Journal of World Architecture, 2023, Volume 7, Issue 6
https://ojs.bbwpublisher.com/index.php/JWA
Online ISSN: 2208-3499
Print ISSN: 2208-3480

A Brief Introduction to Infrastructure Planning for Next-Generation Smart Computing Data Centers

Yun Zhou*

Guangdong Telecom Planning and Design Institute Co., Ltd., Guangzhou 510630, China

***Corresponding author:** Yun Zhou, 13826266337@139.com

Abstract: Globally, digital technology and the digital economy have propelled technological revolution and industrial change, and it has become one of the main grounds of international industrial competition. It was estimated that the scale of China's digital economy would reach 50 trillion yuan in 2022, accounting for more than 40% of GDP, presenting great market potential and room for the growth of the digital economy. With the rapid development of the digital economy, the state attaches great importance to the construction of digital infrastructure and has introduced a series of policies to promote the systematic development and large-scale deployment of digital infrastructure. In 2022 the Chinese government planned to build 8 arithmetic hubs and 10 national data center clusters nationwide. To proactively address the future demand for AI across various scenarios, there is a need for a well-structured computing power infrastructure. The data center, serving as the pivotal hub for computing power, has evolved from the conventional cloud center to a more intelligent computing center, allowing for a diversified convergence of computing power supply. Besides, the data center accommodates a diverse array of arithmetic business forms from customers, reflecting the multi-industry developmental trend. The arithmetic service platform is consistently broadening its scope, with ongoing optimization and innovation in the design scheme of machine room processes. The widespread application of submerged phase-change liquid cooling technology and cold plate cooling technology introduces a series of new challenges to the construction of digital infrastructure. This paper delves into the design objectives, industry considerations, layout, and other dimensions of a smart computing center and proposes a new-generation data center solution that is "flexible, resilient, green, and low-carbon."

Keywords: Smart computing data centers; AI; Dual carbon goals

Online publication: December 26, 2023

1. Data center arithmetic demands and industry changes

In October 2023, the Ministry of Industry and Information Technology, along with six other departments, released a notice outlining the Action Plan for the High-Quality Development of Arithmetic Infrastructure. The plan introduces the concept of "smart computing centers," which are facilities leveraging large-scale heterogeneous arithmetic resources, encompassing general-purpose arithmetic (CPU) and intelligent arithmetic

(Graphics Processing Units [GPU], Field Programmable Gate Arrays [FPGA], Application Specific Integrated Circuits [ASIC], etc.). These centers primarily supply the necessary computing power, data, and algorithms for artificial intelligence applications, such as AI deep learning model development, model training, and model inference. Intelligent computing centers cover facilities, hardware, and software, and can provide full-stack capabilities from bottom-layer computing power to top-layer application enablement [1].

With the continuous development of the digital economy, generative AI, and other new business outbreaks, a new generation of AI applications represented by ChatGPT exploded globally, which drove the demand for smart computing centers. Traditional data centers, considered the key providers of computational power, are undergoing a shift from the "cloud network era" to the "intelligent computing era" to better address the evolving needs of the AI industry.

2.1. Digital economy brings high-speed growth in demand for arithmetic power

On a global scale, computational power serves as the fundamental and crucial foundation for the digital economy. Recognizing its significance, many countries, including China, have embraced the development of this industry as a core component of their national strategy. The Chinese government, in particular, places significant emphasis on advancing computational power and has implemented a range of policies and measures to reinforce investment and support within the computational power sector [2].

In January 2023, the National Information Center and relevant departments jointly issued the "Guide for the Innovation and Development of Intelligent Computing Centers." This guide states that smart computing centers will become a key information infrastructure to support and lead the development of the digital economy, the smart industry, the smart city, and the smart society. It is estimated that the scale of China's smart arithmetic will continue to grow exponentially, and more than 30 cities across the country are currently building or proposing to build smart computing centers. In the next five years, the compound annual growth rate of China's intelligent arithmetic scale is expected to exceed 50%, and the scale of China's artificial intelligence core industry will exceed 400 billion yuan, increasing the scale of related industries to exceed 5 trillion yuan.

The rapid expansion of the digital economy, coupled with the rise of generative AI exemplified by ChatGPT and the emergence of new sectors like the metaverse, anticipates a surge in smart computing application scenarios over the next 3–5 years. This growth is poised to demand substantial computational power. China's smart arithmetic scale is expected to exceed 300 EFLOPS by 2025, and the compound annual growth rate of smart arithmetic scale will be as high as 33.9%, while the compound growth rate of generalized arithmetic scale in the same period is 18.5%. According to China's comprehensive arithmetic index, the proportion of intelligent arithmetic in the country will rise from 25.4% this year to 85% by 2025.

