SYNTHRA CAPITAL

Synthra Capital

AI computing power industry dynamics research report

AI and Big Data Financial Management Fund

*For Synthra Capital investors only

CONTENT

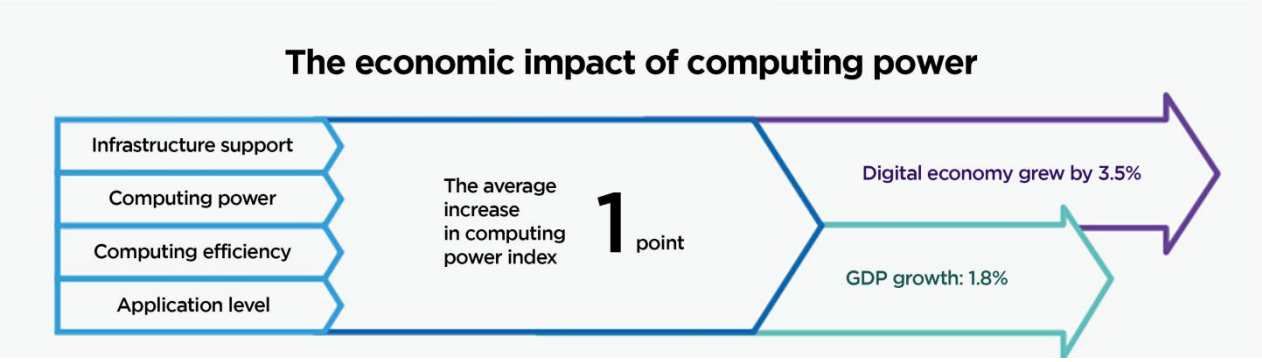
1. Computing Power - The Foundation of AI

Computing power is the foundation for the development of artificial intelligence applications and the core element of artificial intelligence. Computing power, algorithms and data are the three core elements of the development of the artificial intelligence industry. Massive amounts of data are generated every moment. The production of data is no longer a problem, but how to process, analyze and use data is a problem. Algorithms have undergone decades of development. After the emergence of deep learning and accelerated computing, they have been rapidly developed and optimized, and have begun to become a hot topic of social concern. However, among the three major elements, computing power is the most core element. Only the progress of computing power can drive the artificial intelligence industry forward and mature. Computing power is the decisive force to carry and promote artificial intelligence to practical applications.

The rapid growth of data puts higher demands on the development of computing power. With the continuous advancement of informatization and digitalization, the amount of new data generated worldwide is growing rapidly. According to IDC data, the total amount of new data in the world will reach 104.5 ZB in 2021, and it is expected that the total amount of new data in the world will reach 221.2 ZB by 2026, with a compound annual growth rate of 21.22% from 2021 to 2026. Massive data provides fertile soil for the development of artificial intelligence applications, but it requires more powerful intelligent computing power to process it. At the same time, with the emergence of new application scenarios, more and more scenarios have higher requirements for the real-time performance of data. The surge in real-time data has made edge computing capabilities increasingly important, and artificial intelligence applications have become increasingly dependent on edge computing support.

The complexity and massiveness of algorithm models require stronger computing power support. Although the total amount of data is growing rapidly, the amount of data being effectively utilized is less than 1%. How to effectively capture high-quality data and build accurate models depends on the development capabilities of AI algorithms. In recent years, the number of parameters and complexity of algorithm models have shown an exponential growth trend, especially in emerging cognitive intelligence fields such as natural language processing, which require computing power far more than traditional AI fields such as image recognition and speech recognition. The model parameter volume of the self-talk language processing large model GPT-1 launched in June 2018 was 117 million, and the model parameter volume of GPT-2 launched in February 2019 was greatly increased to 1.5 billion. The parameter volume of GPT-3 launched in May 2020 was further increased to 175 billion. The substantial increase in model parameters has also rapidly increased the demand for intelligent computing power. If we use "computing power equivalent (PFLOPS-day, PD)", that is, the total computing power consumed by a computer running for one full day at a trillion times per second, to measure the total amount of intelligent computing power required for artificial intelligence tasks.

The increase in computing power also has extremely strong economic benefits, and thus has become the focus of policy support in various countries. Through regression analysis of the computing power index, digital economy and GDP of 15 key countries in the world, it is found that for every 1 point increase in the computing power index of 15 key countries, the country's digital economy and GDP will increase by 3.5‰ and 1.8‰ respectively. This trend is expected to continue between 2021 and 2025. Further research has found that when a country's computing power index reaches 40 points and 60 points or more, the driving force for GDP growth will increase to 1.5 times and 3 times for every 1 point increase in the computing power index, and the driving effect on economic growth will become more significant. Considering the significant economic benefits of increasing computing power, especially intelligent computing power, supportive policies for the development of computing power infrastructure have become a policy focus of various countries.



The increase in computing power has strong economic benefits

2. GPU is the most valuable component in the AI hardware industry chain

2.1 Demand for AI data centers surges, and AI servers are rapidly increasing in volume

An AI data center is a facility or physical space dedicated to supporting AI computing and data processing tasks. It is specially configured and optimized based on traditional data centers to meet the needs of machine learning, deep learning, and other AI-related tasks. AI data centers usually have a large number of high-performance servers, GPU accelerators, and specialized storage systems to provide powerful computing power and accelerate deep learning. In addition, AI data centers are equipped with high-speed network equipment and optimized software

frameworks to support efficient data transmission and algorithm training. Through these specialized configurations and optimizations, AI data centers are able to provide a reliable and stable computing environment for AI workloads of various sizes and complexities, and meet the needs of large-scale data storage, backup, and analysis. AI data centers play a key role in promoting the development and application of artificial intelligence technology, and provide strong support for AI applications and services in all walks of life.

Figure: AI data centers require better computing power, storage requirements, network bandwidth, and software support than ordinary data centers

Aspect	Need
Calculate ability	High-performance computing devices (such as GPUs and specific AI chips) for large-scale parallel computing and processing of complex machine learning and deep learning algorithms
Storage requirements	High-capacity, high-speed storage for storing large data sets
Internet bandwidth	High network bandwidth, low latency network for fast data transfer and communication to meet the sensitive needs of data transmission speed
software support	Software support for machine learning and deep learning tasks, including specialized AI frameworks, libraries, and tools, as well as optimized software stacks and distributed computing platforms to improve computing efficiency and performance

The industry landscape of AI data centers presents a multi-level supply chain structure, including upstream data center equipment manufacturers and solution providers, midstream cloud service providers and large technology companies, and downstream corporate users and individual developers. Participants in these different links cooperate with each other to build a competitive and diversified ecosystem.

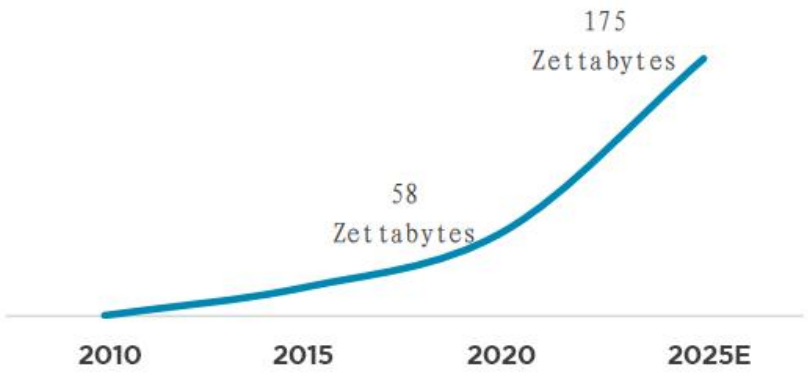
The upstream of AI data centers mainly include chip manufacturers, servers and network equipment suppliers. Currently, many Internet companies engaged in AI research and development, such as Amazon, have deployed their own AI data centers. They use their own technology and resources to provide AI solutions and support for enterprises and research institutions. Third-party IDC service providers also play an important role in the AI data center market. They provide services such as data center hosting, cloud hosting, and network bandwidth to meet the needs of enterprises and institutions for AI data center infrastructure and operation and maintenance. These service providers usually provide flexible solutions that enable enterprises to quickly deploy and expand AI applications.

The downstream user demand of AI data centers mainly comes from enterprises, research institutions and government departments from all walks of life. These users choose suitable

cloud service providers or build private data centers according to their own needs to meet their needs for AI technology. The diversity of downstream user needs makes the AI data center market fragmented.

