Top 10 Trends of Data Center Facility 2024

White Paper

January 2024

HUAWEI

CONTENTS

Foreword

The data center industry is currently in a special period of rapid development and technological changes. While new AI applications are reshaping the entire world, bringing convenience to the society, it also brings new opportunities and challenges to the data center industry. The explosive growth of intelligent computing puts forward new demands for greater computing power and higher performance for data centers, and provides a strong impetus and a broader space for the industry.

With the further development of digital technologies, data centers have made remarkable progress and breakthroughs in terms of scale, architecture, technology, and management. As a result, they are now attached with greater importance. However, safety accidents concerning data centers happened from time to time in recent years, and the resulting social impact and economic losses have been on the rise. In the context, as the most essential element of data centers, safety and reliability emerge as the common topic of the industry.

In addition, the dramatically rising intelligent computing power has brought unprecedented changes to the data center industry, which requires data centers, on the basis of reliability, availability, and cost-effectiveness, to evolve in a flexible manner to adapt to business requirements in different periods. At the same time, the application of AI technology in data center infrastructures has heralded more possibilities for the industry regarding operation and maintenance (O&M) innovation.

After in-depth discussions with industry leaders, technical experts, and customers, Huawei Digital Power released the *White Paper on Top 10 Trends of Data Center Facility* based on its deep insights and long-term practices, providing references for promoting the healthy development of the data center industry.

1 High-Reliability Product and Professional Service are The key to Ensuring Secure and Reliable Data Center Operation

With the development of digitalization, the digital economy has become the main engine of social development. As the foundation of the digital economy, the operation of data center is directly related to social and economic development. High-quality and reliable operation has become the core requirement of data center. The high-reliability products and professional service, as the critical component of data center, are the key to secure and reliable data center operation.

Security and reliability are the essential requirement of data centers

As the digital foundation, the data center is the physical foundation for massive data. It is the core resource for centralized information processing, computing, storage, transmission, exchange and management. Data center is also the guarantee for normal operation of society and economy. Therefore, security is the significant feature of the data center. However, the reliability and security of data center infrastructure are always weak links. A comprehensive end-to-end assurance mechanism is the most reliable base for stable and secure operation of a data center during its lifecycle.

High-reliability products and professional services are the key to secure and reliable data center operation

Each data center is composed of tens of millions of different components. To ensure high reliability and security of data center, an end-to-end full-chain guarantee mechanism is needed, which should be based on the product security, professional design and O&M.

The concept of high-reliability product includes reliable design of product and reliable product manufacture.

Reliable design of product: The design of product has an effect on the safety of the product. Good design can reduce the impact of accidents or even avoid accidents. Take the lithium battery as an example, the design of product, including selection of battery cell, module assembly, pack-level connection and parallel connection between battery system, all affects the safety and failure rate of battery operation. For example, high-reliability lithium iron phosphate cell can greatly reduce the fire risk of battery due to thermal runaway, which improves the security level of data center power backup system.

Reliable product manufacture: The design determines the "gene" of the product. For most of the products with a large number of parts, the manufacture plays a key role in the quality of the product. To ensure the consistency and reliability of product, the impact of uncertainty (such as manual intervention) should be minimized, and an quality control system with authentication and standardized manufacture process should be built. For example, in the manufacturing stage, the automatic standard production line can greatly reduce problems such as poor consistency. In addition, the digital AI and explicit technologies are used to monitor the parameters of the equipment, so as to further identify potential risk factors in the manufacture such as poor welding, loose bolts, insulation damage, liquid leakage. Therefore, the safety and reliability can be ensured from the manufacture stage.

The professional services include professional deployment and professional O&M:

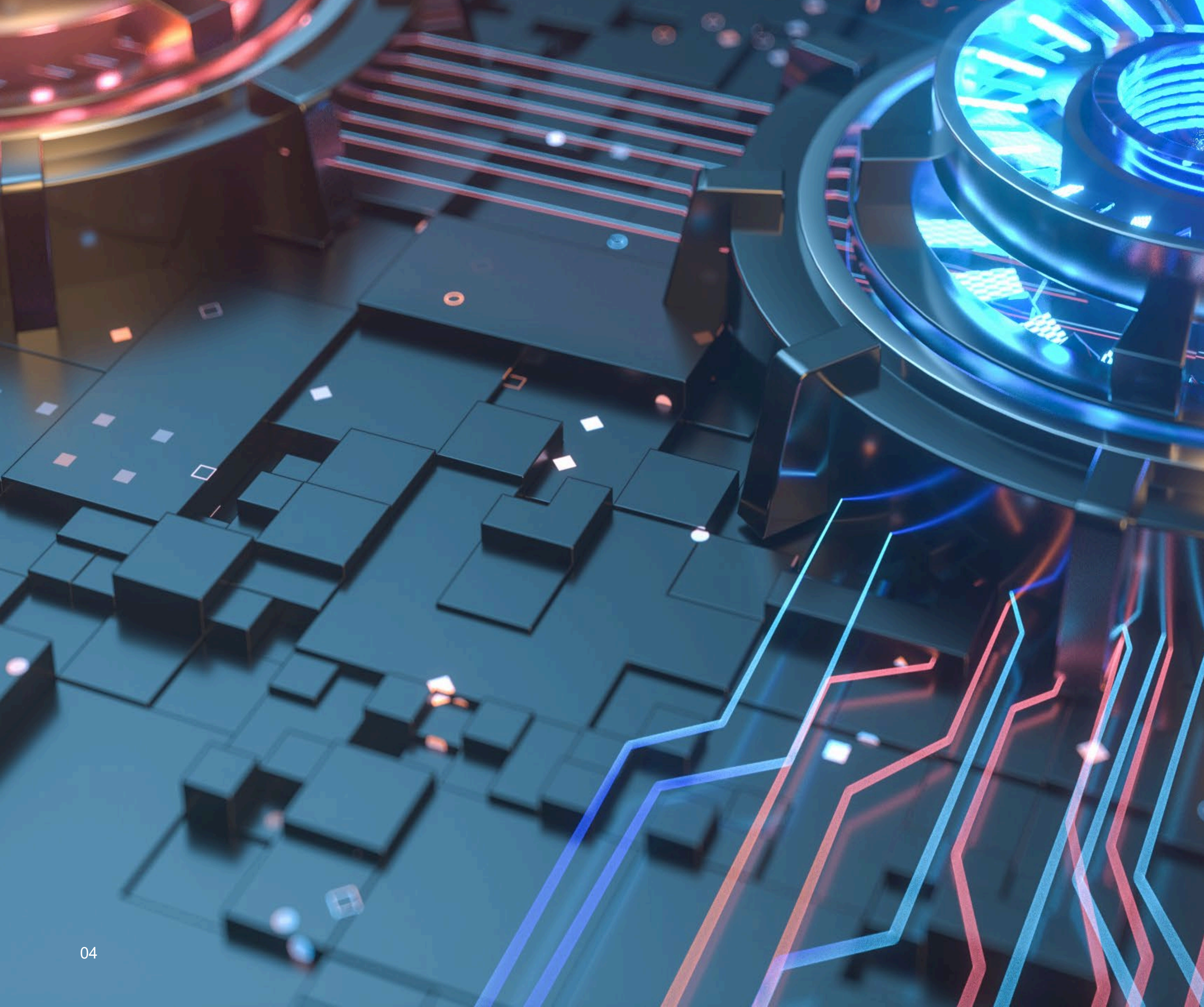
Professional deployment: Data center construction is professional work, including electrical installation, ELV equipment commissioning, cooling system deployment. The professional and standardized construction is critical to the installation quality. For example, torque and resistance measurement during power distribution equipment installation, battery installation, pipeline welding and pressure preservation in the cooling system require careful operation to ensure quality. In addition, technical standards must be complied with to avoid safety risks caused by non-standard operation.

