
BUSINESS

OUR VISION

Architecting the world with AI.

OUR MISSION

Forge AI computing super systems. Transcend AI performance limits.

OVERVIEW

Who We Are

We are the largest independent AI computing cluster provider in China, leading disruptive innovation in AI infrastructure. The rapid evolution of AI technologies has driven growing demand for high-performance AI infrastructure. The bottleneck in designing and delivering such infrastructure lies not only in high-performance computing and networking chips themselves, but also in the ability to turn thousands or even tens of thousands of processors into a high-performance supercomputer through network systems, orchestration and scheduling mechanisms, and parallel computing technologies. Leveraging over two decades of accumulated expertise in distributed system and computer architecture, as well as our distinctive design rationale of hardware-level architecture convergence and system-level cross-layer optimization, we develop and provide full-stack AI computing cluster products and operation services. These capabilities enable us to fully unlock the potential of AI infrastructure and help customers achieve superior AI productivity.

We provide AI computing cluster products under the *Galaxy* brand, comprising AI computing devices and AI computing software. The AI computing devices consist of compute, networking and storage elements. Among them, our self-developed AI computing network system, *Mercury*, enables ultra-high-bandwidth interconnection among various elements and supports efficient, coordinated communication across homogeneous, heterogeneous and cross-region hardware. We have integrated a number of core technologies developed for our AI computing software into *Venus*, our AI computing operating system. *Venus* includes runtime and common component interfaces for executing AI workload tasks, as well as high-speed service engines for pre-training, post-training, inference and agentic workloads. We also provide a suite of AI computing cluster operation services under the *Galaxy Services* brand, primarily including maintenance and technical support services and compute services, enabling customers to access high-quality computing resources in a convenient and flexible manner.

We are among the few market participants in China with full-stack AI computing cluster product and service capabilities, according to Frost & Sullivan. Our offerings cover the full lifecycle of AI infrastructure, from software and hardware development, architecture design, and device and system deployment to performance tuning, fault detection and ongoing operations. We continuously serve key AI stakeholders, including large model companies, cloud service providers, education and research institutions, and telecommunications operators, addressing their evolving requirements across different types of AI workloads. By doing so, we help customers address real-world challenges and improve the utilization and stability of their AI computing clusters, which have earned us broad customer recognition and high customer loyalty. We are also gradually extending our business reach from the large model industry to a wide range of industry verticals through converged AI computing clusters with built-in agentic applications.

Our deep understanding of customer needs and unwavering commitment to technological advancement have contributed to a number of important achievements. According to Frost & Sullivan:

- We are the first independent AI computing cluster provider in China to participate in the deployment of RoCE-based clusters at the 1,000-GPU scale with our self-developed AI computing network system. We have an established system for deploying, operating and maintaining multiple clusters at the 10,000-GPU scale.

BUSINESS

- We demonstrated leading performance in cluster scalability and reliability among AI computing cluster providers in China, as evidenced by cluster MFU improvement by 30% and average annualized SLA of 99.97%.
- We ranked first among independent AI computing cluster providers in China in terms of revenue in 2025.
- As of December 31, 2025, our market share in China’s AI computing cluster market for clusters at the 10,000-GPU scale was approximately 10%, while the remaining leading participants were primarily large internet companies and telecommunications operators.

The following sets forth our business highlights:

Business Highlights	4 AI Computing Clusters at 10,000-GPU Scale Powered by Us	Over 90,000 GPUs included in the AI Computing Clusters Powered by Us ¹	Over 34,000 PFLOPs Computing Capacity Under AI Computing Cluster Operation Services with SLA Guarantee ¹
Cluster Performance	By 30% Cluster MFU Improvement ¹	99.97% Average Annualized SLA ¹	Over 96% Computing Utilization Linearity on Clusters at the 10,000-GPU Scale ¹
Operating Data	304% Revenue CAGR from 2023 to 2025	Over 80% Repurchase Rate for KA Customers ²	2 Consecutive Years Adjusted Net Profit (Non-IFRS measure)

Notes:

1 As of the Latest Practicable Date

2 During the Track Record Period and up to the Latest Practicable Date

Our Market Opportunities and Value Propositions

Large-scale AI workloads typically involve massive parameter sizes, vast volumes of data and intensive computing requirements, and therefore require ultra-large-scale parallel data processing capabilities and highly efficient parameter synchronization. Traditional computing architectures centered on general-purpose computing, which primarily rely on CPU-based centralized and serial processing, are increasingly unable to meet these requirements. In an AI computing cluster, AI computing software splits AI workload tasks into numerous concurrent subtasks, which are processed in parallel across thousands or even tens of thousands of GPUs. The AI computing network then efficiently synchronizes the computation results generated by each GPU. Through this high-concurrency processing model, AI workload tasks can be accelerated efficiently.

BUSINESS

We have observed the following trends shaping the AI computing cluster industry:

- ***Scaling efficiency and enhanced stability:*** As AI computing clusters continue to expand rapidly, AI computing infrastructure is no longer focused solely on increasing cluster scale, but is shifting toward a balanced emphasis on both scale expansion and efficient resource utilization. According to Frost & Sullivan, the current model FLOPs utilization (“MFU”) of AI computing clusters is 35% to 45%, indicating that a substantial portion of parallel computing resources in AI workload tasks remains underutilized. Improving computing resource utilization has therefore become a core priority. At the same time, as AI workloads require rapidly increasing computing power, a growing number of compute nodes need to operate in parallel and coordination. Software and hardware failures may lead to wasted computing resources and task interruptions. As a result, the focus of cluster optimization is shifting from single-node performance enhancement to improved stability at the distributed cluster level.
- ***Open ecosystem and full-stack technology:*** The industry has historically been constrained by single-vendor proprietary technology stacks, which often feature limited architectural compatibility and interoperability, increasing deployment and upgrade costs and restricting the full release of computing performance. As a result, AI infrastructure companies are accelerating the development of highly interoperable architectures, moving from closed and standalone components toward open and interconnected systems that converge diverse computing resources, connect with existing infrastructure, and enable flexible and efficient capacity expansion.

At the same time, as AI computing clusters continue to scale and their architectures become increasingly complex, relying on a single hardware component or standalone service is no longer sufficient to meet evolving needs. The industry requires core capabilities encompassing compute, networking and scheduling. These capabilities are critical to building full-stack solutions that cover the entire cluster lifecycle and support complex, high-end scenarios such as ultra-large model training and inference. Accordingly, the AI computing cluster delivery and service model is shifting from fragmented, point-to-point solutions toward converged software-hardware enablement.

- ***Expansion of AI applications:*** As the AI industry continues to evolve rapidly, AI workloads have expanded from foundational pre-training to a broader range of scenarios, including post-training, inference and agentic workloads. As a result, it is increasingly difficult for static and homogeneous cluster operation models to meet the differentiated requirements of emerging AI workloads. Therefore, the industry urgently needs AI computing cluster architectures capable of supporting diverse AI training, inference and application scenarios. These architectures should precisely match the resource deployment, network transmission and task scheduling requirements of different scenarios.
- ***Rapid technological advancement:*** Current computing architectures are increasingly constrained by bandwidth, power consumption, latency and operations bottlenecks, which materially limit multi-node coordination efficiency and the scaling potential of computing resources. As large models rapidly evolve toward trillion-parameter scale, AI computing clusters are also iterating rapidly across multiple dimensions. In terms of scale, clusters are upgrading toward 10,000-GPU and even 100,000-GPU scale. In terms of technology, computing chips, optical interconnection and storage technologies continue to achieve breakthroughs. Therefore, in line with the development trends of next-generation computing, the industry needs to continuously optimize cluster architectures and resource scheduling models by leveraging multiple technology pathways, including opto-electronic hybrid interconnection, heterogeneous computing convergence and advanced computing architectures.

BUSINESS

Our value propositions are rooted in the following capabilities:

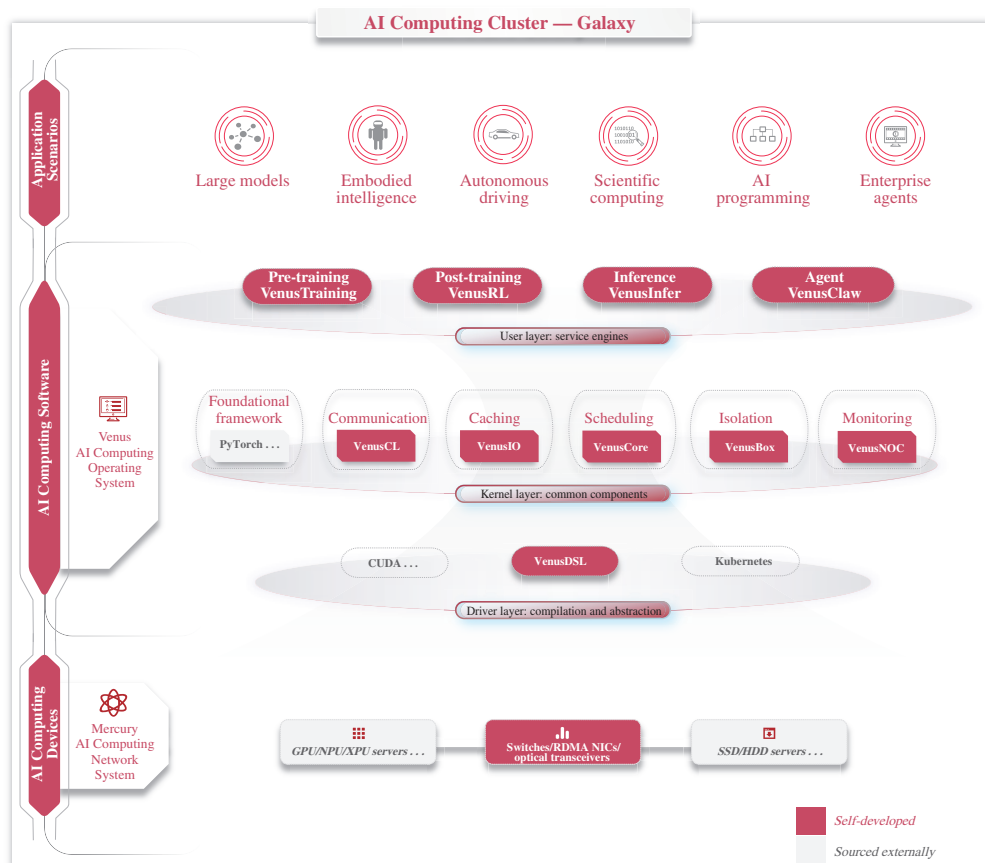
- ***High-performance AI computing cluster products and services:*** In response to these evolving needs, we design and deploy high-performance AI computing cluster products and provide operation services to support customers in achieving high-performance AI computing. By improving how different computing resources are interconnected, operated and orchestrated within AI computing clusters, we help customers achieve greater parallel computing capabilities and higher resource utilization, thereby unlocking stronger cluster performance. We also enhance the operational stability of distributed clusters through rapid fault localization and multi-layer fault-tolerant recovery. In addition, we have established a real-time monitoring and intelligent alerting system covering a comprehensive set of cluster performance metrics, which significantly reduces the probability of network jitter and hardware failures and helps ensure the stable operation of clusters at the 10,000-GPU scale.
- ***Diversified ecosystem with full-stack technology capabilities:*** We adopt an open cluster architecture that organizes resources into distinct but interoperable layers, and supports modular expansion based on customer deployment requirements. This enables compatibility with hardware from different vendors and mainstream AI frameworks, while supporting the unified management and coordinated operation of heterogeneous hardware resources with different specifications. We actively help build a domestic, self-developed AI computing cluster ecosystem, enabling performance comparable to leading international solutions and capturing opportunities driven by the rise of domestic AI computing chips. Through our full-stack technology and optimization system, we are able to conduct coordinated optimization for large-scale and complex AI workloads, enabling us to provide customers with comprehensive one-stop solutions that support streamlined deployment, out-of-the-box usability and a seamless customer experience.
- ***AI computing operating system enabling varied AI workloads:*** Our AI computing operating system, *Venus*, supports a variety of AI workloads across pre-training, post-training, inference and agentic scenarios. For pre-training scenarios, *Venus* provides topology-aware intelligent scheduling, a high-speed unified collective communication library and multi-level fault tolerance mechanisms to support the efficient and stable training of ultra-large models. For post-training scenarios, *Venus* improves resource utilization by decomposing tasks and applying fine-grained scheduling to coordinate different stages of multiple tasks. For inference scenarios, *Venus* supports traffic-based elastic scaling, as well as hardware acceleration mechanisms that coordinate compute, networking and storage resources to improve inference bandwidth. For agentic scenarios, *Venus* incorporates agent sandbox technology within its fine-grained isolation components to provide an efficient and controlled execution environment required for task understanding, planning, tool calling and iterative verification.
- ***Multi-dimensional technology innovation:*** We closely follow industry development trends and are committed to applying advanced technologies to our products and services. Through our AI computing network system, *Mercury*, we have completed validation of OCS-based opto-electronic hybrid switching and implemented NPO solutions to address bottlenecks in ultra-high-speed transmission, while balancing high performance with ease of maintenance. By adopting heterogeneous computing and advanced computing architectures, we leverage the computing strengths of different chips to enable more precise matching and efficient utilization of computing resources, significantly improving the computing efficiency of complex AI tasks and addressing the limitations of single-chip computing capacity.

BUSINESS

Our Products and Services

AI Computing Cluster Products — *Galaxy*

Our AI computing cluster products turn thousands or even tens of thousands of GPUs into a supercomputer, delivering unified AI computing resources with strong scalability, availability, observability and security. Our AI computing cluster products are deployed in high-power data centers comprising numerous racks. This overall architecture enables the AI computing clusters to operate as converged AI infrastructure. Our AI computing cluster products consist of AI computing devices and AI computing software, supporting various types of AI workloads, such as large models, embodied intelligence and scientific computing. The following diagram illustrates the architecture and components of our AI computing cluster product, *Galaxy*, including components developed in-house and those sourced externally, as well as the AI applications it supports:



- AI computing devices:** Within AI computing devices, our self-developed AI computing network system, *Mercury*, includes ultra-high-bandwidth switches, RDMA network interface cards (“NICs”), optical transceivers and other hardware components, as well as *MercuryNOS*, our open-source network operating system responsible for the scheduling and operations of these hardware components. *MercuryNOS* is designed to improve network bandwidth utilization, traffic distribution efficiency and transmission stability under highly concurrent workloads, while providing comprehensive, fine-grained monitoring visibility to support performance tuning and fault diagnosis for AI tasks. In addition, our AI computing cluster products and solutions also cover compute hardware and storage. Leveraging our deep understanding of AI computing clusters and close cooperation with business partners including compute hardware vendors, storage vendors and contract manufacturers, we have the capabilities to design, deploy and optimize compute and storage elements of AI computing clusters.

BUSINESS

- **AI computing operating system:** Our AI computing operating system, *Venus*, functions as the “brain” of the AI computing clusters, and consists of the driver layer, kernel layer and user layer. The driver layer provides the kernel layer with unified interfaces for accessing distributed devices. Within the kernel layer, in addition to vendor-provided parallel programming libraries and open-source cloud-native resource management platforms, our self-developed *VenusDSL* further enhances system portability and chip adaptability. The kernel layer provides the user layer with core runtime components required for AI workload processing, including the AI computing base framework, the high-speed collective communication library (*VenusCL*), the cache input/output (*VenusIO*), the multi-queue task scheduler (*VenusCore*), the resource and task isolation component (*VenusBox*), and the multi-dimensional monitoring component (*VenusNOC*). The user layer provides optimized service engines for different types of AI workloads, including *VenusTraining* for pre-training, *VenusRL* for post-training, *VenusInfer* for inference, and *VenusClaw* for agentic workloads.

To continuously address the performance challenges of AI infrastructure, our AI computing cluster products boast hardware-level architecture convergence and system-level cross-layer optimization, meeting the need to improve AI computing cluster efficiency and overcome bandwidth and power consumption bottlenecks. Unlike approaches that focus on isolated optimization of individual components, our AI computing cluster products are designed to transcend the performance limit brought by single-component optimization. It enables computing utilization optimization for complex, high-value scenarios such as large-scale AI training and inference, allowing further MFU improvement of SOTA solutions. Below are the technology highlights of our core self-developed components within our AI computing cluster products:

- **AI computing devices:** Mercury fully applies opto-electronic coordination in its design. We have validated OCS-based opto-electronic hybrid switching technology at scale and implemented 3.2T near-packaged optics, addressing bottlenecks in ultra-high-speed transmission while balancing system performance with ease of maintenance. These capabilities help overcome the bandwidth and power consumption limitations of traditional electrical interconnection. In addition, at the AI computing device level, we adopt optimized and workload-specific hardware architectures based on the different characteristics of the prefill and decoding stages in inference, as well as the different characteristics of attention computation and expert computation during decoding. This enables software-hardware coordination and heterogeneous acceleration.
- **AI computing operating system:** Venus applies compute-network-storage coordination and software-hardware coordination to achieve cross-layer optimization across multiple dimensions. *VenusDSL* improves computing efficiency by fusing multiple operators into optimized execution units. *VenusCL* improves computing resource utilization and pre-training efficiency by enabling kernel-free scheduling and more efficient parallel execution of computation and communication. *VenusRL* and *VenusCore* improve post-training efficiency by allowing rollout and model computation in reinforcement learning tasks to proceed in parallel where possible, thereby increasing computing resource utilization. *VenusInfer* and *VenusIO* build a hierarchical cache system across GPU memory, network direct access and flash storage, expanding cache capacity and significantly reducing the cost of long-context inference. *VenusClaw* and *VenusBox* reduce sandbox idle time through dynamic optimization of sandbox start-up and shutdown, while maintaining secure isolation, thereby lowering operating costs. *VenusNOC* combines active and passive monitoring methods to enable real-time performance analysis of online tasks without disrupting task execution, thereby strengthening our ability to track and diagnose complex issues. See “— Our Products and Services — AI Computing Cluster Products — AI Computing Software — Venus” for details.

Leveraging our full-stack AI computing cluster product and service capabilities, we are able to serve customers across the cluster lifecycle, including architecture design, hardware selection, software-hardware adaptation, network interconnection, system tuning and performance validation. Depending on

BUSINESS

customer requirements, our delivery may range from full cluster deployment to targeted deployment of selected hardware and software components, cluster-level networking and convergence services, or software and performance tuning based on customers’ existing hardware infrastructure.

AI Computing Cluster Operation Services — Galaxy Services

Built on our high-performance AI computing cluster products, we launched our AI computing cluster operation services under the *Galaxy Services* brand. Powered by our self-developed AI computing network system, *Mercury*, and AI computing operating system, *Venus*, *Galaxy Services* address the limitations of traditional AI computing cluster deployments, which often focus on initial deployment but lack sustained operations support. It helps customers address challenges such as fragmented computing resources, high operation and maintenance requirements, low internal collaboration efficiency, and the burden of large upfront capital expenditure and ongoing operations costs. *Galaxy Services* also reduce the need for customers to maintain a large, dedicated operating team, while preserving data and operational security.

Galaxy Services boast the following key features, designed to help customers improve business agility and reduce operating expenses:

- ***Comprehensive support and SLA guarantee:*** Leveraging *Galaxy*’s native full-stack observability and elastic system architecture, *Galaxy Services* provide comprehensive operations and business adaptation support, covering hardware, systems and tasks, reducing the complexity of day-to-day cluster operations for customers. We also provide systematic maintenance support to guarantee hardware availability. Leveraging *VenusNOC*, we can accurately identify and diagnose hidden and complex issues in AI computing clusters.
- ***Cost-effective, high-quality computing resources:*** *Galaxy Services* provide customers with efficient and cost-effective high-quality computing resources through intelligent workload scheduling, refined cost management, and mechanisms that preserve the value of customers’ infrastructure investment. The system dynamically allocates resources and improves utilization, while supporting detailed billing, cost accounting and optimization recommendations to help enterprises reduce capital and operating expenditures. It also supports the reuse of existing infrastructure and smooth cluster expansion, enhancing performance, cost and long-term value.

Our Results

We achieved significant growth during the Track Record Period. Our revenue increased from RMB31.8 million in 2023 to RMB520.3 million in 2025 at a CAGR of 304.5%. Our gross profit increased from RMB2.4 million in 2023 to RMB65.4 million in 2024 and further to RMB113.2 million in 2025. We became profitable in 2024, and our adjusted net profit (non-IFRS measure) was RMB20.3 million and RMB31.1 million in 2024 and 2025, respectively.