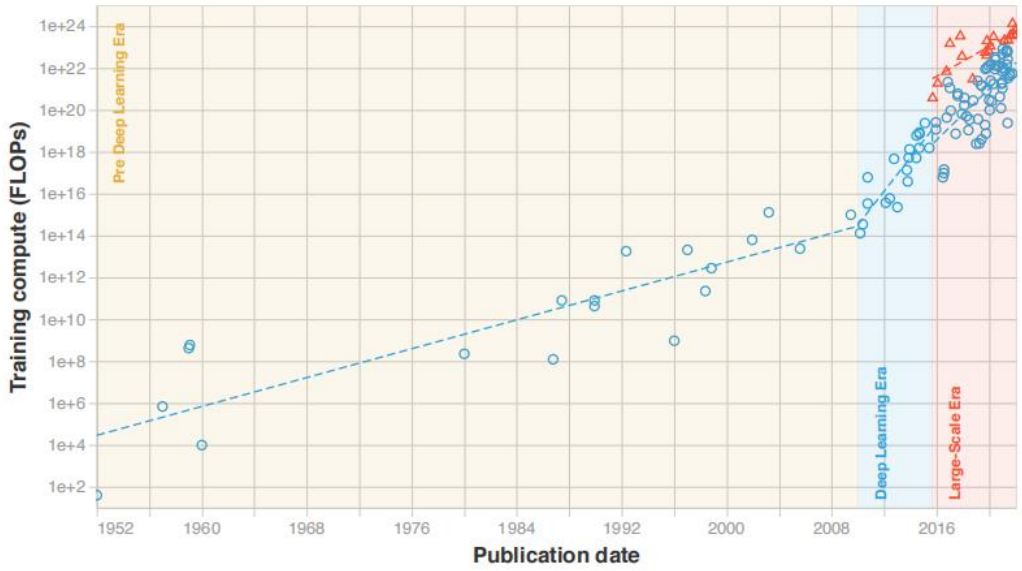
As AI industry applications continue to develop in depth, the number and scale of AI data centers will continue to increase. This trend is driven by the increase in data volume and application demand, the incremental computing power demand brought by ChatGPT, and technological innovation. As a key facility supporting large-scale data processing, deep learning model training and reasoning, AI data centers will play an important role in promoting the widespread application and further innovation of artificial intelligence technology.

The increase in data volume and the expansion of AI applications are important driving factors for the development of AI data centers. With the increase in data sources such as the Internet of Things, social media, and sensor technology, AI data centers need to process and store larger data sets. According to the IDC report, from 2014 to 2020, the average amount of data managed by each IT staff increased from 230GB to 1231GB, more than 5 times. Enterprise data is expected to grow at a rate of 42.2%. At the same time, the widespread application of artificial intelligence technology in various industries and the need to train deep learning models have also led to an increase in demand for AI data centers.



Data is growing rapidly

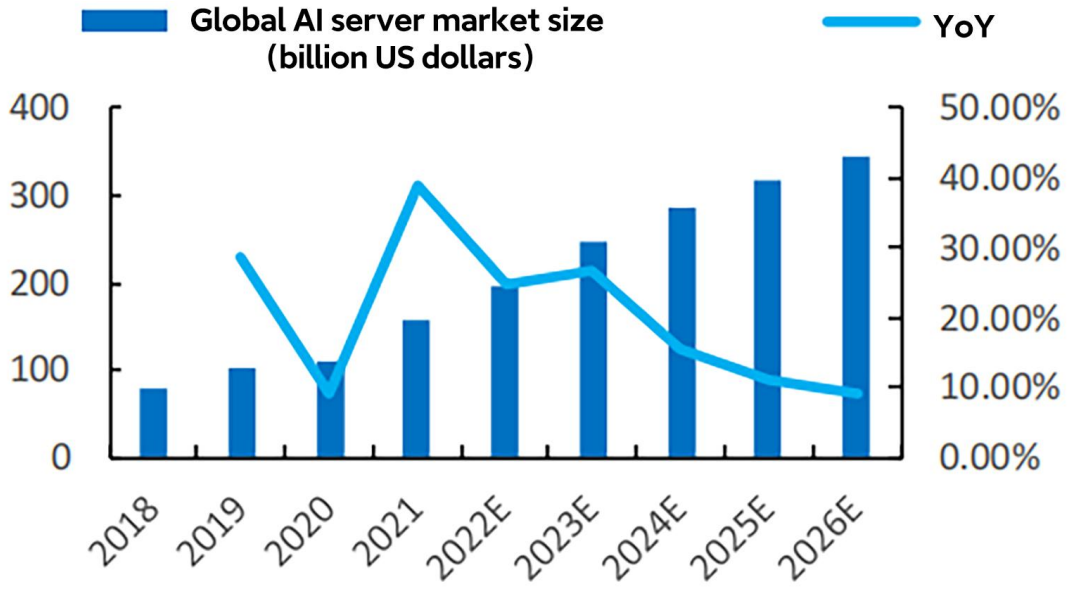
The emergence of large models has brought about an incremental demand for computing power. According to the documents released by OpenAI on the GPT-4 model, it contains 175 billion parameters and requires tens of millions of computing operations to complete an inference task. The total computing power consumption of ChatGPT is about 3640PF-days, and it requires 7 to 8 data centers with an investment scale of 3 billion and a single computing power of 500P to support its operation. Such scale and complexity require high-performance computing equipment and large-scale parallel computing capabilities, which has driven the growth of demand for AI data centers.



The era of large models doubles the demand for computing power

Technological innovation also promotes the development of AI data centers. The emergence of new processor architectures, high-speed networks, storage technologies, and more efficient cooling and energy management systems has improved the performance and efficiency of data centers, providing technical support for the development of AI data centers.

Due to common driving factors such as data volume growth, computing power requirements, technical infrastructure needs, and economic benefits, the growth trend of professional AI data centers is expected to be the same as that of large-scale data centers, and AI data centers will show a straight upward trend in the next few years.



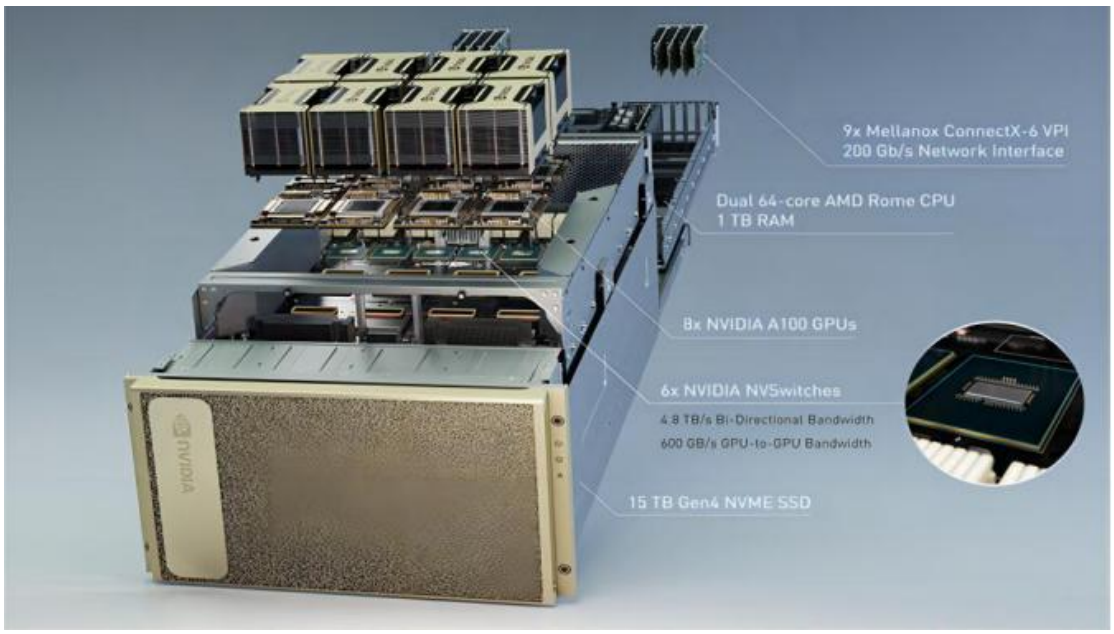
Hyperscale data center market size forecast (in billion U.S. dollars)

AI data centers mainly include servers, storage devices, network devices, cooling systems, and power supply devices. These components together form an efficient and reliable infrastructure that provides data centers with powerful computing and storage capabilities, supporting the operation of AI algorithms and data processing.

Servers are the core of data centers, used to perform complex computing tasks and run artificial intelligence algorithms. The computing power of a data center depends on the size and number of servers. Larger servers mean more processing units and computing resources, thus providing more powerful computing power. Storage devices play an important role in AI data centers, used to store and manage large data sets and models. These storage devices include hard disk arrays (RAID), network attached storage (NAS), etc., which provide support for data storage and access. The capacity and performance of storage devices determine the scale of data that a data center can process and store. Storage devices with larger capacity and higher speed can support the storage and access of more data, thus providing necessary support for the use of computing power.