Professional O&M: Reliable products and deployment are the basis for high-quality data centers. Professional O &M is the shield for reliable running of data centers. Good O&M work requires well-established process, professional skills and plans for emergency. In this way, abnormal situations can be detected and be handled in a timely manner. The quick response in case of emergencies can reduce the impact and ensure the long-term stability of data centers.

Only products that strictly abide by the end-to-end assurance mechanism can ensure the secure, stable, reliable and long lasting operation of data centers.

2 Distributed Cooling Architecture Will Become a Better Choice for Ensuring Cooling safety

According to a survey of Uptime Institute in 2023, cooling systems account for 19% of data center accidents or outages, making them the second largest source of failure after power supply and distribution system. Among key factors defining the reliability of data centers, in addition to the failure rate of cooling devices, the cooling architecture design of a data center is essential to the reliability of its cooling system.



The risk of single-point failures exists in the centralized cooling architecture

Currently, most large data centers adopt the centralized chiller plant cooling system, which consists of seven subsystems including the chiller, cooling tower, chilled water tank, cooling device, cooling water pump, plate heat exchanger, and management system, involving a myriad of equipment connected through hundreds to thousands of meters of water pipes with numerous adapters and valves. As a result, problems including numerous fault points and large fault domain keep haunting the system. Once a single-point failure occurs, multiple equipment rooms or buildings in the data center may break down on a large scale, posing great challenges on service stability of the data center.

In recent years, a number of head data center vendors in Hong Kong, Singapore, Guangzhou and other regions have been questioned by the local Ministry of Industry and Information Technology due to the failure of centralized chilled water system which triggered breakdowns of over 10h, resulting in level-1 safety incidents. Meanwhile, services of multiple websites and apps were interrupted, bringing great economic losses. In December 2022, water leakage and air intake occurred in the cooling pipeline of a large data center in Hong Kong, leading to a complete shutdown of the chiller. Rising temperatures in the equipment room thus triggered a secondary fire accident, with servers halted for over 15h. As a consequence, services of multiple major companies were severely impacted, with which came inestimable economic losses. In South China, air lock occurred in the cooling water system of a data center due to the lack of water and air intake in the main pipe. As a result, the entire cooling system failed, and the cooling of the whole building was interrupted. In 2023, a large data center service provider in Singapore failed to start the cooling system due to improper software upgrade and optimization of the chiller. As a result, many servers broke down due to overtemperature and services were interrupted. Online services of a head bank running in the data center were unavailable for a long period.

With independent subsystems, the distributed cooling architecture has a better performance in terms of reliability

The distributed cooling system has a flexible architecture, with its subsystems independent from each other. The failure of a single device does not affect other devices, which delivers better performance in securing cooling safety.

In the distributed cooling architecture, cooling sources are configured for a single data hall and architecture redundancy is configured based on service importance. If a single device is faulty, only a single subsystem is affected, and services in the equipment rooms stay intact. The architecture safeguard important services better and services in other equipment rooms are not affected. It greatly improves the reliability of data centers and emerges as a preferred choice in the intelligent computing era.

On top of that, the distributed cooling system is more easily fabricated, which reduces the burden of onsite engineering and thus the risks brought by inferior construction quality. Besides, a distributed cooling system features easy O&M. Take the indirect evaporative cooling system as an example: compared with the chiller, an indirect evaporative cooling air conditioning unit has a very simple structure, consisting of only one main equipment and several auxiliary devices. It has fewer connection points and only 1/10 pipes are required when compared with a chilled water system. Therefore, errors are less likely to occur during emergency handling and the O&M is greatly simplified. This ensures the cooling efficiency and stability of data centers to the maximum extent.

As data centers grow in scale, there is a corresponding increase in the disadvantages of centralized cooling. With its flexible architecture and high reliability, the distributed cooling system will be more widely used in new data centers and gradually replace centralized cooling as the mainstream solution. Extensive market demands have also driven the technological breakthroughs and progress in the industry. A number of major vendors have started to promote vigorously the distributed cooling architecture, the most representative of which is the indirect evaporative cooling solution. Currently, racks supported by the indirect evaporative cooling solution have exceeded 300,000. In areas with various climate conditions, the solution has been verified through implementation. It is believed that with the further penetration and promotion of new energy-saving technologies represented by the indirect evaporative cooling solution and distributed cooling architecture, the data center industry will usher in a new era of low-carbon, energy-saving, safe and reliable development.