COMPETITIVE STRENGTH

Pioneering Specialist for High-Performance AI Computing in China and Driving Continuous Industry Innovation

We have one of the leading AI infrastructure teams in China, bringing together a deep pool of experts in AI computing clusters. We have accumulated over 20 years of expertise in distributed system and computer architecture, and our R&D team members have collectively published more than 90 papers in leading international conferences and journals, including USENIX Symposium on Networked Systems Design and Implementation, ACM Special Interest Group on Data Communication, ACM International Conference on Architectural Support for Programming Languages and Operating Systems and Optical Fiber Communication Conference. Combining academic insight with deployment experience, we continue to advance the AI computing technologies and drive innovation in the AI infrastructure industry. We have demonstrated industry-leading scalability and reliability: we have improved computing resource

BUSINESS

utilization for training of sparse large models by 30% on clusters at the 10,000-GPU scale; achieved computing utilization linearity of over 96% on clusters at the 10,000-GPU scale; and maintained industry-leading stability across clusters under our operation, with an annualized hardware failure rate of below 0.3% and an average annualized SLA of 99.97%.

We have developed distinctive rationale of hardware-level architecture convergence and system-level cross-layer optimization for AI computing optimization. Our R&D team consists of engineers and scientists who remain focused on AI and compute system technologies. Leveraging our team’s accumulated experience, we have implemented design mechanisms such as compute-network-storage coordination, software-hardware coordination, opto-electronic coordination and heterogeneous acceleration. While different AI applications have distinct characteristics, the underlying infrastructure supporting these workloads can generally be categorized into pre-training, post-training, inference and agentic applications. To date, we have achieved a series of R&D results in terms of AI computing devices and software with respect to these workload types, including:

- **Network infrastructure:** Through opto-electronic coordination, *Mercury* builds a hybrid optical-electrical AI computing network and incorporates intelligent traffic scheduling mechanisms. This has reduced collective communication time by 50.3%, increased bandwidth utilization by 47.1%, and reduced power consumption for large-model training by 32.1%.
- **Pre-training infrastructure:** Through cross-layer optimization between computation and communication, *VenusCL* reduces the communication workload handled by GPUs by offloading such tasks to CPUs, allowing computation and communication to run more concurrently in pipeline workflows. According to Frost & Sullivan, this improves computing resource utilization for training workloads by 2% to 6% above the leading industry performance level in clusters at the 1,000-GPU scale, and can reduce associated capital expenditure by over RMB100 million based on the average capital investment for a cluster at the 10,000-GPU scale.
- **Post-training infrastructure:** Through cross-layer optimization between computation and input/output, *VenusRL* allows the generation of rollout for reinforcement learning to proceed in parallel with model computation where possible. Compared with conventional post-training implementation approaches commonly used in the industry, our *VenusRL* architecture improves overall task efficiency by 37%. Through cross-layer optimization between computation and security isolation, *VenusBox* optimizes the management of secure agent sandboxes by reusing sandbox environments. This improves the efficiency of on-demand sandbox start-up and shutdown, reduces sandbox idle time, and lowers agent sandbox operating costs by over 50%.
- **Inference infrastructure:** Through cross-layer optimization across computation, networking and storage, together with heterogeneous hardware acceleration, *VenusInfer* builds a tiered cache system that allows GPUs to access flash storage directly through the network. This increases cache capacity by 10 to 30 times and significantly reduces the cost of scaling long-context inference. *VenusInfer* also decouples attention computation from expert-layer computation, enabling fine-grained dynamic scaling of inference resources, and reduces latency by 49.6% through communication offloading.

We have built a robust R&D system that closely integrates product innovation with advanced technology planning. Through deep industry-academia-research collaboration, we have worked with multiple leading research institutions, universities and cloud service providers in China to implement a number of advanced AI infrastructure research projects, effectively shortening the cycle from technology development to product deployment. Our industry-academia-research collaboration has delivered meaningful results, including:

- We completed the full process of in-house development and deployment of switches based on domestically developed technologies within eight months, compared to an industry average of approximately 18 months. In collaboration with researchers from leading universities, we optimized the RDMA congestion control algorithm and shortened the relevant development cycle from eight months to less than four months.

BUSINESS

- Our intelligent operation system, *VenusNOC*, originated from a proof-of-concept developed in an academic research setting and has been further validated through our published research. Building on this foundation, we have translated this concept into an operational system deployed in production environments. *VenusNOC* can accurately identify and diagnose hidden and complex issues in ultra-large-scale computing clusters at the 10,000-GPU scale. It has been adopted by multiple leading large model companies, where it has helped identify and analyze numerous online issues, while only causing less than 0.1% degradation in terms of MFU with online real-time, function-level monitoring.

Pioneering AI Infrastructure Capabilities and Continuously Setting Industry Benchmarks

As an early mover in China’s AI computing cluster industry, we identified market opportunities at an early stage and leveraged our first-mover advantages to establish a strong market position and earn customer recognition. Through continued participation in frontline projects, we have accumulated deep industry insights and practical experience, which have strengthened our long-term competitive barriers. We have also been deeply involved in the formulation of industry standards and have led the development of key technical specifications and project technical requirements, building strong competitive barriers. According to Frost & Sullivan, we have recorded a series of milestone achievements, which have significantly contributed to the development of AI infrastructure in China:

- ***Pioneering cluster scale-out capabilities:*** We are the first independent AI computing cluster provider in China to participate in benchmark projects of AI computing cluster delivery at increasing levels of scale, including the first RoCE-based clusters at the 1,000-GPU scale in 2023, and followed by one of the first clusters at the 4,000- and 10,000-GPU scales in 2024. We have developed core capabilities in ultra-large-scale deployment and resource optimization, while building industry-leading solution flexibility and hands-on deployment experience.
- ***Pioneering cluster scale-across capabilities:*** We are the first independent AI computing cluster provider in China to deploy long-distance distributed RoCE-based AI computing clusters. In March 2024, we further upgraded our networking capabilities to support clusters spanning 30 and 50 kilometers, achieving over 98% effective computing performance compared with a locally deployed cluster. In July 2025, we completed real-world testing of cross-regional heterogeneous distributed training over a distance of 1,500 kilometers, demonstrating reliable system scalability and full-stack compatibility with domestic hardware and software ecosystems.
- ***Pioneering AI networking autonomy and breakthroughs:*** We are the first independent AI computing cluster provider in China to independently develop and mass-produce a fully domestic high-throughput Ethernet switch based on a domestic 25.6T switch chip, and to complete optoelectronic hybrid interconnection for AI computing clusters, with in-depth experience in tuning and deployment validation. We have also self-developed a 100T ultra-high-throughput switch based on domestic NPO technology, with overall switching capacity comparable to leading international flagship equipment. In 2025, we took the lead in building a fully self-developed AI computing cluster equipped with domestic CPUs, GPUs, switches and RDMA NICs, and demonstrated its commercial viability through large-scale deployment.
- ***Pioneering enterprise-grade AI cluster delivery:*** We are one of the earliest companies to deploy large-scale, enterprise-grade converged AI computing clusters for commercial use. We have developed benchmark agentic applications and built a fully domestic AI infrastructure foundation, providing core support for enterprises’ intelligent transformation.

BUSINESS

Building a Business Expansion Flywheel and Empowering Benchmark Customers

Deploying and operating high-performance AI computing clusters involve substantial technical barriers and require extensive accumulated experience. Leveraging our early-mover advantages, we have continued to expand the scale of the AI computing clusters we power. As of the Latest Practicable Date, we had delivered and serviced AI computing clusters comprising over 90,000 GPUs in aggregate, with more than 34,000 PFLOPS of computing capacity under our AI computing cluster operation services with SLA guarantee. We have developed a mature and highly scalable operation model that continues to create greater value for our customers:

- ***Proven deployment experiences:*** We have developed a strong understanding of the software and hardware characteristics of AI computing clusters and standardized our delivery process. As a result, for clusters at the 10,000-GPU scale, we are able to complete software tuning and bring the cluster into production within two weeks after hardware deployment. This significantly shortens the time from hardware deployment to stable cluster operation, supports customers’ rapid business launch, and improves the utilization efficiency and return on investment of their infrastructure.
- ***Extensive operation data and intelligent agents:*** Leveraging extensive operation data and a comprehensive fault database, together with our reinforcement learning capabilities developed for post-training scenarios, we have built AI computing cluster operation agents that can identify potential operational risks in advance, accurately locate root causes of faults, conduct systematic in-depth analysis and perform basic self-recovery functions. These capabilities comprehensively improve the stability and reliability of AI computing clusters, as well as the efficiency of our operation services.
- ***Distinct value proposition across the industry chain:*** As our operations continue to expand, we have developed procurement advantages in key networking equipment, including optical transceivers, white-box switches and high-speed interconnect cables. This has significantly enhanced our bargaining power with suppliers, allowing us to pass cost efficiencies on to customers and service AI computing clusters in a more cost-effective manner.

We focus on benchmark customers with demanding AI requirements, including large model companies, cloud service providers, education and research institutions, and telecommunications operators. These customers are major drivers of AI infrastructure investment and industry growth in China. By serving them over the long term, we have built deep technical expertise that enables us to quickly adapt to diverse application scenarios and meet the growing computing needs of a broader customer base. Our long-standing strategic partnerships with benchmark customers also provide us with stable, recurring growth opportunities as industry concentration increases. During the Track Record Period and up to the Latest Practicable Date, the repurchase rate of our KA customers exceeded 80%. By continuously deepening cooperation with benchmark customers, we are strengthening a self-reinforcing growth flywheel that drives our business expansion.

Full-Stack AI Computing Cluster Offerings Enabling One-Stop, High-Performance Computing Services

The development of AI computing clusters involves high entry barriers, mainly due to the high technical requirements, the broad range of technologies involved, the need to coordinate various players along the industry chain and the long deployment cycle. A high-performance and stable AI computing cluster requires not only hardware deployment, but also network interconnection, intelligent scheduling, workload adaptation, and operations support, among other key capabilities. It also requires coordination among various participants along the industry value chain, including hardware vendors, software providers and deployment service providers. As the components of AI computing clusters are highly interconnected, any inefficiency in coordination may affect cluster performance, delivery efficiency and long-term stability. As a result, customers often face practical challenges in building AI computing clusters,

BUSINESS

including complex system convergence, prolonged ramp-up periods, difficult ongoing operations and inability to fully realize system performance. This creates strong demand for a one-stop AI computing cluster solution that covers the entire process, enables efficient deployment, and supports continuous optimization.

According to Frost & Sullivan, we are one of the few independent AI computing cluster providers with full-stack AI computing cluster product and service capabilities. We provide a true one-stop offering that addresses customers’ comprehensive needs, from software and hardware development, architecture design and system deployment to performance tuning, fault detection and ongoing operations. Through coordinated optimization, we fully unlock system performance potential and address the key challenges associated with AI computing clusters of efficiency, performance and cost.

Unlike traditional computing service models that provide limited support after initial delivery, we treat AI computing cluster operation services as an important part of our full-stack capabilities to support the long-term and efficient operation of customers’ businesses. Leveraging unified data visibility across the full technology stack, we provide customers with evolving cluster operation services supported by granular monitoring, diagnosis, optimization and management. These services enable automatic identification and rapid diagnosis of software and hardware issues, supported by 24/7 dedicated operations support. We also continuously provide operator optimization, parallelization strategies and acceleration technologies for customers’ AI computing clusters to help maintain optimal cluster performance. In addition, through cost analysis and scheduling optimization recommendations, we help customers reduce overall capital and operating expenditures.

Building Open Ecosystem to Empower the AI Industry Chain

The value chain of AI industry currently faces challenges arising from fragmented upstream and downstream collaboration and closed ecosystems. Upstream hardware innovation often cannot quickly adapt to the workload requirements of frontier models, while downstream large models may find it difficult to fully unlock the performance potential of diverse hardware resources. These constraints limit the overall efficiency of industry innovation. We serve as an “industry bridge” connecting upstream hardware suppliers and downstream large model developers. By aligning frontier model workload requirements with infrastructure innovation, we strengthen our full-stack advantages and help foster a more open and collaborative AI ecosystem.

For upstream hardware, our products and services have completed compatibility validation and deployment with the compute hardware, RDMA NICs, optical transceivers and switches offered by mainstream vendors. This allows customers to select hardware components from different vendors and flexibly build AI computing clusters that meet their own needs, without being tied to specific hardware architecture of a single vendor. As a result, we are also able to fully unlock the performance potential of domestic hardware and help clusters incorporating domestic hardware achieve performance comparable to leading international standards.

For downstream models, we keep pace with the rapid iteration of large models and enable fast deployment and efficient operation of models with different architectures through our open and compatible adaptation layer. Mainstream models based on the CUDA ecosystem, as well as self-developed open-source models based on domestic ecosystems, can run efficiently on *Venus* without requiring changes to their core code. *Venus* provides instant inference adaptation for mainstream open-source models. We work closely with leading domestic open-source model companies, helping them accelerate large model research, development and iteration.

We serve as a key player in the AI infrastructure industry. Unlike traditional solutions that only provide basic adaptation and integration, we identify bottlenecks and requirements in AI development and cluster deployment and use our AI computing cluster offerings to empower both downstream model developers and upstream hardware providers. By doing so, we help build a truly open ecosystem platform and unlock greater AI productivity in commercial applications. This allows upstream hardware providers to focus on chip innovation and downstream model developers to focus on algorithm innovation, while we

BUSINESS

handle the connectivity and optimization across different layers of the industry chain to fully realize each participant’s capabilities. Together, these efforts support a vibrant and open AI computing ecosystem and accelerate the upgrade of the broader industry.

GROWTH STRATEGIES

Continuously Iterate AI Computing Cluster Products and Consolidate Market Leadership

As AI agents experience explosive growth, market attention has shifted from model innovation to inference services, and AI infrastructure is evolving from research-oriented clusters built around GPUs into AI factories that provide production-grade AI computing power. Against this backdrop, we will continue to strengthen and enhance our industry influence as a global leading independent AI computing cluster provider, in line with the development trends of AI infrastructure.

We will continue to iterate our AI computing cluster products with increasing emphasis on heterogeneous convergence. For our next-generation AI computing cluster products, we will adopt a SuperPod architecture supported by opto-electronic coordinated interconnection, aggregating computing resources with larger memory capacity through SuperPod and improving interconnection efficiency through optical technologies. In the future, the architecture of our AI computing cluster products is expected to further evolve into a hierarchical heterogeneous acceleration architecture by introducing more domain-specific processors, driving clusters toward greater efficiency and specialization and better supporting large-scale, production-grade AI workloads.

Specifically, we will focus on the following key technologies and business initiatives:

- ***Silicon photonics and optical switching:*** We will continue to advance our strategic initiatives in interconnection technologies. Leveraging opto-electronic technologies such as silicon photonic engines, PICs and EICs, we aim to develop interconnection solutions suitable for high-bandwidth GPU scale-up and address signal integrity challenges in 224G and next-generation high-speed SerDes technologies. We also plan to develop large-capacity, ultra-low-latency silicon photonic OCS optical switching technologies and opto-electronic high-performance network scheduling optimization technologies. Together with advanced optical interconnection capabilities such as NPO and OCS, these initiatives will enable us to build high-performance networks with ultra-high bandwidth and ultra-low latency, and further strengthen our core advantages in AI computing networks.
- ***Hardware acceleration:*** We will further implement a hierarchical heterogeneous acceleration architecture to optimize the full-stack cluster architecture. Drawing on the design logic of dedicated acceleration architectures for large models, we plan to co-develop LPU with ecosystem partners which can independently handle lightweight specialized tasks such as large-model decoding and inference, context processing and sequence scheduling. This will allow us to decompose the computing chain, decouple different workload layers and create differentiated computing roles for training and inference. For heterogeneous collaboration among different types of processing units and accelerator chips, such as CPUs, GPUs, LPUs and other specialized accelerators, we will dynamically allocate computing units based on the workload characteristics of different tasks and combine the capabilities of different acceleration chips on demand. This is expected to address the performance limitations of single-hardware architectures and comprehensively improve overall cluster throughput, concurrent processing capability and resource utilization efficiency.
- ***Computing software:*** We will develop the next-generation AI computing operating system and iterate on our *Venus* operating system based on the characteristics of high-density compute nodes in SuperPod and massive high-bandwidth scale-up and scale-out interconnections. We are optimizing its architecture and enhancing its functions in areas such as collective

BUSINESS

communication libraries, unified memory management, and elastic scheduling and failover for training and inference workloads, so as to better unlock the performance of SuperPod and SuperPod clusters and ensure cluster stability.

As these technological capabilities continue to iterate and upgrade, we also plan to further expand the scale of AI computing clusters that we design, deploy and own based on customer needs, so as to continuously enhance our supply of high-quality computing services and better meet customers’ diversified computing requirements.

Strategic Vertical Integration and Value Chain Expansion

We plan to pursue targeted vertical integration around key segments of the AI infrastructure industry chain, particularly in core hardware and advanced technology system design. Through strategic investments and acquisitions, we aim to strengthen our control over key technologies and enhance whole-process capabilities from core design, scalable production to project deployment. These strategic initiatives are expected to improve product delivery efficiency, reinforce full-process quality control and support long-term cost optimization.

Through such strategic acquisitions or investments, we may rapidly obtain access to advanced interconnection technologies, as well as core liquid cooling processes and proven production capacity in these regards, without going through lengthy technology development and customer certification cycles. This is expected to significantly shorten the time required for technology implementation and mass production. It will also further improve our coordination efficiency from system design to scaled deployment, continuously enhance hardware capabilities for AI infrastructure, and strengthen our core competitiveness in the industry.

We also intend to expand into AI application scenarios and make targeted acquisitions or investments in companies focused on agent development. Leveraging our strong AI computing infrastructure capabilities, we aim to promote deeper synergy between value chain consolidation and application-layer innovation, enabling more scenario-specific, enterprise-grade AI computing cluster products and solutions to be efficiently developed and deployed at scale, and building an integrated ecosystem combining AI computing power and applications.

Deepening Collaboration with Benchmark Customers and Expanding Market Penetration

Leveraging our ongoing collaboration with benchmark customers in the AI industry, we provide deeply optimized AI computing infrastructure for leading large model companies to support the continued exponential growth of token traffic. We will deepen customer collaboration through agile product iteration, technology upgrades and long-term stable cooperation, thereby further strengthening customer engagement and stickiness. We also intend to participate more deeply in a broader range of benchmark projects, such as next-generation model development and ultra-large-scale computing deployments, in order to accumulate more strategically valuable practical experience and further expand the depth and breadth of our cooperation with benchmark customers.

In addition, we plan to expand our business presence among customers operating in specific, innovative industry sectors such as scientific computing, embodied intelligence, autonomous driving and enterprise services. Our standardized and modular product architecture, open cluster architecture compatible with mainstream computing, networking and chip ecosystems, and refined operations capabilities are well aligned with the needs of these customers. By providing locally deployed AI computing cluster products and solutions tailored to customer requirements, we aim to drive further business growth among these customers of specific industry sectors and broaden our market coverage.

BUSINESS

Expanding Overseas Presence and Enhancing Commercialization Capabilities

Compared with the rapid development of China’s AI infrastructure market, overseas markets still present meaningful opportunities, as independent AI computing cluster providers remain relatively scarce and market demand is largely served by major international suppliers. Leveraging our accumulated operating experience in China, full-stack technical capabilities, and standardized and modular product architecture, we are advancing the international expansion of our technologies, products and services, and supply chain capabilities. Through these efforts, we aim to build comprehensive global delivery capabilities, support the cross-regional deployment of AI computing clusters, steadily expand our overseas presence and strengthen our global commercialization capabilities.

We plan to expand our presence in Southeast Asia, with a particular focus on Singapore, Malaysia and neighboring markets, to address regional demand for AI computing resources. For AI computing cluster projects in the region led by our domestic business partners, we will provide comprehensive support covering software and hardware deployment, performance tuning, resource management and remote operations. These initiatives will help us establish a foundation in overseas markets and accumulate international project experience.

Looking ahead, we also plan to expand into Europe and other high-growth overseas regions to broaden our international business coverage. We will continue to provide locally deployed AI computing cluster products and solutions tailored to the core needs of different regions and customer types, thereby driving overseas business growth, enhancing our global commercialization capabilities and building a global AI infrastructure service system.