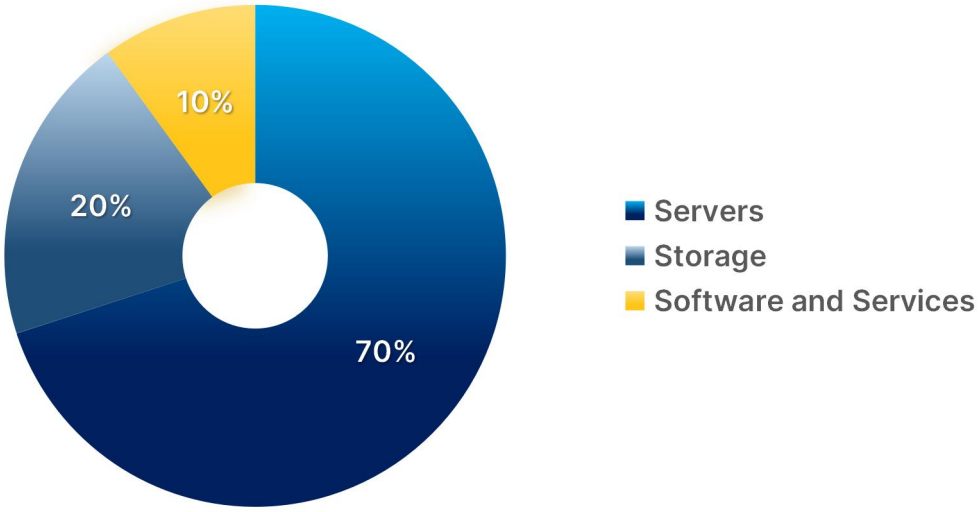
Network devices play a key role in connecting and transmitting data in data centers. AI data centers require high-speed and reliable network devices to connect servers and storage devices to achieve fast data transmission, as well as efficient communication between various components within the data center. Fast data transmission and efficient communication can improve the utilization efficiency of computing power and the speed of data processing. Since a large number of servers and computing equipment generate a lot of heat, data centers need a powerful cooling system to maintain the normal operating temperature of the equipment. The cooling system ensures that the temperature of the data center is appropriate to prevent the equipment from overheating and affecting performance and reliability. An efficient cooling system can ensure the stable operation of the data center and ensure the continuous use of computing power.

Power supply equipment is one of the infrastructures of the data center, providing a stable power supply to ensure the stable operation of the data center. Power supply equipment such as generators, UPS (uninterruptible power supply system) and power distribution systems provide a continuous and stable power supply to ensure that the data center will not be interrupted due to power problems and ensure the continuous availability of computing power.



All components work together to build a high-computing AI data center

In AI data centers, the value share of servers, storage devices, network equipment, cooling systems, and power supply equipment varies depending on factors such as the size, demand, and architectural design of the data center. Generally, servers account for a large proportion. Servers are the core part of AI data centers and are used to perform complex data processing and machine learning tasks. According to research data from the China Academy of Information and Communications Technology, their value accounts for about 70% of the total. Storage devices play a key role in AI data centers, used to store large-scale data sets, training data, and model parameters, accounting for about 15% to 30% of the total. According to data from the Prospective Research Institute, network equipment provides high-speed, reliable data transmission and connection, and its proportion is relatively low, accounting for about 15% to 20% of the total, but it is still an indispensable part. The proportion of cooling systems and power supply equipment is relatively low, accounting for 5% to 15% of the total, but they are critical to the reliability and stable operation of data centers. The specific value share varies depending on factors such as demand, industry differences, and technology.



NVIDIA data center acquisition budget composition

2.2 Among AI servers, GPUs have the greatest value

AI servers are an important part of AI data centers. AI servers are servers designed and configured specifically for artificial intelligence applications. They have powerful computing power and efficient data processing capabilities. They are key components for executing AI tasks and processing large-scale data. They provide computing resources and computing power for data centers to execute complex AI algorithms and models. There are two main architectures for AI servers: hybrid architecture and cloud-based architecture. Hybrid architecture allows data to be stored locally, while cloud-based architecture uses remote storage technology and hybrid cloud storage for data storage, that is, the joint use of local storage and cloud storage technology. AI servers usually adopt heterogeneous architectures, combining different types of processors, such as CPUs and accelerator cards (such as GPUs, TPUs, etc.) to provide higher computing performance. This heterogeneous architecture enables servers to process parallel computing and specific AI computing tasks at the same time, accelerating data processing and model training processes. As the core equipment of the data center, AI servers carry computing, storage, and network functions, providing stable support for the operation of the data center to meet the needs of tasks such as model training, reasoning, and data processing. Its scalability enables it to adapt to the growing computing and storage needs and maintain the performance and efficiency of the data center.

Computing performance is a key indicator for measuring the performance of AI servers. It depends on the performance and number of processors. For AI tasks, processors with highly parallel computing capabilities are usually used, such as central processing units (CPUs),

graphics processing units (GPUs), and tensor processing units (TPUs). Among them, floating-point operation performance (FLOPS), GPU core count and frequency, memory bandwidth, storage system performance, network performance, and latency will have an impact on the computing power of AI servers. Floating-point operation performance measures the server's ability to handle deep learning tasks, and the number and frequency of GPU cores determine the speed of parallel computing. More cores and higher frequencies can provide faster computing speeds and higher parallel processing capabilities, thereby accelerating the execution of AI algorithms and models. Memory bandwidth and storage system performance affect data read and write efficiency, network performance is related to data transmission between servers, and latency is related to real-time response requirements. These indicators can be considered together to evaluate the computing performance of AI servers, and the selection and trade-offs depend on specific application scenarios and requirements. AI applications need to process large-scale data sets and models, so storage capacity and speed are also important attributes of AI servers. AI servers need to provide sufficient storage capacity to store data and model parameters. At the same time, the read and write speed of storage devices is also crucial, which affects the data access efficiency and the speed of model training. Traditional hard disk drives (HDDs) can provide large storage capacity, while solid-state drives (SSDs) have faster read and write speeds and lower access latency, which are essential for fast data access and model loading.
Larger memory capacity and faster memory bandwidth can support efficient data processing and computing of AI servers. Memory capacity and memory modules (RAM) affect data storage and transmission capabilities. The performance and bus architecture of the memory controller determine the bandwidth and efficiency of the memory. The number and frequency of memory channels also affect the bandwidth and response speed of the memory. In addition, the use of ECC (Error-Correcting Code) memory can improve data integrity and server reliability.

AI servers need to have high-speed and stable network connections to communicate and transfer data with other servers, clients, and external data sources. The network connection speed of AI servers is affected by the performance of multiple components. The network interface card (NIC) determines the physical connection and data transfer rate between the server and the network. Different network connection media (such as Ethernet cables or optical fibers) have different transmission capabilities and bandwidths, which affect the connection speed of the server. The performance of network switches and routers plays a key role in the routing and forwarding of data packets. Their processing power and forwarding capabilities directly affect the network connection speed of the server. The optimization of network protocols and protocol stacks, as well as network topology and bandwidth management strategies, will also affect the network connection speed of the server. To ensure the stability and reliability of the server, AI servers require effective heat dissipation and cooling systems. The efficiency of the radiator, fan, and cooling technology determines the heat dissipation capacity of the server under high load conditions. Good heat dissipation and cooling design can reduce temperature and improve system performance and reliability.

The high power efficiency of AI servers can provide better performance-to-power ratio, reduce energy consumption and operating costs. The power supply unit provides stable and reliable power supply and reduces energy loss through stable voltage output and efficient power

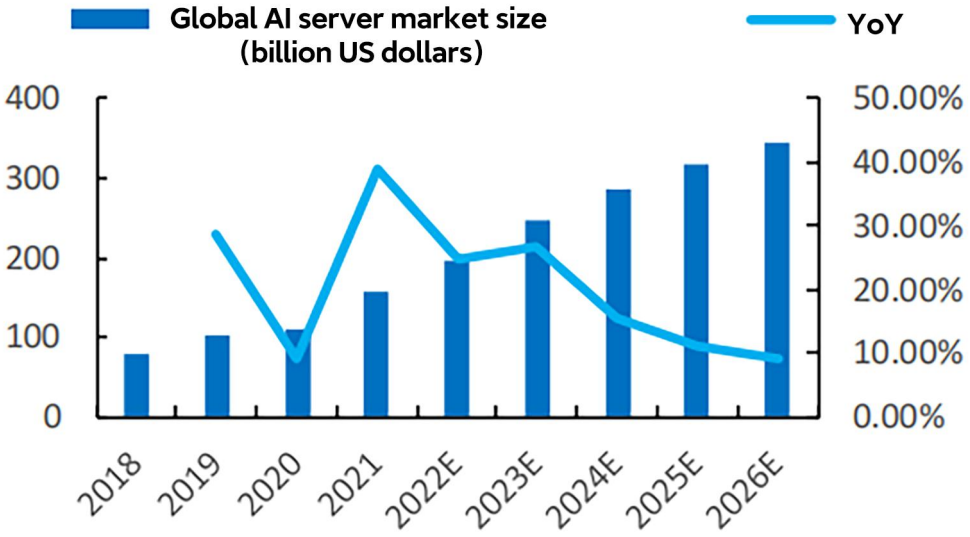
conversion. The power management module controls and monitors the operating status of the power supply, providing precise power regulation and energy management to optimize the energy efficiency performance of the server.

