GLOSSARY OF TECHNICAL TERMS

In this document, unless the context otherwise requires, explanations and definitions of certain terms used in this document in connection with our Company and our business shall have the meanings set out below. The terms and their meanings may not always correspond to standard industry meaning or usage of these terms

"Agentic infra"	Agentic infra comprises a suite of infrastructure designed to build a cloud environment in which AI agents can operate autonomously, reflecting a future in which AI agents, rather than humans, will be the primary users of cloud services
"AI"	artificial intelligence
"AI cluster"	a group of computers or servers working together to perform tasks related to artificial intelligence
"AI cloud computing"	the delivery of computing resources, storage, and AI development tools over the cloud to support the training, inference, and deployment of artificial intelligence models. It combines the scalability of cloud infrastructure with the high-performance requirements of AI workloads
"AI-native"	products, services or systems that are built from the ground up with artificial intelligence as a core component, rather than integrating AI as an add-on
"AIDC"	AI data center, a specialized type of data center designed and optimized to handle the immense computational demands of AI workloads. AIDCs rely extensively on GPUs and are designed to support both the training of AI models and the execution of AI inference tasks
"AIGC"	AI-generated content, content created using AI technologies
"AI fine-tuning"	the process of adapting a pre-trained model to perform specific tasks or use cases
"AI infrastructure"	the combination of hardware, software and frameworks necessary to develop, deploy and manage AI applications and solutions
"AIOps"	an advanced approach to managing and maintaining our infrastructure by using artificial intelligence and machine learning technologies
"API"	a set of defined protocols, routines and tools that enables different software applications to communicate with each other. In the context of AI, an API provides a standard way for external applications to access AI-powered services or models without the user needing to understand the underlying algorithms or infrastructure

GLOSSARY OF TECHNICAL TERMS

"bandwidth"	the maximum rate at which data can be transferred between devices, edge nodes and central data centers over the network
"CDN"	content delivery network, a network of geographically distributed nodes close to end users designed to deliver content (such as web pages, images and videos) to users more quickly. By caching content on servers closer to users, CDNs reduce data transmission distances, enhancing load speeds and stability
"CPU"	central processing unit, the primary component of a computer that performs most of the processing inside the computer
"CPU core"	an individual processing unit within a CPU capable of executing its own computational tasks
"compute node"	a physical or virtual server that provides computing resources, such as CPU cores, GPUs, memory and storage
"cross-cloud scaling"	the ability to dynamically allocate resources across multiple cloud environments to handle varying workloads and ensure high availability
"concurrency"	the ability of the cloud infrastructure to handle a large number of AI requests, sessions or agents running at the same time
"edge bare metal"	physical, standalone servers deployed at the edge that provide dedicated hardware resources for computing. Unlike virtualized environments where multiple users or processes share the same physical machine, edge bare metals deliver direct and unshared access to the infrastructure, offering high performance, low resource overhead and precise control
"edge CDN"	a network of servers together with software and applications distributed geographically to optimize the distribution of servers as the servers are located closer to the end users and data origins (compared with traditional CDNs), resulting in faster transfer of data and shorter buffering time
"edge cloud"	a cloud computing model where cloud resources, such as data storage and processing power, are moved closer to the location where data are generated and where end-users are located
"edge computing"	a distributed computing paradigm that brings computing and storage closer to the sources of data, with benefits such as improved response times and better bandwidth utilization compared to cloud computing

"edge container"	containers are virtualized units that allow developers to package their applications into a single portable unit. Edge containers are functionally similar to traditional containers but are deployed on edge devices or nodes. They allow applications to run efficiently in the online edge compute nodes which are at or near where data are generated, reducing latency and improving performance for real-time applications.
"edge node"	computing resource that composed of one or more devices located close to the data source at the edge of a network, designed to facilitate data processing and analytics near the data's origin
"generative AI"	a subfield of AI that focuses on creating new content, such as text, images, videos or audio, based on patterns learned from existing data, such as multimodal content generation, sound cloning and creating virtual hosts
"GPU"	a specialized processor originally designed for rendering graphics, but now widely used to accelerate parallel computations in fields like artificial intelligence, especially for training and inference of deep learning models
"IDC"	internet data center, is a specialized physical facility designed to house a large number of computer systems, network equipment and associated infrastructure (such as power supply, cooling systems and security systems). Primarily relies on CPUs and is designed to support a broad range of general-purpose IT workloads. The broader definition of an IDC now includes both traditional CPU-based IDCs and GPU-based AIDCs.
"LLM"	large language model, a deep learning model trained on large-scale text data to understand, process and generate human language. It specializes in natural language tasks and forms the foundation of many language-based AI applications
"latency"	the time delay between sending data or a request from a user device or AI agent to the cloud and receiving the model's output or response
"MB"	megabyte, a measure of computer storage. 1MB equals approximately 1 million bytes
"model API"	an interface that allows developers to access and use pre-trained AI models via standardized API calls, enabling tasks such as text generation, image recognition, or speech processing without managing the underlying infrastructure

"model training"	the process of optimizing a machine learning model's parameters using data, enabling it to learn patterns and perform specific tasks. It typically involves forward passes, loss calculation and backpropagation
"model inference"	the process of applying a trained machine learning model to new input data to generate predictions, decisions, or structured outputs. It represents the deployment-time execution of the model's learned parameters, often involving complex reasoning, pattern recognition, or language understanding in real-world scenarios
"multi-cloud platform"	the use of cloud computing services from multiple cloud providers within a single architecture
"multimodal model"	an AI model trained to process and understand multiple types of input data—such as text, images, audio, or video—enabling it to perform reasoning and generation across different modalities in a unified way
"open-source model"	a model whose source code is made freely available for possible modification and redistribution
"PUE"	power usage effectiveness, a metric used to measure the energy efficiency of a data center
"QoS"	quality of service
"SDK"	software development kit, a set of software development tools that can be used to create and develop applications
"SLA"	service-level agreement, a contract between a service provider and a customer that outlines the services to be provided and the expected level of performance
"token"	the basic unit of data processed by AI models
"VPC"	virtual private cloud, a secure, isolated segment of a public cloud that allows customers to run code, store data, host websites and perform other tasks as they would in a private cloud, but with the infrastructure hosted by a public cloud provider