OUR PRODUCTS AND SERVICES

We provide AI computing cluster products under the *Galaxy* brand to meet customers’ evolving demand for high-performance computing across a diverse range of AI tasks. Leveraging our extensive expertise and practical experience in powering AI computing clusters, we also offer a suite of AI computing cluster operation services under the *Galaxy Services* brand, enabling a broader range of customers to access our technical know-how and high-quality computing services. Our offerings cover the full lifecycle of AI infrastructure, from software and hardware development, architecture design, and device and system deployment to performance tuning, fault detection and ongoing operations. The following table sets forth our revenue by business line for the periods indicated:

	Period from February 20, 2023 to December 31, 2023		Year ended December 31,			
			2024		2025	
	RMB	%	RMB	%	RMB	%
(RMB in thousands, except for percentages)						
AI computing cluster products — <i>Galaxy</i> . . .	31,802	100.0	307,717	94.8	436,624	83.9
AI computing cluster operation services — <i>Galaxy Service</i>	—	—	16,955	5.2	83,637	16.1
Total	31,802	100.0	324,672	100.0	520,261	100.0

As a company with full-stack AI computing cluster product and service capabilities, we can undertake end-to-end product and service delivery across the full lifecycle of AI computing clusters. Depending on customer requirements, such end-to-end delivery may take the form of a full cluster deployment, or a more targeted deployment involving selected hardware and software components together with cluster-level networking and convergence services. In addition, we may, in certain cases, build on a customer’s existing hardware infrastructure and provide software and performance tuning services. Regardless of the delivery model, our goal and responsibility is to deliver improved AI

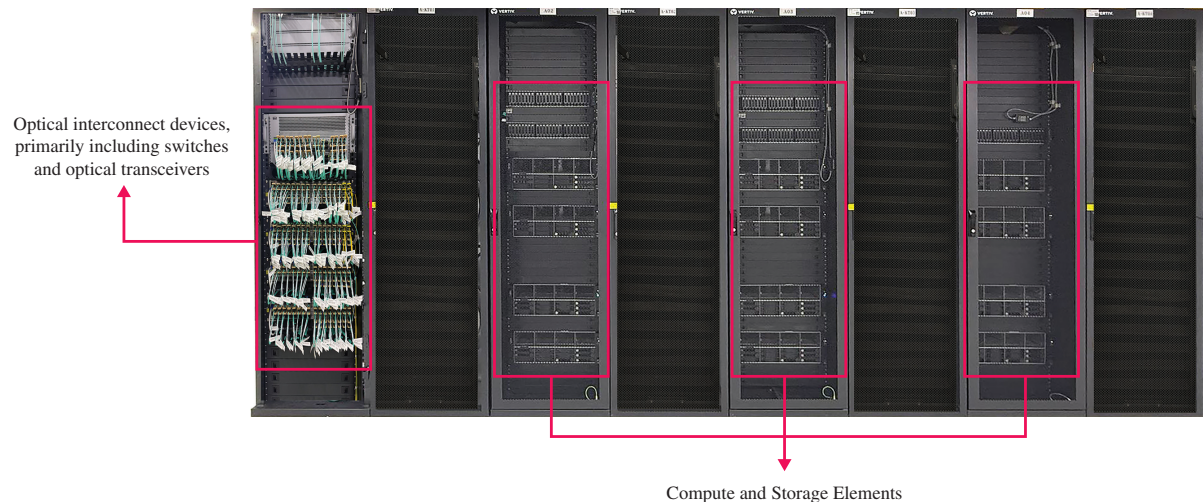
BUSINESS

computing cluster performance through our products, technical expertise, and practical know-how. *Galaxy Services* help customers address challenges such as fragmented computing resources, high operations and maintenance requirements and inefficient vendor collaboration. *Galaxy Services* also reduce the need for customers to maintain a large, dedicated operations team, while achieving refined management and cost optimization.

As of December 31, 2025, we had completed 66 projects under our AI computing cluster product business, including two ultra-large-scale AI computing clusters built at 10,000-GPU scale each. As of the Latest Practicable Date, we provided AI computing cluster operation services with guaranteed SLA with respect to an aggregate computing capacity of over 34,000 PFLOPS, including two 10,000-GPU AI computing clusters.

Our AI Computing Cluster Products

An AI computing cluster is a converged AI infrastructure designed to support large-scale AI workloads. It uses thousands of GPUs to build a distributed system that delivers AI computing power as a form of productive capacity. Through system-level engineering, an AI computing cluster enables deep coordination among compute, networking, storage and software elements, allowing users to conveniently access high-performance AI computing resources. A typical AI computing cluster consists of AI computing devices and AI computing software. AI computing devices include compute, networking and storage elements, and AI computing software abstracts underlying hardware resources and provides training and inference services to users. The following image illustrates an AI computing cluster delivered by us:



Given the complexity of AI infrastructure, we have leveraged our deep expertise in distributed system and computer architecture, and together with close collaboration with upstream and downstream partners, we provide our AI computing cluster products under the *Galaxy* brand. We offer these products, together with our practical know-how, as an integrated part of AI computing clusters, helping customers effectively improve cluster performance, efficiency, scalability, availability and observability. Our products and solutions cover both AI computing devices and AI computing software, giving us full-stack capabilities to serve the full lifecycle of AI computing clusters.

BUSINESS

- **AI computing devices:** The AI computing devices within an AI computing cluster comprise compute, networking and storage elements. Notably, for an AI computing cluster, one of the most critical capabilities lies in effectively connecting its underlying hardware components, including compute, networking and storage devices, into a unified system. Without such interconnection and coordination, the cluster would not be able to fully realize its performance potential. We have therefore developed strong capabilities in network architecture and system interconnection, enabling these hardware components to communicate efficiently and operate in a coordinated manner across the cluster. This is essential to unlocking the overall performance, efficiency and scalability of an AI computing cluster. For the AI computing devices within our AI computing cluster products, we have self-developed the network elements — our AI computing network system, *Mercury*. Leveraging our deep understanding of AI computing clusters and close cooperation with business partners including compute hardware vendors, storage vendors and contract manufacturers, we have the capabilities to design, deploy and optimize compute and storage elements of an AI computing cluster.
- **AI computing software:** For the AI computing software within our AI computing cluster products, we have developed a number of core technologies and integrated them into our AI computing operating system, *Venus*. As the brain of our AI computing cluster products, *Venus* is composed of the driver layer, kernel layer and user layer. The driver layer provides the kernel layer with unified interfaces for accessing distributed devices. The kernel layer provides the user layer with core runtime components required for AI workload processing. The user layer provides optimized service engines for different types of AI workloads, including pre-training, post-training, inference, and agentic workloads.

Built on a modular model, our AI computing cluster offerings combine a broad range of readily available hardware options and procurement sources, self-developed software offerings and a mature library of design and deployment solutions for mainstream application scenarios and customer needs. This modular approach enables us to assess each customer’s computing scale, application scenarios and performance requirements, define the overall cluster configuration, and select suitable and cost-effective combinations of AI computing cluster device and software components.

AI Computing Devices

AI computing devices comprise compute, networking and storage elements, which are turned into a unified system that together form the underlying hardware infrastructure for executing large-scale AI workloads:

- **Compute elements** typically include GPU servers, and provide the core processing capacity required for large-scale AI workloads. The performance of GPUs define the upper bound of cluster performance.
- **Networking elements** typically include switches, optical transceivers and RDMA NICs, as well as network operating system. They connect the compute elements through high-bandwidth, low-latency interconnects, enabling distributed communication that underpin AI workloads such as large-model training and inference. The performance and architecture of the networking elements directly determine the efficiency, throughput and cost performance of the entire AI computing clusters.
- **Storage elements** typically include SSDs or HDD servers, as well as distributed storage systems. They supply the data foundation of an AI computing cluster, supporting the storage and access of large-scale datasets required for AI workloads, including model parameters, training data and intermediate results.

BUSINESS

Compute

With full-stack AI computing cluster capabilities, for the compute elements, our role principally lies in designing and configuring the compute architecture for each customer’s intended AI workloads. We take into account factors such as model type, scale requirements and application scenarios, in order to achieve an appropriate balance between performance and cost. This includes defining the appropriate mix of accelerators, determining how compute nodes should be organized and interconnected, and planning a scalable structure that can accommodate future workload growth. In determining compute composition for customers’ target AI workloads, we evaluate factors such as the computing power of GPUs, node interconnection topology, and memory and power consumption constraints. Through our compute-element design, as well as hardware selection and convergence capabilities, our AI computing cluster products help customers establish a computing foundation that aligns with their AI workloads while delivering strong performance and efficient resource utilization.

Our compute-element architecture design supports heterogeneous compute configurations by selecting and combining GPUs, CPUs and specialized accelerators in a flexible manner according to different workload requirements. We have designed and delivered compute setups for a wide range of model and workload types, including Dense, MoE and multimodal training, as well as high-throughput inference. Our AI computing cluster products are currently compatible with mainstream compute hardware from both international and domestic vendors, thereby giving customers greater flexibility to configure and scale AI computing clusters under different technical and regulatory requirements.

AI Computing Network System — Mercury

Networking is critical for an AI computing cluster, as it determines how effectively computing resources communicate, coordinate and operate as a unified system rather than as a simple collection of individual GPUs. Its performance has a direct impact on data transmission efficiency, task synchronization and overall cluster utilization, and is therefore a key factor in whether an AI computing cluster can deliver high and stable performance at scale. Recognizing its importance, we have made sustained investments in this area and developed substantial proprietary technologies and R&D results.

For the networking elements of our AI computing cluster products, we provide holistic solutions, covering network design and deployment, performance tuning, as well as ongoing monitoring, operations and maintenance. In the design and deployment stage, we build high-performance network architectures for AI computing clusters, covering key components such as switches, optical transceivers and cabling, as well as component interconnection technologies. We then tune network performance for different AI workloads and optimize data exchange and synchronization across nodes to improve efficiency of large-scale training and inference. Through these capabilities, we aim to deliver network infrastructure with high stability, low latency and high availability for large-scale AI workloads, with an average annualized SLA of 99.97%.

Elastic design is also a core consideration in cluster design. For large-scale computing clusters, the key bottlenecks in scaling from dozens of accelerators to tens of thousands of accelerators lie in network interconnection and cluster-level reliability. To address these challenges, we optimize network architecture through modular resource networking, standardized communication boundary definitions and fine-grained fault isolation domain design. These capabilities help address the challenges of large-scale node interconnection and enable seamless cluster scale-out without reconstructing the underlying network topology. At the same time, built on this optimized network foundation, our clusters can support dynamic scheduling and centralized management of heterogeneous resources, as well as cross-regional training collaboration, elastic inference scheduling and on-demand allocation of computing resources across departments. This ensures stable network performance and continuous, reliable business operations as AI computing clusters continue to scale.

To meet the networking requirements of AI computing clusters for high performance, low latency, stability and scalability, we developed our proprietary AI computing network system, *Mercury*, comprising a network operating system, *MercuryNOS*, and hardware components, including switches, optical transceivers and RDMA NICs, to enable efficient and reliable communication across the cluster.

BUSINESS

MercuryNOS

Mercury operates with support from our proprietary network operating system, *MercuryNOS*, which serves as an underlying software layer for the coordination and operations of its hardware components. *MercuryNOS* is intended to improve bandwidth utilization, traffic distribution efficiency and transmission stability under high-concurrency workloads, while also providing enhanced monitoring visibility to support AI performance tuning and fault diagnosis. Its key features include the following:

- **Dynamic traffic balancing:** *MercuryNOS* supports a range of traffic load-balancing mechanisms and can select the most suitable approach based on network conditions and the characteristics of the switching chips, including adaptive routing or flowlet-based dynamic traffic balancing. This helps make more efficient use of available bandwidth across the network topology and reduces localized congestion caused by uneven traffic across multiple links.
- **Enhanced congestion management:** *MercuryNOS* incorporates advanced congestion detection and control mechanisms designed to reduce packet loss and transmission interruptions. By managing network traffic and buffering more effectively, it helps maintain communication stability and reduce performance loss caused by repeated data transmission.
- **Millisecond-level monitoring visibility:** AI workloads often involve sudden spikes in network traffic. To support performance tuning and fault diagnosis in these scenarios, *MercuryNOS* works with the switching chips to capture traffic conditions at intervals as short as one millisecond. By comparison, the minimum monitoring interval supported by advanced international AI computing network systems is typically 100 milliseconds.

MercuryNOS is deployed on our self-developed switches and has also been successfully adapted to mainstream international white-box switches.

Mercury Hardware

Hardware components of *Mercury* consist of our self-developed switches, as well as optical transceivers and RDMA NICs, which together provide the physical interconnect foundation for AI computing clusters.

- **Switches:** Our self-developed AI computing switches provide high-capacity, low-latency interconnection for AI computing clusters. Each switch supports 400G connectivity and delivers an aggregate switching capacity of 51.2 Tbps with a packet forwarding rate of 21.6 Gbps. This helps the network handle intensive, highly parallel data exchanges across thousands of GPUs without becoming a performance bottleneck. The switches support deployment in both centralized and distributed environments and can scale with different computing cluster sizes as computing demand increases. They are designed to maintain stable and reliable network performance under demanding AI workloads, helping reduce packet loss, support smoother data transmission during heavy traffic, and improve the overall stability of cluster operations. The switches also include real-time monitoring and telemetry functions, which provide visibility into network conditions and support more effective traffic management and coordination with our software systems. The following image illustrates our AI self-developed computing switch:



Our self-developed AI computing switch

BUSINESS

- **Optical transceivers:** Our 400G/800G high-speed optical transceivers are integrated with our in-house developed technologies to perform real-time bit error rate monitoring, estimate remaining service life of optical transceivers, provide high-speed optical connectivity for AI computing clusters, and support stable data transmission across network links. They are designed to meet the bandwidth and reliability requirements of large-scale AI workloads. Their telemetry functions provide real-time visibility into the health and stability of underlying physical links and enable early detection of potential abnormalities and intelligent analysis of interconnect faults, thereby supporting more proactive operations and maintenance.
- **NIC:** Our RDMA NIC, integrated with our in-house developed congestion control algorithm, provides high-bandwidth, low-latency connectivity for data to be transferred in and out of each compute node. Each NIC supports dual 200 Gbps ports, delivering aggregate bandwidth of 400 Gbps with overall latency of less than 10 microseconds. Our RDMA NICs enable data to be transferred directly between compute nodes with minimal CPU involvement, which helps reduce communication delay and improve overall computing efficiency in AI workloads. They also support congestion control tuning for different workload conditions and hardware-level encryption to improve network reliability and security.

Storage

For the storage element, our role principally lies in designing the storage configuration for each customer’s intended AI workloads. We determine the storage hardware setup, capacity and performance parameters, and organize the storage elements to work in coordination with the compute and networking elements. This helps support the bandwidth, latency and I/O performance required for AI workloads. The storage element is also designed to support disaster recovery and flexible capacity scaling, helping maintain stable and continuous cluster operations as workload demands evolve.

AI Computing Operating System — Venus

In large-scale AI computing clusters, hardware serves as the physical foundation that carries computing power, while AI computing software acts as the central orchestrating layer that unlocks hardware potential and drives the efficient, coordinated operation of the entire cluster. It is responsible not only for abstracting and scheduling vast and heterogeneous hardware resources, but also for enabling seamless coordination across compute, networking and storage elements, allowing tens of thousands of GPUs to operate as a unified system and fully realize the performance potential of large-scale computing power. Historically, the industry has often relied on fragmented open-source components or multi-vendor software stacks assembled on an ad hoc basis to build cluster software systems. Such approaches not only limit the ability to achieve system-level, cross-layer optimization, but also give rise to a number of challenges, including complex hardware adaptation, high operation and maintenance costs and inefficient resource coordination. To address these challenges, we have unified our self-developed core software technologies into our AI computing operating system, *Venus*. It enables comprehensive management and control of the entire cluster, overcoming the performance limitations associated with fragmented components while delivering a consistent user experience and standardized interfaces. This approach simplifies the deployment and operation of complex large-scale AI clusters, making them more efficient, reliable and easier to use.

Venus comprises a driver layer, a kernel layer and a user layer. The driver layer provides the kernel layer with unified interfaces for accessing distributed devices. The kernel layer provides the user layer with core runtime components required for AI workload processing. The user layer provides optimized infrastructure services for different types of AI workloads, including pre-training, post-training, inference, and agentic workloads. Through these layers, *Venus* enables AI computing clusters to operate as a supercomputer comprising tens of thousands of GPUs, capable of achieving efficient and stable operation over time.

BUSINESS

Venus boasts the following key features:

- **Comprehensive functionality:** *Venus* provides a complete set of foundational system capabilities for AI computing cluster operation, including scheduling, communication, caching, isolation, monitoring, fault diagnosis, recovery and workload acceleration, thereby supporting both AI workload execution and long-term cluster stability.
- **User-friendly interaction:** *Venus* provides intuitive, visualized and concise user interfaces through which customers can manage and monitor the AI computing cluster without interacting directly with underlying hardware and system components.
- **Strong adaptability:** *Venus* supports a range of AI workloads and application scenarios and can adapt to different models, workload patterns and heterogeneous hardware environments through its underlying capabilities in component orchestration, task scheduling and abstraction.

Driver Layer

The driver layer is mainly responsible for compilation and abstraction. It is the software layer that works most directly with underlying hardware and translates general system instructions into commands that specific hardware components can execute. Its main role is to shield upper-layer applications from differences across hardware components, so that they can use computing resources through unified interfaces without needing to handle the technical details of the underlying hardware. This makes hardware resources easier to abstract, manage and coordinate.

Within the driver layer of *Venus*, key technology components mainly comprise CUDA, Kubernetes and our self-developed *VenusDSL*. CUDA serves as the software interface for GPU computing power and provides parallel computing APIs for a wide range of large model training and inference workloads. Kubernetes abstracts distributed heterogeneous computing resources into a unified resource pool, which can then be allocated to various large model training and inference tasks, helping ensure the stable operation of distributed workloads. Our self-developed *VenusDSL* is a proprietary GPU computing interface written in Python, designed to enable efficient parallel computation for large model training and inference. Its key optimization mechanisms include:

- ***Overlapping computation and communication to accelerate inference:*** In large model inference, communication operations such as AllReduce in tensor parallel (TP) groups and token distribution and aggregation in expert parallel (EP) groups are required. By processing two batches in parallel, the system overlaps computation for one batch with communication for another, effectively hiding communication latency. To avoid GPU resource contention that could degrade performance, communication tasks are offloaded to CPUs, further improving inference performance.
- ***Operator optimization for Prefill and Decode:*** The Prefill stage involves compute-intensive large matrix operations, whereas the Decode stage is memory-intensive and focuses on single-token attention computation, with significantly different computational characteristics. We design dedicated operators for each stage and optimize their performance through techniques such as warp-level specialization, pipeline parallelism and custom kernels implemented with Triton, reducing overhead and improving overall efficiency.
- ***Operator fusion to reduce memory overhead:*** The Decode stage is constrained by memory bandwidth. We apply operator fusion to reduce CPU overhead and unnecessary memory access. Intermediate tensors are passed directly to subsequent computation without being written to and read from high-bandwidth memory (“HBM”), thereby improving efficiency. The Triton programming model is used to achieve single-kernel fusion, improving memory access efficiency.
- ***Efficient MoE token routing and reordering:*** During inference in MoE models, tokens need to be reordered based on Top-K routing decisions. We implement a fused Triton operator that integrates Top-K routing with token placement calculation and performs token reordering in parallel. The entire process is executed within a single large-scale parallel kernel, enabling efficient execution while allowing flexible trade-offs between resource usage and computational performance.

BUSINESS

Kernel Layer

The kernel layer serves as the central control system of *Venus*. Built on top of the driver layer, the kernel layer is responsible for managing critical system resources and supporting the stable operation of AI computing clusters. Based on various foundational frameworks such as PyTorch, it provides the core runtime environment for AI workloads to run efficiently and reliably. In *Venus*, the kernel layer’s most critical components include: the high-speed collective communication library (*VenusCL*), the cache input/output (*VenusIO*), the multi-queue task scheduler (*VenusCore*), the resource and task isolation component (*VenusBox*) and the multi-dimensional monitoring component (*VenusNOC*).

- **Communication (*VenusCL*):** The communication component of *Venus* is supported by our self-developed *VenusCL*, a unified collective communication library, and is the foundation for distributed large-model workloads. It enables data to be synchronized efficiently and accurately across multiple compute nodes, which is critical to the performance of training and inference tasks. Compared with commonly used industry solutions, *Venus* delivers better synchronization performance through mechanisms such as deep synchronization of computing and communication and zero-copy data transfer.
- **Caching (*VenusIO*):** The caching component of *Venus* is supported by our self-developed *VenusIO*. Caching is one of the key factors affecting how efficiently AI workloads run in an AI computing operating system. In model inference, cache is used to reduce redundant token computation and accelerate autoregressive generation, making it one of the most important means of improving inference performance. *VenusIO* employs large-capacity SSD caching and GDS technology to address a common industry bottleneck in long-context inference, where KV Cache is limited by GPU memory or RAM capacity.
- **Scheduling (*VenusCore*):** The scheduling component of *Venus* is supported by our self-developed *VenusCore* and enables our AI computing operating system to assign different AI workloads to suitable GPU compute nodes through Kubernetes runtime interfaces. Beyond basic functions such as model and data management, it also supports topology-aware scheduling, multiple resource selection strategies and automatic elastic scaling, improving resource utilization across large-scale AI computing clusters and dynamically adjusting resources in response to changes in inference demand.
- **Monitoring (*VenusNOC*):** The monitoring component of *Venus* is supported by our self-developed *VenusNOC* and serves as the “health-check center” and “dashboard” of an AI computing cluster. It collects data across compute, networking and storage resources to support performance analysis, fault diagnosis and resource scheduling. Through proactive operation and maintenance capabilities, such as optical link quality analysis and real-time online task performance analysis, as well as millisecond-level traffic monitoring, it ensures the efficient and stable operation of AI computing clusters.
- **Isolation (*VenusBox*):** The isolation component of *Venus* is supported by our self-developed *VenusBox* and helps ensure that multiple tenants and workloads can share resources within an AI computing cluster without interfering with one another. It supports resource isolation through Kubernetes-based quota and node resource management, as well as GPU virtualization and partitioning. We have also developed more fine-grained, lightweight and high-concurrency sandbox isolation technology for reinforcement learning and agentic tasks that require large numbers of virtualized runtime environments, significantly improving task execution efficiency and AI computing cluster resource utilization.