International Data Corporation expects the global market size for AI servers to grow from $19.5 billion in 2022 to $34.7 billion in 2026, with the market size for servers used to run generative AI growing from 11.9% of the overall AI server market in 2023 to 31.7% in 2026. This means that an average of 1 in 3 AI servers will be used for generative AI services in the next few years. The liquid-cooled server market in China is expected to grow at a compound annual growth rate of 54.7% from 2022–2027, and reach a market size of $8.9 billion in 2027.

Based on the analysis above, it can be concluded that the skyrocketing demand for intelligent computing will make the new generation of smart computing data centers, primarily featuring liquid-cooled servers, a central focus in the future digital infrastructure layout and construction.

2.2. Analysis of data center customer groups

Over the past two decades, data centers have evolved from traditional, small to medium-sized, energy-intensive, self-constructed, and self-maintained server rooms to a model dominated by large and ultra-large data centers or

data center clusters. This model is complemented by edge server rooms, emphasizing green energy efficiency, low-carbon environmental protection, rapid deployment, and customized requirements. The industry's main players in development and operation now include telecom carriers, internet enterprises, private enterprises, and more.

Over the past two decades, data centers have evolved from traditional, small to medium-sized, energy-intensive, self-constructed, and self-maintained server rooms to a model dominated by large and ultra-large data centers or data center clusters. This model is complemented by edge server rooms, emphasizing green energy efficiency, low-carbon environmental protection, rapid deployment, and customized requirements. The industry's main players in development and operation now include telecom carriers, Internet enterprises, private enterprises, and more. The three major telecom operators, serving as the primary drivers of domestic digital infrastructure construction, control half of the eight computing power hubs and ten data center clusters planned and constructed by the state. Apart from constructing server rooms for their internal use, they also lease a significant number of server rooms to external entities or operate them in collaboration with others. Regular customers mainly fall into the following categories:

(1) Internet customers: Internet customers mainly comprise domestic Internet companies. These companies typically have corporate standards for their server rooms, with well-defined technical regulations governing data center layout design, electromechanical support, and equipment specifications. The predominant demand for cabinets is generally for high-power cabinets, and some customers have specific requirements for liquid-cooling and high customization capacity.

(2) Governmental customers: Governmental customers mainly include education, public security, and other departments. Their demand for cabinets primarily involves low and medium-power cabinets, with a greater emphasis on data center security, ease of operation, and maintenance. They also require a high level of customer service for the data center.

(3) Financial customers: Financial customers refer to banks, insurance, securities, and other financial institutions. Their demand for cabinets is primarily for medium-power cabinets, with high requirements for network latency, network security, and server room-level certification.

(4) Enterprise customers: Enterprise customers have diversified demands, both high volume and low volume, primarily requiring medium and low-power cabinets. They are more sensitive to price and have relatively lower requirements for server room level and network delay.

There are also stark differences in data-center-bearing service functions depending on the customer group and arithmetic requirements. In terms of digital infrastructure construction, data centers can be mainly divided into four types, such as cloud data centers, supercomputing centers, smart computing centers, and City Brains, etc. In order to reduce costs and improve operational efficiency, the aforementioned types of infrastructures are often placed in a single arithmetic center in some cities.

(1) Cloud data center: Most existing data centers are cloud data centers, catering to a diverse range of service targets and providing a wide array of services. They primarily handle tasks related to cloud computing, big data, and other requirements, addressing the challenges of handling massive data or application loads. These centers offer general arithmetic services to support the digital transformation of various industries.

(2) Supercomputing center: Supercomputing centers primarily cater to the demands of high-performance computing and massively parallel computing. They offer potent computing resources for enterprises and scientific research institutions, supporting major projects or collaborative research and development initiatives.

(3) Smart computing center: Smart computing centers are composed of AI chips, large-scale storage systems, high-performance computing units, and other infrastructure. Their primary focus is on research in artificial intelligence, machine learning, and related fields. They typically adopt distributed computing to process large volumes of data for efficient data storage and management.

(4) City Brain: City Brains are generally invested in, constructed, and operated by the government. They encompass a digital system and modern urban infrastructure that includes elements such as systems, platforms, application scenarios, etc. This framework is based on and supported by data, arithmetic power, algorithms, and other components, applying new technologies like big data, cloud computing, blockchain, etc. The goal is to comprehensively achieve the modernization of the urban governance system and governance capacity.

