GLOSSARY OF TECHNICAL TERMS

This glossary of technical terms contains definitions of certain terms used in this document in connection with our Company and our business. The terms and their meanings may not always correspond to standard industry meaning or usage of these terms and may not be directly comparable to similarly titled terms adopted by other companies operating in the same industries as our Company.

"abstraction"	the process of simplifying complex system details by providing a higher-level interface that allows users or applications to interact with underlying resources without needing to understand their technical intricacies
"accelerator"	a hardware component that enhances computing performance by transferring specific tasks from the central processing unit to specialized processors
"adaptive routing"	a routing method under which network paths are dynamically selected or adjusted based on real-time network conditions in order to improve transmission efficiency and overall network performance
"agentic"	relating to AI systems that can autonomously plan, make decisions and execute multi-step tasks with minimal human intervention to achieve a given objective
"AI"	artificial intelligence, an area of computer science that focuses on simulating human intelligence by machines
"AI agent"	an intelligent system capable of autonomously performing specific tasks on behalf of a user to achieve a proposed goal
"AI chip"	semiconductor chips specifically designed or optimized for performing AI-related computations, such as machine learning training and inference
"AI computing"	the computational resources and processing capability used to develop, train, deploy and run AI models and applications
"AI computing cluster"	a cluster of interconnected computing resources designed to work together to provide computing power for AI workloads, including model training, fine-tuning, inference and deployment
"AI infrastructure"	the hardware, software and system resources that support the development, training, deployment and operation of AI models and applications. It typically includes computing, networking, storage and related platform software and system capabilities
"AI workloads"	the computing tasks performed in connection with AI models and applications, including data processing, model training, fine-tuning, inference and deployment

GLOSSARY OF TECHNICAL TERMS

"AllReduce"	a collective communication operation in parallel computing in which data from all participating processes are combined using a specified operation (such as summation) and the result is distributed back to all processes, commonly used to synchronize model parameters during distributed AI training
"bit error rate"	the number of received bits of a data stream over a communication channel that have been altered due to noise, interference, distortion or bit synchronization errors per unit time
"caching"	the temporary storage of frequently accessed data in a high-speed memory layer to reduce retrieval time and improve processing efficiency
"checkpoint"	a saved snapshot of a model's training state at a given point, enabling training to be resumed from that point in the event of a system failure or interruption
"CNP"	Congestion Notification Packet, a network control message sent to a data source to signal that congestion has been detected along the transmission path, prompting the source to reduce its sending rate
"coefficient of variation"	a standardized measure of dispersion calculated as the ratio of the standard deviation to the mean, typically expressed as a percentage, and used to assess the relative variability of a data set
"collective communication library"	a software library that provides standardized routines for coordinating data exchange and synchronization among multiple processors or nodes in a distributed computing system
"compilation"	the process of translating high-level programming code into lower-level instructions that can be executed by hardware or a runtime environment
"compute node"	a single server or machine within a computing cluster that provides processing power, memory and storage resources for executing computational tasks
"computing resource utilization"	the extent to which available computing resources, such as CPU, memory, storage and network bandwidth, are actively being used relative to their total capacity, serving as a key metric for assessing system efficiency, capacity planning and cost optimization
"computing utilization linearity"	a metric that quantifies how close a system's performance scales to the ideal proportional increase when computational resources are added
"containerization"	a technology that packages software applications and their dependencies into isolated, portable units called containers, enabling consistent deployment and execution across different computing environments

