Magic Quadrant for Cloud AI Infrastructure

6 July 2026 - ID G00851200 - 62 min read

By Dennis Smith, Alessandro Galimberti, and 4 more

The cloud AI infrastructure market comprises cloud service providers delivering optimized infrastructure for model training, inference and agentic AI. This Magic Quadrant helps organizations navigate this fast-moving market of major hyperscaler vendors and specialized cloud vendors.

Market Definition/Description

Gartner defines cloud AI infrastructure as cloud service providers that focus on delivering infrastructure optimized for AI workloads including AI model training, inference and servicing. This market includes major hyperscaler vendors and specialized cloud vendors delivering AI-optimized infrastructure.

Cloud AI infrastructure offerings address the need to enable AI and machine learning (ML) activity. The offerings include a combination of compute, storage and networking components, along with the requisite enablement tooling, middleware and libraries.

Mandatory Features

Cloud AI infrastructure must have infrastructure and/or infrastructure-related tools that support building, deploying and/or operating AI workloads. This includes the enablement of:

- Automated aspects of building, deploying and maintaining AI models
- Data processing and preparation of data for model training and inference
- Integration of AI/ML libraries

- Tools enabling the deployment and operation of AI/ML workloads
- Security, including encryption, access controls and compliance
- Governance of the data and the underlying infrastructure
- Server clusters with AI accelerators (e.g., CPUs, ASICs [TPUs, Trainium, Maia, Ascend, Kunlunxin, Zixiao, Zhenwu etc.], FPGAs, graphics processing units [GPUs])
- High-throughput, low-latency and high-capacity storage necessary to support the large datasets and high-speed data transfers that keep GPUs engaged with large quantities of data
- Managed AI platform capabilities — model training platform capabilities, inference engines, APIs for pretrained models, AutoML services and MLOps tools

Optional Features

- A combination of compute, storage and networking infrastructure optimized for AI workloads. This includes additional compute modes such as serverless compute and container services, as well as object, block and file storage, and data lakes.
- Software that enables the operation of the underlying infrastructure that supports AI workloads. This includes a combination of automated provisioning, optimization, observation and/or security capabilities for the underlying infrastructure.
- Dedicated networks for training (e.g., a supercomputing network with lossless architecture) and inference (lower cost but low-latency). Additional networking capabilities include virtual private cloud and load balancing.
- Support for frameworks, AI engineering tools and computing libraries for building AI models.
- High reliability and regulatory compliance that serve diverse industries (SOC, PII, HIPAA, FEDRAMP etc.) to safely and securely integrate with systems of record.
- Advanced data centers supporting efficient power, high-performance networking and advanced cooling technologies to help high-density AI-optimized server deployments.
- Orchestrating large-scale AI model training and inference (e.g., Kubernetes, Slurm).

- Low-latency and lossless networking with high throughput — including near-range networks for data transfer between AI accelerators and/or CPUs (e.g., NVLink and NVLink Switch).
- High-speed, lossless networks interconnecting AI servers (e.g., InfiniBand, RDMA over Converged Ethernet [RoCE]).
- Observability and monitoring: cloud monitoring, cost management and budgeting, logging, and error tracking.
- Security and access controls: Identity and access management, key management service, security posture management.
- Analytics and data pipelines capabilities: data ingestion tools, ETL, big data processing, BI and visualization.
- Model garden or model hub.
- Fault tolerance for training, such as automated failure detection, checkpointing and job resumption.

Magic Quadrant

Figure 1: Magic Quadrant for Cloud AI Infrastructure

Gartner

Vendor Strengths and Cautions

Alibaba Cloud

Alibaba Cloud, headquartered in Hangzhou, China, offers a comprehensive spectrum of cloud AI infrastructure services, encompassing everything from high-performance infrastructure to AI model and agent development platforms, and pretrained foundation models. Its infrastructure features a mix of internally developed and third-party compute resources, underpinned by a variety of pricing models.

A key offering is PAI-Lingjun, a specialized compute platform optimized for large-scale training and inference workloads, powered by high-performance GPUs and its proprietary PPU chips. This platform boasts high-speed networking capabilities through

eRDMA and supports a wide range of storage options, such as CPFS and EBS. Alibaba Cloud delivers an end-to-end platform for AI life cycle management, with key capabilities of AI data annotation, interactive modeling via Jupyter Notebook and distributed training for trillion-parameter models. It also offers a dedicated Model Studio as a model-as-a-service platform for its entire Qwen family of models (such as Qwen3.7) and third-party integration (such as DeepSeek, MiniMax and GLM). The overall portfolio is a full-stack offering, also integrating essential data services including databases and data warehouses.