BUSINESS

The key functions of the kernel layer include:

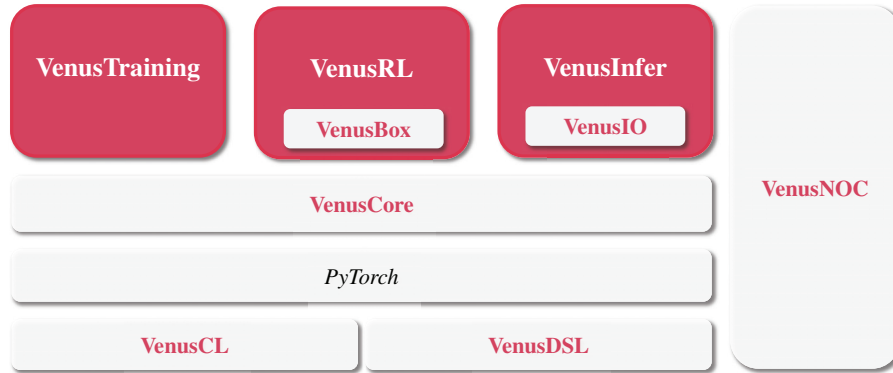
- **Topology-aware workload placement:** *Venus* deploys workloads based on inter-node network topology and intra-node connectivity, including the interconnection of GPUs and RDMA NICs, to optimize data synchronization paths across nodes, alleviate bottlenecks and improve resource utilization. This helps reduce idle time and significantly improve MFU across ultra-large-scale computing clusters. In customer deployments reduced idle GPU time and lowered compute cost per training run. *Venus* also supports multiple scheduling modes to accommodate different workload requirements more flexibly.
- **Cross-region compute scheduling:** *Venus* supports unified management and scheduling of AI computing cluster resources distributed across different locations. By pooling available resources and routing workloads based on availability, latency and cost, it helps customers make better use of idle capacity within their own deployments. This improves resource utilization, reduces delay and response time, and supports multi-tenant environments and diverse industrial AI use cases. In customer deployments, *Venus* improved cross-team collaboration efficiency by up to 45%.
- **Multi-tenant workload isolation and scheduling:** *Venus* allows multiple teams, business units and external users to share the same AI computing cluster while reducing interference across workloads. It supports this through resource quotas, priority-based scheduling, workload isolation controls and containerization. This helps provide more predictable resource availability and service quality for each tenant, while reducing resource contention and unfair scheduling in shared environments.
- **Fault-tolerant task continuity:** *Venus* continuously monitors network traffic, device conditions and workload execution to detect faults and performance issues. When a data synchronization link issue occurs, it switches in real time to a pre-established backup path to ensure uninterrupted task execution. When a node issue arises, it isolates the affected nodes or GPUs, shifts workloads to backup resources and uses checkpoints to support faster recovery. Through this design, MTTR was reduced to within minutes and recovery time from training interruptions was shortened by up to 70%.
- **Task-focused monitoring and fault management:** *Venus* supports monitoring and fault management across the entire AI task execution process. Before important tasks begin, it can perform cluster-wide system checks covering node health, network stability, storage throughput and GPU conditions to confirm resource readiness and identify potential risks in advance. During task execution, *Venus* continuously monitors task status, GPU health, network performance and storage responsiveness to detect abnormalities early and reduce the risk of long-running tasks failing after extended computation. When issues arise, *Venus* uses its expert knowledge base and analysis algorithms to identify root causes and provide remediation suggestions. *Venus* also supports automated and manual recovery mechanisms, including checkpoint-based recovery, node isolation and rapid hardware replacement workflows, to minimize disruption and restore system capacity efficiently.

User Layer

At the user layer, the core capabilities of *Venus* are designed to support the execution of a wide range of AI workloads. Although different AI applications, such as model training, autonomous driving and embodied intelligence, have distinct characteristics, the underlying infrastructure can generally be categorized into pre-training, post-training, inference and agent-based applications. While the service engines for pre-training, post-training and inference differ in their specific task execution workflows, their overall architectural frameworks and core components are largely consistent.

BUSINESS

Training and Inference Service Engine



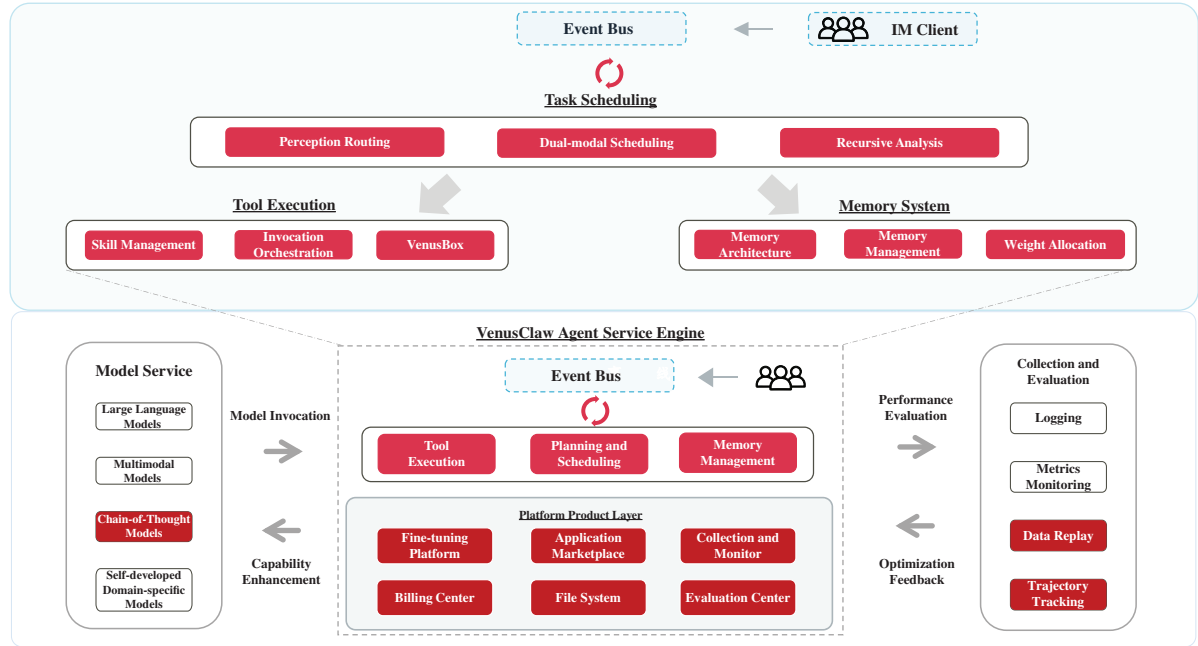
VenusCore serves as the unified entry point of our AI computing clusters, dynamically scheduling pre-training, post-training and inference workloads to appropriate resources based on cluster conditions, and orchestrating three dedicated service engines, namely *VenusTraining*, *VenusRL* and *VenusInfer*. Built on top of *VenusCore*, these engines provide foundational capabilities for pre-training, reinforcement learning-based post-training and inference, respectively. All three engines are supported by shared kernel components, including *VenusCore* for resource scheduling, *VenusCL* for collective communication, *VenusDSL* for operator optimization and *VenusNOC* for monitoring and operations. In addition, *VenusBox* provides dedicated sandbox execution environments for reinforcement learning workflows in *VenusRL*, while *VenusIO* supports *VenusInfer* through hierarchical caching to enhance inference efficiency.

- ***VenusTraining***: The pre-training service engine of *Venus* supports pre-training of various large models. Pre-training tasks are mainly used for the training of foundation models, including Dense models, MoE models, VLA models and world models, for advanced application scenarios such as embodied intelligence and autonomous driving. Such workload is characterized by extremely large model sizes, massive training datasets and highly intensive distributed computation, and typically require thousands or even tens of thousands of GPUs, with high demands on distributed coordination and robust fault monitoring and tolerance. Our *VenusCL* high-performance collective communication, *Mercury* high-performance network and *VenusNOC* components effectively address these requirements.
- ***VenusRL***: The post-training service engine of *Venus* supports various post-training tasks conducted through fine-tuning or reinforcement learning, such as fine-tuning of vertical industry models in sectors including healthcare and finance, or model quantization and distillation using reinforcement learning approaches. Post-training tasks typically involve two closely connected stages. First, a pre-trained foundation model is further adapted using a relatively small volume of task-specific data through fine-tuning. Second, the fine-tuned model is used to perform inference on designated datasets, and the outputs are evaluated against corresponding validation data. Through repeated rounds of fine-tuning, inference and result validation, the model is progressively optimized to achieve the desired inference quality and application performance. As post-training involves a significant number of validation processes, it relies heavily on lightweight, high-efficiency and high-concurrency isolation environments. Our *VenusRL*, in coordination with *VenusCore* through trajectory sampling and scheduling, together with lightweight and high-concurrency sandbox capabilities of *VenusBox*, provides efficient post-training infrastructure.

BUSINESS

- VenusInfer:** The inference service engine of *Venus* supports inference tasks for various models. Inference tasks are primarily used to support the deployment and operation of trained models across a range of real-world inference scenarios, including low-latency, high-concurrency real-time interactions, long-context and logic-intensive decision-making and planning, content generation and multimodal inference. To address the characteristics of such workloads, including significant traffic fluctuations and heterogeneous load patterns, a layered architecture with Prefill-Decode decoupling is adopted, computation and caching, it delivers high elastic, low-latency inference services, effectively accommodating fluctuating request traffic and intra-day load variations.

Agent Service Engine



VenusClaw, as the agent service engine of our AI computing operating system, serves as an infrastructure supporting the large-scale deployment and stable operation of AI agents. Unlike traditional inference services, *VenusClaw* not only provides computing capacity for model inference, but also delivers orchestration, execution and optimization capabilities across the entire agent workflow, supporting a closed-loop process from task understanding to tool invocation, result validation and iterative refinement. The control plane of *VenusClaw* enables declarative configuration of agents, version management and health monitoring. At the same time, *VenusClaw* provides a unified entry point through its gateway and policy modules, handling request routing and distribution, traffic management, security authentication and protocol adaptation.

VenusClaw provides advanced reasoning and planning capabilities, including task understanding and decomposition, chain-of-thought reasoning, long-horizon hierarchical planning and reflective validation, enabling effective decision-making for complex tasks. It also supports efficient interaction between agents and external environments through capabilities such as tool registration and discovery, dynamic scheduling and orchestration, and multi-modal perception. In addition, *VenusClaw* enables task scheduling, multi-agent collaboration and memory enhancement, supporting cross-session contextual continuity and knowledge accumulation.

BUSINESS

Building on this foundation, *VenusClaw* establishes a layered agent platform architecture. Leveraging the core capabilities of the kernel layer of *Venus*, it provides comprehensive support for agents and enables unified orchestration of multiple model types, including large models, multimodal models and domain-specific fine-tuned models. Through a unified model access layer, it supports load balancing, streaming output and fault-tolerant fallback mechanisms. At the middle layer, *VenusClaw* provides key capabilities including agent scheduling, low-code orchestration, high-performance secure sandboxing, memory management, token-based billing, API gateway services and an event bus. In particular, its millisecond-level secure sandboxing mechanism provides an isolated and controlled execution environment for code execution and tool invocation. At the upper layer, *VenusClaw* supports comprehensive evaluation across performance, security and cost dimensions. Combined with trajectory capture and replay capabilities, it enables closed-loop optimization of agent behavior and supports continuous iteration of business applications.

The *VenusClaw* architecture is well aligned with the core characteristics of agent-based applications. To address the iterative nature of agent tasks and fluctuating traffic patterns, *VenusCore* enables dynamic resource allocation through elastic scaling and unified scheduling of synchronous and asynchronous tasks, significantly improving system throughput and stability in high-concurrency scenarios. At the same time, through standardized interfaces and full-stack observability, *VenusClaw* substantially lowers the barriers to agent development and operations, enabling enterprises to rapidly build, deploy and operate industry-specific agent applications and fully unlock the productivity gains from automation.

We have deployed converged AI computing clusters embedded with office agents and the *VenusClaw* agent service engine for enterprise services, significantly improving employees’ daily work efficiency. Tasks such as analytical report preparation, which previously required up to one week of manual effort, can now be completed within half a day. Our converged AI computing clusters are built on open-source models and fully domestic hardware components, including CPUs, GPUs, NICs and switches, to meet enterprises’ requirements for private deployment.

User Interfaces

Venus provides a suite of user-friendly, intuitive graphical interfaces for managing AI workloads, including training, inference and agentic tasks, as well as monitoring the status of AI computing clusters and the real-time execution status of individual tasks.

- **Pre-training task management:** *Venus* offers a highly configurable task management framework that allows users to flexibly customize the training process for different models and workloads. Specifically, during task initialization, users can specify the number of GPUs to be allocated based on model scale, configure shared storage paths to efficiently manage large-scale training datasets and select from a range of container images to adapt the environment to different models. In addition, *Venus* provides topology-aware intelligent scheduling, which can be configured through a dedicated interface to optimize resource allocation across the computing cluster. Users may also enable the high-efficiency *VenusCL* unified collective communication library and activate backup channels for multi-layer fault tolerance by setting the relevant environment variables within the task configuration interface. Furthermore, both active and passive monitoring mechanisms are enabled by default across the system to provide continuous oversight of the cluster. Together, these capabilities help ensure the efficient and stable operation of ultra-large-scale AI computing clusters and allow *Venus* to meet the demanding requirements of pre-training workloads. The following screenshot illustrates the pre-training task scheduling function of *Venus*:

BUSINESS

资源配置

实例配置 * MPI PyTorch Simple CPU Deployment

☐ 启用MIMO模式 ☐ 整机调度 ☐ vGPU

Roles	实例数量	资源规格	GPU数量	CPU(核)	内存(GB)
master	1	选择资源规格	1		1500
worker	1	选择资源规格	1		1500

+ 添加角色 (0/6)

优先级 * 6 5 4 3 2 1

☐ 数值越大, 优先级越高

回收策略 ☐ 不可回收

自动重试 ☐

存储 * + 存储卷 + 临时卷 + S3

☐ 存储卷, 临时卷, S3: 每个任务独立不能重复。
☐ 存储卷: 可通过“库→存储”页面管理的持久卷。
☐ 存储路径: 持久卷在容器内部挂载的路径。
☐ 临时卷: 由 EmptyDir 提供支持的临时卷, 其生命周期与 Pod 绑定。
☐ 主机路径: 由 Host Path 提供支持的卷, 用于将宿主节点文件系统中的目录或文件挂载到 Pod 内。

环境配置

镜像 * 预置镜像 自定义镜像

请选择镜像地址

工作目录 * 请输入命令执行目录

命令模式 * 自定义模式 shell模式

入口命令 *

```
1 /bin/bash
2 -c
3 sleep inf
```

① 自定义模式可输入 Linux 命令 (参数以空格分隔, 按回车换行); Shell模式支持直接粘贴完整的 Shell 脚本。

Setting task type and GPU parameters

Customizing environment variables

Pre-training task scheduling

- Post-training task management:** *Venus* is designed to address these post-training requirements through an integrated and configurable platform capability. *Venus* enables customers to select an appropriate pre-trained foundation model from the model repository as the basis for post-training, and supports multiple mainstream finetuning approaches, including LoRA and RLHF, together with parameter configuration tailored to the selected method. It also allows customers to allocate the computing resources required for each task and configure customized environment variables to support specific runtime requirements and functional options. In addition, *Venus* supports the coordinated management of the inference dataset and the corresponding validation dataset required for post-training, thereby providing an integrated framework for fine-tuning, inference, validation and iterative optimization. Through these capabilities, *Venus* helps customers improve the efficiency, flexibility and reliability of post-training workflows and better achieve the target model performance for downstream applications. The following screenshot illustrates the post-training task scheduling function of *Venus*.

BUSINESS

< 创建模型微调任务 - 常规模式

基础信息

任务名称 *

描述

模型仓库 * +

训练配置

微调方式 *

超参数	参数名	参数值
截断长度	truncate_len	1024
学习率	learning_rate	0.000001
日志间隔参数	logging_steps	5
最大样本数	max_samples	1000
NEFTune内建噪声参数	neftune_noise_alpha	0.00
训练轮数	num_train_epochs	1.0
批处理大小	per_device_train_batch_size	1
学习率预热比率	warmup_ratio	0.00

资源配置

实例配置 * ☐ 应用MIG模式 ☐ 整机调度 ☐ vGPU

资源规格 GPU数量 CPU核数 内存(GB)

环境变量 [+ 新增环境变量](#)

数据集配置

训练集 * +

验证集 * +

Choosing foundation model

Choosing fine-tuning method and configuring fine-tuning parameters

Designating post-training and validation dataset

Post-training task scheduling

- Inference task management:** *Venus* provides a comprehensive set of configuration and management capabilities that enable dynamic and efficient allocation of inference computing resources. Customers can select a trained model from a centralized model repository for deployment as an inference service, and choose an inference framework that best suits the requirements of the inference task. Customers can also configure whether to adopt distributed inference, which allows a single inference instance to leverage multiple compute nodes working in coordination. In addition, customers can specify the number of inference replicas to be deployed. Since a single inference instance can only handle a limited volume of inference requests, running multiple replicas in parallel enables proportional scaling of inference throughput. Critically, *Venus* supports elastic scaling that dynamically adjusts the number of active inference instances in response to the fluctuating inference requests, deploying more instances during peak daytime periods to accommodate higher request volumes and scaling down during off-peak nighttime hours to reduce resource consumption and avoid unnecessary waste. Through these capabilities, *Venus* enables customers to efficiently deploy and manage inference services while optimizing resource utilization across varying demand patterns. The following screenshot illustrates the inference task scheduling function of *Venus*.

BUSINESS

环境配置

分布式推理

每个实例都在多台机器上进行部署。例如需要 4 个实例，单实例部署的机器数为 2，则一共需要 4x2=8 台机器来部署该服务

推理框架

请选择推理框架

模型仓库

请选择模型仓库

环境变量

+ 新增环境变量

资源配置

实例配置

启用 MIG 模式

整机调度

vGPU

资源规格

GPU 数量

CPU (核)

内存 (GB)

副本数

+ 存储卷

+ 临时卷

+ S3

高级配置

扩缩容

选择方式

规则

Configuring inference model

Designating multiple inference replicas to meet throughput targets

Configuring elastic scaling for changing traffic conditions

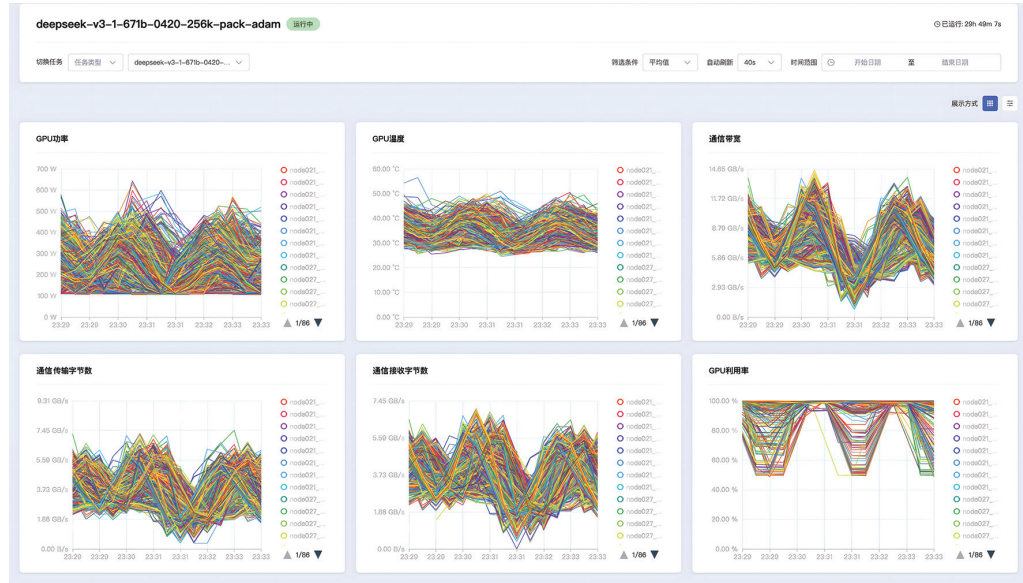
Inference task scheduling

- Visualized monitoring:** Venus enables monitoring of the overall resource status of AI computing clusters, as well as highly visual real-time monitoring of the resources utilized by each task, such as GPU utilization and temperature, helping users assess performance and identify abnormalities during workload execution.