In training AI servers, the CPU value accounts for 9.8%, the GPU accounts for 72.8%, the memory accounts for 8.7%, and the others account for 8.7%. In inference AI servers, the CPU value accounts for 10%, the GPU accounts for 50%, the memory accounts for 10%, the storage accounts for 5%, and the others account for 25%. Taking Nvidia DGX H100, a high-performance AI system for training, reasoning and analysis, as an example, the GPU value accounts for a higher proportion, accounting for 86.66%, while the CPU value accounts for 2.31%, memory accounts for 3.49%, storage accounts for 1.54%, network interface cards account for 4.85%, and others account for 1.15%.

3. The global AI server market has broad potential

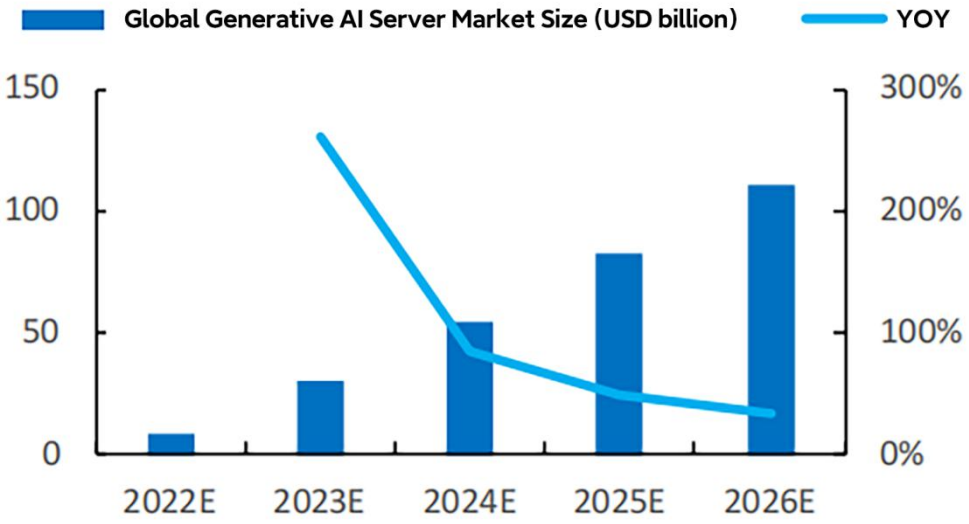
3.1 Generative AI drives continued growth in the global AI server market

As the core infrastructure for the development of the artificial intelligence industry, the global AI server market continues to grow. According to IDC statistics, the global AI server market size reached US$15.63 billion in 2021, with a year-on-year growth rate of 39.06%. Thanks to the strong growth in demand for various AI application segments, the AI server market growth rate has recovered rapidly from the impact of the 2020 COVID-19 pandemic, with a compound annual growth rate of 25.01% from 2018 to 2021, achieving rapid growth. IDC predicts that by 2026, the global AI server market size will reach US$34.71 billion, with a compound annual growth rate of 17.30% from 2021 to 2026, and will continue to maintain a relatively fast growth trend. At the same time, the proportion of the global AI server market size to the overall server market size will increase from 15.25% in 2021 to 21.69% in 2026. The growth rate of the AI server market size is higher than that of the overall server market size, becoming the core driving force for the global server industry to maintain a booming growth.



The global AI server market continues to grow

Generative AI will become a new driving force for the continued growth of the global AI server market. With the rapid iteration of large generative AI models represented by the GPT model, and the surge in the use of generative AI applications represented by ChatGPT and Microsoft 365 Copilot, the demand for AI servers from generative AI is exploding. According to IDC statistics, the new demand for AI servers from global generative AI in 2022 will be $820 million, and it is expected to reach $10.99 billion by 2026, with a compound annual growth rate of 91.34% from 2022 to 2026, while the compound annual growth rate of other types of AI servers during the same period is only 6.15%. It is expected that the proportion of demand for AI servers from generative AI will increase rapidly from 4.21% in 2022 to 31.66% in 2026, becoming a new driving force for the continued growth of the global AI server market.



The global generative AI server market is expected to grow explosively

Inference servers will gradually become the mainstream of AI servers worldwide. In the early stages of the development of generative large models, AI server demand was mainly for model training, so training servers occupied the dominant position in the market. With the rapid development of subsequent generative AI applications, AI servers will mainly meet the needs of data analysis and model output, so inference servers will gradually become the mainstream of the market. According to IDC statistics, 57.33% of the global AI server market in 2021 were training servers, but it is expected that the market size of inference servers will surpass training servers for the first time in 2024, and the market share of inference servers will reach 53.01% in 2026, and the gap with training servers will continue to widen.

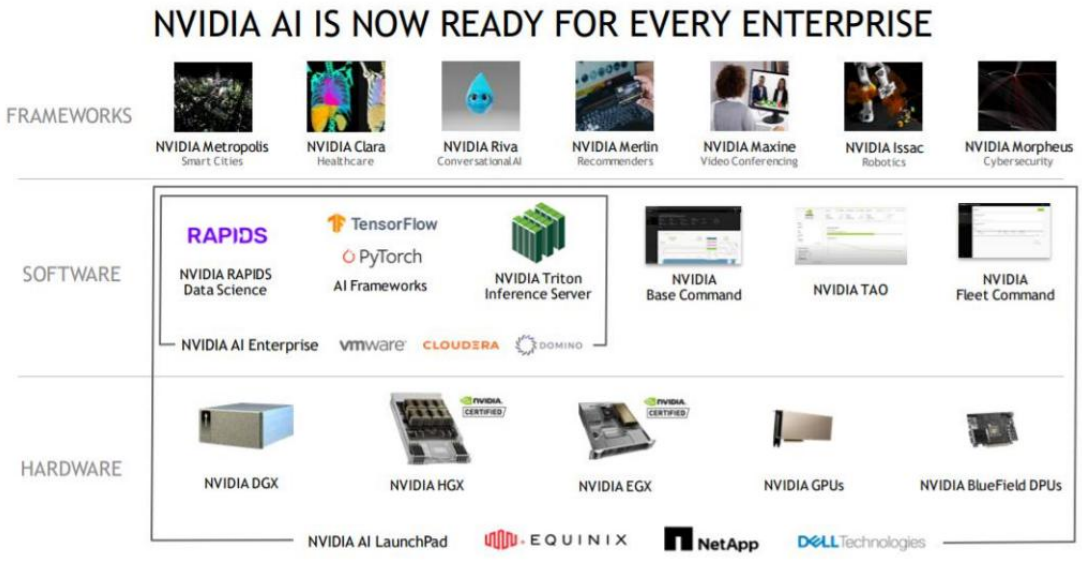
AI cloud computing and edge computing will become the fastest-growing deployment methods for AI servers worldwide. On the one hand, as the model parameters of generative AI large models are getting larger and larger, the AI computing power required for their training is growing rapidly, and traditional locally deployed computing centers are increasingly unable to meet the computing power requirements of large models, resulting in a rapid growth in demand for AI cloud computing services based on cloud deployment. On the other hand, as the growth rate of data generated by various IoT devices exceeds the growth rate of network bandwidth, the demand for direct data computing and analysis closer to the data generation is growing rapidly, which leads to faster growth of AI servers based on edge computing. According to IDC statistics, from 2021 to 2026, the annual compound growth rate of the AI server market based on local centralized deployment, cloud deployment and edge computing deployment is expected to be 8.2%, 19.8% and 26.1% respectively; it is estimated that in 2026, 56.6% of the global AI servers will be based on cloud deployment, which is the most mainstream deployment method, and 20% will be based on edge computing deployment, which is the fastest growing deployment method.

Accelerated computing AI servers are more in line with the needs of large-scale AI computing and have become the mainstream choice of AI servers. Accelerated computing AI servers refer to AI servers with one or more coprocessors, including GPGPU, FPGA or ASIC coprocessors, which are more suitable for processing deep learning AI models with increasingly larger scale and more complex algorithms, and have therefore become the mainstream choice of AI servers. The market size of non-accelerated computing AI servers that mainly use CPUs for computing will still grow, but they are mainly used for reasoning of small artificial intelligence models and some training loads. According to IDC statistics, the market size of accelerated computing AI servers in the global AI server market will be US$9.1 billion in 2021 and will grow to US$24.5 billion in 2026, with a compound annual growth rate of 22%. The compound annual growth rate of the market size of non-accelerated computing AI servers during the same period is only 9.3%.