Arithmetic has emerged as a crucial driver for promoting the high-quality development of the digital economy. The infrastructure of data centers is transitioning from a general arithmetic-based pattern to a model that includes general arithmetic and smart arithmetic. This shift has attracted various entities such as carriers and Internet companies into competition. Supercomputing centers, being a national strategic asset invested in by the state, have elevated the design of data centers into the era of smart arithmetic centers from the traditional Internet Data Center server room.

2.3. Difficulties and challenges in smart computing data design

Shifting from cloud data centers to smart computing data centers brings changes in process design due to differences in infrastructure composition. This is because the layout of cloud data center layout focuses on the cabinet rate, scalability, flexibility, and reliability of the server room; while smart computing data centers focus on AI-specific computing needs and the safe configuration of the server room's power supply and air-conditioning system.

Smart computing centers utilize liquid-cooled intelligent arithmetic equipment with a single cabinet power exceeding 35 kW, which results in a power density increase of over 20 times compared to general-purpose equipment in standard server rooms. Therefore, this necessitates innovative cooling methods such as using two-phase immersion cooling or cold plates, thereby significantly impacting the electromechanical process scheme of traditional data centers.

While considering the operational efficiency and cost-effectiveness of the server room, smart computing centers also need to accommodate the needs of certain customer groups requiring 4–8 kW low-power cabinet layouts. This diversity in power supply solutions and cooling mode requirements necessitates a flexible process planning program for smart computing centers. Acknowledging the limitations of the commonly used modular layout approach in cloud data center designs, the smart computing center design emphasizes high flexibility, a wide range, and full compatibility to address these challenges.

3. New-generation smart computing center design solutions

To address the challenges posed by high computing power scale, high power density, increased elasticity demand, and rapid deployment requirements for smart computing centers, while also accommodating the low-density cabinet arrangement needs of regular customers, we adopt the construction concept of "unit module, dynamic expansion, extreme energy efficiency, and intelligent operation." This approach aims to resolve the limitations of the deterministic construction mode observed in traditional data centers, making it more adaptable to the uncertainties of the market.

3.1. Applying the concept of "green, low-carbon, open and integrated construction"

The new generation of smart computing centers should embody the advantages and characteristics of being "green, low carbon, open, flexible, and integrated." The aspect of "green and low carbon" is demonstrated through high-standard construction and low-energy operation. This involves employing advanced equipment at the front end, such as distributed power supply, high-efficiency modularized integrated cooling stations, or high-efficiency cooling technology products. The front end incorporates advanced equipment such as distributed power supply, high-efficiency modular integrated cooling stations, or high-efficiency cooling technology products. The back end integrates submerged two-phase immersion cooling or cold plates to achieve a 4A level in the "Low Carbon Assessment of Data Centers." The lowest power usage effectiveness (PUE) can be below 1.10.

The concept of being "open, flexible, and integrated" is manifested in the adaptable mechanical and electrical equipment layout of the server room. The spatial arrangement can be adjusted flexibly based on the requirements of mechanical and electrical equipment with expandability. This adaptability caters to the diverse needs of multiple industries, enabling resource pooling for elastic power supply, refrigeration, and flexible scheduling. The expansive and modular design of the space accommodates rapid deployment options like container rooms.

3.2. Smart computing data module unit program

The market analysis reveals three distinct business forms for intelligent computing centers. Firstly, there is the pure artificial intelligence machine room, exemplified by models like ChatGPT, with cabinet power typically exceeding 35 kW and employing full liquid-cooled setups. Secondly, the cloud-intelligence fusion centers target public cloud and major Internet enterprises, featuring cabinet power ranging from 8–15 kW or exceeding 35 kW. These centers utilize liquid-cooled cooling methods, with supplementary low-density rooms employing air-cooled or water-cooled systems, as seen in Telecom Tianyi Cloud and similar platforms. Thirdly, there is the general-purpose smart computing server room, catering to private cloud, government, enterprise departments, small and medium-sized digital customers, etc. Cabinet power in this category usually ranges from 4–8 kW, 8–15 kW, or 35 kW or more, adopting air-liquid air conditioning with air-cooling as the primary cooling method. In practice, some owners often integrate these infrastructure types into a single computing center to attract customers, reduce production costs, and improve operational efficiency.

The new-generation smart computing center is designed with the objective of super-pooling and open integration of resources. It adopts a flexible expansion plan of modular units, accommodating N modular units in one machine room. Different cooling processes such as full air-cooling, air-liquid cooling, and full liquid-cooling can be employed, and layouts can be dynamically adjusted and flexibly transformed, achieving a minimum PUE of less than 1.10. Additionally, flexible capacity expansion is achieved through Lego-style stacking, enabling one computing center to accommodate multiple modularized server rooms. Multiple server rooms can form computer center clusters, meeting various scale demands efficiently.