As a Leader in this Magic Quadrant, Alibaba Cloud should be considered by enterprises seeking a comprehensive, full-stack AI provider with a strong and established presence across China and the broader Asia region.

Strengths

- **Full-stack AI portfolio:** Alibaba Cloud offers a comprehensive spectrum of cloud AI infrastructure services, including raw high-performance infrastructure, specialized development platforms, and pretrained foundation models. Its portfolio is appealing to enterprises seeking a unified, full-stack AI provider that delivers integrated solutions for seamless operations.

- **Optimized AI supercomputing platform for large-scale workloads:** The PAI-Lingjun specialized compute platform is engineered for large-scale training and inference workloads with heterogeneous computing infrastructure, and is proven by the Alibaba Group's own advanced AI workloads. This specialized platform supports organizations with demanding AI applications that require highly efficient, low-latency performance essential for scaling model training and high-speed inference.

- **Open-source innovation:** Alibaba Cloud offers a set of open-source technologies that offers strong community support; for example: EasyNLP, EasyCV, PAI-DSW, SRv6 and SONiC. Some of the Qwen open-weight models are among the most popular in the world. This will appeal to enterprises wanting technological sovereignty and to avoid lock-in to vendor proprietary technology.

Cautions

- **Steeper learning curve for users:** Some Gartner users have noted that the interface and documentation for Alibaba Cloud are more complex to master, compared to some of those from other U.S. hyperscale competitors. This complexity might increase

time to market for new projects or necessitate hiring staff with specific Alibaba Cloud expertise.

- **Technology embargo impact:** While Alibaba Cloud has initiated efforts to develop proprietary technology, it still lacks access to some of the most advanced technology currently available in the market due to embargoes. Enterprises relying on the absolute cutting edge of AI accelerator technology (e.g., specific GPUs or specialized components) might face potential performance gaps or limitations, compared to competitors that have unrestricted access to the latest hardware innovations.
- **Regulatory and data sovereignty concerns:** Storing data on a Chinese-owned and operated cloud provider can face regulatory hurdles within certain regions of the world, particularly in globally regulated industries such as healthcare and finance. Organizations in these regulated sectors must conduct a rigorous compliance audit regarding data governance, sovereignty, and regulatory requirements before committing to Alibaba Cloud.

Amazon Web Services

Amazon Web Services, headquartered in Seattle, Washington, U.S., leverages its resilient and scalable infrastructure to provide a comprehensive suite of cloud AI infrastructure services that supports extensive training and inference use cases. The infrastructure is built on purpose-built silicon (theirs and third-party), high-speed networking, and highly performant storage, which is combined with a set of management services.

The compute capabilities include AWS' proprietary silicon, Trainium and Inferentia, along with compute resources from a wide variety of other vendors (e.g., NVIDIA, Intel and Cerebras). AWS Elastic Fabric Adapter (EFA) enables networked distributed training, while EC2 UltraClusters enables training across large clusters. AWS provides high-performance object and file storage services. Services include Amazon Bedrock for inference, Amazon SageMaker AI for training and custom inference, Bedrock AgentCore for agentic AI development, and Nova Forge for model development. On-premises AI offerings are provided through Outposts and AWS AI Factories, and AWS also offers edge capabilities. AWS uses a tiered model, including on-demand, Savings Plans, and Spot Instances.

AWS is a Leader in this Magic Quadrant, offering a broad and deep service portfolio able to handle every cloud AI infrastructure use case.

Strengths

- **Foundational cloud capabilities:** AWS brings one of the industry's most extensive and mature cloud infrastructures to AI workloads, combining broad global reach, multi-availability-zone resiliency, and deep security/compliance capabilities. This ensures that organizations can confidently deploy mission-critical, large-scale AI workloads, knowing the underlying infrastructure meets the highest standards for uptime, data integrity and compliance.

- **Deep security and governance for enterprise AI:** AWS combines hardware-rooted isolation in the Nitro System with a broad compliance and governance posture, including ISO/IEC 42001 certification support. This strengthens AWS' appeal for enterprises running sensitive or regulated AI workloads that require high assurance around security, operational control and responsible AI governance.

- **Reduced operational overhead:** AWS offers multiple levels of managed AI infrastructure, from serverless foundation model access in Amazon Bedrock to large-scale model development in SageMaker HyperPod. This gives enterprises more choice in how much infrastructure they want to manage directly, helping reduce operational overhead and accelerate time to value for AI initiatives.