4. Nvidia leads the global AI chip development trend

4.1 The rise of AI applications will continue to drive high growth in Nvidia's data center business

NVIDIA is the inventor of GPU. In 1999, NVIDIA was listed on NASDAQ and proposed the concept of GPU in the same year, releasing GeForce 256, which is regarded by the industry as the beginning of modern computer graphics technology. GPU was originally used mainly for PC games and console games such as Sega, Xbox and PS3. It can support T&L (Transform and Lighting) from the hardware, because T&L is an important part of 3D image rendering. Its function is to calculate the 3D position of polygons and process dynamic light effects, providing detailed 3D objects and advanced light effects. 3D image rendering scene is a parallel computing task. Since there is no connection or dependency between the regions in the image, this task can be easily decomposed into several independent tasks, each of which can be parallelized at the same time, which can also speed up the process. This parallel computing capability makes GPU unexpectedly become the hardware infrastructure for AI computing: in AI computing, the most common task is deep learning. Deep learning models usually require a lot of matrix calculations, which is the strength of GPU. GPUs can perform a large number of matrix operations at the same time, thereby accelerating the training and reasoning process of deep learning models. Due to the winner-takes-all effect in the chip industry, NVIDIA, as a global GPU giant, currently occupies the main global GPU market and has ushered in new development opportunities in the AI era.



GPUs become the infrastructure of the AI era

At present, games and data centers account for the majority of Nvidia's revenue, and the data center business has just surpassed the traditional game business to become the company's mainstay. Nvidia is targeting four major markets: games, data centers, professional visualization, and automobiles. The game business is the traditional business on which Nvidia was founded. High-performance gaming graphics cards are Nvidia's forte. It used to be the business with the highest revenue share, but it has been surpassed by the data center business in 2022. The data center business is currently the company's largest source of revenue. Cloud service providers are using graphics processing unit (GPU) technology to process massive amounts of data created by data users, including videos, photos, and messages that are ultimately stored on cloud servers. Therefore, the demand for GPUs is very strong, especially the large amount of computing power brought by AI applications, which further promotes the development of the data center business. Professional visualization products play an important role in design and manufacturing, digital content creation, and enterprise image vision, and can improve image display effects. The smart car business may become Nvidia's core business in the future, including GPUs and SoC chips sold to OEMs and suppliers, as well as corresponding development platforms.

4.2 Nvidia's AI chip architecture continues to evolve

As the development of artificial intelligence requires more and more computing power, NVIDIA is also constantly improving its chip architecture. Every one to two years, NVIDIA will propose a new chip architecture to adapt to the upgrade of computing needs. Among them, it proposed GPUDirect technology in the Kepler architecture released in 2012, which can bypass CPU/System Memory and complete direct data exchange with other GPUs on the local machine or GPUs of other machines; in addition to considering deep learning and adding DP unit in the Pascal architecture in 2016, NVLink is also used for point-to-point communication between multiple GPUs in a single machine, with a bandwidth of 160GB/s; in 2017, the Volta architecture was proposed, which is basically based on Deep Learning and introduced Tensor Core; in 2020, the Ampere architecture doubled the number of FP32 shader operations that can be executed per clock, and RT Cores provided twice the throughput for ray/triangle intersection tests. The new Tensor Core can process sparse neural networks at twice the rate of Turing Tensor Core, while Turing GPU does not support sparsity; in 2022, NVIDIA introduced the FP8 Tensor Core of the new generation of streaming multiprocessors in the Hopper architecture to accelerate AI training and reasoning. Compared with the previous generation architecture, the new Transformer The engine is combined with the Hopper FP8 Tensor Core to provide up to 9 times the AI training speed and 30 times the AI inference speed on large NLP models.

With the continuous evolution of chip architecture, the computing performance of chips is also rapidly improving. Taking the representative chips V100, A100 and H100 of the three typical architectures Volta, Ampere and Hopper in recent years as examples, there have been significant improvements in the number of cores, computing speed, process technology, etc. Compared with V100, A100's single-precision floating-point computing power has increased from 15.7TFLOPS to 19.5TFLOPS; and double-precision floating-point operations have increased from 7.8TFLOPS to

9.7TFLOPS. In terms of price, the price of GPU chips has also increased rapidly with the improvement of performance. According to a report by Fast Technology in May 2023, the V100 accelerator card currently costs about $10,000, the A800 costs about $12,000, the A100 costs about $15,000, and the H100 costs about $36,500. Among them, A800 was launched by NVIDIA on the basis of A100 in order to meet the compliance requirements of the US government for the Chinese market, reducing the bandwidth of the NVLink high-speed interconnect bus from 600GB/s to 400GB/s.

The rise of large AI models has catalyzed the demand for computing power. With the launch of the ChatGPT application by OpenAI worldwide, the industry has been chasing after large models. As we all know, the training and reasoning of large models require a lot of computing power support. As the main "arms supplier" in the computing power field, NVIDIA has fully enjoyed this wave of dividends.

The consumption of computing power by large models can be divided into three links. The first link is the model pre-training link. According to the paper "Language Models are Few-Shot Learners" published by the OpenAI team in 2020, the computing power required to train a GPT-3 model with 174.6 billion parameters is about 3640 PFlop-day. Calculated based on the A100 FP16 computing speed of 156TFLOPS (floating point operations per second, 1TFLOPS=10^12 FLOPS, 1PFLOPS=1024TFLOPS), one card needs to run for more than 20,000 days, and 10,000 cards only need to be trained for two days. The second link is the model iteration and tuning link. The third link is the daily operation reasoning link. According to Semianalysis analyst Dylan Patel's analysis of factors such as model parameters, daily active users, and hardware utilization, OpenAI needs 3617 HGX A100 servers (8 GPU configuration) to maintain operation. That is to say, companies that are interested in large-scale model training and daily operations need to configure about 10,000 A100 GPUs. In order to cope with the large-scale model competition triggered by chatGPT, global technology giants have begun to reserve computing resources, and Nvidia has directly benefited from this wave of computing power arms race. According to Business Insider, Twitter purchased about 10,000 GPUs in April for one of the company's two data centers; Microsoft's artificial intelligence supercomputer in Azure was built in cooperation with OpenAI and has 285,000 CPUs and 10,000 GPUs; Google announced an AI supercomputer with about 26,000 Nvidia H100s in May.

In response to the strong market demand, Nvidia is also constantly launching a more powerful product portfolio. At the pre-Computex 2023 press conference, NVIDIA CEO Jensen Huang officially launched the new GH200 Grace Hopper superchip and the NVIDIA DGX GH200 supercomputer with 256 GH200 superchips driven by the NVIDIA NVLinkSwitch System. The GH200 superchip uses NVIDIA NVLink-C2C chip interconnect to integrate the Arm-based NVIDIA Grace CPU with the NVIDIA H100 Tensor Core GPU to provide a CPU+GPU consistent memory model, eliminating the need for traditional CPU-to-GPU PCIe connections. This also increases the bandwidth between the GPU and CPU by 7 times, reduces interconnect power consumption by more than 5 times, and provides a 600GB Hopper architecture GPU building block for the DGX GH200 supercomputer compared to the latest PCIe Gen5 technology.

4.3 NVIDIA attaches great importance to the supporting software ecosystem

4.3.1 Launching AI models to assist application development

NVIDIA announces new custom AI model foundry service. On May 29, 2023, NVIDIA announced the launch of a new custom AI model foundry service, NVIDIA ACE Game Development Edition (NVIDIAAvatar Cloud Engine (ACE) for Games), which uses AI-driven natural language interaction technology to bring intelligence to non-player characters (NPCs) in games, thereby changing the gaming experience.
Middleware, tool and game developers can use "ACE Game Development Edition (ACE for Games)" to build and deploy customized speech, dialogue and animation AI models in their games and applications. Developers can integrate the entire "NVIDIA ACE Game Development Edition (NVIDIA ACE for Games)" solution or use the components they need separately.

Based on NVIDIA Omniverse, "ACE Game Development Edition (ACE for Games)" provides optimized AI basic models for speech, dialogue and character animation, including:

- NVIDIA NeMo™ Large Language Model (LLM): Build, customize and deploy language models using proprietary data. Customize LLM based on the world view and character background of the game story, and use NeMo Guardrails to protect the dialogue from counterproductive or unsafe content.
- NVIDIA Riva: Used for automatic speech recognition (ASR) and text-to-speech to enable real-time voice dialogue.
- NVIDIA Omniverse Audio2Face: Used to create facial expression animations for game characters in real time with voice tracks. Audio2Face is paired with the Omniverse Connector for Unreal Engine 5, and developers can directly add facial animations to MetaHuman characters.