In order to maintain the orderly stacking of modular units and efficiently utilize pooled resources, it is important to implement artificial intelligence and digital operation and maintenance management in the smart computing center.

3.3. Key points for implementing the modular unit program

The modular unit program is designed with the maximization of resource pooling in mind. It allows for flexible and rapid deployment, utilizing adaptable power supply systems and compatible cooling systems to meet the diverse needs of customers.

(1) Resource pooling: Water, electricity, air conditioning, arithmetic, and other resources required for data center operation are integrated to form a unified resource pool for dynamic allocation and management.
(2) Power supply energy: Intelligent mini-bus and other high-efficiency product technologies are adopted to reduce energy consumption in the server room and realize flexible scheduling of power resources.
(3) Refrigeration technology: The implementation of modular liquid cooling technology and inter-column air-conditioning enables plug-and-play server room deployment, ladder cooling, and efficient energy utilization.

Additionally, by comprehensively optimizing the airflow organization within the server room, employing closed channels, and utilizing CFD simulation, efficient cooling is achieved, aiming to reduce the PUE to below 1.10. In summary, the smart computing center's construction goal of being "green, low-carbon, open, flexible, and integrated" is realized through the optimization of the server room layout and the implementation of the modular unit scheme.

4. New-generation smart computing centers help achieve national dual-carbon goals

Driven by the national "dual-carbon" strategic goal, China's data centers are accelerating their green and low-carbon development. By 2025, the average PUE of newly built large-scale and ultra-large-scale data centers will be reduced to below 1.3, and that of national hub nodes will be reduced to below 1.25, and their green and low-carbon grade will be 4A or above. Smart computing centers are particularly power-consuming due to the large number of high-density, high-power cabinets, coupled with the surging demand for computing power, which will lead to the increasing construction of smart computing centers. The increasing constraints on the country's energy resources and carbon emissions have become a central challenge in the development of the smart computing industry. Consequently, the establishment and operation of environmentally friendly next-generation computing centers are crucial in ensuring the high-quality development of the digital economy and achieving China's strategy of "carbon peak and carbon neutrality" by 2060.

Building a new-generation green computing center includes four main directions:

Firstly, the smart computing center should optimize land site layouts, prioritizing construction within the country's eight computing hubs and 10 data center clusters. This strategy leverages intensification, incorporates the coordination of source, network, load, and storage, and substantially boosts the utilization of renewable energy sources like wind power and photovoltaic power in Internet access.

Secondly, the process design of the smart computing center should constantly be optimized, predominantly embracing natural cooling methods alongside promoting liquid cooling solutions. Notably, water-saving refrigeration technologies like indirect evaporative cooling, recognized for efficiently utilizing natural cold sources and reducing energy consumption, have emerged as the industry's mainstream choice.

Thirdly, the integration of advanced green technology products should be prioritized to boost power utilization efficiency. This involves advocating for the adoption of cutting-edge cooling systems like two-phase immersion cooling in the air-conditioning system of smart computing centers, emphasizing both high efficiency and flexibility.

Fourth, the operation and management of smart computer centers should be digitalized. The complexity of operation and maintenance management rises significantly as the scale of the data center increases. Therefore, it is imperative to enhance the data center infrastructure. With the widespread adoption of high-definition video and image recognition technologies, artificial intelligence is finding broader applications in data centers. These digitalization and artificial intelligence technologies play a crucial role in transforming the focus of data center

operation and maintenance management from energy consumption to carbon emissions, contributing to the overarching goal of achieving carbon neutrality.

5. Conclusion

To meet market demands and adhere to the principles of "green, low-carbon, open, flexible, and integrated," we have innovated beyond traditional construction methods to establish a new generation of smart computing centers. Leveraging the modularity and prefabrication of modular units, we achieve optimal resource utilization, flexible power supply system scheduling, and adaptable cooling system compatibility. This approach ensures agile responsiveness to customers' diverse needs, delivering enhanced green energy efficiency and operational benefits for the smart computing center.

Disclosure statement

The author declares no conflict of interest.

References

[1] Zhong J, Fu L, Ding L, et al., 2021, Planning and Design of New Infrastructure Data Center, Press of Electronics Industry, Beijing.

[2] Lan J, Tu J, Niu C, et al., 2021, Foundation of 5G Network Technology Planning and Design, People's Posts and Telecommunications Press, Beijing.