3 Predictive Maintenance Will Become a Basic Feature of Data Center Infrastructure

The emergency response time to faults decreases greatly with the improvement of power density of data centers, which places stricter demands for the maintenance of data centers. Thanks to the advancement of AI technologies, it is now possible to predict risks and manage data center infrastructure with the aid of AI technologies. AI algorithms can learn from historical and real-time data to predict and identify abnormal working conditions. In this way, the safety management of data centers changes from passively targeted maintenance to proactively predictive maintenance, improving data center reliability in terms of O&M.

Elevated power density of data centers has reduced significantly the emergency response time to failures

As intelligent computing technologies evolve, the power of a single cabinet in data centers will soar from 6–8 kW up to 30–40 kW, considerably boosting the data processing capability. This leap not only optimizes computing efficiency, but also fuels the innovation of power supply and cooling technologies for data centers, as high density requires greater supply power, backup batteries with higher energy density, and more efficient heat dissipation performance.

Whereas this also brings the risk of larger failure domains. For example, lithium batteries have a place in the energy storage field of data centers thanks to their high energy density and long service life, but there is also a risk of overheating, especially in abnormal situations such as overcharging, internal defects, and improper use. Public research data shows that from the thermal runaway trigger temperature T2 (150°C–250°C) to the peak temperature T3 (generally no more than 500°C), it takes only 30–60s.

In power-intensive scenarios with IT equipment, cooling system faults may be rapidly escalated, resulting in overheating cabinets. In the event of a failure, the heat dissipation burden increases dramatically, considering that IT equipment generates four to five times more heat per unit of time than conventional computers. In emergency response to faults, temporary measures including direct ventilation and deploying dry ice fans may be taken in a conventional data center. However, in high-density scenarios with a liquid cooling system, conventional methods may no longer be applicable. Generally, for a 30 kW cabinet, if the plate liquid cooling plus direct ventilation solution is used, the emergency response time is limited to only 30s to 1 min when the secondary pipe is faulty.

With the increase of equipment operating time, the electrical connection contacts in the power transformation and distribution system are subject to surface corrosion and loosening under the joint influence of construction quality, moisture and dust corrosion, and vibration stress, leading to abnormal contact temperatures. Such a problem is not easily detectable at low loads, but will pop up instantaneously when loads increase, posing a serious threat to the power safety of data centers.

In these cases, relying entirely on manual emergency handling will reduce us into a passive position. Therefore, it is urgent to develop predictive maintenance technologies to detect potential faults in advance and handle them in time.

Predictive maintenance enables proactive fault prevention in data centers

In a data center, predictive maintenance is a strategy that uses big data and AI algorithms to monitor and analyze the operating status of devices in real time to predict and diagnose faults in advance.

For example, large-scale lithium battery data is accumulated for a long term based on technologies such as big data and cloud computing, which helps capture changes of safety risks, model and identify safety characteristics and quality defects, and monitor parameters such as temperature, voltage, and current of lithium batteries. In this way, it is made possible to predict the health status and remaining service life of batteries, manage the charging and discharging of batteries, and replace batteries in a timely manner, preventing safety accidents caused by overheating or overdischarge.

In high-density scenarios with a liquid cooling system, parameters such as the flow and pressure of liquid cooling pipes are monitored and abnormal parameter warnings are provided to remind O&M personnel to rectify faults timely, preventing high temperature caused by liquid leakage.

In the power transformation and distribution system, the reasonable temperature under the current load is obtained by using the temperature rise model combined with the periodically collected data of the copper bar contact current, ambient temperature, and temperatures of adjacent contacts. When the measured temperatures of contacts exceed the threshold, indicating that the temperatures of certain contacts point are abnormal. Power outage caused by fires due to high temperature can be prevented by alerting O&M personnel to make timely corrections through overtemperature warnings.

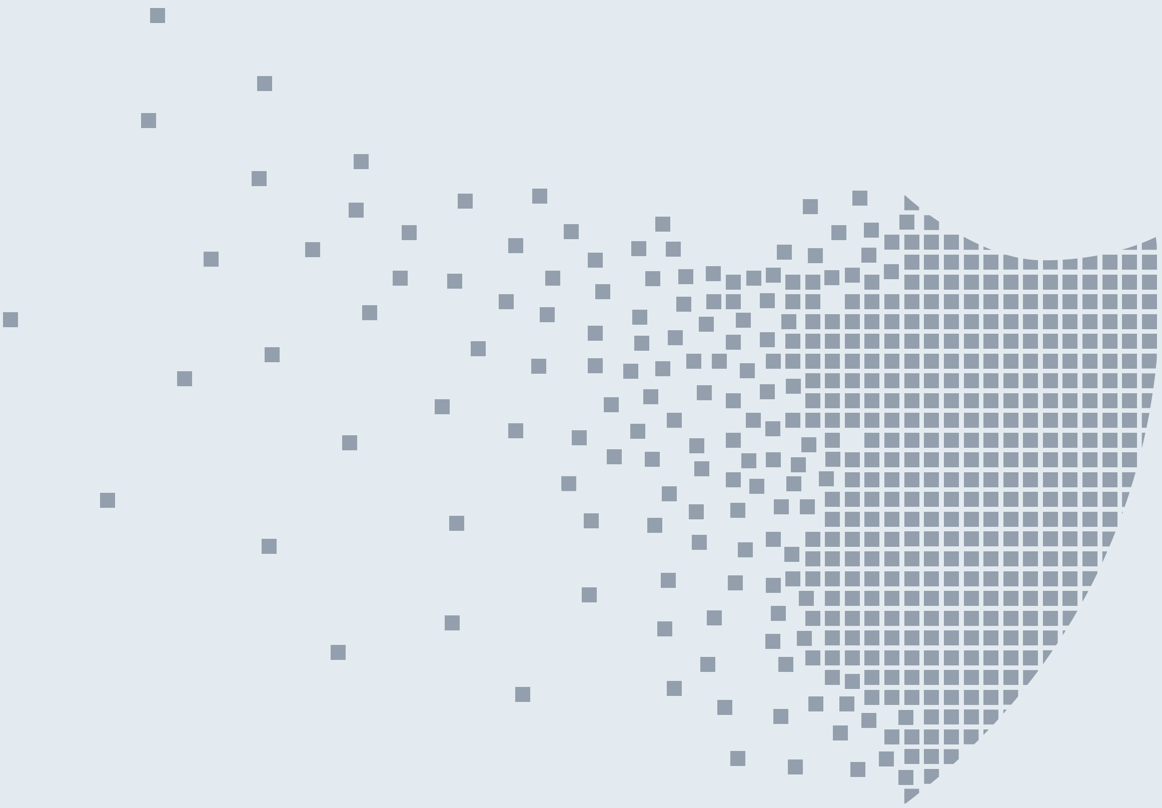
Through these measures, the safety management of data centers changes from passively targeted maintenance to proactively predictive maintenance, considerably shortening the emergency response time to faults and thus improving the reliability of data centers.