"converged"	the integration and unification of multiple computing elements, such as compute, networking and storage, within a single AI computing cluster or its underlying architecture, enabling coordinated operation, improved efficiency and simplified system design
"convergence"	the process or capability of bringing together and coordinating different components, systems or resources
"CPU"	central processing unit, a complex set of electronic circuitry that runs the machine's operating system and apps
"Decode"	the subsequent stage of AI inference in which the model generates output tokens sequentially, typically one token at a time, based on the processed input context and previously generated tokens
"Dense"	an AI model architecture in which all or substantially all model parameters participate in the processing of each input or token
"distributed system"	a computing architecture in which multiple interconnected machines work together to perform tasks, share resources, and coordinate processes as a unified system
"DRAM"	Dynamic Random-Access Memory, a type of volatile memory that stores data in capacitors, requiring periodic refreshing, and is commonly used as the main memory in computing systems
"ECN"	Explicit Congestion Notification, a network mechanism that marks data packets with a congestion indicator rather than dropping them, allowing endpoints to detect and respond to network congestion without packet loss
"EP"	Expert Parallelism, a distributed computing technique that assigns different expert sub-networks of a MoE model across multiple devices, enabling parallel processing and communication during training and inference
"fine-tuning"	the process of further training a pre-trained AI model on a smaller, task-specific dataset to adapt its capabilities to a particular application or domain
"flowlet-based dynamic traffic balancing"	a network traffic management method that divides a data flow into smaller flowlets separated by brief idle intervals and dynamically distributes such flowlets across multiple network paths based on real-time network conditions, so as to improve load balancing, reduce congestion and enhance transmission efficiency
"foundation model"	a large-scale, pre-trained model developed on broad and diverse datasets designed to serve as a general-purpose model that can be used for solving a wide variety of tasks

GLOSSARY OF TECHNICAL TERMS

"framework"	a structured software platform that provides pre-built tools, libraries, and interfaces to facilitate the development, training and deployment of AI models and applications
"Gbps"	gigabits per second
"GDR"	GPUDirect RDMA, a data transfer technology that enables direct communication between GPU memory and compatible network or peripheral devices without unnecessary data copying through host memory or extensive CPU involvement
"GDS"	GPUDirect Storage, a technology that enables direct data transfer between storage devices and GPU memory, bypassing the CPU and system memory to reduce latency and improve data throughput for GPU-accelerated workloads
"GPU"	graphics processing unit, a specialized electronic circuit designed to manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device
"high-bandwidth memory"	a type of high-performance memory that uses vertically stacked DRAM and wide data interfaces to provide significantly higher data transfer bandwidth and energy efficiency, commonly used to support data-intensive workloads such as AI computing
"HDD"	hard disk drive, a data storage device that uses spinning magnetic disks to read and write digital data
"herd"	a phenomenon in which a large number of requests or processes simultaneously converge on the same resource, causing sudden spikes in demand that can lead to congestion, contention or performance degradation
"hot switching"	the replacement or reconfiguration of hardware or software components while a system remains operational, without requiring downtime or interruption to ongoing processes
"inference"	the process by which a trained AI model uses its learned parameters to process input data and produce output, as distinct from the training stage in which such parameters are developed or optimized
"inference instance"	a single deployed copy of an AI model allocated with dedicated computing resources to process inference requests
"inference replicas"	multiple identical instances of a deployed AI model running concurrently to handle parallel inference requests, improving throughput and availability
"input/output" or "I/O"	input/output, the interface processes that handle data transfer between a computing system and external devices

"jitter"	the variation in latency or delay of data packets transmitted across a network, which can affect the consistency and performance of data communication
"KA customer"	a customer with revenue contribution of over RMB10 million in any period/year during the Track Record Period
"kernel"	a function that executes computation in parallel across a number of GPU threads
"KV Cache"	Key-Value Cache, a memory mechanism used during model inference to store previously computed key and value pairs from attention layers, reducing redundant computation and improving inference speed and efficiency
"KV storage"	Key-Value Storage, a data storage system that organizes and retrieves data as pairs of unique keys and their associated values, enabling fast and efficient data access
"long-context"	the ability to process and retain large amounts of input data or text within a single inference operation
"LoRA"	Low-Rank Adaptation, a parameter-efficient fine-tuning technique that updates only a small subset of a model's parameters by introducing low-rank matrices, enabling faster and less resource-intensive adaptation to specific tasks
"lossless"	referring to data transmission or compression in which no data is lost or degraded, ensuring complete fidelity of the original information
"MEMS-based OCS"	Micro-Electro-Mechanical Systems-based Optical Circuit Switch, an optical circuit switch that uses tiny mechanical mirrors or movable components fabricated through MEMS technology to physically redirect optical signals between ports
"MFU"	Model FLOPs Utilization, a metric indicating the effective utilization of a GPU's floating-point operation capacity when running AI models, reflecting the degree of hardware and software optimization
"MoE"	Mixture-of-Experts, an AI model architecture that uses a group of sub-networks, or "experts," where only a subset is activated for each input. This allows the model to scale efficiently by allocating computational resources based on the nature of the task, improving performance while maintaining efficiency
"MTTR"	mean time to repair
"multimodal"	an AI model architecture designed to support the processing and integration of two or more data modalities, such as text, images, audio or video, enabling such model to perform tasks involving multiple types of inputs and outputs