4.4 Realizing resonance between cloud and edge

4.4.1 NVIDIA Launches DGX Cloud

At the 2023 GTC conference, NVIDIA announced that it would cooperate with several cloud service providers to launch NVIDIA DGX Cloud. This allows enterprises to use DGX Cloud infrastructure hosted by cloud service providers without purchasing and owning servers. Microsoft Azure and Google Cloud will gradually start to provide services to obtain supercomputer-level AI computing capabilities through the browser. The charging standard for NVIDIA DGX Cloud starts at US$36,999 per instance per month. Each instance includes eight Nvidia H100 or A100 80 GB GPUs, and each GPU node has up to 640 GB of memory, which enables dedicated computing resources and is not shared with other tenants in the cloud.

On top of DGX Cloud, NVIDIA's AI platform also includes AI Enterprise and AI Foundations. NVIDIA AI Foundations is a set of cloud services that promote enterprise-level generative AI and support customization of use cases across fields such as text, visual content, and biology. AI Enterprise is the software layer of NVIDIA's AI platform, which provides end-to-end AI frameworks and pre-trained models to simplify the development and deployment of production AI.

Through DGX Cloud, NVIDIA can release the power of AI to more small and medium-sized enterprises. As mentioned above, NVIDIA's AI chips and DGX supercomputers are very expensive, and it is difficult for small and medium-sized enterprises to deploy them on a large scale. With DGX CLOUD, small and medium-sized enterprises can also enjoy the dividends of the AI era. This is also one of the core advantages of cloud computing, which can make expensive IT services popular and accelerate the popularization of advanced technologies.

4.4.2 NVIDIA Jetson edge computing platform covers more scenarios

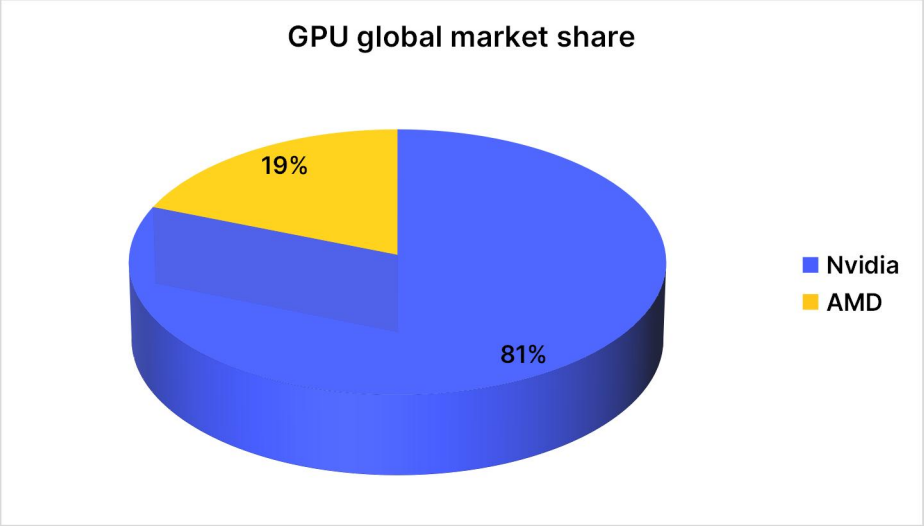
NVIDIA's Jetson edge computing platform can be applied to application scenarios such as robots, smart driving, and smart manufacturing. The Jetson platform includes Jetson modules (compact high-performance computers), JetPack SDK for accelerating software, and an ecosystem of sensors, SDKs, services, and products to speed up development. The Jetson platform has the advantages of small size, low power consumption, both software and hardware can be tailored, high degree of customizability, and open source system. The Jetson series modules are small in size but can provide powerful AI computing power. The size is within 100mm, and the AI computing power can reach up to several hundred TOPS. Among them, Xavier NX is 70mm x 45mm, and the AI computing power is 21TOPS. NVIDIA released the new Jetson AGX Orin industrial-grade module at COMPUTEX 2023, which can provide higher levels of computing power in harsh environments-up to 248 TOPS of AI performance in a power range of 15-75W.

Nvidia to launch Orin-based modules in 2023

Jetson creates a software isolation layer to reduce software migration costs. While providing software tools and solution SDKs for developers, NVIDIA has created a software isolation layer to minimize the cost of software migration on embedded platforms. NVIDIA's software tools and SDKs are universal on all Jetson hardware platforms, so migration on the Jetson hardware platform only requires recompilation without any code-level changes.

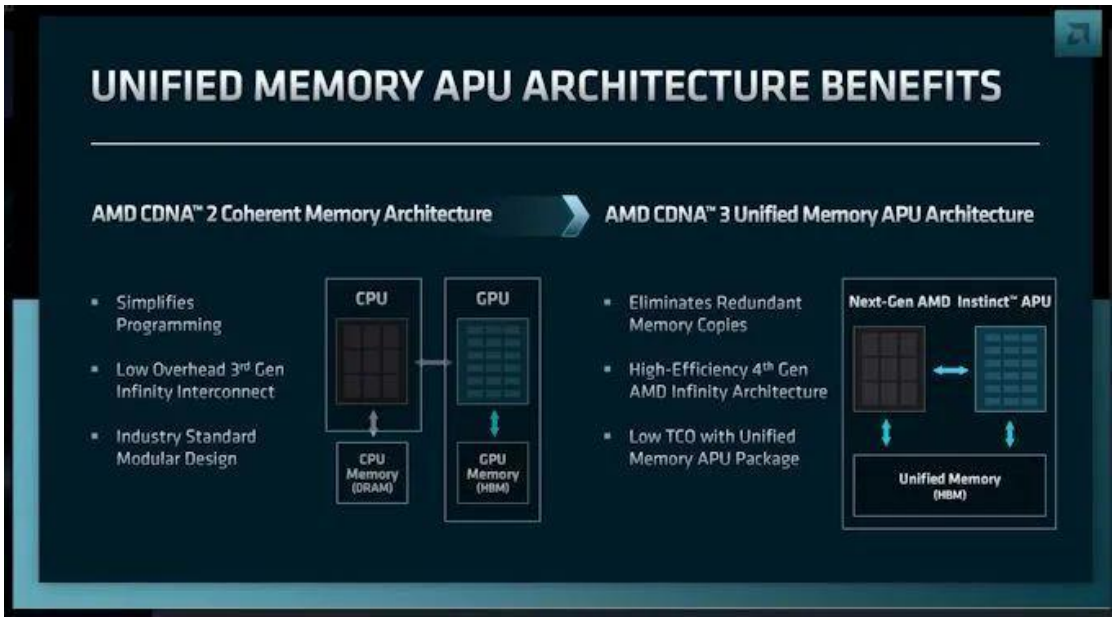
4.5 AMD is trying its best to catch up with Nvidia

The GPU market will continue to grow rapidly, with Nvidia enjoying the market growth dividend and AMD being the only competitor. The global GPU industry market size was $33.47 billion in 2021 and is expected to reach $477.37 billion in 2030, with an estimated CAGR of 34.4% from 2021 to 2030. In the field of independent GPUs, Nvidia has an absolute leading position, with a market share of about 81%, while AMD accounts for about 19%.
GPU global market share



Nvidia and AMD form an oligopoly in the global GPU market

Faced with the historic computing power opportunities brought by AI, AMD is taking a different approach to catch up with Nvidia. As Nvidia's only significant competitor in the GPU field, AMD is trying to package the CPU and GPU and a large amount of high-speed memory on the same chip, called APU. It achieves unified memory by allowing the CPU and GPU to share a unified physical memory. The advantage of physical memory sharing is that the CPU can store data in HBM and the GPU can read the data directly. The bandwidth of HBM is much higher than the Infinity (or PCIe) link between the CPU and GPU, allowing new designs to improve performance. This also simplifies socket-level HPC and AI programming because both processor types can directly access the same memory pool.



AMD creates APU architecture

The MI300, based on the APU architecture, is expected to be used in supercomputers by the end of the year. A new supercomputer (El Capitan with a performance of 2 exaflop) at the 2023 International Supercomputing Conference (ISC) will be launched at the end of this year, when it will replace the current Frontier as the world's fastest supercomputer. The supercomputer is supported by AMD and the chip to be used is the latest Instinct MI300. MI300 is a data center APU that integrates 13 small chips, most of which are 3D stacked. MI300 includes 24 Zen 4 CPU cores, 1 CDNA 3 graphics engine and 8 HBM3 memories, with a total memory of 128GB. With 146 billion transistors, MI300 is the largest chip AMD has put into production. Among them, the nine compute dies are a mixture of 5nm CPU and GPU, 3D stacked on four 6nm base dies. During the ITF World 2023 Semiconductor Conference, AMD proposed a "30x25" goal to increase chip energy efficiency 30 times by 2025. The key to this plan is Instinct MI300.