Cautions

- **AI training network latency constraints:** AWS' networking architecture and cluster design are less performant than some other vendors in this evaluation, specifically for ultralarge distributed AI training. Although AWS is more flexible than more performant offerings that are vertically integrated and has strong partnerships with the frontier AI labs, organizations using AWS Trainium chips for massive training must factor in AWS' performance gaps compared to competitors, which may lead to higher total cost of ownership (TCO), software optimization needs and greater environment complexity.

- **Offering strategy:** While AWS has one of the broadest sets of cloud AI infrastructure services, it often fails to present these services in a consistent and clear narrative. This complexity can lead to longer deployment cycles, increased reliance on expensive external consulting resources, and potential overengineering of solutions, increasing time to market and TCO for AI initiatives.

- **Overlapping AI service paths can complicate standardization:** AWS provides multiple ways to build and run AI workloads — including Amazon Bedrock, SageMaker

HyperPod, EC2, Outposts, and AI Factories — each with different levels of abstraction, control and operational responsibility. While this gives customers flexibility, it can also make enterprisewide standardization and reference architecture development more difficult, especially for organizations trying to enforce common patterns across teams.

Cloudflare

Cloudflare, headquartered in San Francisco, California, U.S., is a global cloud network that provides a suite of edge AI infrastructure services focused specifically on AI inference at the edge to meet stringent latency requirements by running AI close to user activity.

Cloudflare's offerings include a serverless platform for inference (Workers AI) and a distributed vector database (Vectorize) that supports retrieval-augmented generation (RAG) requirements. These are supported by an AI Gateway that acts as a proxy between applications and AI models. Capabilities include providing analytics/logging, caching, rate limiting, request retries and model fallback. AI Gateway also frontends Workers AI, major third-party model providers and custom providers. Cloudflare also provides S3-compatible object storage (R2).

Cloudflare is a Niche Player in this Magic Quadrant, and is well-positioned to capitalize on inference use cases as a strategic edge partner alongside centralized cloud providers that handle large-scale large language model (LLM) training. Cloudflare does not charge per-token fees for its AI Gateway; it uses a freemium model where core features are free, while high-volume usage is billed via Cloudflare Workers or measured in "Neurons" for its Workers AI product.

Cloudflare declined requests for supplemental information. Gartner's analysis is therefore based on other credible sources.

Strengths

- **Low-latency inference handling:** Cloudflare operates GPU-enabled servers in over 300 locations, often delivering responses with latency lower than 50ms, depending on the model. This is essential for highly responsive conversational AI, and instant content moderation, where a delay of even a few milliseconds can significantly impact user experience and business outcomes.

- **Desirable pricing:** Pricing is based on actual inference usage (Neurons, which abstract the underlying GPU compute across models) rather than idle GPU capability. This consumption-based model allows enterprises to control costs more effectively, focusing investment on optimization and governance for production inference use cases, rather than paying for underutilized infrastructure.
- **Zero egress fees:** Cloudflare's R2 object storage eliminates data egress fees. This significantly reduces the cost barrier for enterprises to adopt multicloud strategies, enabling them to move data freely between cloud environments for redundancy, best-of-breed services, and negotiation of leverage without incurring prohibitive network charges.

Cautions

- **Focused only on inference:** Cloudflare does not currently support large-scale AI model training, meaning customers would need to partner with other providers for training services. Enterprises pursuing a full life cycle AI strategy, from initial model training to deployment and inference, typically would use a centralized provider for the resource-intensive training phase, while utilizing Cloudflare for distributed edge inference.
- **Limited model customization:** Workers AI historically focused on curated open-source models. Cloudflare has recently acquired Replicate (closed January 2026) and plans to integrate the Replicate model catalog and fine-tuning pipeline, so organizations with unique datasets or regulatory requirements that necessitate the fine-tuning of proprietary models will benefit from tracking this ongoing product convergence.
- **Resource limits:** Workers AI has fixed memory and CPU limits, potentially hindering its ability to support complex, memory-intensive AI tasks (Cloudflare intends to leverage Replicate to address this issue). The platform limits target individual model size boundaries rather than operational volume, and it regularly supports high-throughput enterprise inference for optimized models.

CoreWeave

CoreWeave, headquartered in Livingston, New Jersey, U.S., is a pure-play cloud AI infrastructure vendor that offers purpose-built infrastructure for compute-intensive AI

workloads. This infrastructure supports large-scale model training, fine-tuning and high-throughput inference. CoreWeave operates at over 40 data centers.