Technology application

A large data center in Shenzhen can provide about 156,000 racks after its completion to meet the development requirements for building a smart city and digital government in the next 5 to 10 years. Concerning the power supply and distribution system, the data center adopts Huawei's PowerPOD solution, which supports full-link temperature detection, AI-based low load and high temperature warnings, and early warnings for maintenance, and meets the requirements of high reliability and fast deployment.

4 The Lifecycle Network Security Protection System Will Become a Shield of Data Center Facility

With the advancement of global digital, networking and intelligent technologies, the number and severity of network security vulnerabilities are soaring dramatically, attracting a great deal of attention. According to the security dynamics statistics released by Chinese National Vulnerability Database (CNNVD), the number of vulnerabilities has been on a continuous growth for five consecutive years from 2018 to 2022, and the number of new ultra-high-risk vulnerabilities in 2022 has increased by 52% compared to 2018. The critical infrastructure of data centers is the carrier of massive data, and their safe and reliable operations directly affects the national economy and people's livelihood.



Network security develops into a weak link undermining data center infrastructure

With the rapid development of digital and AI technologies, data center infrastructure, as a digital foundation, shoulders the important responsibility of processing, computing, storing, exchanging, and managing massive information. It is of great significance for the economy, society, and security of a country, and also an important part of all walks of life. Building and developing critical data center infrastructure can ensure national security and promote national prosperity.

However, going forward, there will be an exponential growth of interconnected devices that come from different supply chains and use technologies provided by many ICT solution providers. Such a complex, intertwined ecosystem makes it possible for people to steal relevant technologies and exploit them in unintended ways, including tampering with and sabotaging the infrastructure. When hackers cannot digitally break into servers or applications, they may sabotage the power system, cooling system and other critical infrastructure to disrupt data center operations. For example, manipulate a data center's cooling system by accessing the monitoring system or hacking into the intranet, causing servers to overheat and suffer damage, or disrupt the backup process and upload a malicious backup file, or even shut down a data center's UPSs. All these pose unforeseen risks to data centers.

Security serves as the foundation of data centers, and network security of data center infrastructure has always been a weak link.

Mature ICT network security technologies can be reused for data center infrastructure

It is the cornerstone of network security to build an end-to-end control process from software selection, design, development, validation, and release to achieve software supply chain security where software information can be displayed, software development can be evaluated, suppliers are trustworthy, and risk monitoring is uninterrupted. Based on the intrinsic security design concept and best practices in the industry, build an in-depth defense architecture for product solutions through access control, integrity protection, minimum system, and data security. During data center O&M, the incorrect or missing configuration of configuration items is an important factor leading to attacks. Vulnerability exploitation is the major means of network attacks. Therefore, the system must detect malicious attack traffic in real time, quickly identify network attacks, and respond to attacks in a targeted manner, which is fundamental to ensure the security of important network assets. Therefore, we need to establish situational awareness, security configuration, certificate management, and vulnerability management capabilities, complete organizations, and formulate quick response procedure to ensure visible and controllable O&M security, thus greatly reducing network security risks.

Moving forward, software supply chain security, an in-depth defense architecture for product solutions, and O&M security will together form a network security protection system for data centers throughout the lifecycle.

Technology application

Wuhan Supercomputing Center is a science and technology landmark project located in East Lake High-tech Development zone in Wuhan City, Hubei Province, which is the largest containerized supercomputing center in China. According to the overall planning and design, its computing power is expected to reach 200 petaflops in total, and 50 petaflops in its first phase. The supercomputing center mainly provides high-performance computing services for high-tech fields and advanced technology research. The project employs Huawer's PowerPoD, SmartLi, pre-fabricated modules, monitoring system, and AI energy-saving solution to safeguard the infrastructure of the supercomputing center in an all-around manner.

Wuhan Supercomputing Center

5 Prefabricated and Modular Will Become an Optimal Choice for High-Quality and Fast Delivery

Data center infrastructure has a complicated component system and demonstrates strong engineering attributes, and the construction pace and delivery quality of data center infrastructure hold the key to the rollout and stable running of customers' services. Prefabricated engineering is especially suitable for areas with a weak foundation of data center infrastructure as it boasts technologies such as product modularization and prefabrication.

Emerging Internet markets are in urgent need of simplified deployment for DC delivery

China's Internet industry has matured and experienced a slowdown in recent years. To expand overseas presence for new opportunities has become a consensus of the industry. In addition, with the rollout of the support policies and mechanisms for Internet enterprises by authorities, going global is now a strategic choice for these enterprises to pursue growth. Cloud computing, as a major infrastructure for Internet enterprises, can enable them to deploy services overseas more flexibly. Fast deployment of cloud computing data centers means more opportunities for enterprises.

However, in emerging markets with huge potentials and rapid growth (such as the Middle East, North Africa, Latin America, and Southeast Asia), the large data center industry starts late and the data center infrastructure needs to be improved. The limited industry scale, weak engineering capability, and low construction level contain the growth of traditional data centers in a short period of time. Insufficient practitioners, uneven experience, and short supply constitute bottlenecks of data center construction. In traditional construction mode, a project is subcontracted to multiple vendors using multiple products, making it difficult to ensure the quality of data center construction. Against this backdrop, emerging Internet markets are in urgent need of simplified deployment for DC delivery.