Cluster information overview

BUSINESS



Task runtime status monitoring

Case Studies

Case Study A

In 2024, we deployed and delivered an AI computing cluster to a domestic large model company to support its model training, inference and application deployment needs. As the customer had established a dedicated team for large model development, it required a tailored AI computing cluster with sufficient computing capacity, system stability and scalability to support its evolving development roadmap.

At the initial stage of the project, we worked closely with the customer to understand its specific requirements, including the parameter scale of the models under development, expected training time cycle, anticipated usage patterns of the cluster, number of teams requiring access, isolation requirements among different workloads, and preferred computing resource sharing model. Based on such requirement analysis, we proposed an overall solution comprising an AI computing cluster at the 2,000-GPU scale together with the corresponding storage configuration, AI computing network architecture, data center space planning, and supporting power cabinet and cooling system arrangements for the cluster deployment. For the software element, the cluster was deployed with our AI computing operating system, *Venus*, to support unified management, resource scheduling and cluster operations. Following the customer’s approval of the proposed solution, we organized and completed the procurement, deployment and performance tuning of the relevant hardware and software elements, and delivered the entire AI computing cluster as a converged product. Upon delivery, we also undertook the ongoing operation and maintenance of the cluster to support its stable operation after deployment. During the subsequent usage of the cluster, we further assisted the customer in deploying inference services and agentic applications on the cluster through *Venus*, enabling the customer to extend the cluster’s use cases beyond model training and into downstream application scenarios. As of the Latest Practicable Date, the cluster had maintained stable operation with strong quality performance. In 2025, the cluster achieved an annualized hardware failure rate of less than 0.3% and an average annualized SLA of 99.97%.

BUSINESS

Case Study B

In 2025, we provided software and performance tuning services for an AI computing cluster at the 10,000-GPU scale used by a leading large model company. During model training, the cluster experienced significant and unexpected performance degradation, accompanied by severe fluctuations in training performance. The training process also often became unresponsive, resulting in poor overall training efficiency and rendering the cluster largely unable to operate normally for its intended workloads.

In response, we first deployed our AI computing operating system, *Venus*, on the cluster. Through its scheduling and monitoring functions, we observed that data synchronization errors occurred on servers during model training and identified the underlying cause as a program failure on a specific GPU of a server in the cluster. Based on our accumulated experience in large-scale cluster optimization and our established troubleshooting expertise, we systematically identified and replaced suboptimal GPU by running function-level online performance profiling tools, and updated the relevant server parameter configurations, and updated the relevant server parameter configurations.

Following our optimization, the customer’s cluster demonstrated stable performance in subsequent training runs. Overall training performance, as measured by MFU, improved by more than 30% as compared to the pre-optimization level.

Case Study C

In 2025, leveraging our *VenusBox* technology, we deployed our agent sandbox product for a leading large model company. The project focused on two core scenarios: large-scale post-training and agent applications. Leveraging *VenusBox*’s lightweight sandbox instances and elastic resource pool mechanism, the solution optimized across compute, networking and storage elements, and, at the high-availability deployment architecture level, enabled lossless hot switching of traffic in the event of a control-node failure.

The sandbox product is designed for agent use cases involving high-frequency calls, batch execution and real-time interactions. It adopts a scheduling architecture that supports fast local visibility into resource availability and efficient updates to resource allocation, thereby improving scheduling efficiency and cluster-level load balancing under high-concurrency conditions. Following key algorithm upgrades to the API scheduling module, the cluster load imbalance coefficient of variation decreased significantly from 0.941 to 0.021, approaching near-perfect balance. In addition, its template-based sandbox warm-up system effectively reduced latency fluctuations associated with cold starts, thereby improving first-response speed and service continuity. As a result, our product stably supported over 10,000 concurrent sandbox instances and helped the customer reduce average monthly sandbox costs by 40%, without any major system failures or service interruptions during the period.

Case Study D

In 2026, we deployed and provided ongoing operation and maintenance for an AI computing cluster with a computing capacity of 250 PFLOPs, featuring our inference service engine, *VenusInfer*, for a large model company in China. As the customer progressively rolled out AI agents, inference traffic has grown rapidly, creating a need to build a stable and cost-effective inference service to provide its self-developed model APIs. This inference service, based on the customer’s proprietary models, has been connected to OpenRouter.

VenusInfer provides inference service capabilities spanning gateway routing, load balancing, inference engine management, usage metering, node management, task scheduling, and fault migration. *VenusInfer* is optimized for the “long input, short output” workloads typical of AI agent applications. It achieves dynamic KV Cache management through our self-developed *VenusIO* and high-performance computing resource scheduling through *VenusCore*, enabling robust elasticity and high resource utilization even under significant traffic fluctuations guided design principles of cross-layer optimization and heterogeneous acceleration, we introduced two types of computing hardware options: one with high

BUSINESS

computing capacity and large cache, and another offering lower computing capacity but high GPU memory bandwidth, tailored to the prefill and decoding stages of inference, respectively. Compared to other solutions using homogeneous GPU, our software-hardware coordinated and heterogeneous-accelerated approach improved the supported inference traffic by 100% at the same level of capital expenditure.

Delivery of Our AI Computing Clusters

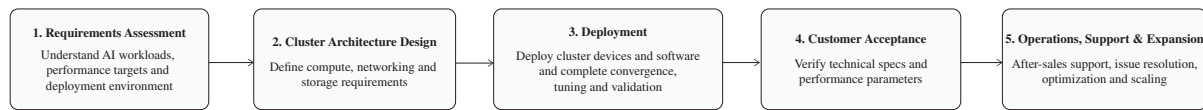
As a company with full-stack AI computing cluster capabilities, we can undertake end-to-end projects across the full lifecycle of AI computing clusters, including cluster architecture design, hardware selection, software-hardware adaptation, network interconnection, system tuning and performance validation. Built on a modular model, our AI computing cluster offerings combine a broad range of readily available hardware options and procurement sources, self-developed software offerings and a mature library of design and deployment solutions for mainstream application scenarios and customer needs. This modular approach enables us to assess each customer’s computing scale, application scenarios and performance requirements, define the overall cluster configuration, and select suitable and cost-effective combinations of AI computing device and software components.

Depending on customer requirements and existing infrastructure, our end-to-end product delivery may take the form of a full cluster deployment, or a more targeted deployment involving selected hardware and software components together with cluster-level networking and convergence services. In addition, we may, in certain cases, build on a customer’s existing hardware infrastructure and provide software optimization, network deployment and performance enhancement services. In all of these delivery scenarios, regardless of the specific scope of delivery in a particular project, we leverage our products, technical expertise and practical know-how to improve cluster-level interconnection, communication efficiency and overall system performance. This flexible delivery model enables us to meet a wide range of customer needs while maintaining responsibility for the overall availability, performance and networking effectiveness of the cluster.

During the Track Record Period, we generated all revenue under this business line from delivering and servicing AI computing clusters. We recognized revenue upon customer acceptance. In 2023, 2024 and 2025, we completed four, 10 and 52 projects under our AI computing product business, respectively, for four, 10 and 48 customers, respectively. The increase in the number of our AI computing cluster product projects during the Track Record Period was primarily driven by growing customer demand for AI computing resources and AI computing cluster-related solutions. As AI adoption continued to expand, customers increasingly required not only compute hardware itself, but also the ability to effectively connect and coordinate compute, networking and storage resources into a unified cluster that could operate reliably and efficiently. In this context, the key value of our offerings lies not merely in the particular hardware or software components delivered in an individual project, but in our ability to enable cluster-level interconnection, communication efficiency and overall system performance. Accordingly, whether a project involved software, hardware or a combination of both, and regardless of the scale of the specific delivery scope, our delivery formed an important part of enabling customers to use AI computing clusters effectively.

BUSINESS

Set forth below is a typical workflow starting from customer requirements assessment to post-delivery maintenance:



- **Requirement assessment and architecture design:** The delivery process typically begins with an initial requirements assessment, during which we work with customers to understand their intended AI workloads, target performance metrics and deployment environments. Based on this assessment, we define the overall cluster configuration, including computing, networking and storage requirements, and propose a converged cluster architecture aligned with such requirements.
- **Cluster deployment:** Following confirmation of the cluster configuration, we proceed with deployment of the AI computing cluster product. We are responsible for the deployment of all required AI computing device components, ensuring that the cluster is provisioned with the computing, networking and storage resources necessary to meet the defined workload and performance requirements. We also deploy and configure AI computing operating system, and perform product convergence, tuning and validation to ensure that the AI computing cluster operates as a unified system.
- **Customer acceptance:** Upon completion of deployment, the AI computing cluster product is subject to acceptance and verification by the customer to confirm that the delivered cluster products meet the agreed technical specifications and performance parameters. After acceptance, the AI computing cluster enters operational use in accordance with the customer’s deployment plans.
- **After-sales support:** Following product delivery, we provide after-sales and maintenance support to assist customers in operating and evolving their AI computing clusters over time. Our after-sales services include responding to customer inquiries and operational requests, supporting issue investigation and resolution, and assisting with system adjustments and optimizations as customer workloads, operating conditions or deployment requirements evolve. We also support customers in capacity planning, system expansion and configuration updates following initial deployment, enabling AI computing cluster products to adapt to changing usage patterns while maintaining stable operation.

The delivery cycle for our AI computing cluster products is approximately one month, subject to variation depending on factors such as hardware procurement lead time, cluster size, overall deployment complexity, the customer’s existing infrastructure, software-hardware adaptation requirements, network interconnection arrangements and on-site deployment conditions.

Our AI Computing Cluster Operation Services

Leveraging our extensive expertise and practical experience in designing and delivering AI computing clusters, we provide AI computing cluster operation services under the *Galaxy Services* brand. Our AI computing cluster operation services enable a broader range of customers to access our technical know-how and high-quality computing services. Powered by our self-developed AI computing network system, *Mercury*, and AI computing operating system, *Venus*, *Galaxy Services* address the limitations of traditional AI computing cluster deployments, which often focus on initial deployment but lack sustained operations support. It helps customers address challenges such as fragmented computing resources, high operations and maintenance requirements, low internal collaboration efficiency, and the burden of large upfront capital expenditure and ongoing operations costs. *Galaxy Services* also reduce the need for customers to maintain a large, dedicated operations team, while preserving data and operational security. The number of customers of our AI computing cluster operation services increased from two in 2024 to 16 in 2025.

BUSINESS

Galaxy Services boast the following key features, designed to help customers improve business agility and reduce operating expenses:

- ***Comprehensive operation support:*** Leveraging *Galaxy*’s thorough observability and elastic system architecture, *Galaxy Services* provide an operations and business adaptation framework covering hardware, systems and tasks throughout the cluster lifecycle, helping customers reduce the complexity of day-to-day cluster operations.
 - ***Fault self-recovery and proactive operations:*** *Galaxy Services* have built-in task fault-tolerant recovery and automatic hardware fault migration capabilities. Through operations agents, it conducts proactive hardware health checks and multi-dimensional, multi-scale monitoring and alerting, helping identify potential risks in advance and address issues before they lead to service interruptions. Operations agents can also automatically handle common issues such as node downtime and network jitter, achieving average annualized SLA of 99.97%.
 - ***24/7 technical expert support:*** *Galaxy Services* are supported by a dedicated expert operations team for AI computing clusters, providing round-the-clock technical support.
- ***Enhanced cost efficiency:*** *Galaxy Services* provide comprehensive cost management capabilities, helping customers track computing resource usage more precisely, optimize costs and better meet enterprise requirements.
 - ***Intelligent cost optimization:*** *Galaxy Services* can generate usage bills for different project teams, enabling customers to conduct detailed internal accounting of computing resources. By analyzing historical usage data, it can automatically identify idle resources and waste, and provide optimization recommendations to help enterprises reduce hidden computing costs.
 - ***Compatibility and adaptability:*** *Galaxy Services* support the integration and reuse of customers’ existing computing infrastructure and enable progressive cluster upgrade and expansion. As customers’ AI businesses grow, they can scale cluster capacity without replacing their existing infrastructure, thereby improving long-term resource utilization and preserving the value of their infrastructure investment.

Maintenance and Technical Support Services

We offer maintenance and technical support services to help ensure the effective operation of AI computing clusters, including clusters delivered by us as well as clusters deployed by other customers that require professional operation and maintenance support. These services are supported by the underlying capabilities of *Venus*, particularly its kernel-layer functions in monitoring, fault detection, diagnosis and recovery, which enhance our ability to identify issues, analyze root causes and support timely remediation. Our maintenance services include system monitoring, routine inspection, performance optimization, troubleshooting and issue resolution, with technical support provided in response to customers’ operational needs.

BUSINESS

Case Study

We provide maintenance and technical support services for a leading research institute for its AI computing clusters with more than 20,000 GPUs. Our services included 24/7 operation and maintenance support, as well as customized software development and maintenance for fault tolerance, fault migration, intelligent operations, performance analysis and monitoring. These services improved cluster stability and service quality, reducing weekly training interruptions caused by hardware failures from more than 15 times to approximately three times, with an average annualized SLA of 99.97% in 2025. The clusters supported more than 700 users, with the largest single task using thousands of GPUs and running continuously for up to three days, across a total deployed cluster scale of more than 28,000 GPUs.

Compute Services

We provide compute services through the computing capacity of AI computing clusters designed, built and operated by us. These services are underpinned by the same technology, expertise and operational know-how embodied in our AI computing cluster products, including our AI computing device capabilities and the software functions of *Venus*, such as scheduling, resource coordination, monitoring and AI workflow adaptation. As a result, our compute services enable customers to access high-quality and efficient computing resources that are closely aligned with their AI workload requirements. This offering is particularly suited to customers seeking access to high-performance computing resources without the need to own GPUs or bear the significant capital expenditure, operational complexity and ongoing maintenance burden associated with building and operating dedicated AI computing clusters. Under this model, customers can access computing resources in a more flexible and cost-efficient manner, while we charge fixed fees based on the amount of computing resources provided. Our compute services primarily target large model companies, cloud service providers, and education and research institutions with stable demands for computing resources and services.

OUR TECHNOLOGIES

We have established a comprehensive technology system that spans hardware, communication protocols, scheduling algorithms, and system software, enabling our converged AI computing clusters to operate with high performance, reliability, and scalability across large-scale environments. Our core technologies span several key areas. At the AI computing device level, these include opto-electronic integration technologies, high-throughput switching hardware and network congestion control technologies. At the AI computing operating system level, our technologies include high-performance collective communication at the kernel layer, large-capacity caching, ultra-lightweight, high-efficiency and highly concurrent sandbox, and multi-dimensional, multi-scale active and passive monitoring. At the user layer, we also develop high-efficiency post-training frameworks. Looking ahead, we will continue to advance next-generation AI infrastructure through software-hardware co-design and cross-layer optimization, enhancing hardware programmability, software intelligence and system-level coordination. Through these efforts, we aim to deliver higher-quality AI computing cluster operation services and further strengthen our leading position in the AI computing cluster industry.

BUSINESS

AI Computing Device Technologies

At the AI computing device level, we focus on hardware and *MercuryNOS*, which provides the underlying capabilities that support the functions and performance of our AI computing network system, *Mercury*. They cover key aspects of networking hardware of AI computing clusters, together with critical features designed for AI computing cluster scenarios, including NPO-based optical interconnection, OCS opto-electronic switching technologies, and traffic congestion control. These technologies are intended to support efficient, scalable and reliable communication within large-scale AI computing clusters.

Opto-Electronic Integration Technologies

With the rapid increase in SerDes speeds of AI switching chips and the growing number and bandwidth of the scale-up interconnection links of GPUs, the commonly adopted architecture based on pluggable optical transceivers combined with high-speed SerDes PCB routing is increasingly constrained by signal integrity challenges and power limitations. Optical interconnection technologies based on NPO and CPO, together with OCS optical switching technologies, address these limitations by shortening electrical signal transmission distances and replacing electrical switching with optical switching. These approaches can fundamentally reduce system power consumption, latency and signal attenuation, and are emerging as preferred solutions for next-generation AI computing clusters.

For OCS, we have conducted technical research and validation of opto-electronic integrated architectures combining electrical switches and OCS optical switches in AI computing clusters. This includes validation of real-time traffic scheduling between optical and electrical switching paths through *Venus*, as well as comparative analysis of energy consumption. For NPO, we focus on evaluating differences among silicon photonics solutions from various vendors, including their compatibility with AI computing cluster architectures and their support for operations and monitoring capabilities. Specifically, we have achieved the following milestones:

- ***OCS opto-electronic integration:*** In collaboration with our partners, we have completed validation of an opto-electronic convergence architecture in AI computing clusters, combining MEMS-based OCS optical switches with our self-developed high-throughput switches. This has resulted in a 50.3% reduction in AllReduce time, a 47.1% increase in bandwidth utilization and a 32.1% reduction in energy consumption for large model training. The related research has been accepted for publication at OFC 2026. We have also initiated further research on silicon photonics-based OCS switching technologies.
- ***NPO:*** We have selected state-of-the-art 3.2T NPO components and commenced deployment on our self-developed 100T high-throughput switches.

Going forward, we will continue to increase our investment in silicon photonics technologies and accelerate the integration of AI computing cluster requirements into silicon photonics products, in order to better meet the evolving needs of AI computing clusters.

Switch Hardware Technologies

An AI computing cluster relies on a high-speed network to connect its distributed compute nodes into a converged system. Such high-speed networking is achieved through AI switches that provide traffic forwarding and interconnection with high bandwidth and low latency. The aggregate bandwidth of the switches determines the scale of compute nodes that the AI computing cluster can support, and therefore the cluster’s overall computing capacity. The aggregate switch bandwidth is determined by two main factors: (i) the forwarding bandwidth of each individual switch chip used in the switch; and (ii) the number of switch chips integrated into a single box switch. As the number of switch chips increases, power consumption grows proportionally, placing more stringent and predictable requirements on thermal design. Higher bandwidth per switch chip is achieved by increasing the number of forwarding lanes supported by the chip and by doubling the SerDes rate per lane. This, however, significantly increases chip power consumption. In addition, once the SerDes rate is doubled, maintaining signal integrity requires higher-performance PCB materials and shorter transmission distances in hardware design.

BUSINESS

We have developed proprietary switch hardware technologies for high-throughput AI networking. Through an orthogonal backplane design, our switch hardware supports the integration of multiple high-bandwidth switch chips to increase the aggregate bandwidth of each switch. We also address the air-cooling bottleneck caused by high power consumption through an optimized combination of heat pipes and high-density fins. Going forward, we plan to further iterate our liquid-cooling solutions to support switch hardware with even higher bandwidth and power requirements. We also adopt advanced 3.2T near-packaged optics technology to address the challenges of ultra-high-speed signal transmission and support the high-bandwidth optoelectronic interconnection performance required by next-generation AI computing clusters. At the operating system level, we use a primary-secondary architecture to manage multiple switch chips more efficiently, which helps control the substantial increase in resource requirements that may otherwise arise in multi-chip deployment scenarios.

Network Congestion Control Technologies

For RoCE-based AI computing networks, lossless networking is achieved through a feedback-based mechanism in which the data receiver detects network congestion through ECN markings in packets, sends CNP notifications to the sender, and the sender reduces its transmission rate upon receiving such CNP notifications. Under this mechanism, the receiver determines the proportion and frequency of CNP packets to be sent based on factors such as the number of ECN-marked packets received and the packet intervals within a given period of time. The sender, in turn, determines the appropriate rate reduction level based on multiple parameters, including the number of CNP packets received and the intervals between such packets. As this mechanism depends on numerous factors on both the sending and receiving sides, selecting an appropriate set of parameters is particularly important. We have developed a set of advanced congestion control technologies to address network congestion issues in large-scale distributed AI training:

- **Distributed adaptive congestion control algorithm:** We use a distributed architecture with enhanced adaptive tuning algorithms, together with traffic balancing and orchestration, to adjust traffic dynamically across the AI computing cluster. This approach optimizes system design, algorithm efficiency and traffic scheduling strategies to reduce RoCE-related packet congestion, lower communication latency and maintain high effective bandwidth utilization, thereby directly improving large-model training performance.
- **Large-model-aware congestion control (“LLM-CC”):** LLM-CC is a congestion control communication protocol designed for RoCE-based AI computing cluster networks. It supports unified traffic management both within and across data centers, including long-distance deployments across multiple AI data centers. Because RoCE is highly sensitive to congestion, especially in large-scale and long-distance deployments, LLM-CC works with the network operating system on our switches to monitor congestion and adjust traffic across the computing cluster. By dynamically controlling traffic rates and congestion responses across distributed nodes, it helps reduce the performance impact of higher latency and congestion in long-distance deployments. As a result, AI computing clusters deployed across distances of up to 30 kilometers can achieve approximately 99% of the effective computing performance of comparable non-long-distance deployments, while deployments spanning up to 1,500 kilometers can maintain approximately 95%.