GLOSSARY OF TECHNICAL TERMS

"node"	an individual computing unit or device within a network or system, such as a server, processor, or workstation, that can process, send or receive data
"NPO" or "near-packaged optics"	an optical interconnect architecture in which optical transceivers or optical engines are placed in close physical proximity to a switch chip, processor or other high-performance computing component, so as to reduce electrical transmission distance, improve bandwidth and power efficiency and support high-speed data communication
"NVMe SSD"	Non-Volatile Memory Express Solid-State Drive, a high-performance storage device that uses flash memory and connects via the NVMe protocol to deliver faster data read and write speeds compared to traditional storage interfaces
"OCS"	Optical Circuit Switch, a network switch that establishes dedicated optical signal paths between ports without converting signals to electrical form, enabling high-bandwidth, low-latency data transmission
"optical transceiver"	a device that converts electrical signals into optical signals, and vice versa, for the transmission of data over optical fiber networks
"orthogonal backplane"	a circuit board architecture in which line cards and switch cards are connected perpendicularly to one another, enabling high-density, high-bandwidth interconnections within a network switch or computing chassis
"packet"	a unit of data formatted for transmission across a network
"packet forwarding rate"	the rate at which a network device, such as a switch, can process and forward data packets, typically measured in packets per second
"parallel computing"	a form of computation in which multiple calculations or processes are carried out simultaneously, typically across multiple processors or cores, to increase processing speed and efficiency
"parameter"	internal numerical variables or configurations that are learned and adjusted by the foundation model during its training process. These parameters effectively encapsulate the knowledge, patterns and relationships extracted by the model from the extensive datasets it has been trained on
"partitioning"	the division of a single physical hardware resource, such as a GPU, into multiple isolated segments that can be independently allocated to different tasks or users
"PCB"	printed circuit board, a flat board made of insulating material with conductive pathways etched or printed onto it, used to mechanically support and electrically connect electronic components

GLOSSARY OF TECHNICAL TERMS

"PFLOPS"	PetaFLOPS, a unit of computing performance equal to one quadrillion floating point operations per second, commonly used to measure the processing power of high-performance computing systems
"post-training"	the stage following pre-training in which an AI model is further refined through techniques such as fine-tuning, alignment or reinforcement learning to improve its performance on specific tasks or desired behaviors
"Prefill"	the initial stage of AI inference in which the model processes the input prompt or context, computes the corresponding internal representations and cache data, and prepares for subsequent output generation
"pre-training"	the initial phase of training an AI model on large-scale datasets to learn general patterns and representations before it is fine-tuned for specific tasks
"Python"	a widely used, general-purpose programming language known for its readability and versatility, commonly used in artificial intelligence, data science, and software development
"PyTorch"	an open-source machine learning framework widely used for building, training, and deploying AI models, known for its flexibility and dynamic computation capabilities
"RDMA"	remote direct memory access
"reinforcement learning"	a machine learning approach in which an AI model learns to make decisions by receiving feedback in the form of rewards or penalties based on the outcomes of its actions
"repurchase rate of KA customers"	the percentage of KA customers who had made at least two purchases of any products or services from the Group as of the Latest Practicable Date, calculated by dividing (i) the number of such KA customers by (ii) the total number of KA customers during the Track Record Period
"RLHF"	reinforcement learning from human feedback, a training technique in which an AI model is refined using human evaluative feedback as a reward signal, guiding the model to produce outputs that better align with human preferences and expectations
"RoCE"	RDMA over converged ethernet, a network protocol which allows RDMA over an Ethernet network