Prefabricated engineering facilitates simplified and rapid DC construction

With ripening technologies of modular and prefabricated products, engineering prefabrication will enable simplified and rapid construction of data centers. Most complicated civil engineering and electrical & mechanical engineering are finished in the factory. Inspired by this practice, major components such as cooling devices and power supply and distribution devices or even the whole data center can be integrated and pre-installed through a modular design in the factory, and then they can be assembled on site like Lego blocks. In this way, a data center can be built quickly. Prefabricated engineering cannot only lower the requirements for engineering construction in areas where data centers are located, but also slash overall delivery time with concurrent engineering procedures.

Providing corresponding prefabricated solutions will become the new normal for data center construction. According to Omdia, 99% of carriers of enterprise data centers consider prefabricated modular data centers a part of their future data center strategies. Take a 200-rack data center as an example. It takes about 24 months to complete such a project in traditional construction. By contrast, the prefabricated modular design can cut the construction period to about 10 months, shortening the TTM by 50%. For example, a PowerPOD is prefabricated and adopts a modular design. Core components are pre-installed and pre-commissioned in the factory. The onsite delivery time will be shortened from two months to two weeks, accelerating service rollout.

Prefabrication at scale ensures high-quality DC delivery

High-quality engineering prefabrication requires strong engineering integration capabilities and rich experiences in the prefabrication industry. It poses a huge test for all service providers to integrate a large number of mechanical and electrical equipment in tons in standard-sized containers while making sure reliable transportation and excellent structural safety and performance after assembly. Standardized manufacturing technology, advanced auxiliary production equipment, sufficient workers, rigorous testing and quality inspection system, supply continuity and diversity, and rich prefabricated experiences are decisive factors in prefabricated engineering. Since the prefabrication industry is large in size and in a leading position, it has all the advantages mentioned above. This will improve the quality of prefabricated engineering, and minimize the impact of DC construction capabilities on DC delivery.

6 Professional Management Platform Makes Data Center O&M More Secure and Efficient

As the scale and complexity of the data center infrastructure increases, the overall management complexity increases greatly. At the same time, the data center infrastructure is becoming intelligent and digital. Device vendors use AI and other technology to enhance their device management capabilities through cloud service. It will be a new direction to efficiently use the professional management platform build on the cloud to reduce the O&M complexity of data centers, which improves operation efficiency and reliability of the infrastructure.

The complexity of cloud data centers increases the complexity of maintenance

As the power density of servers increases, data center infrastructure are gradually integrated, and functions and features are becoming more intelligent. This poses higher requirements on O&M personnel’s skill. At the same time, the data center scale has gradually evolved from 1000 cabinets to 10,000 cabinets, and the overall O&M complexity has increased accordingly. In this context, data center managers and O&M teams are facing unprecedented challenges. The O&M architecture of data center infrastructure should keep flexible and agile in the ever-changing environment to meet the intelligent computing power requirements of higher performance and higher power density in the future.

Professional management platform makes O&M more secure and reliable

With the development of cloud computing technologies, more and more data center equipment vendors use professional management platforms built on the cloud to assist routine management and O&M, enhance the service capabilities and value-added features of their equipment, and further help customers improve O&M efficiency and reliability. With the AI, big data, and IoT technologies, the professional management platform helps users automatically diagnose device faults based on the in-depth understanding of device structures, working principles, and maintenance methods. The more professional and efficient guidance on the maintenance can extend the service life of the equipment.

AI continuously boost the fault prediction and diagnosis capabilities

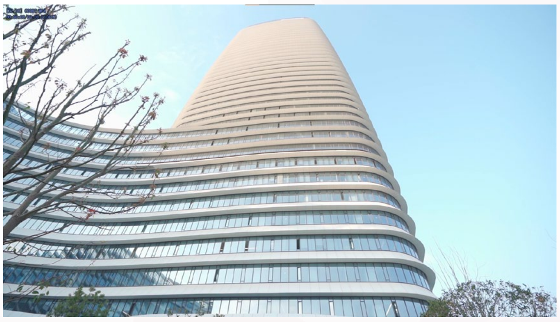
With the technologies such as big data, AI and IoT, fault tree modeling can be performed on a professional management platform. When a fault occurs, intelligent fault diagnosis can be automatically performed, invalid secondary alarms can be shielded in real time, and faults of various devices can be quickly located. In addition, the platform continuously accumulates massive fault handling experience, which makes the model more accurate and has stronger diagnosis and prediction capabilities. With the evolution of the AI, more device fault scenarios, such as rPDU faults, mains power failures, UPS power module faults, and diesel generator faults are supported, greatly shortening the fault response and rectification time.

Devices are directly connected to vendors, providing quicker response and guidance

During the running of the data center, various alarms and emergencies may occur. Currently, alarm notification such as email and SMS are configured in the local management system. Alarms are triggered based on defined alarm rules. However, missing and false alarms may occur. Through a professional management platform, the devices are directly connected to vendors. For example, when an emergency such as smoke alarm, water leakage or high temperature alarm occurs in the data center, the customer service personnel of the vendor can accurately identify the emergency and contact the user immediately for troubleshooting. In addition, remote log sending and OTA upgrade notification can be used to quickly rectify faults and improve the reliability and stability of the data center.

Technology and application

The data center of Hubei Radio & Television Media Building adopts Huawei iManager-M solution, which not only implements remote mobile O&M of the equipment room, but also utilize the professional maintenance experience, proactive alarm notification capability and fault prediction and diagnosis capability of the original manufacturer to make the data center O&M more secure and efficient.



Hubei Radio & Television Media Building

7

The Convergence of Air and Liquid Cooling Becomes the Preferred Architecture in Uncertain Service Requirements Scenarios

The continuous development of AI technology gives birth to the demand for high-performance and high-density servers, and the liquid cooling technology is required to ensure stable hardware operation. As we shift from general-purpose computing to intelligence computing, projects vary in terms of the proportion of the two. Therefore, data center infrastructure must be flexible and can adapt to service development in the future.

