Computer Architecture, Communications of the ACM, 2019

Appendix B: Acronyms and Abbreviations

Acronyms and Abbreviations	Full Spelling
AIoT	Artificial Intelligence of Things
CBA	CMOS-bonded Array
CFET	Complementary Field-Effect Transistor
CIM	Computing In-Memory
CMOS	Complementary Metal-Oxide-Semiconductor
CNA	CMOS Near Array
COP	Coefficient of Performance
CUA	CMOS Under Array
CUE	Carbon Use Efficiency
DNA	Deoxyribonucleic Acid
DPU	Data Process Unit
DRAM	Dynamic Random Access Memory
DTCO	Design-Technology Co-Optimization
EDA	Electronic Design Automation
HAMR	Heat-assisted Magnetic Recording
HDD	Hard Disk Drive
HPDA	High Performance Data Analytics
IoT	Internet of Things
LCA	Life Cycle Assessment

Acronyms and Abbreviations	Full Spelling
LTFS	Linear Tape File System
LTO	Linear Tape-Open
MAMR	Microwave Assisted Magnetic Recording
MRAM	Magnetoresistive Random-Access Memory
PB	Petabyte
PLC	Penta-Level Cell
PUE	Power Usage Effectiveness
QLC	Quad-Level Cell
SCI	Storage Compute Integrated
SCM	Storage Class Memory
SSD	Solid-State Drive
STCO	System Technology Co-Optimization
STT-MRAM	Spin-Transfer Torque MRAM
TEE	Trusted Execution Environment
UB	Unified Bus
Wafer Level	Wafer Level
YB	Yottabyte
ZB	Zettabyte
ZB	Zettabyte

Appendix C: Acknowledgment

We express our heartfelt gratitude to the numerous experts from Huawei and over 100 distinguished scholars from various fields who actively engaged in the discussion and shared their insights towards shaping the future of the data storage during the compilation of Data Storage 2030. Your invaluable input is instrumental in identifying the development direction and technical characteristics of the sector. Thank you for your valuable contribution.

(This list is sorted by the initial letter of the scholar's name.)

- Bao Yungang (Researcher Fellow, Institute of Computing Technology, Chinese Academy of Sciences)
- Cui Heming (Associate Professor, the University of Hong Kong)
- Chen Mingyu (Researcher Fellow, Institute of Computing Technology, Chinese Academy of Sciences)
- Feng Dan (Distinguished Professor of the Changjiang Scholars Program, Huazhong University of Science & Technology)
- Gu Rong (Distinguished Research Fellow, Nanjing University)
- Guo Minyi (Professor, IEEE Fellow, Shanghai Jiao Tong University, Member of Academia Europaea)
- Huang Qin (Professor, Beihang University)
- Jiang Dejun (Associate Researcher, Institute of Computing Technology, Chinese Academy of Sciences)
- Jin Hai (Distinguished Professor of the Changjiang Scholars Program, IEEE Fellow, Huazhong University of Science & Technology)
- Li Yi (Associate Professor, Huazhong University of Science & Technology)
- Liu Xianming (Professor, Harbin University of Technology)
- Lu Youyou (Associate Professor, Tsinghua University)
- Miao Xiangshui (Professor, Huazhong University of Science & Technology)
- Ren Kui (Professor, ACM Fellow, IEEE Fellow, Zhejiang University)
- Shu Jiwu (Distinguished Professor of the Changjiang Scholars Program, IEEE Fellow, Tsinghua University)
- Tang Zhuo (Professor, Hunan University)
- Wang Cong (Professor, City University of Hong Kong)
- Wang Zeke (ZJU 100 Young Professor, Zhejiang University)
- Wang Zhaoguo (Associate Professor, Shanghai Jiao Tong University)
- Wu Hequan (Academician of the Chinese Academy of Engineering)
- Xie Changsheng (Professor, Huazhong University of Science & Technology)
- Zhao Shizhen (Associate Professor, Shanghai Jiao Tong University)
- Zhou Ke (Distinguished Professor of the Changjiang Scholars Program, Huazhong University of Science & Technology)



Notes on the update:

Huawei collaborates with industry experts, customers, and partners to explore the intelligent world. The progress towards an intelligent world has accelerated significantly, with new technologies and scenarios emerging constantly, and industry-related parameters changing exponentially. As a result, Huawei has updated the *Intelligent Automotive Solution 2030* report released in 2021, providing insights into the scenarios and trends towards 2030, and adjusting the relevant forecast data.

HUAWEI TECHNOLOGIES CO., LTD.

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

Trademark Notice

 HUAWEI, HUAWEI,  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

The information in this document may contain predictive statement including, without limitation, statements regarding the future financial and operating results, future product portfolios, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

Copyright © 2024 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.