Kernel Layer Technologies

Based on our deep understanding of AI computing cluster requirements and our focus on performance optimization, we have adopted a cross-layer optimization approach and carried out targeted technology development in areas such as collective communication, large-capacity caching, sandbox isolation and proactive operations.

BUSINESS

VenusIO Large-Capacity Inference Caching Technology

As inference workloads increasingly require support for longer context, demand for KV cache continues to grow. Existing KV storage and management systems, typically represented by memory-centric architectures, are constrained by limited capacity, high cost and relatively high failure rates in distributed environments.

To address these challenges, we have developed our proprietary *VenusIO* technology. *VenusIO* adopts a storage-centric KV cache management architecture, leveraging the capacity and cost advantages of local NVMe SSDs as an alternative to DRAM, while utilizing GDS technology to accelerate data transfer between GPUs and storage. At the engine level, *VenusIO* employs storage prefetching and compute-I/O overlap pipelines to mask storage I/O latency. As a result, *VenusIO* is able to achieve latency performance comparable to DRAM while significantly improving capacity and cost efficiency, and supports high-throughput computation during the prefill stage. We expect *VenusIO* to increase cross-request KV cache hit rates in the prefill stage by approximately 30% and improve throughput by approximately 20%.

VenusBox High-Efficiency Sandbox Technology

In high-concurrency agentic environments, a large number of sandbox requests may arrive at the same time. If resource information is updated with delay, too many requests may be directed to a small number of nodes, leading to temporary overload and reduced scheduling efficiency. To address this, we have developed our proprietary *VenusBox* technology and upgraded our scheduling algorithms by maintaining a more up-to-date local view of resource usage and pre-adjusting available capacity once a scheduling decision is made, rather than waiting for monitoring data to catch up. This helps prevent too many sandbox requests from being directed to the same nodes at the same time, thereby reducing load skew and herd effects caused by delayed metrics.

Furthermore, in agentic environments, many requests need a sandbox to be available before tool execution can begin. If a sandbox is created only after a request arrives, the first response may be delayed. To address this, we developed a scheduling algorithm based on local shadow state and optimistic resource update mechanisms, which reduced cluster load imbalance (measured by the coefficient of variation, or CV) from 0.941 to 0.021 in production environments. We developed a template-based sandbox warm-up framework. The system prepares a pool of ready-to-use sandbox instances in advance based on expected demand and usage patterns. When a request arrives, it can allocate a ready instance directly from the pool, which helps reduce waiting time, improve response speed and eliminate latency variability associated with cold starts. Once an instance is taken from the pool, the system automatically prepares a replacement in the background and adjusts the overall pool size as demand changes. This helps our system maintain low-latency performance during traffic spikes, while improving throughput stability and resource utilization.

VenusNOC Fault Monitoring and Proactive Maintenance Technologies

In large-scale AI computing clusters, the number of interconnection links and optical transceivers can reach tens of thousands. Optical transceivers are indispensable components in the interconnection of compute nodes, and any link or transceiver failure may cause link flapping or downtime, thereby affecting AI workloads such as training and inference running on the cluster. Based on our extensive experience in the delivery and operation of large-scale clusters, we have found that link failures involving optical modules and cables are among the principal sources of system faults.

To address these issues, we have added enhanced feature support to our optical transceivers, including indicators such as real-time bit error rate and cumulative operating time, and have established corresponding proactive detection and monitoring capabilities at the system level. This enables us to identify potential issues at an earlier stage and plan remedial measures in advance, thereby reducing task interruptions and other disruptions through a more forward-looking maintenance approach. By combining

BUSINESS

proactive operation and maintenance technologies, coordinated fault monitoring and scheduling fault-tolerance capabilities, our AI computing cluster products have achieved a monthly network jitter rate of below 0.2%, enabling AI computing clusters at the 10,000-GPU scale or above to achieve an average annualized SLA of 99.97%.

User Layer Technologies

In post-training and agent-based scenarios, reinforcement learning has become a frequently used foundational capability. To improve its efficiency, our *VenusRL* technology introduces innovations in reinforcement learning frameworks, significantly enhancing overall training efficiency.

VenusRL Post-Training Framework Technologies

Post-training is a critical stage in the development of large models. It involves fine-tuning pre-trained models using relatively small amounts of labeled data to enhance specific capabilities. With the rapid development of AI applications, reinforcement learning from human feedback (RLHF) has become one of the most widely adopted approaches. Taking RLHF in AI programming as an example, the workflow typically follows an iterative loop from labeled data training, inference to generate new data, data validation and selection to further training, continuously improving the model’s coding capabilities. In this process, the most resource- and time-intensive stage is the generation and validation of inference outputs. Since external tools are often required to verify and filter the generated results, the workflow involves multiple iterative cycles of inference, result validation, insufficient outcome, updated input and re-inference until the output meets the expected criteria. The validated outputs are then used as training data for subsequent rounds, and the process repeats across different tasks. As a result, the validation stage requires a large number of isolated execution environments, while GPUs used for inference often remain idle during these steps. These factors together lead to relatively low efficiency in post-training workflows.

For post-training workloads, which are characterized by iterative alternation between training and inference and relatively short training duration in each iteration, we have developed our *VenusRL* post-training frameworks as an efficient engine designed to improve resource utilization. Under our architectural framework, we use interface abstraction and scheduling orchestration technologies to decompose integrated training-and-inference tasks into smaller workflow units that can be independently orchestrated and scheduled. Through pre-execution algorithms and online profiling and tracing technologies, we calculate the resource requirements of each workflow stage and conduct fine-grained scheduling and orchestration accordingly. This allows us to schedule the training stages of multiple post-training tasks onto the same computing resources, thereby reducing idle capacity in training resources during inference stages and improving overall resource utilization.

In addition, inference-stage workflows typically follow an iterative process of generating outputs, validating them and refining the results to produce improved outputs. Validation is usually performed using general-purpose computing resources such as CPUs, during which GPUs allocated for inference may be underutilized. Through our asynchronous input/output orchestration technology, we are able to schedule lower-priority compute tasks onto inference GPUs during the validation stage, and preempt and remove such tasks once validation is completed. This enables more efficient utilization of GPU computing resources without blocking inference tasks. At the same time, for the management of isolated execution environments used in the inference validation process, we leverage our lightweight and efficient *VenusBox* technology to improve the efficiency of environment creation and teardown, thereby reducing validation time. Through these two technologies, compared with conventional post-training implementation approaches commonly used in the industry, our *VenusRL* architecture improves overall task efficiency by 37%.

BUSINESS

RESEARCH AND DEVELOPMENT

Our R&D Focus

During the Track Record Period, our R&D primarily focused on improving technological capabilities of our AI computing cluster products. See “— Our Technologies”. Looking ahead, we intend to continue increasing our investment in R&D to further enhance our technology capabilities across our converged AI computing cluster products. See “— Strategies — Continuously Iterate AI Computing Cluster Products and Consolidate Market Leadership.”

We have established an R&D framework that integrates industrial application with academic research. Through close collaboration between our research and product teams, we systematically evaluate emerging technological demands and advance solutions from conceptual development to commercial deployment where market potential is identified. In parallel, we actively engage with leading domestic hardware manufacturers and research institutions to conduct forward-looking joint research aligned with next-generation hardware architectures and performance requirements. This industry-academia integration model has contributed to accelerated innovation and development efficiency. For instance, we completed the full development cycle of our proprietary switches, from core research to product deployment, within eight months, significantly shorter than typical industry timelines. In addition, through joint research initiatives with academic partners, we reduced the development cycle of congestion control algorithms for our RDMA NICs. We also participate in collaborative research programs with domestic universities, research institutes, and cloud service providers, including projects involving intelligent simulation platforms and open-source operational systems. These initiatives support advancements in AI infrastructure technologies while reinforcing our technological capabilities, innovation efficiency, and industry collaboration ecosystem.

Our R&D Process

We have established an integrated R&D framework encompassing demand evaluation, concept development, architecture design, technology development, and ongoing optimization.

- ***Demand evaluation and technology assessment:*** We systematically evaluate emerging technological trends and market demands, including identifying target AI workloads, performance benchmarks, and deployment scenarios. Our management and R&D team determine the priorities.
- ***Core research and conceptual design:*** We translate research insights into preliminary technology concepts and design frameworks either in-house or through forward-looking joint research with leading hardware manufacturers, research institutions, and academic partners. At this stage, we formulate initial design approaches and evaluate their feasibility across key technical domains, including high-performance networking, collective communication, scheduling algorithms, and system software. This conceptual design work establishes the foundational technical direction that guides subsequent R&D initiatives.
- ***Architecture design and detailed planning:*** Based on assessed requirements, we define the overall technology architecture. We formulate detailed plans to allocate resources, set milestones, and mitigate risks throughout the R&D process.
- ***Technology development and engineering implementation:*** We execute our development plans and perform internal integration testing, tuning, and validation. This includes rigorous stress testing and performance benchmarking to verify that the technology under development meets target specifications.
- ***Ongoing optimization and iteration:*** Following internal validation, we channel feedback and performance data back into the R&D process to drive continued technology iteration, system optimization, and capability enhancements.

BUSINESS

Our R&D Team

The AI computing cluster industry features a multi-layered value chain spanning chips, hardware, cluster scheduling and application services, each involving distinct and complex technical requirements. Technology iteration cycles are extremely short across the AI computing cluster industry value chain, and each specialized segment is subject to high technical barriers. We have one of the leading AI infrastructure teams in China, bringing together a deep pool of experts in AI computing clusters. We have accumulated over 20 years of expertise in distributed system and computer architecture, and our R&D team members have collectively published more than 90 papers in leading international conferences and journals.

As of the Latest Practicable Date, we had a research and product development team comprising 80 members, representing 66.1% of our total employee base. Our R&D team possesses a solid educational background and extensive industry experience across system architecture, software engineering, and high-performance computing solutions, supporting our continuous innovation in AI computing clusters.

We focus on building and maintaining a large pool of talented researchers to drive our R&D efforts through internship programs and campus recruitment. As a participant in one of the world’s most cutting-edge industries, where few professionals possess directly relevant experience and expertise, we place strong emphasis on cultivating and developing our own talent base through our internship programs. In the recruitment process, we highlight our technical ambitions, engineering challenges, and long-term growth vision to attract candidates who share our passion for innovation.

SALES AND MARKETING

Sales

The technological strength and advanced capabilities of our AI computing cluster products and operation services form the foundation of our customer acquisition and expansion strategy. By delivering high-performance and scalable products that address the evolving computing demands of enterprises and public institutions, we establish deep technical credibility and long-term trust with our customers. Building on successful deployments with benchmark customers, we continue to expand our market penetration through demonstrated results and measurable performance improvements, which validate the reliability and competitiveness of our offerings while serving as key reference cases that attract new customers across adjacent sectors.

We adopt a direct sales model, enabling our sales and technical teams to engage closely with customers throughout the entire sales cycle from technical consultation and product design to deployment and post-delivery support. During the customer engagement and negotiation stage, our sales team gathers customers specifications and relay them to our product teams as the basis for proposal preparation and product refinement. Throughout project deployment, on-site service personnel and technical experts maintain communication with customers to collect feedback on workflow, customer experience, and new feature requirements. Such close and continuous engagement not only strengthens customer stickiness, but also enhances our ability to identify follow-on demand, support customers in expanding from training to inference and agentic application scenarios, and generate recurring business opportunities.

Pricing

For AI computing cluster products, we charge customers an one-off fixed fee and recognize revenue upon customer acceptance. We price our AI computing cluster products based on a range of factors, including the scope, scale and complexity of customer workload requirements, product configuration, delivery model, customer profile, procurement costs for certain hardware components, and prevailing market conditions.

For maintenance and technical support services, we charge customers on an annual basis based on the scale of the AI computing cluster under maintenance, the time worked and labor assigned. For compute services, we charge customers a fixed fee based on the amount of computing resources provided.

BUSINESS

Marketing

Our market expansion strategy is resource-driven and reputation-led. At inception, we leveraged our management team’s and investors’ extensive alumni and industry networks to gain access to potential customers through referrals and introductions from strategic partners. In the early stages of market development, we focused on acquiring representative, high-profile customers whose successful deployments could serve as benchmark cases. We actively publicize these projects to demonstrate our technological capabilities and the replicability of our products, thereby enhancing market awareness and credibility. As our customer base has grown, we have capitalized on existing relationships with benchmark customers to pursue broader business development opportunities. Through these established networks, we are able to reach additional enterprises, government entities and public-sector organizations with similar AI computing needs, facilitating scalable customer acquisition and deeper market penetration.

CUSTOMERS AND CUSTOMER SUPPORT

Customers

Our customers primarily comprise large model companies, cloud service providers, education and research institutions, and telecommunications operators.

In 2023, we served four customers, to whom we provided AI computing cluster products. In 2024 and 2025, revenue generated from our five largest customers in each year during the Track Record Period amounted to RMB321.1 million and RMB294.2 million, representing approximately 98.9% and 56.6% of our total revenue for the corresponding year, respectively. In 2023, 2024 and 2025, revenue generated from our largest customer amounted to RMB15.5 million, RMB191.5 million and RMB86.3 million, respectively, representing approximately 48.6%, 59.0% and 16.6% of our total revenue for the same periods/years, respectively. The typical credit period we grant our customers, including our top five customers in each period/year during the Track Record Period, was within 30 days. Our customers primarily pay us by bank transfer.

The following table sets forth details of our five largest customers in each period/year during the Track Record Period:

Customers	Revenue amount <i>(RMB in thousands)</i>	Percentage of total revenue <i>(%)</i>	Background	Product and services purchased	Year of commencement of business relationship
<i>Period from February 20, 2023 to December 31, 2023</i>					
Customer A	15,460	48.6%	A private company headquartered in Zhejiang Province, China, primarily engaged in the provision of software development, cloud computing and AI-related technical services and solutions	AI computing cluster products	2023
Customer B	15,419	48.5%	A public company listed on the Hong Kong Stock Exchange, primarily engaged in the R&D of cognitive LLM and the provision of related technical support and application services	AI computing cluster products	2023

BUSINESS

Customers	Revenue amount <i>(RMB in thousands)</i>	Percentage of total revenue <i>(%)</i>	Background	Product and services purchased	Year of commencement of business relationship
Customer C	883	2.8%	A public company listed on the Nasdaq Stock Exchange, primarily engaged in the operation of data centers, intelligent computing centers and the provision of comprehensive cloud computing services	AI computing cluster products	2023
Customer D	40	0.1%	A private company headquartered in Beijing, China, primarily engaged in the R&D, production and sales of imaging solutions	AI computing cluster products	2023
Total	<u>31,802</u>	<u>100.0%</u>			
<i>For the year ended December 31, 2024</i>					
Customer A	191,548	59.0%	A private company headquartered in Zhejiang Province, China, primarily engaged in the provision of software development, cloud computing and AI-related technical services and solutions	AI computing cluster products	2023
Customer E	53,294	16.4%	A public company listed on the Shanghai Stock Exchange, primarily engaged in deploying communication systems and developing communication technologies, together with one of its subsidiaries	AI computing cluster products and AI computing cluster operation services	2024
Customer B	52,311	16.1%	A public company listed on the Hong Kong Stock Exchange, primarily engaged in the R&D of cognitive LLM and the provision of related technical support and application services	AI computing cluster products	2023

BUSINESS

Customers	Revenue amount (RMB in thousands)	Percentage of total revenue (%)	Background	Product and services purchased	Year of commencement of business relationship
Customer F	13,983	4.3%	A private company headquartered in Shanghai, China, primarily engaged in the provision of digital technology services, data processing services and digitalized industrial park construction and operation services	AI computing cluster products	2024
Customer G . . .	9,942	3.1%	A private company headquartered in Shanghai, China, primarily engaged in the provision of data processing and storage support services, cloud computing equipment services and AI-related technical services	AI computing cluster operation services	2024
Total	321,078	98.9%			
<i>For the year ended December 31, 2025</i>					
Customer E	86,258	16.6%	A public company listed on the Shanghai Stock Exchange, primarily engaged in deploying communication systems and developing communication technologies, together with one of its subsidiaries	AI computing cluster products and AI computing cluster operation services	2024
Customer H . . .	63,717	12.2%	A private company headquartered in Guangxi Province, China, primarily engaged in the provision of high-end manufacturing, new energy and leasing services	AI computing cluster products	2025

BUSINESS

Customers	Revenue amount <i>(RMB in thousands)</i>	Percentage of total revenue <i>(%)</i>	Background	Product and services purchased	Year of commencement of business relationship
Customer I	51,096	9.8%	A subsidiary of a listed company, headquartered in Jiangsu Province, China, primarily engaged in the provision of software and information technology services, cloud computing equipment sales and technical development services	AI computing cluster products	2025
Customer J	48,894	9.4%	A subsidiary of a listed company, headquartered in Beijing, China, primarily engaged in the construction and operation of clean energy power plants	AI computing cluster products	2025
Customer K . . .	44,284	8.5%	A private company headquartered in Shanghai, China, primarily engaged in the provision of technology promotion and application services, information system integration and software development services	AI computing cluster product and AI computing cluster operation service	2025
Total	<u>294,248</u>	<u>56.6%</u>			

Our historical customer concentration was primarily attributable to the characteristics of the AI computing cluster market and our customer development strategy during our early stage of development. Demand for AI computing cluster is generally concentrated among leading market participants, such as major large model companies, cloud service providers, education and research institutions, and telecommunications operators, whose computing needs are typically larger in scale, more complex in configuration and more representative of advanced AI computing cluster deployment scenarios. From the outset, we adopted a customer development strategy focused on serving benchmark customers, as we believed that working with such customers would enable us to address the most demanding use cases, validate our products and technical capabilities, and build reference cases for broader market expansion.

As we were only established in 2023, we were also in the process of refining our products and strengthening our technical and delivery capabilities during the early stage of our operations. Given our strategic focus on benchmark customers, we devoted substantial resources to product adaptation, service optimization and project execution for a limited number of leading customers, while our service and deployment bandwidth was still relatively constrained. As a result, during the first two years of our operations, we served only a relatively small number of customers, which contributed to a higher level of customer concentration.

BUSINESS

We do not consider our historical customer concentration to be unusual in light of the characteristics of the AI computing cluster market, as confirmed by Frost & Sullivan. Our major customers are typically key AI stakeholders, whose demand for AI computing infrastructure is continuous and evolving as their AI workloads expand and their requirements for computing performance, cluster efficiency and operational stability continue to increase. Users of AI computing clusters in China often face constraints in access to high-performance compute hardware and must operate across hardware resources with different architectures and ecosystems. In this context, one of the most critical capabilities for an AI computing cluster lies in effectively interconnecting and coordinating compute, networking and storage resources into a unified and efficiently operating system. Our AI computing cluster products and AI computing cluster operation services are designed to address these needs by enabling cluster-level interconnection, communication efficiency and overall system performance. As a result, our relationships with such benchmark customers are mutually beneficial: they provide us with opportunities to serve demanding and representative use cases, while our products and services support their continuing need for high-performance and effective AI computing infrastructure, which in turn enables us to generate recurring revenue from such customers.

At the same time, we have been progressively diversifying both our service offerings and our customer base. Our experience in serving benchmark customers has already translated into tangible benefits, including further enhancement of our products and technical capabilities, stronger market recognition and broader customer acquisition. The number of customers of our AI computing cluster products increased from four in 2023 to 10 in 2024 and further to 48 in 2025. In addition, we have continued to expand our offerings and operation model in response to customer needs. While we initially provided only AI computing cluster products in 2023, we began to offer AI computing cluster operation services in 2024. Although these two business lines differ in delivery format, both are aimed at enabling customers to access AI computing resources in a reliable and efficient manner. The expansion of our service offerings has allowed us to address a wider range of customer requirements and deployment scenarios, which has also contributed to the diversification of our customer base. The number of customers for our AI computing cluster operation services increased from two in 2024 to 16 in 2025. Therefore, we do not consider ourselves dependent on any particular customer.

As of the Latest Practicable Date, to the best knowledge of our Directors, none of our Directors or their respective close associates, and none of our Shareholders who own more than 5% of the Shares in issue had any interest in any of our five largest customers in each year during the Track Record Period.