The rapid explosion of intelligent computing will bring great uncertainty to data centers

As the infrastructure in the information age, data centers are seeing constantly changing computing demands as services grow. This is inevitably followed by requests for infrastructure evolvement. we stand in the transition from general-purpose computing to intelligent computing. With the rapid development of generative AI technologies, the demand for intelligent computing will register an explosive growth, with a compound annual growth rate of 80%, far exceeding the average computing growth of data centers. This presents huge opportunities for data centers and brings great uncertainty to the service requirements of data centers.

In mainstream data centers, general-purpose servers are widely used. The power density of a single rack does not exceed 15 kW, and air-cooled devices will suffice to ensure stable operation. In contrast, the intelligent computing involves tons of deduction algorithms, the built-in intelligent computing chips require a high power density (≥ 30 kW/rack), and liquid cooling technology is usually used in this scenario. At the early stage of data center construction, it is difficult for users to accurately predict what share general-purpose and intelligent computing will take up respectively and how they will evolve in the future. That's why we need to design a solution based on the existing computing while not ignoring the need for intelligent computing growth during construction. It is imperative to develop a data center architecture that supports future evolution.

Future Evolution: The Convergence of Air and Liquid Cooling Will Become a Preferred Architecture in Scenarios with Uncertain Service Requirements

With the introduction of intelligent computing, there will be a mixture of medium and low power density (≤ 15 kW/rack) and high power density (≥ 30 kW/rack) in one data center, making it much harder to plan and construct the cooling system. This means users need to satisfy current service requirements while taking into account of service evolution.

In this context, the air-liquid convergence architecture will serve future data centers better. The core idea is to dynamically allocate the cooling capacity of the data center based on different characteristics of air cooling and liquid cooling. Starting from the cooling sources, a system is installed with air cooling and liquid cooling solutions. The design of air channels and ducts makes it possible to select a proper cooling mode for severs according to power density and service characteristics.

The key is the capability to adjust the ratio of air cooling to liquid cooling, or to dynamically adjust the cooling capacity between the two based on actual requirements with a given total cooling capacity of the data center, thus achieving the optimal cooling effect. For example, when there is an increase in the demand for intelligent computing, the proportion of air cooling can be reduced and that of liquid cooling can be raised, and vice versa.

What distinguish the air-liquid convergence is that the architecture can adapt to changes in requirements of data centers and make data centers more efficient and flexible. To summarize, this converged architecture has the following advantages:

Energy conservation: It can dynamically adjust the ratio of air cooling to liquid cooling based on the actual requirements of data centers, thus maximizing cooling efficiency. Compared with a single cooling mode, this mixed mode can cut energy consumption and OPEX of data centers.

Adaptability: It can adapt to changes in requirements of data centers and find a suitable cooling mode for both general-purpose computing and intelligent computing. Compared with air cooling or liquid cooling, to converge these two modes can enhance the adaptability of data centers and avoid over-or under-design.

Future evolution: It can flexibly allocate the cooling capacity for air cooling or liquid cooling as data centers evolve. After the two cooling modes are integrated, the evolvability of data centers is improved, providing support for future-oriented evolution.

8 Indirect evaporative cooling is still the best refrigeration scheme now and in the future

As AI draws wide attention in the industry, data center infrastructure faces new challenges and requirements. For example, the liquid cooling technology gains popularity nowadays. Does it mean that this technology will be used in most or all of the data centers in the following years? At what pace will it evolve in the future? What changes will indirect evaporative cooling experience?

Despite the need for high-density intelligent computing data centers, the market will concentrate on providing medium- and low-density data centers for general-purpose computing in the following three years.

The rapid development of AI technologies leads to the boom of intelligent data centers, posing higher requirements on power density of data centers. Although the number of high-density data centers continue to increase, cloud data centers will dominate the market in the short term in terms of the number of data centers under construction and growth. It is estimated that in the next three years, more than 90% of new data centers will still be medium- and low-density cloud data centers, with power density of a single cabinet not exceeding 15 kW, which mainly use the air cooling solution. This shows that traditional cloud data centers will remain in a dominant position in the short term despite the rise of high-density data centers.

The AHU performs one-time heat exchange to maximize the use of natural cooling sources, lowering PUE and enhancing cost-effectiveness.

For non-intelligent data centers, the indirect evaporative cooling solution outperforms others in meeting data center requirements.

From the perspective of architecture, the indirect evaporative cooling solution adopts a distributed architecture, which can mitigate the risk of system breakdown caused by the single-point failure compared with the centralized chilled water system, thus improving the reliability of equipment room.

As for heat exchange efficiency, the indirect evaporative cooling solution can maximize the use of natural cooling sources using the heat exchange core design for one-time heat exchange. It has advantages of saving power and water over traditional chilled water system in which heat is exchange four times. In particular, the indirect evaporative cooling solution mainly uses natural cooling sources at low temperature, eliminating the need for mechanical cooling. This can lower power usage effectiveness (PUE) and water usage effectiveness (WUE) significantly.

When it comes to delivery and maintenance, many of components of this product-based solution are prefabricated in factories, which reduces the on-site construction workload, shortens the delivery time, and ensures the quality. Besides, thanks to the simplified structure of the solution, maintenance is much less complicated, reducing the cost of routine maintenance.

Having withstood the test of market for over five years, this solution is mature in design, construction, and O&M. It boasts a better business logic as it has a 20% lower overall cost than the traditional chilled water system. Given the development trend of the data center industry, the indirect evaporative cooling solution is predicted to be the most cost-effective solution with a low PUE.

Technology application

In a 1000-rack data center in Ulanqab, a total of 368 pre-fab. modules are installed on five floors, and the indirect evaporative cooling solution is deployed on the second to fifth floors. The data center has an annual PUE of as low as 1.15, cutting the annual electricity cost by 12.2%.