"runtime environment"	the underlying software infrastructure that provides the necessary resources, libraries, and services for executing and managing AI workloads during operation
"sandbox"	an isolated and controlled environment used to carry out activities in a manner designed to limit unintended effects on other environments, systems, resources or data
"SerDes"	Serializer/Deserializer, a hardware component that converts data between serial and parallel formats, enabling high-speed data transmission over communication links
"skew"	an uneven distribution of workload or data across computing resources, which can lead to performance imbalances where some resources are overutilized while others remain underutilized
"SLA"	service level agreement, formal contracts between a service provider and a customer that define the expected level of service, including performance metrics such as uptime, response time and support quality, as well as remedies if standards are not met
"SOTA"	state of the art, the highest level of performance or technical achievement attained by a model, algorithm, or system on a recognized benchmark or evaluation metric at a given point in time
"supercomputer"	a high-performance computing system capable of processing extremely large volumes of data and performing complex calculations at significantly higher speeds than conventional computers
"SuperPod"	a large-scale AI computing unit consisting of multiple interconnected compute servers, GPUs or other accelerators, high-performance networking devices, storage systems and related software, designed to operate as an integrated cluster to support large-scale AI training and inference workloads
"switch"	a network device that interconnects servers, accelerators, storage systems and other computing infrastructure and directs data traffic among them within an AI computing cluster
"Tbps"	terabits per second
"telemetry"	an automated collection, transmission and reporting of operational and performance data from network devices and systems for monitoring, analysis and management purposes
"tensor"	a tensor is a multi-dimensional array of numerical values used to represent and process data within machine learning models
"token"	a unit of text as the fundamental element for input processing and output generation for models
"tool"	an external function, service or resource that an AI agent can invoke to perform specific tasks, such as retrieving data, executing code or interacting with external systems

"tool calling"	the capability of an AI model to identify and invoke external tools or functions during its reasoning process in order to complete a given task
"Top-K routing"	a selection mechanism in MoE models that directs each input token to the K highest-scoring expert sub-networks for processing, balancing computational efficiency and model performance
"topology"	the arrangement or interconnection pattern of nodes, devices or components within a network or computing system
"TP"	Tensor Parallelism, a distributed computing technique that partitions individual model tensors across multiple devices, enabling parallel computation and communication to train or run large-scale AI models that exceed the memory capacity of a single device
"training"	the process of developing an AI model by using data and computational resources to optimize the model's parameters so that it can perform designated tasks or generate outputs based on given inputs
"Triton"	an open-source programming language and compiler d that enables developers to write efficient GPU kernel code for deep learning workloads without requiring low-level hardware programming expertise
"validation"	the process of evaluating a model's outputs against a designated set of reference data to assess its accuracy and performance
"virtualization"	a technology that abstracts physical hardware resources, such as GPUs, into multiple virtual instances, allowing them to be shared, allocated and managed independently across different users or workloads
"VLA model"	Vision-Language-Action Model, an AI model that integrates visual perception, language understanding and action planning capabilities, enabling it to interpret visual and textual inputs and generate corresponding actions
"warp"	a hardware execution unit (commonly in GPUs) comprising a group of threads that execute the same instruction in parallel in a lockstep manner, which may be optimized or specialized for specific workloads or operations
"world model"	an AI model that learns and maintains an internal representation of the physical or virtual environment, enabling it to simulate, predict, and reason about how the world behaves
"white-box"	hardware or system architectures based on open and customizable designs, with transparent specifications that allow users to flexibly select, modify and optimize software components without reliance on proprietary solutions
"zero-copy data transfer"	a data transfer technique that eliminates unnecessary intermediate copying of data between memory buffers during communication, thereby reducing latency and CPU overhead by enabling data to move directly between source and destination