Major Terms with Our Customers

Set forth below is a summary of the major terms in our contracts for AI computing cluster products:

- **Products and services:** We provide comprehensive products and services covering the supply of hardware equipment, deployment of proprietary software systems and/or performance tuning. The service scope is specified in the customer contracts.
- **Payment:** The customer is typically required to pay a portion of the contract amount after contract execution, and the remaining payments are made based on service delivery milestones.
- **Delivery and acceptance:** We are responsible for delivering our products and services to the customer’s designated site and performing installation, testing, and trial operation, as required. The customer conducts acceptance according to the agreed procedures.
- **Intellectual property:** We own proprietary intellectual property rights to our self-developed software and systems. We do not own the underlying IP of hardware procured from third parties.

BUSINESS

- **Liability:** If the customer fails to make payments on time, it shall bear the corresponding default liability. If we fail to deliver on schedule or the products or services do not meet agreed standards, we shall be responsible for repair, replacement, or compensation. Either party may be exempted from liability in the event of force majeure.
- **Termination:** Either party may terminate the contract early by written notice in the event of a material breach, force majeure, or by mutual written agreement. Termination does not affect the liabilities that arose prior to termination.

Our contracts with AI computing cluster operation service customers typically address terms on the scope of services, the obligation of each party, payment terms, and termination provisions. Maintenance and technical support services generally include inspections, fault diagnosis, and trouble shooting. For compute services, we provide customers with dedicated AI compute resources, to which the customer has exclusive access for the specified duration of the contract. The customer is obliged to make upfront and periodic payments. We are obliged to deliver AI computing cluster operation services in accordance with the specified service scope, performance benchmarks and SLA objectives set out in the contract. Either party may terminate the contract early by written notice in the event of a material breach, the other party’s insolvency or entry into liquidation proceedings, the expiry of the service term, or as required by applicable law.

After-Sales Services

Our sales and technical teams maintain close coordination to ensure high-quality customer experience. We provide comprehensive after-sales services tailored to our product and service offerings. See “— Our Products and Services — Our AI Computing Cluster Products — Delivery of Our AI Computing Clusters” for details. The warranty period of the hardware components of our AI computing cluster products is typically three years. We also assign dedicated support specialists who deliver ongoing consulting and technical assistance to ensure seamless customer experience.

During the Track Record Period, we did not experience any material service interruption, contract termination, or customer dispute relating to project delivery or after-sales services.

SUPPLIERS AND PROCUREMENT

During the Track Record Period, our suppliers primarily comprised providers of hardware components integrated into our products, including primarily compute hardware, networking equipment and storage units, as well as contract manufacturers for optical transceivers, RDMA NICs, and our self-developed switches. All of our suppliers during the Track Record Period were based in China.

Supply Chain Management

We maintain stringent protocols for selecting and managing suppliers of critical components and services. Our selection criteria emphasize technological capabilities, production stability and quality assurance standards, with key suppliers required to maintain relevant certifications and demonstrate consistent quality management systems. Our supplier management framework includes regular performance evaluations, focusing on quality metrics, delivery reliability, and technical capabilities, with immediate corrective action plans required for any issues identified.

To mitigate potential supply chain disruption risks, we have implemented a comprehensive risk management system. We maintain active engagement with a diverse pool of qualified suppliers and, where feasible, ensure that at least one alternative supplier is available for each key supplier category. As we continue to scale our business at a rapid pace, we consider it essential to preemptively maintain sufficient inventory levels to meet the growing demand for our products and ensure continuity of supply. During the Track Record Period and up to the Latest Practicable Date, we had not experienced any suspension or termination of relationships with our major suppliers, nor had we encountered any material disputes or business disruptions arising from supply chain issues, and there had been no material changes to the terms

BUSINESS

of our agreements with our major suppliers. Through this comprehensive approach, we maintain strong relationships with our supplier base while ensuring consistent access to high-quality components and services necessary to meet our business needs.

Hardware

When we undertake an end-to-end approach to deliver converged AI computing cluster products, which could take the form of a full cluster deployment or a more targeted deployment involving selected hardware and software components together with cluster-level networking and convergence services, we may need to procure compute hardware, networking equipment and storage units from qualified suppliers for incorporation into our products to support such deployment. Set forth below is a summary of the salient terms in our contracts with key hardware suppliers:

- **Products and services:** Suppliers are responsible for providing products as specified in the relevant purchase orders.
- **Pricing and payment:** Prices and payment arrangements are determined in accordance with the contract and related purchase orders. For hardware procurement, payments are typically made in instalments based on delivery and acceptance schedules agreed between the parties. For certain hardware components, payments may also be made in full within a specified period after contract execution, and the supplier delivers the products only after full payment has been settled.
- **Delivery and acceptance:** Suppliers are required to deliver goods to the designated locations within agreed timeframes. We conduct inspection and acceptance upon receipt to verify quantity, appearance, and functionality. In the event of non-conformity, damage, or performance issues, suppliers are responsible for repair, replacement, or re-delivery.
- **Quality assurance and after-sales service:** Suppliers must ensure that their products meet applicable national and industry standards and provide warranty support for a specified period. During the warranty period, suppliers bear the costs for repair, replacement, and relevant maintenance arising from quality issues.
- **IP and confidentiality:** The IP rights of supplied products and related technical materials remain with the suppliers or original manufacturers. We are granted the right to use such products within the agreed scope. Both parties are bound by confidentiality obligations with respect to proprietary and commercial information, which remain effective after the termination of the agreements.
- **Termination:** Either party may terminate the agreement early by written notice in the event of material breach, force majeure, or other circumstances rendering performance impracticable.

Contract Manufacturing

We engage third party independent contract manufacturers for the production of our self-developed switches, as well as optical transceivers and RDMA NICs. As the network system is critical to AI computing clusters, certain network hardware needs to be customized for our AI computing cluster products. Accordingly, instead of directly procuring standardized products from suppliers, we engage contract manufacturers to manufacture such networking equipment for us. During the Track Record Period, we engaged seven contract manufacturers.

In selecting our contract manufacturers, we generally cooperate with leading manufacturers to ensure the quality and reliability of our products. We also implement quality control measures over our contract manufacturers. Prior to shipment, products are subject to testing and corresponding test records are maintained. For example, we generally require our contract manufacturers of switches to complete a series of pre-shipment tests, including functional testing, full-port stress testing and high-temperature burn-in testing, and to maintain comprehensive test records throughout the process.

BUSINESS

To safeguard our intellectual property rights, we have implemented a series of contractual measures. In particular, our contracts with contract manufacturers contain intellectual property and confidentiality clauses, which impose obligations on them to maintain the confidentiality of our proprietary technologies and not to use such information in any manner for any purpose other than as permitted under the relevant contracts. Set forth below is a summary of the salient terms in our contracts with contract manufacturers:

- **Products and services:** Suppliers are responsible for manufacturing and delivering hardware products in accordance with our technical requirements and specifications and for providing necessary system integration and technical support services. The specific products and service requirements are determined by individual purchase orders.
- **Pricing and payment:** Prices and settlement arrangements are specified in the relevant contracts or purchase orders. Payments are generally made by bank transfer within the agreed period after acceptance of the relevant products.
- **IP and confidentiality:** Suppliers are required to ensure that all products and technologies involved in their products and services do not infringe any third-party IP rights. Both parties are subject to confidentiality obligations regarding proprietary technologies and business information obtained during the course of cooperation. Such obligations remain in effect after the termination or expiration of the agreements.
- **Quality assurance and after-sales service:** Suppliers must ensure that the products conform to applicable national and industry standards and provide a specified quality warranty period after acceptance. During the warranty period, suppliers are responsible for repair, replacement, and associated costs in the event of product quality issues.
- **Technical support:** Suppliers shall provide necessary technical support and maintenance services during the cooperation period to assist us in ensuring stable operation of the relevant products and systems.
- **Termination:** Either party may terminate the agreement upon the occurrence of significant breach, insolvency, regulatory changes, or other circumstances rendering performance impracticable, subject to the notice requirements stated in the contract.

Major Suppliers

In 2023, 2024 and 2025, the purchases from our five largest suppliers in each period/year during the Track Record Period amounted to RMB47.1 million, RMB254.8 million and RMB284.7 million, respectively, representing approximately 90.7%, 75.5% and 54.2% of our total purchases for the corresponding period/year, respectively. In 2023, 2024 and 2025, the purchases from our largest suppliers in each period/year during the Track Record Period amounted to RMB28.3 million, RMB122.4 million and RMB151.6 million, respectively, representing approximately 54.5%, 36.2% and 18.8% of our total purchases for the corresponding period/year, respectively. We primarily pay our suppliers by bank transfer. The credit period granted to us by our suppliers generally ranged from two to 60 days during the Track Record Period.

BUSINESS

The following table sets forth details of our five largest suppliers in each period/year during the Track Record Period:

Suppliers	Purchase amount <i>(RMB in thousands)</i>	Percentage of total purchase <i>(%)</i>	Background	Year of commencement of business relationship	Products and services provided
<i>Period from February 20, 2023 to December 31, 2023</i>					
Supplier A . . .	28,271	54.5%	A private company headquartered in Beijing, China, primarily engaged in the provision of compute hardware, networking equipment and technical services	2023	Networking equipment
Supplier B . . .	8,173	15.8%	A private company headquartered in Beijing, China, primarily engaged in the provision of communication, telecommunications and networking equipment-related business and technical services	2023	Compute hardware and networking equipment
Supplier C . . .	6,097	11.7%	A private company headquartered in Shanghai, China, primarily engaged in the provision of communication technology, network technology and chip technology-related technical development and services	2023	Networking equipment

BUSINESS

Suppliers	Purchase amount <i>(RMB in thousands)</i>	Percentage of total purchase <i>(%)</i>	Background	Year of commencement of business relationship	Products and services provided
Supplier D . . .	2,798	5.4%	A private company headquartered in Beijing, China, primarily engaged in the provision of information technology professional services, information system integration and end-to-end data center implementation services	2023	Networking equipment
Supplier E . . .	1,714	3.3%	A private company headquartered in Guangdong Province, China, primarily engaged in the provision of software and information technology services and related technical development and consulting services	2023	Networking equipment
Total	47,053	90.7%			
<i>For the year ended December 31, 2024</i>					
Supplier A . . .	122,375	36.2%	A private company headquartered in Beijing, China, primarily engaged in the provision of compute hardware, networking equipment and technical services	2023	Networking equipment and associated warranty services
Supplier F . . .	39,679	11.8%	A private company headquartered in Beijing, China, primarily engaged in AI intelligent computing infrastructure technology services and information system integration	2024	Networking equipment

BUSINESS

Suppliers	Purchase amount <i>(RMB in thousands)</i>	Percentage of total purchase <i>(%)</i>	Background	Year of commencement of business relationship	Products and services provided
Supplier C . . .	36,908	10.9%	A private company headquartered in Shanghai, China, primarily engaged in the provision of communication technology, network technology and chip technology-related technical development and services	2023	Networking equipment
Supplier G . . .	33,496	9.9%	A private company headquartered in Guangdong Province, China, primarily engaged in the provision of software and information technology services and related technical development and consulting services	2024	Networking equipment
Supplier E . . .	22,292	6.6%	A private company headquartered in Guangdong Province, China, primarily engaged in the provision of software and information technology services and related technical development and consulting services	2023	Networking equipment
Total	254,750	75.5%			

BUSINESS

Suppliers	Purchase amount <i>(RMB in thousands)</i>	Percentage of total purchase <i>(%)</i>	Background	Year of commencement of business relationship	Products and services provided
<i>For the year ended December 31, 2025</i>					
Supplier H . . .	151,649	18.8%	A private company headquartered in Zhejiang Province, China, primarily engaged in the R&D of cloud computing technology, and the provision of cloud computing, big data and artificial intelligence-related products and services	2025	Compute hardware, Networking equipment, security software and cloud services
Supplier I . . .	132,036	16.4%	A private company headquartered in Shanghai, China, primarily engaged in the provision of information system integration, network technical services and communication equipment sales.	2025	Compute hardware
Supplier J . . .	63,009	7.8%	A private company headquartered in Guangdong Province, China, primarily engaged in the operation of computing power centers, and the provision of computing power rental, computing power scheduling and related technical services	2025	Compute hardware
Supplier K . . .	48,451	6.0%	A private company headquartered in Zhejiang Province, China, primarily engaged in the R&D of graph database technology, and the provision of graph computing and data intelligence-related solutions and services	2025	Compute hardware

BUSINESS

Suppliers	Purchase amount <i>(RMB in thousands)</i>	Percentage of total purchase <i>(%)</i>	Background	Year of commencement of business relationship	Products and services provided
Supplier L . . .	41,239	5.1%	A private company headquartered in Guangdong Province, China, primarily engaged in the provision of intelligent Internet-of-Things solutions, software development and information system integration services	2025	Networking equipment
Total	284,735	54.2%			

Purchases from our five largest suppliers in 2023 and 2024 accounted for 90.7% and 75.5%, respectively, of our total purchases for the corresponding years. The relatively high supplier concentration in 2023 and 2024 was primarily attributable to our decision to source from reputable suppliers to ensure consistent product quality during the early stage of our business. As our business grew, we gradually diversified our supplier base, as reflected in the decrease from 90.7% in 2023 to 54.2% in 2025. Multiple alternative suppliers in China are available for the products and services that we procure, therefore, we do not consider ourselves to be dependent on any single supplier.

To the best knowledge of our Directors, none of our Directors, their respective close associates, and none of our Shareholders who own more than 5% of the Shares in issue, had any interest in any of our five largest suppliers in each year during the Track Record Period.

OVERLAP BETWEEN MAJOR CUSTOMERS AND SUPPLIERS

During the Track Record Period, certain of our five largest customers during each period/year of the Track Record Period were also our suppliers, and certain of our five largest suppliers during each period/year of the Track Record Period were also our customers, details of which were explained as below:

Customer E, one of our five largest customers in 2024 and 2025, was also our supplier in 2025. Revenue generated from providing AI computing cluster products and AI computing cluster operation services to Customer E accounted for nil, 16.4% and 16.6% of our total revenue in 2023, 2024 and 2025, respectively. As part of our AI computing cluster product offerings, we also procure and provide certain ancillary cluster management systems to support customers’ cluster operating environments, which was purchased from Customer E, and accounted for nil, nil and 0.2% of our total purchases in 2023, 2024 and 2025, respectively.

Some of our major suppliers also purchase from us. As participants along the AI computing cluster value chain, these suppliers also demand AI computing cluster products and operation services. According to Frost & Sullivan, overlap between major customers and suppliers is common in the AI computing cluster industry. Supplier A, one of our five largest suppliers in 2023 and 2024, was also our customer in 2025. We primarily procured networking equipment and associated warranty services from Supplier A, which accounted for 54.5%, 36.2% and 1.9% of our total purchases in 2023, 2024 and 2025, respectively. Revenue generated from sales of AI computing cluster products to Supplier A accounted for nil, nil and 7.4% of our total revenue in 2023, 2024 and 2025, respectively. Supplier C, one of our five largest suppliers in 2023 and 2024, was also our customer in 2025. We primarily procured networking equipment from Supplier C, which accounted for 11.7%, 10.9% and 0.005% of our total purchases in 2023, 2024 and

BUSINESS

2025, respectively. Revenue generated from sales of AI computing cluster products to Supplier C accounted for nil, nil and 0.003% of our total revenue in 2023, 2024 and 2025, respectively. Supplier E, one of our five largest suppliers in 2023 and 2024, was also our customer in 2025. We primarily procured networking equipment from Supplier E, which accounted for 3.3%, 6.6% and 4.7% of our total purchases in 2023, 2024 and 2025, respectively. Revenue generated from sales of AI computing cluster products to Supplier E accounted for nil, nil and 1.8% of our total revenue in 2023, 2024 and 2025, respectively.

Our Directors confirm that (i) such overlapping customer-supplier relationships represented normal commercial arrangements in the ordinary and usual course of business, (ii) the terms of these transactions were fair and reasonable, (iii) negotiations of the terms of our sales to and purchases from these overlapping customers and suppliers were conducted on an individual basis and the sales and purchases were neither inter-connected nor inter-conditional with each other, and (iv) such arrangements did not have any adverse impact on our operational independence, pricing mechanisms or business arrangements. We are not reliant on such overlapping relationships.

DATA PRIVACY AND SECURITY

During the Track Record Period and up to the Latest Practicable Date, our users are enterprises rather than individuals, and our products are deployed on our customers’ own on-premises infrastructure and operated exclusively within customer-controlled environments, with all data processing activities being carried out under our customers’ exclusive management. Given that our core business focuses on product delivery and necessary technical support, with our products being deployed in a privatized manner on the customer side, we do not collect, store, or use the data that customers input or generate while using the delivered products. The data we collect and generate in the course of our business operations primarily include: our self-developed source code; technical plans we formulate and deliver; our internal financial data and human resource information; as well as customer contact details.

Data protection and cybersecurity are among our highest priorities. In this regard, we have designed strict data protection and cybersecurity policies to ensure strict compliance with applicable laws, regulations and prevalent industry practice. We have implemented comprehensive internal policies on protecting data privacy and security, with the purpose to ensure data and information security, optimize data governance, protect the benefits of our customers, employees and other third parties, and ensure compliance with all applicable laws and regulations in material respects.

We implement a robust internal authentication and authorization system to ensure that our confidential data can only be accessed for authorized use and by authorized personnel. We have clear and strict authorization and authentication procedures and policies in place. Our employees only have access to data which is directly relevant and necessary for their responsibilities and for limited purposes, and are required to verify authorization upon every access attempt.

Our core product features a multi-dimensional security protection system, which includes: (i) a role-based access control mechanism that assigns different permissions to various end-users, strictly limiting their access to and operations on customers’ on-premises resources, thereby ensuring the security of end-user data and systems; (ii) support for integration between customer systems and third-party authentication systems to achieve unified end-user identity management; rigorous security protection for end-user data, supporting encrypted data storage and transmission to prevent data leakage risks; (iii) restrictions on cluster network access permissions, allowing only necessary service ports to be exposed externally to prevent external malicious attacks; support for configuring network allowlists to further enhance network security; (iv) implementation of user-level permission injection and policy control through a multi-tenant isolation architecture, coupled with the adoption of soft isolation via resource quotas and hard isolation via scheduling zones, establishing a multi-dimensional, fine-grained resource security boundary.

During the Track Record Period and up to the Latest Practicable Date, (i) we had not received any claim from any third party against us on the ground of infringement of such party’s right to data and privacy protection as provided by any applicable laws and regulations, (ii) there had been no investigation or other legal proceeding pending or threatened against us initiated by competent government authorities

BUSINESS

or third parties with respect to cybersecurity, data and personal information protection, and (iii) we had not experienced material leakage of personal information in relation to laws and regulations of cybersecurity, data protection and personal information protection.

Based on the above, as well as that (i) we currently do not directly obtain or store customers’ data in its primary business; (ii) we have implemented internal policies on protecting cybersecurity, data security and personal information; and (iii) we will continuously pay close attention to the legislative and regulatory development in cybersecurity, data protection and personal information protection, maintain ongoing communication with relevant government authorities and implement all necessary measures in a timely manner to ensure continuous compliance with the relevant laws and regulation, our PRC Data Compliance Legal Adviser is of the view that, during the Track Record Period and up to the Latest Practicable Date, we have implemented compliance measures concerning cybersecurity, data protection and personal information protection in accordance with the requirements of relevant cybersecurity, data and personal information protection laws and regulations in material aspects. However, as the laws and regulations in the cybersecurity, data protection and personal information protection are still developing, we cannot assure you that we can always timely adapt to all the aspects of such laws and regulations. See “Risk Factors — Risks Related to Our Business and Industry — Our business is subject to a variety of laws, rules, policies and other obligations regarding privacy and data protection. Any losses or unauthorized access to or unauthorized releases of confidential information and personal data could subject us to significant reputational, financial, legal and operational consequences.”

COMPETITION

We face intense competition in China’s AI computing cluster industry. The principal competitive factors in our industry include technological capabilities, functionality, scope and performance of products, scalability and reliability of services, delivery expertise, marketing and sales capabilities, customer experiences, pricing, brand recognition and reputation. In addition, new and enhanced technologies and new market entrants may further increase competition in our industries. We believe that we are well positioned to compete effectively on the basis of the foregoing factors. See “— Competitive Strength.”

Nevertheless, some of our competitors may have greater brand recognition, longer operating histories, larger customer bases as well as greater financial, technical and other resources. See “Risk Factors — Risks Related to Our Business and Industry — If we fail to compete effectively, our business, financial condition and results of operations may be materially and adversely affected.” For more information on the competitive landscape of our industry, see “Industry Overview.”

INTELLECTUAL PROPERTY

Intellectual property rights are fundamental to our business. We currently hold many intellectual properties related to our core technologies, and we devote significant time and resources to their development and protection. We rely on a combination of patent, trademark, copyright, domain name, trade secret and other proprietary rights protection laws in the PRC as well as confidentiality procedures and contractual provisions to protect our intellectual properties.