9 To Further Reduce PUE, the Optimal Solution is to Shift the Focus on Efficient Components to System Engineering Optimization

The emerging of the AI foundation model technology drives data centers to enter the era of intelligent computing. On one hand, computing power surges, and data center energy consumption increases continuously. On the other hand, the carbon peak and neutrality goals pose higher requirements on data center energy consumption. Relying solely on the selection of efficient products and components is not enough to mitigate the increase in high energy consumption. To reduce PUE of data centers, we need to change our mindset from improving component efficiency to optimizing system engineering.

The component efficiency is approaching the bottleneck. The time and cost of slight improvements are far beyond the computing power demand

With the great progress in the development and application of foundation models represented by ChatGPT, the demand for computing power, specially intelligent computing power, rises prominently. According to the *White Paper on Intelligent Computing Development (2023)* released by China Academy of Information and Communications Technology, the growth rate of global intelligent computing power in 2022 was 25.7%, while China, in particular, saw a 41.4% increase. It is estimated that the global computing power will increase by more than 50% in the next five years. The essence of a data center is to convert electricity into computing power. On one hand, the sharp increase in computing power results in a surge in energy consumption. On the other hand, to achieve the carbon peak and neutrality goals, the requirements for green and low-carbon data center development are increasing, and the PUE regulation is becoming stricter. The cooling and power supply systems account for more than 40% of the energy consumption of a data center. In traditional mode, data centers adopt efficient devices to improve the efficiency of components, such as efficient chillers, air conditioners, and UPSs, to reduce PUE. After years of development, the dual-conversion efficiency of UPSs reaches 97%, and the coefficient of performance (COP) of chillers exceeds 8. The COP of chilled water air conditioners is close to 4. As the efficiency of a single component is approaching the bottleneck, many industry vendors choose to develop micro-innovation. However, slight improvement of component efficiency requires a large amount of R&D investment and time. The business and time costs required far exceed the return on investment (ROI) of the computing center. Therefore, to improve the energy efficiency of data centers, we need to explore a potential way to reduce PUE.

Reducing PUE requires systematical review on the actual conditions and technical levels of each component

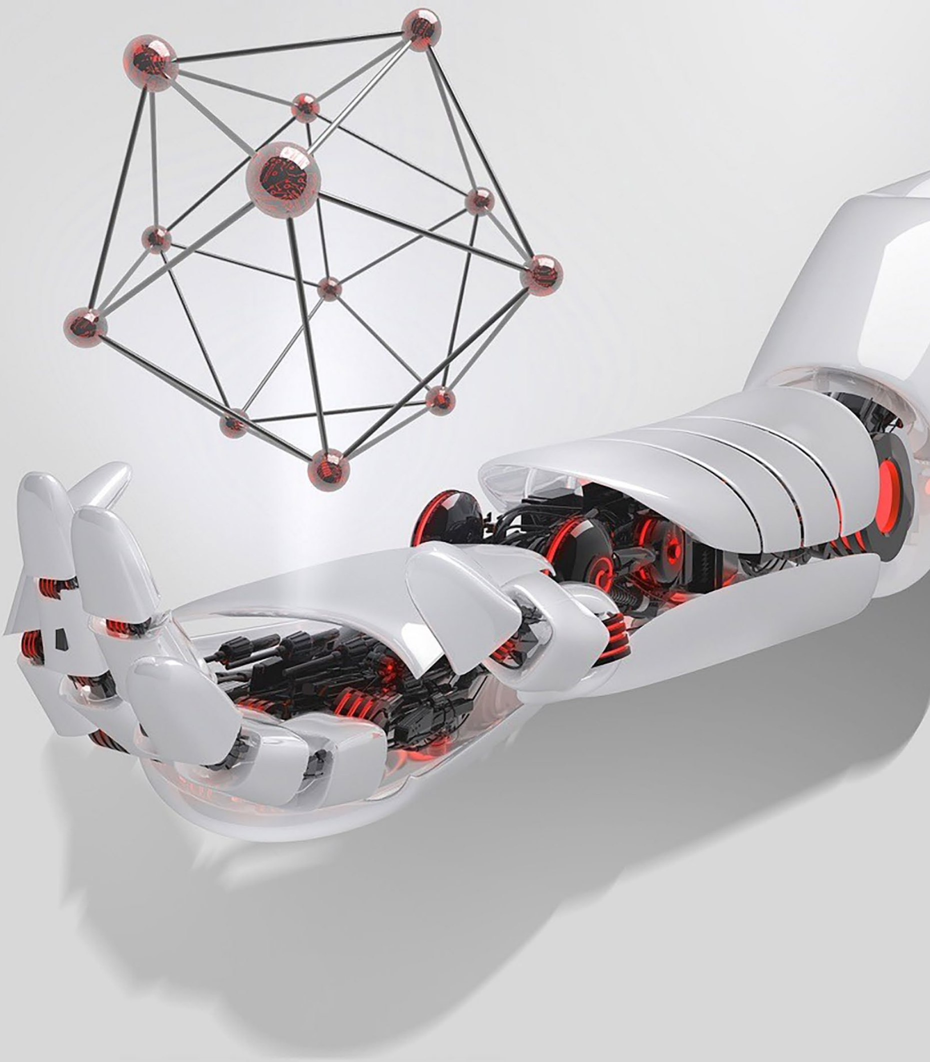
A data center is a systematic project involving multiple subsystems, such as IT, cooling, power supply, and network. There are many factors that affect the energy efficiency of a data center, such as the technical architecture, device selection, operational strategy, operating environment, IT working environment, and natural conditions. These factors affect each other. To reduce PUE, we need to comprehensively review PUE with a system engineering mindset to realize the optimal balance between the actual system conditions and the component technical level. We need to change the focus from component efficiency to link efficiency, upgrade the operating mode, and improve system efficiency. For example, the 2N power supply system adopts one mains supply and the S-ECO mode to ensure reliability and improve power supply efficiency. We need to focus on operating environment instead of component efficiency, and increase the supply air temperature and the difference between the supply and return air temperatures within the permitted range of servers, which can reduce mechanical cooling, increase the proportion of natural cooling sources, and reduce the energy consumption of the cooling system. In addition, the wider operating temperature range helps improve the deployment density and load rate of IT servers, achieving optimal computing power under the same energy consumption. Moreover, the AI optimization technology can be used to optimize the operating configuration of each system, achieving a comprehensive balance between computing power and energy consumption, and upgrading from optimal PUE to optimal petaflops PUE (PFPUE).