CoreWeave's range of cloud AI infrastructure services includes a wide array of compute services mainly based on NVIDIA GPUs. CoreWeave utilizes InfiniBand to enable high-speed connectivity among its GPUs. It provides various storage capabilities: object, file, dedicated and local. For object storage, CoreWeave offers the Local Object Transport Accelerator (LOTA) to handle stringent performance needs. By leveraging LOTA and SUNK Anywhere, CoreWeave has established a hybrid and multicloud AI strategy, abstracting hardware complexities to orchestrate workloads across diverse cloud providers and on-premises infrastructure.

A number of management services are also offered, including a Kubernetes-native environment and SUNK, which combines Slurm and Kubernetes. They also provide Mission Control, a common place to manage all aspects of operation, along with Weights & Biases integration (for model and agent development) that assists with experiment tracking, model evaluation, and monitoring.

CoreWeave is a Visionary in this Magic Quadrant. It employs a granular pricing model where users pay separately for GPU, CPU, RAM and storage. In early 2026, it introduced Flex Reservations and Spot to balance predictable training needs with variable inference spikes.

Strengths

- **Purpose-built AI infrastructure:** CoreWeave's cloud is designed specifically for high-density, liquid-cooled racks, offering higher GPU cluster performance than generalized clouds. This specialized infrastructure minimizes thermal constraints and maximizes the efficiency of large-scale, distributed AI workloads, leading to faster training times.

- **Cost-efficiency:** CoreWeave offers lower costs for GPU instances compared to traditional hyperscalers, including variable pricing models and no data egress charges, along with a data ingress solution to support customer data migration at reduced or zero cost. This financial model reduces TCO for resource-intensive AI projects and provides predictable operational budgeting.

- **Kubernetes-native platform and HPC integration:** CoreWeave provides a native Kubernetes environment and offers SUNK (which combines Slurm and Kubernetes).

This expertise enhances the training platform and lowers latencies for inference workloads, allowing organizations to leverage familiar container orchestration and high-performance computing (HPC) scheduling tools for optimized AI workflows.

Cautions

- **High customer concentration:** Historically, a large portion of CoreWeave's revenue has been concentrated among a few major clients, including some hyperscalers. This high concentration creates financial volatility and risk of service disruption if a major client departs, posing a risk to the long-term stability and resource planning for other customers.
- **Extreme capital intensity:** The business is heavily reliant on debt to finance the expansion and purchase of expensive GPU infrastructure during a time of rapid technology obsolescence. This high reliance on debt in a rapidly changing market creates financial risk, which could constrain future resource availability or lead to unpredictable cost structures for users.
- **NVIDIA dependency:** While the partnership is a strength, the heavy reliance on NVIDIA for supply of high-demand GPUs and financial support introduces supply chain risk and potential price volatility, as CoreWeave's ability to scale is tied to NVIDIA's allocation decisions. CoreWeave neither has any custom silicon offerings, nor has near-term plans to develop its own silicon to differentiate its AI compute offerings.

Crusoe

Crusoe, headquartered in Denver, Colorado, U.S., offers a vertically integrated AI infrastructure platform, specializes in building high-density, sustainable data centers, and takes an energy-first design approach. Its services are designed for large-scale training and inference, providing lower costs compared to traditional cloud providers via a flexible pricing model.

Crusoe offers GPU resources from both NVIDIA and AMD, available as virtualized instances and as a managed Kubernetes service, which is now powering more than half of its fleet. These are delivered through comprehensive training and inference management services. Low-latency networking (InfiniBand) and high-performance storage (file and object) are also offered. All of the infrastructure is supported by high-density cooling technology, including liquid cooling. It also provides distributed edge

compute units designed for training and low-latency inference that can be deployed as dedicated regions. Additionally, it offers managed services that support Kubernetes and Slurm, enabling scalable large-scale workloads. Embedded observability provides automated node swapping and real-time monitoring.

Crusoe is a Visionary in this Magic Quadrant and will appeal to enterprises that need high-performing AI infrastructure with strong sustainability requirements.

Strengths

- **Sustainability:** Crusoe prioritizes clean energy resources (such as solar, geothermal, hydro and wind) and stranded or otherwise low-cost energy resources to power its data centers, which are designed for high-density AI compute, allowing it to pass on cost savings to clients. This approach appeals to enterprises seeking to meet environmental, social and governance (ESG) commitments while securing high-performance AI infrastructure.