As of December 31, 2025, we had seven registered patents, 39 trademarks, 44 software copyrights, and one domain name registered in the PRC. See “Appendix IV — Statutory and General Information — Further Information About Our Business — Intellectual Property Rights” for details of our material intellectual property rights.

We have used our best efforts to ensure compliance with applicable intellectual property laws. Our Directors confirmed that, during the Track Record Period and up to the Latest Practicable Date, we were not involved in any intellectual property infringement actions brought by third parties. Our Directors further confirm that we were not involved in any intellectual property infringement actions that,

BUSINESS

individually or in the aggregate, would have a material and adverse effect on our business, financial condition and results of operations. See “Risk Factors — Risks Related to Our Business and Industry — We may be subject to claims by third parties for alleged infringement of their intellectual proprietary rights.”

We protect our intellectual property through strategic planning, such as submitting intellectual property registrations and applications, anti-counterfeiting mechanism, especially for trademarks and designs, and litigation mechanism to defend against any existing or potential intellectual property infringement. Despite our efforts, third parties may still obtain and misappropriate our intellectual property without authorization. As of the Latest Practicable Date, we did not find any of such misappropriations of our intellectual property rights. However, unauthorized use of our intellectual property by third parties and the expenses incurred in protecting our intellectual property rights may adversely affect our business and results of operations. For details of related risk, see “Risk Factors — Risks Related to Our Business and Industry — Our intellectual property rights are critical to our success and infringement of our intellectual property rights by any third party may materially and adversely affect our business, financial conditions and results of operations.”

EMPLOYEES

We had 121 employees as of the Latest Practicable Date. The following table sets forth a breakdown of our employees by function as of the Latest Practicable Date. All of our employees are based in the PRC.

Function	Number of Employees	% of Total
Research and development	80	66.1
Sales and marketing	16	13.2
General and administration	25	20.7
Total	121	100.0

Our success depends on our ability to attract, retain and motivate qualified personnel, and we believe that our high-quality talent pool is one of the core strengths of our company. We adopt high standards and strict procedures in our recruitment, including campus recruitment, online recruitment, internal referral and recruitment through executive search, to satisfy our demands for different types of talents.

As required under PRC labor laws, we enter into individual employment contracts with our employees covering matters such as wages, bonuses, employee benefits, workplace safety, confidentiality obligations, non-competition and grounds for termination. In compliance with PRC regulations, we participate in various employee social security plans that are organized by applicable local municipal and provincial governments, including housing, pension, medical, work-related injury and unemployment benefit plans.

As of the Latest Practicable Date, we had not established a labor union. We believe that we maintain a good working relationship with our employees, and we have not experienced any significant labor disputes or any difficulty in recruiting staff for our operations during the Track Record Period.

INSURANCE

We consider our insurance coverage to be adequate as we have in place all the mandatory insurance policies required by PRC laws and regulations and in accordance with the commercial practices in our industry. Our employee-related insurance consists of pension insurance, maternity insurance, unemployment insurance, work-related injury insurance and medical insurance, as required by PRC laws and regulations.

BUSINESS

In line with general market practice, we do not maintain any business interruption insurance or product liability insurance, which are not mandatory under PRC laws. We do not maintain key man life insurance, insurance policies covering damages to our network infrastructures or information technology systems or any insurance policies for our properties. See “Risk Factors — Risks Related to Our Business and Industry — We have limited insurance coverage, and any claims beyond our insurance coverage may result in our incurring substantial costs and a diversion of resources.” During the Track Record Period, we did not make any material insurance claim in relation to our business.

ENVIRONMENTAL, SOCIAL AND GOVERNANCE MATTERS

ESG Governance

We have established comprehensive environment, social and governance (“ESG”) policies which outline our environmental initiatives, social responsibility commitments, and internal governance frameworks. Following the [REDACTED], we will comply with the requirements of ESG reporting and publish ESG report on an annual basis in accordance with the requirements of Appendix C2 to the Listing Rules.

The Board holds primary accountability for our ESG governance, overseeing the identification, assessment, and approval of major ESG and climate-related issues, objectives, strategies, risks, and opportunities, and ensuring the integrity of ESG data and external disclosures. Our management team oversees day-to-day ESG implementation and provides the Board with annual updates on major ESG matters, while proactively monitoring how such risks and opportunities may impact business operations, financial performance and sustainability commitments.

Environmental Responsibility

Air Pollutant and Waste Management

The primary source of our air pollutant emissions is waste from daily office activities, with the remainder of our environmental footprint consisting of GHG emissions from purchased electricity. We had no direct GHG emissions during the Track Record Period.

The waste we generate is classified as non-hazardous or hazardous. Non-hazardous waste mainly comprises general office waste, while hazardous waste includes waste batteries, lamp tube and ink cartridges. We have implemented a range of measures to reduce waste across our operations, including defaulting to double-sided and ink-saving print settings, prioritizing recyclable or reusable consumables, preventing office supply over-stocking, and encouraging paperless workflow.

Resource Consumption

Our energy consumption consists primarily of purchased electricity and heat for office operations. To improve resource efficiency, we service air conditioning systems regularly to minimize refrigerant leakage and limit office lighting in low-traffic areas. Employees are required to switch off air conditioning, lighting, and electronic equipment when leaving the office. Water consumption is attributable primarily to routine office activities and is sourced in compliance with all applicable national and local regulations. Conservation measures include infrared-sensing faucets and water-saving fixtures, reduced water pressure, and regular water meter inspections to detect and prevent hidden leaks. We remain committed to improving resource efficiency.

BUSINESS

The table below sets forth our energy and water consumption during the Track Record Period:

	Unit	Period from February 20, 2023 to December 31, 2023	Year ended December 31,	
			2024	2025
Total Energy Consumption	MWh	46.89	134.83	218.67
Direct energy consumption	MWh	—	—	—
Indirect energy consumption	MWh	46.89	134.83	218.67
Intensity – Total Energy Consumption	MWh/million RMB of revenue	1.47	0.42	0.42
Total Water Consumption	m³	191.63	369.56	1,478.25
Intensity – Water Consumption	m ³ /million RMB of revenue	6.03	1.14	2.84

Responding to Climate Change

We recognize the risks presented by climate change. To minimize potential environmental and climate-related risks arising from our daily operations, we are committed to fostering green practices across the Group. We strictly comply with applicable PRC environmental laws and regulations governing air emissions, wastewater discharge, and waste disposal. Our environmental management strategy is informed by quantitative data, including monitoring of greenhouse gas (GHG) emissions and resource consumption.

The table below sets forth our GHG emissions during the Track Record Period:

	Unit	Period from February 20, 2023 to December 31, 2023	Year ended December 31,	
			2024	2025
Total GHG Emissions (Scope 2 + Scope 3)	tCO₂e	30.00	148.05	274.31
Scope 2 Energy Indirect GHG Emissions ⁽¹⁾	tCO ₂ e	22.54	60.70	100.67
Scope 3 Other Indirect GHG Emissions ⁽²⁾	tCO ₂ e	7.46	87.35	173.64
Intensity – Total GHG Emissions	tCO ₂ e/million RMB of revenue	0.94	0.46	0.53

Notes:

- (1) Scope 2 indirect emission primarily includes the GHG emissions from our usage of purchased electricity and heat.
- (2) Scope 3 other indirect emission includes the GHG emissions from business travel and electricity for the treatment of water and wastewater.

We have established a series of environmental targets across four key aspects. With respect to GHG emissions, we aim to reduce our total greenhouse gas emission intensity by 5% by 2030, using the year 2023 as the base year. In relation to electricity consumption, we have similarly committed to reducing our electricity consumption intensity by 5% by 2030, again measured against the 2023 base year. For water consumption, the target is to reduce water consumption intensity by 5% by 2030, with the year 2023 serving as the base year. Finally, regarding waste generation, we intend to strictly implement an office supplies registration system for office suppliers and to commit to reducing the generation of office waste.

BUSINESS

Employee Responsibility

Employees are fundamental to our success. Recognizing their vital role in driving sustainable growth, we are dedicated to maintaining a fair, safe, and inclusive work environment underpinned by structured policies and robust management systems.

We strictly comply with all applicable labor laws and regulations, with our employment practices formally documented in our standardized Employee Handbook. Every member of our team is employed under a formal contract, and we uphold the principles of non-discrimination, fair treatment, and equal opportunity throughout every phase of the employment lifecycle. We organize regular compliance training sessions. We are committed to fostering a diverse workforce, and as of December 31, 2025, 35.2% of our employees are female.

We place the highest priority on occupational health and safety. All new employees, as well as those transitioning into new roles, receive comprehensive safety training covering our safety protocols and emergency response procedures. Fire drills are conducted regularly across our office premises. We maintain a robust occupational health program, which includes periodic health check-ups and accident insurance coverage, to safeguard the overall well-being of our employees. During the Track Record Period, there were no material occupational health and safety incidents, accidents, or claims in connection with our Group’s business operations.

Community Engagement

We are committed to creating positive impact in the communities in which we operate. Through active engagement and meaningful contributions, we seek to support local development and address social needs. Going forward, we intend to deepen our community involvement by strengthening local partnerships, enhancing our support programs, and identifying new opportunities to contribute to the well-being and sustainable development of surrounding communities.

Anti-corruption

We are firmly committed to anti-corruption and ethical business conduct, and have established a robust compliance framework underpinned by key internal policies, including the Anti-fraud, Anti-bribery and Anti-money Laundering Management System, which prohibits acts such as bribery, embezzlement, and unauthorized disclosure of confidential information. All our employees are required to sign the Anti-corruption Compliance Commitment Letter upon joining, and we have also put in place a confidential whistleblowing mechanism to enable the reporting of misconduct without fear of retaliation.

PATH TO PROFITABILITY

We achieved significant business growth during the Track Record Period. Our revenue increased from RMB31.8 million in 2023 to RMB324.7 million in 2024 and further to RMB520.3 million in 2025, representing a CAGR of 304.5%. We recorded a profit of RMB11.5 million in 2024, while we incurred losses of RMB2.2 million and RMB355.6 million in 2023 and 2025, respectively. As a company founded in 2023, we were at a nascent stage of our business operations that year, actively exploring and refining our business model and product offerings. Our loss in 2023 was primarily attributable to substantial initial investments in research and development activities necessary to establish our technological foundation and product offerings. Since then, we have achieved significant revenue growth, driven by the increasing maturity of our technological capabilities and product and service offerings, as demonstrated through successful deliveries to benchmark customers. This strong foundation has enabled us to capitalize on favorable market tailwinds, with customer demand for our AI computing cluster products and AI computing cluster operation services increasing significantly amid rapid advancements in the AI industry. Our net loss in 2025 was primarily due to changes in the fair values of financial liabilities on shares with

BUSINESS

preferential rights. We recorded adjusted net profit (non-IFRS measure) of RMB20.3 million in 2024 and RMB31.1 million in 2025. Our shares with preferential rights will be automatically converted into ordinary shares upon the [REDACTED] and the related financial instruments will be reclassified from liabilities to equity.

Looking ahead, we intend to achieve and sustain profitability through several strategic initiatives. We will continue to refine our technological capabilities and iterate our products to maintain our market position and address evolving customer needs. We plan to pursue targeted vertical integration with a view to improving product delivery efficiency, reinforcing full-process quality control and supporting long-term cost optimization. We will also continue to execute our sales and marketing strategy, which focuses on serving benchmark customers while expanding our customer base through successful demonstrative deliveries that demonstrate our capabilities. In addition, we plan to further strengthen our supply chain position and improve our supply chain management, enabling us to maintain competitive cost structures. We also aim to enhance cost and operating efficiency through the adoption of digital systems, automation tools, and standardized processes across our operations.

PROPERTIES

As of the Latest Practicable Date, we operated our business through seven leased properties with a total gross floor area of over 1,751.1 square meters in the PRC, including Beijing, Shanghai, Hangzhou and Nanjing. Such properties primarily serve as our office facilities. All such properties have been used for non-property activities as defined under Rule 5.01(2) of the Hong Kong Listing Rules and are primarily used as office premises for our business operations.

Our lease agreements in respect of the abovementioned seven leased properties generally have expiration dates ranging from May 20, 2026 to November 30, 2028. We plan to renew our leases or negotiate new terms when the existing leases expire. All lessors are Independent Third Parties. We did not experience material difficulties in negotiating renewal of our leases with our landlords during the Track Record Period and up to the Latest Practicable Date. We believe that there is sufficient supply of properties in China.

Pursuant to the applicable PRC laws and regulations, property lease agreements must be registered with the local branch of the Ministry of Housing and Urban-Rural Development of the PRC (中華人民共和國住房和城鄉建設部). The registration of such leases will require the cooperation of our lessors. As of the Latest Practicable Date, we had not obtained lease registration for all of our seven leased properties, primarily due to the difficulty of procuring our lessors' cooperation to register such leases. We will take all practicable and reasonable steps to ensure that such leases are registered. As advised by our PRC Legal Advisors, the lack of registration of the lease agreements will not affect the validity of such lease agreements. According to the relevant PRC laws and regulations, we may be ordered by the relevant government authorities to register the relevant lease agreements within a prescribed period, failing which we may be subject to a fine ranging from RMB1,000 to RMB10,000 for each non-registered lease. As of the Latest Practicable Date, we had not received any such request or suffered any such fine from the relevant government authorities. We undertake to cooperate fully to facilitate the registration of lease agreements once we receive any requirements from relevant government authorities.

As of the Latest Practicable Date, the lessors for three of our leased properties had failed to provide property ownership certificates, which may adversely affect our ability to continue to use them in the future. The aggregate leased area of these defective properties is approximately 950.2 square meters. Should disputes arise due to title encumbrances to such properties or government action, we may encounter difficulties in continuing to lease such properties and may be required to relocate. We do not expect to incur significant time for identifying, or incur significant cost to relocate our operations to, comparable alternative properties in proximity. Our Directors believe that relocation will not have a material adverse effect on our business, results of operations and financial condition.

As of the Latest Practicable Date, we were not aware of any challenge being made by a third party or government authority on the titles of any of these leased properties that might have a material adverse effect on our current occupation. In addition, as discussed above, our Directors do not anticipate any

BUSINESS

material practical difficulty in or significant costs relating to identifying comparable alternative premises for any of the defective premises above. There are no rules or regulations requiring the lessee to obtain the ownership certificate or regulatory punishment on the lessee for not doing so. Accordingly, our PRC Legal Advisors has advised that we, as lessee, are not subject to any material administrative penalty for the lessors’ failure to fulfill its obligation to rectify the aforementioned title defects in the leased properties. Moreover, according to relevant PRC laws and regulations, the lessee has the right to claim compensation if the lease agreement is invalid due to the lessor’s fault. In case where our ability to continue leasing such properties is affected by a third-party objection, we may seek indemnity from the lessor in accordance with relevant PRC laws and regulations. As a result, our Directors believe that these title defects would not materially and adversely affect our business, results of operations and financial condition.

As of the Latest Practicable Date, the actual land use of one of our leased properties used for our business operations was inconsistent with its approved land use as specified in its land use right certificate. If the owner of this property is required by government authorities to rectify such land use, we may have to relocate and bear relocation costs and other additional expenses. As of the Latest Practicable Date, we were not aware of any such rectification request by government authorities. As advised by our PRC Legal Advisors, under relevant PRC laws and regulations, it is the lessor (not the tenants) that will be subject to administrative punishment or penalties due to the lessors’ failure to fulfill the responsibility to ensure that the actual use is consistent with the approved use. We would not be subject to any penalty therefrom, but we may not be able to continue leasing such property.

As of December 31, 2025, none of the properties leased or owned by us had a carrying amount of 15% or more of our consolidated total assets. Therefore, according to Chapter 5 of the Hong Kong Listing Rules and section 6(2) of the Companies (Exemption of Companies and Prospectuses from Compliance with Provisions) Notice (Cap. 32L of the Laws of Hong Kong), this document is exempted from compliance with the requirements of section 342(1)(b) of the Companies (Winding Up and Miscellaneous Provisions) Ordinance in relation to paragraph 34(2) of the Third Schedule to the Companies (Winding Up and Miscellaneous Provisions) Ordinance which requires a valuation report with respect to all our Group’s interests in land or buildings.

LEGAL PROCEEDINGS AND COMPLIANCE

During the Track Record Period and up to the Latest Practicable Date, we, or any of our Directors, had not been involved in any actual or pending legal, arbitration or administrative proceedings (including any bankruptcy or receivership proceedings) that we believe would have a material adverse effect on our business, results of operations, financial condition or reputation and compliance.

During the Track Record Period and as of the Latest Practicable Date, we did not commit any material non-compliance of the laws and regulations, and we did not experience any systemic non-compliance incident. Our Directors are of the view that, we had complied, in all material respects, with all relevant laws and regulations in the jurisdictions where we operated during the Track Record Period and up to the Latest Practicable Date.

BUSINESS

LICENSES AND PERMITS

Our PRC Legal Advisors have advised that we had obtained all licenses, permits, and certificate necessary to conduct our operations in all material respects from the relevant government authorities in China, and such licenses, permits and approvals remained in effect as of the date of this document. We also do not anticipate any material difficulty in renewing such licenses, permits and certificate in the future. The following table sets forth material license, permit, and approval currently held by us:

Holder	Licenses/Permits	Issuing Authority	Issued Date	Expiration Date
The Company	Telecommunications Equipment Network Access License	Ministry of Industry and Information Technology of the People’s Republic of China	April 3, 2025	April 3, 2028
The Company	Telecommunications Equipment Network Access License	Ministry of Industry and Information Technology of the People’s Republic of China	April 14, 2026	April 14, 2029
Shanghai Infrawaves Guanglian	Telecommunications Equipment Network Access License	Ministry of Industry and Information Technology of the People’s Republic of China	October 31, 2025	October 31, 2028
Shanghai Infrawaves Guanglian	Value-Added Telecommunications Services License	Ministry of Industry and Information Technology of the People’s Republic of China	April 16, 2026	April 16, 2031

RISK MANAGEMENT AND INTERNAL CONTROL

It is the responsibility of our Board to ensure that we maintain sound and effective internal control and risk management system to safeguard our Shareholders’ investment and our assets. The objectives of our risk management system are to: (i) identify potential events that may impact us, ensuring relevant risks are controlled within acceptable and appropriate levels relative to our objectives; (ii) facilitate accurate and reliable communication of information to both internal and external stakeholders; (iii) ensure compliance of our business operations; (iv) enhance the efficiency and effectiveness of our business activities, thereby reducing uncertainties in achieving operational goals; and (v) establish crisis management plans for significant risks to protect our Company from substantial losses due to catastrophic risks or human errors.

Our Audit Committee is responsible for reviewing the regulations and primary objectives related to risk management, submitting an annual comprehensive risk management report to the Board, reviewing risk management strategies and solutions for significant risks, and addressing other matters related to comprehensive risk management as authorized by the Board. Our internal audit department is tasked with implementing our risk management policies and systems. Other departments and business units are supervised by the internal audit department in their risk management efforts.

The typical operating procedures of our risk management work include the following: (i) collecting information related to risk management, including historical data and future forecasts. Specifically, in terms of strategic risk, we extensively gather cases where domestic and international companies have

BUSINESS

suffered losses due to uncontrolled strategic risks; for financial risk, we collect cases of financial crises caused by uncontrolled financial risks; for market risk, we gather cases where companies have suffered losses due to neglecting market risks and lacking responsive measures; for operational risk, we collect information relevant to us and the industry; and for legal risk, we gather cases where companies have suffered losses due to neglecting legal and regulatory risks and lacking responsive measures; (ii) assessing the significance and likelihood (or frequency and probability) of risks; (iii) senior management assists the Board in determining a unified risk appetite and tolerance based on business characteristics, and establishes risk warning lines and corresponding countermeasures; (iv) senior management formulates risk management solutions for significant risks based on risk management strategies; and (v) senior management supervises the implementation of risk management, making timely improvements based on changes and any deficiencies. Our approach also involves the continuous enhancement of internal control mechanisms and regular reviews of our risk management policies and internal control measures to maintain their effectiveness and adequacy, ensuring that our operations align with industry standards and regulatory requirements.

AWARDS AND RECOGNITION

The following table sets forth major awards and recognitions we received as of the Latest Practicable Date.

Year	Award/Recognition	Awarding Institution/Authority
2025 . .	Future Network Leading Innovation and Technology Achievements “Autonomous and controllable AI computing telecommunication system”	China Institute of Communications Future Network Leading Innovation and Technology Achievements Expert Committee
2024 . .	AI Cloud Platform Demonstration Case for Large-scale AI Computing Centers	The Computing and Network Convergence Industry and Standards Promotion Committee of the China Communications Standards Association
2024 . .	China Computational Power Outstanding Enterprise Award	China Intelligent Computing Industry Alliance
2024 . .	Innovation Engine Award	DPU & AI Networking Conference Committee
2023 . .	China Computational Power Advanced Technology Award	China Intelligent Computing Industry Alliance