Technology application

Guangzhou Unicom IDC, with a total area of 199,100 square meters, adopts the system engineering approach to improve overall energy efficiency. A high-temperature chilled water system is used for cooling, which increases the water inlet temperature from 12°C to 18°C, greatly improving the cooling efficiency. In addition, the supply air temperature is set to 24°C–25°C and the return air temperature is set to 36°C to improve the cooling and IT system efficiency. The power supply system adopts the S-ECO mode. In this mode, the electric energy conversion efficiency reaches 99.1%, which is more than 3% higher than the traditional mode. The PUE test result of the data center is 1.298 under 30% full design load, which means that the overall energy consumption is reduced by more than 20%.

10 AI Optimization Will Become the Optimal Choice for Intelligent Optimization of Energy Efficiency for Existing Data Centers

The national carbon peak and neutrality goals impose stricter requirements on data center energy consumption, which makes energy-saving modernization of existing data centers a top priority. However, the traditional hardware transformation faces many difficulties and challenges. It is worth noting that with the rapid evolution of new AI technologies, simple hardware modernization and AI software optimization are expected to become a large-scale application trend of energy-saving modernization in data centers. This trend will provide a more feasible way to data center energy saving and is expected to become the mainstream choice in the future.



Data centers are large power consumers, and emission reduction is in urgent need

Data centers play a crucial role in the process of informatization and digitalization, and are the strong support for cloud computing, 5G, and AI. According to the annual data of 2022, the power consumption of data centers nationwide has reached an astounding number of 270 billion kWh, accounting for about 3% of the total power consumption of the society, an increase of 25% compared to 2021 (216.6 billion kWh). With the acceleration of Internet digitalization, it is estimated that by 2025, the proportion of data center power consumption in the whole society will increase to 5%. By 2030, the power consumption of data centers nationwide is expected to be close to 400 billion kWh. As can be seen, reducing emissions in data centers is in urgent need.

"Carbon peak and neutrality" policy, stricter PUE supervision, and great difficulties in traditional data center modernization

By the end of 2022, the number of racks in data centers nationwide has reached 6.5 million, of which more than 50%, that is, data centers of more than 3 million racks, have a PUE higher than 1.5. Since 2021, newly constructed projects of large and ultra-large data centers have been regulated: PUE cannot be higher than 1.3. In 2022, the integrated large data center construction of the "east-to-west computing resource transfer project" requires that PUE of data centers in a cluster be lower than 1.25 in the eastern regions and lower than 1.2 in the western regions. Data centers in the advanced demonstration projects need to reduce PUE to 1.15. In the same year, China's national mandatory standard GB 40879 *Maximum allowable values of energy efficiency and energy efficiency grades for data centers* was officially released, indicating that future supervision and management will be based on this mandatory standard and PUE supervision will be stricter. In addition to releasing guiding policies on data center energy efficiency, departments such as Commission of Development and Reform and Ministry of Industry and Information Technology of some major energy consumption provinces have formulated more punitive rules, such as differential electricity prices, dismissal of the unqualified, and online energy consumption monitoring. Data centers that do not meet the PUE requirements may not only face high electricity costs, but also be closed for rectification.

Energy-saving modernization of traditional data centers mainly involves adding, deleting, and modifying existing old devices, such as replacing constant-frequency devices with variable-frequency devices, replacing low-efficiency devices with high-efficiency devices, adding flow meters, and changing channels. These modernizations need to stop services in the data centers, which greatly affect and cause loss to services. To achieve the modernization goal, the cooling field should be the main focus. The industry often uses better cooling devices, such as in-row air conditioners for local cooling, indirect evaporative cooling, high-temperature chilled water fan walls, and refrigerant air conditioners. However, there are always bottlenecks existing in the advancement of energy-saving technologies concerning hardware alone, and more innovative solutions are needed.

AI intelligent optimization becomes the best solution for energy-saving soft modernization of data centers

It is common in the industry to manually optimize the software of the cooling system. However, manual optimization relies heavily on expert experience. The cooling system is complex, and there are many devices and parameters. It is difficult to manually select the optimal combination. Second, manual optimization cannot be performed in real time based on environment parameters and load rates. In addition, manual optimization is usually performed at the component level or partial system level. The actual cooling requirement changes caused by IT load changes are not considered. Therefore, linked optimization of the overall cooling system in data centers cannot be realized. Hence, manual optimization has limited energy saving effects and is highly dependent on manual experience, which cannot be replicated.

With the rapid development of AI technologies, AI energy saving has been widely used in data centers. From the first-generation white-box algorithm, the second-generation data-driven AI black-box algorithm, to the third-generation knowledge + AI growth algorithm, the collaborative learning architecture is used. AI models support transfer learning and can be preset at target sites. For example, HVAC architecture and HVAC device parameters are shared to make up for insufficient data and reduce the number of sensors. In addition, the AutoML capability is supported to keep models updated and ensure optimal model training parameters. The application of these new AI energy-saving technologies facilitates quick delivery of existing data center modernization, performs continuous intelligent optimization without service interruption, and improves the overall O&M and optimization level onsite.

Technology application

The Jinqiao data center of Shanghai Stock Exchange involves multiple types of devices, such as large and small chillers and constant-frequency/variable-frequency devices, and multiple working modes. The system is complex and modeling is difficult. Huawei iCooling energy efficiency optimization is innovatively introduced to the data center. The Jinqiao data center is the first large-scale data center in the financial industry in China that implements AI optimization for dual-cooling systems. It effectively reduces PUE by more than 10% and achieves intelligent cooling.

Learn More

Disclaimer

This document may contain predictive information, including but not limited to information about future finance, operations, product series, and new technologies. 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 purposes only, and constitutes neither an offer nor a commitment. Huawei may change the information at any time without notice.

Huawei Digital Power Technologies Co., Ltd.

Huawei Digital Power AntoHill Headquarters,

Xiangmihu Futian, Shenzhen 518043, P. R. China

digitalpower.huawei.com