5. Policy and demand drive

5.1 AIGC drives the demand for AI computing power, and AI chips will become the future technological oil

As AI enters the "big model" era, the training data continues to grow, the algorithm complexity continues to increase, and the demand for computing power by domestic artificial intelligence manufacturers has risen sharply. As the computing power foundation for big models and AI applications, AI chips are becoming increasingly important.

In a broad sense, AI chips refer to modules that are specifically used to handle a large number of

computing tasks in artificial intelligence applications, that is, chips for the field of artificial intelligence are called AI chips. In a narrow sense, AI chips are chips that have been specially designed for artificial intelligence algorithms. Compared with traditional chips (such as CPUs), the performance advantage of narrow AI chips is mainly reflected in the emphasis on specialization.

AI chips are mainly divided into three types: general-purpose (GPU), semi-custom (FPGA), and custom (ASIC). The three types of chips are represented by NVIDIA's GPU, Xilinx's FPGA, and Google's TPU. GPU has the strongest computing power, but high cost and high power consumption; FPGA is programmable and the most flexible, but the computing power is not strong; ASIC is small in size, low in power consumption, suitable for mass production, but the R&D time is long, and it cannot be edited, the initial investment cost is high, and it brings certain technical risks.

Global manufacturers' layout in different technology paths

- GPU

Global GPU manufacturers have launched products in many fields, and the global GPU market is in its golden period of development. Global GPU giants such as NVIDIA and AMD have large revenue scales and more mature technologies. The demand for global GPU manufacturers in application fields such as data centers, smart cars, and games continues to increase, which has promoted the rapid development of GPUs. For example, NVIDIA's A100 and H100 GPUs have been widely used in desktops, notebooks, servers, self-service terminals and other devices. AMD's Radeon series GPUs also occupy an important market share in various devices. In addition, NVIDIA's CUDA platform has become an industry standard, providing powerful development tools and ecosystem support for global developers.

- CPU+FPGA

Global technology giants have deployed CPU+FPGA hybrid heterogeneous acceleration of AI computing. FPGA chips can be flexibly deployed when used to accelerate AI computing, and have the characteristics of programmability, high parallelism, low latency, and low power consumption. They have great potential in the field of AI inference. Global technology giants such as Amazon AWS and Microsoft Azure have launched FPGA-based services to promote the development of cloud FPGA ecology. FPGA manufacturers such as Altera and Xilinx acquired by Intel are also constantly launching new FPGA chips and solutions to support a wider range of AI applications. DeePhi Technology, acquired by Xilinx, designed a deep learning accelerator architecture based on FPGA, which improved the efficiency of AI computing.

- ASIC

Global ASIC manufacturers are strong and actively catching up with chip giants. As a dedicated integrated circuit, ASIC is widely used in fields such as artificial intelligence equipment. According to the terminal function, it can be divided into TPU chips, DPU chips and NPU chips. The first TPU launched by Google in 2015 greatly improved the performance of AI reasoning. After rapid development in recent years, global ASIC manufacturers have reached the international leading level in terms of process, computing power, and overall performance.

Google's TPUv4 has powerful performance in BF16 floating-point computing power. In addition, other ASIC manufacturers such as Graphcore and Cerebras have also launched AI chips with excellent performance. In the future, the world's leading ASIC manufacturers are expected to continue to maintain their technological advantages, break the monopoly pattern, and achieve "hardware optimization" of dedicated algorithms. Their long-term growth in the field of AI is worth looking forward to. Although ASIC cannot be reprogrammed and has a high initial investment cost, it has the advantages of stronger performance, smaller size, lower power consumption, lower cost, and higher reliability in large-scale mass production. As artificial intelligence technology matures and algorithms gradually converge, ASIC will have greater competitive advantages.

5.2 Global AI chip companies are booming

Driven by both policies and demand, global AI chip manufacturers are developing rapidly, especially in the ASIC direction, with huge opportunities. Overall, the global chip giants are still in the leading position, occupying most of the market share, and have a near-monopoly in GPUs and FPGAs. The following are several of the world's leading AI chip companies and their latest developments:

Intel
Intel has made significant progress in CPU+FPGA hybrid heterogeneous computing. Through the acquisition of Altera, Intel integrated FPGA technology into its product line, providing powerful heterogeneous computing capabilities. Intel's FPGA products are widely used in cloud computing, data centers and edge computing, supporting flexible deployment and efficient AI computing.

Graphcore
Graphcore is an AI computing-focused startup dedicated to developing high-performance AI accelerators. Its IPU (Intelligence Processing Unit) is uniquely designed to provide efficient parallel computing capabilities and significantly improve the performance of AI training and inference. Graphcore's IPUs are used by several leading AI research institutions and enterprises.

Cerebras Systems
Cerebras Systems has made breakthroughs in AI computing by developing the Wafer Scale Engine (WSE), the world's largest AI chip. With trillions of transistors, WSE can provide unprecedented computing power and bandwidth, greatly improving the training speed of deep learning models.

Xilinx
Xilinx is the world's leading FPGA manufacturer, providing flexible and efficient solutions for AI computing through its Versal series of adaptive computing acceleration platforms (ACAP). Xilinx's FPGA products are widely used in cloud computing, data centers and edge devices to support efficient AI inference and training.

Intel Habana Labs

Habana Labs, acquired by Intel, focuses on developing high-performance AI accelerators and has launched the Gaudi and Goya series of AI processors for AI training and inference respectively. Habana's AI accelerators excel in performance and energy efficiency and are adopted by multiple cloud service providers and enterprises.

6. Synthra Capital Investment Advice

The birth of AIGC will have a revolutionary impact and is expected to empower thousands of industries. We have sorted out three roadmaps and actively recommend the following three investment lines:

1) Manufacturers with computing power foundation, beneficiary targets:

Groq
Financing amount: US$300 million
Financing round: Round B
Investors: Tiger Global Management, D1 Capital Partners, T. Rowe Price, Synthra Capital, etc.

Groq was founded in 2016 and is headquartered in Palo Alto, California, USA. The company focuses on developing high-performance computing chips, especially accelerators for artificial intelligence and machine learning. Groq's Tensor Streaming Processor (TSP) architecture is designed for efficient processing of AI and machine learning workloads, providing extremely high computing performance and low latency.

Development progress:
Groq released its second-generation processor in 2021, further improving performance and energy efficiency. The company has established cooperation with many companies and scientific research institutions to apply its technology in fields such as autonomous driving, financial analysis and large-scale data processing.

Investment advice:
Maintain profit forecast and maintain "buy" rating. The company's current performance is steadily improving, and its profit forecast is maintained. It is expected that the company's net profit attributable to shareholders in 2023-2025 will be US$4.1/6.2/8.5 billion, up 75.0%/51.2%/37.1% year-on-year, respectively. EPS in 23-25 will be US$0.39/0.59/0.81, respectively, and the "buy" rating is maintained.

SambaNova Systems
Amount of financing: US$676 million
Funding round: C round
Investors: SoftBank Vision Fund, GV (Google Ventures), BlackRock, Synthra Capital, etc.

SambaNova Systems was founded in 2017 and is headquartered in Palo Alto, California, USA. The company is committed to developing the next generation of AI computing systems and providing end-to-end artificial intelligence and data flow acceleration solutions. SambaNova's Reconfigurable Dataflow Architecture (RDA) is designed for large-scale AI and machine learning workloads, providing high performance and flexibility.

Development progress:
In 2021, SambaNova released the DataScale system, which combines hardware and software to provide powerful AI computing capabilities. The company cooperates with many large enterprises and scientific research institutions to promote the application of AI technology in the medical, financial, energy and other industries.

Investment advice:
Maintain profit forecast and maintain "buy" rating. The company's current performance has steadily improved, and the profit forecast is maintained. It is expected that the company's net profit attributable to shareholders in 2023-2025 will be US$5.0/7.5/10.0 billion, up 82.0%/50.0%/33.3% year-on-year, respectively. Earnings per share in 23-25 are US$0.42/0.63/0.84, respectively, and the "buy" rating is maintained.

2) Manufacturers with commercial implementation of AI algorithms, beneficiary targets:

Runway
Financing amount: US$150 million
Financing round: Round B
Investors: Accel, Coatue Management, Amplify Partners, Synthra Capital, etc.

Runway was founded in 2021 and is headquartered in New York, USA. The company focuses on developing AI algorithms and tools to help creators and companies easily implement AI in content creation and marketing. Runway's platform provides a series of powerful AI tools, including image and video editing, automated design, text generation, etc., to help users improve work efficiency and creativity.