- **High-performance focus:** Crusoe delivers purpose-built AI infrastructure, featuring specialized hardware like high-density cooling technology and low-latency networking (e.g., InfiniBand), alongside software optimizations tailored for large-scale model training and inference. This is ideal for AI-native organizations, researchers and enterprises that require maximum parallel processing and highest performance computing capabilities.

- **Operational expertise:** Crusoe provides a high-reliability environment, guaranteeing 99.5% uptime for production AI workloads, complemented by a responsive technical support team that has demonstrated proficiency in rapid problem detection and resolution, minimizing disruption. The high-uptime guarantee and expert support ensure business continuity for critical AI applications and training pipelines.

Cautions

- **Niche infrastructure focus:** Crusoe operates as an AI infrastructure specialist and, consequently, does not provide the extensive, integrated ecosystem of complementary services (such as proprietary platform services, databases or traditional cloud compute options) that are available from large hyperscaler vendors included in this evaluation. Enterprises with broad, hybrid IT requirements will need to integrate Crusoe's compute resources with other cloud providers or internal systems

for non-AI workloads, potentially leading to increased operational overhead and fragmentation of the IT environment.

- **Technical requirements:** Crusoe's compute instances and managed Kubernetes/Slurm offerings are tailored for high-performance, large-scale workloads and presuppose a specialized user base (AI researchers, data engineers and skilled ML teams). Organizations lacking internal, highly skilled AI/MLOps talent may face friction in adopting and optimizing the platform, limiting appeal for general enterprise users seeking ease of use or abstract, low-code AI development environments.
- **Data center coverage:** While Crusoe is actively expanding its physical data center footprint, its geographic distribution remains limited, compared to the global presence of hyperscale cloud providers. Enterprises with strict data sovereignty mandates or a need for ultra-low latency inference in specific global markets may find their deployment options restricted, potentially necessitating a multivendor strategy or limiting the reach of edge deployments.

Google

Google Cloud, headquartered in Mountain View, California, U.S., provides a comprehensive AI infrastructure stack designed for high-performance training, tuning and serving of large models, built around first- and third-party hardware and open-source software integration. The core of the ecosystem is the AI Hypercomputer architecture and advanced networking.

Google provides a varied pricing model, including consumption-, value- and commitment-based options. Google offers Tensor Processing Units (TPUs), its custom-designed ASICs optimized for machine learning. It also partners with NVIDIA for GPUs. For data, Google offers high-performance storage (e.g., Managed Lustre) and BigQuery for direct data training. Google Kubernetes Engine is optimized for scaling nodes that enable AI workloads. Finally, Gemini Enterprise Agent Platform, the evolution of Vertex AI, is Google Cloud's higher-level platform for model and agent development. It provides access to Model Garden (a marketplace of AI models), low-code and code-first development tools, runtime and governance services, and the broader tooling needed to build, deploy, and manage enterprise AI and agent workflows.

Google Cloud is a Leader in this research and should be considered by organizations seeking massive-scale training, inferencing and a full AI stack spanning from the

hardware layer to AI agents and applications.

Strengths

- **Proprietary, scalable compute:** Google's custom-designed TPUs, designed to support Google DeepMind, are purpose-built for massive-scale, distributed model training and low-latency inference. This proprietary silicon allows enterprises and researchers to achieve greater efficiency and scale for foundation model development compared to generalized cloud infrastructure, and reduces the dependency on the clogged supply chain of NVIDIA.
- **Integrated AI Hypercomputer architecture:** The AI Hypercomputer combines TPUs/GPUs with optimized high-performance networking and storage (like Managed Lustre) into an integrated architecture. This deep integration minimizes bottlenecks and maximizes the efficiency of large-scale distributed AI workloads, resulting in faster training times and higher throughput.
- **Large owner of AI compute:** Gartner estimates that Google accounted for approximately one-quarter of worldwide cumulative AI compute capacity by 4Q25, making it one of the largest individual owners in the market. Unlike other hyperscalers, Google's AI compute footprint is predominantly based on proprietary TPU silicon, in addition to NVIDIA GPU infrastructure.