Development progress:
Runway launched a new generation of AI content creation platform in 2023, enabling users to generate and edit images, videos and other digital content more efficiently. The platform has been widely used in advertising, film and television production, social media content creation and other fields, helping users to significantly improve their creative efficiency.

Investment advice:
Maintain profit forecasts and maintain a "buy" rating. The company's current performance has steadily improved, and the profit forecast is maintained. The company's net profit attributable to shareholders in 2023-2025 is expected to be US$280 million/430 million/600 million, up 68.0%/53.6%/39.5% year-on-year, respectively. Earnings per share in 23-25 are US$0.24/0.37/0.52, respectively, and the "buy" rating is maintained.

Hugging Face
Funding amount: US$175 million
Funding round: C round
Investors: Sequoia Capital, Lux Capital, Addition, Synthra Capital, etc.

Hugging Face was founded in 2021 and is headquartered in New York, USA. The company is committed to providing open source natural language processing (NLP) tools and platforms to help companies and developers apply AI algorithms to actual business scenarios. Hugging Face's Transformers library has become a standard tool in the field of NLP and is widely used in tasks such as text classification, translation, and generation.

Development progress:
Hugging Face released the latest version of the Transformers library in 2023, further improving the performance and ease of use of the model. The company cooperates with many leading global companies and research institutions to promote the application of AI technology in finance, medical care, education and other fields. Hugging Face's platform provides developers and companies with powerful tools that enable them to quickly deploy and optimize AI models.

Investment advice:
Maintain profit forecasts and maintain a "buy" rating. The company's current performance is steadily improving, and its profit forecast is maintained. It is expected that the company's net profit attributable to shareholders in 2023-2025 will be US$320 million/500 million/720 million, up 80.0%/56.3%/44.0% year-on-year, respectively. EPS in 2023-2025 will be US$0.28/0.44/0.63, respectively, and the "buy" rating is maintained.

3) Application manufacturers with AIGC-related technology reserves, the beneficiary targets are:

Jasper
Amount of financing: US$100 million
Funding round: Round A
Investment institutions: Insight Partners, Foundation Capital, Bessemer Venture Partners, Synthra Capital, etc.

Jasper was founded in 2022 and is headquartered in California, USA. The company focuses on providing technical solutions for AI-generated content (AIGC), especially in copywriting and marketing content generation. Jasper's platform combines natural language processing (NLP) and generative AI to help companies and creators quickly generate high-quality text content, including advertising copy, blog posts, social media posts, etc.

Development progress:
Jasper launched its latest content generation platform in 2023, further optimizing the performance and ease of use of AI models. The company works with a number of leading marketing agencies and brands to help them improve the efficiency and effectiveness of content creation. Jasper's technology has been widely used in e-commerce, digital marketing and media.

Investment advice:
Maintain earnings forecasts and maintain a "buy" rating. The company's current performance is steadily improving, and its earnings forecast is maintained. It is expected that the company's net profit attributable to shareholders in 2023-2025 will be US$250 million/380 million/540 million, up 65.0%/52.0%/42.1% year-on-year, respectively. Earnings per share in 23-25 are US$0.22 /0.34/0.48, respectively, and the "buy" rating is maintained.

Synthesia
Funding amount: US$125 million
Funding round: Round B
Investors: Accel, FirstMark Capital, GV (Google Ventures), Synthra Capital, etc.

Founded in 2022 and headquartered in London, UK, Synthesia focuses on developing AI-generated video (AIGV) technology and providing automated video production solutions. Synthesia's platform uses generative AI technology to generate high-quality video content from text scripts without the need for traditional video shooting and editing processes. Its technology is widely used in corporate training, marketing videos, e-learning and other fields.

Development progress:
Synthesia released a new version of its video generation platform in 2023, further improving video quality and generation speed. The company works with many large companies and educational institutions to help them reduce video production costs and improve content creation efficiency. Synthesia's technology shows great potential in marketing, corporate training and online education.

Investment advice:
Maintain earnings forecasts and maintain a "buy" rating. The company's current performance is steadily improving, and the profit forecast is maintained. The company's net profit attributable to shareholders is expected to be US$300 million, US$450 million, and US$650 million in 2023-2025, up 70.0%/50.0%/44.4% year-on-year, respectively. Earnings per share in 23-25 are US$0.26/0.39/0.56, respectively, and the "buy" rating is maintained.

7. Disclaimer

Analyst Statement

Synthra Capital guarantees that the data used in the report are all from compliant channels; the analysis logic is based on the author's professional understanding, and the conclusions are drawn through reasonable judgment, striving to be independent, objective and fair, and the conclusions are not instructed or influenced by any third party; the author has not received any direct or indirect remuneration for the specific suggestions or opinions expressed in his research report in the past, present or future, and hereby declares.

Synthra Capital Investment Rating

CATEGORY	LEVEL	ILLUSTRATE
Stock investment rating	Buy	Stock price outperformed the market index by more than 20%
	Overweight	The stock price outperformed the market index by 10%-20%
	Neutral	The stock price performance is between the market index and ±10%
	Sell	The stock price underperformed the market index by more than 10%
Industry investment rating	Overweight	Industry index outperformed the market index by more than 10%
	Neutral	The performance of the industry index is between ±10% of the market index
	Underweight	The industry index underperformed the market index by more than 10%

Important Statement

This report is prepared by Synthra Capital; the copyright of the report belongs to Synthra Capital. This report is for the use of Synthra Capital clients only. The company will not regard the recipient as a client because he/she receives this report. No organization or individual may use, copy or disseminate it in any form without written permission. Any summary or excerpt of this report does not represent the official and complete views of this report. Everything shall be subject to the complete version of this report released by Synthra Capital to its clients.

This report is based on publicly available data or information, but Synthra Capital does not guarantee the completeness and accuracy of such data and information. The information, materials, suggestions and speculations contained in this report only reflect Synthra Capital's judgment on the day this report is publicly released. At different times, Synthra Capital may write and publish reports that are inconsistent with the information, suggestions and speculations contained in this report. Synthra Capital does not guarantee that the information and materials contained in this report are up to date. Synthra Capital may supplement, update and revise the relevant information and materials at any time. Investors should pay attention to the relevant updates and revisions. Synthra Capital or its affiliates may hold and trade securities issued by the companies mentioned in this report, and may also provide or seek to provide investment banking, financial advisory or financial products and other related services to these companies. The Company's asset management department, proprietary department and other investment business departments may independently make investment decisions that are inconsistent with the opinions or recommendations in this report. This report is for reference only and does not constitute an offer or invitation to sell or purchase securities or other investment targets. In any case, the information and opinions in this report do not constitute investment advice to any individual. Any form of written or oral commitment to share securities investment returns or share securities investment losses is invalid. Investors should judge whether to adopt the content and information contained in this report based on their own investment goals and financial situation and bear their own risks. Synthra Capital and its employees shall not bear any legal responsibility for any consequences caused by investors' use of this report and its contents.

Description of investment consulting business

The Company is qualified for investment consulting business. Securities investment consulting

refers to the activities of institutions engaged in securities investment consulting business and their investment consulting personnel providing securities investors or clients with direct or indirect paid consulting services such as securities investment analysis, forecasts or suggestions in the following forms: accepting the entrustment of investors or clients to provide securities investment consulting services; holding lectures, reports, analysis meetings, etc. on securities investment consulting; publishing articles, comments, and reports on securities investment consulting in newspapers and periodicals, and providing securities investment consulting services through public media such as radio and television; providing securities investment consulting services through telecommunications equipment systems such as telephone, fax, and computer networks; other forms recognized by the China Securities Regulatory Commission. Publishing securities research reports is a basic form of securities investment consulting business, which refers to the behavior of securities companies and securities investment consulting institutions analyzing the value, market trends or related influencing factors of securities and securities-related products, forming investment analysis opinions such as securities valuation and investment rating, preparing securities research reports, and publishing them to customers.

About Synthra Capital

Initiated by the famous angel investor Reid Hoffman, it is a professional private equity investment institution jointly established by well-known companies such as NVIDIA, Microsoft, Open AI, Tesla, as well as emerging industry experts and financial veterans. As a top fund company based in the United States, we are committed to providing excellent financial services to global investors through cutting-edge investment strategies and deep industry knowledge. We firmly believe that artificial intelligence and big data not only represent the future technological trends, but also the core force driving global economic growth and industrial transformation. Through AI and big data financial management funds, investors can not only participate in this trend of the times, but also obtain rich investment returns in the wave of technological revolution. Synthra Capital was established in 2021, with clients and offices all over the world.

- **US$62 billion** in discretionary asset management scale
- **500+** IPOs and M&A exits
- Deeply cultivate **17** AI industry tracks
- Provide financial support and incubation solutions for **187** AI companies
- **1000+** customer relationships
- **12+** global offices