Cautions

- **Complex cost management across platform and infrastructure layers:** Google Cloud's AI pricing spans multiple service layers, including Gemini Enterprise Agent Platform, generative model usage, TPUs, GPUs and AI Hypercomputer consumption options. While pricing is documented, the combination of token-based charges, infrastructure costs, deployment choices and commitment models can make TCO forecasting and chargeback governance more difficult for large-scale AI initiatives, particularly when workloads span both managed platform services and custom infrastructure.
- **Optimization trade-offs in a TPU-led infrastructure strategy:** Google's differentiated TPU footprint can provide strategic supply and performance advantages, but enterprises that optimize heavily for TPU-based environments may take on portability trade-offs. Although Google has expanded support for open frameworks such as PyTorch, JAX and vLLM, customers pursuing maximum performance may still need to align tooling, runtime behavior, and engineering practices to Google-specific

infrastructure and software paths, which can complicate migration across heterogeneous or strictly standardized multicloud environments.

- **Scarcity of high-demand resources:** Despite offering its own proprietary hardware, access to third-party, high-demand NVIDIA GPU instances can still experience availability constraints, a common challenge across the market. This constraint can disrupt large-scale training projects, particularly for critical model development, leading to missed deadlines, stalled innovation and dependence on capacity planning, which reduces operational flexibility and responsiveness to market needs.

Huawei Cloud

Huawei Cloud, headquartered in Shenzhen, China, offers a wide range of cloud AI infrastructure services, buttressed by its self-developed Ascend processors, Ascend compute architecture for Neural Networks (CANN) and its MindSpore computing framework. Its strategy combines engineered hardware and software in solutions deployed in the public, private cloud and edge environments, enabling training and inference requirements.

Among Huawei Cloud's offerings is an AI development platform (ModelArts) for the full life cycle of AI, including data processing algorithm development, model training, deployment and management. It provides Elastic Cloud Service (ECS) as virtualization compute environments and also offers Bare Metal Servers (BMS) equipped with Ascend 910 chips for both training and inference workloads. CloudMatrix enables large scale training by pooling CPU, DPU, NPU and storage resources. Huawei Cloud offers pretrained industry-specific AI models (Pangu models) for a number of industries and also provides open-source LLM models, such as Qwen and DeepSeek, and multimodality models. Additionally, Huawei Cloud offers AI Data Lake and DataArts Studio for processing and governing data. Finally, CloudPond and HiLens enable AI deployment on-premises and to edge.

Huawei Cloud is a Leader in this Magic Quadrant, with strengths in China that are also applicable for other regions. It offers a flexible pricing model for its AI infrastructure services.

Strengths

- **Internally developed full stack:** Huawei Cloud provides a vertically integrated solution, pairing its self-developed Ascend NPU and Kunpeng processors with

optimized CANN, and the MindSpore computing framework and software stack. This control over the hardware and software stack enables deep optimization and resource efficiency, resulting in faster training times and a better cost-performance ratio for enterprises running large-scale AI workloads.

- **Comprehensive AI life cycle support:** The product portfolio covers the entire AI life cycle, anchored by ModelArts, a development platform that also functions as an industry AI foundry. It enables specialized industry AI use cases for enterprises through a set of pretrained Pangu models, specialized industry datasets, agent development platforms and CloudMatrix for large-scale training and inference needs. This breadth of services simplifies vendor consolidation, allowing enterprises to manage the entire AI pipeline from development to deployment within a single ecosystem, and reducing MLOps complexity.
- **Strong hybrid and private cloud offerings:** Huawei Cloud's strategy to deploy solutions in both public and private cloud environments directly addresses enterprise demand for hybrid AI infrastructure. This provides deployment flexibility and is essential for organizations with data sovereignty needs or those that require localized inference on-premises.

Cautions

- **Geopolitical and adoption headwinds:** Due to ongoing geopolitical tensions and increased regulatory scrutiny, the adoption of Huawei Cloud's infrastructure, particularly by Western-based organizations, may be limited, restricting its global expansion and access to North America and Europe. This exposes organizations, particularly multinational companies, to supply chain risks and compliance limitations, potentially hindering the ability to standardize on Huawei's platform globally and requiring costly multivendor strategies for AI deployment.
- **Proprietary technology lock-in:** Heavy reliance on the proprietary Ascend/MindSpore/CANN ecosystem means enterprises must invest in specialized skills and porting efforts, which can create vendor lock-in and increase the learning curve for development teams accustomed to industry-standard frameworks like NVIDIA CUDA. This elevates TCO due to the need for specialized training, limits the available talent pool, and creates migration challenges and extra migration efforts if the enterprise later decides to shift AI workloads to another cloud AI infrastructure provider.

- **Limited global footprint for low latency:** While Huawei Cloud is a global company, its core AI infrastructure footprint and support ecosystem is less extensive outside of its primary operating regions, compared to global hyperscalers. This limits the ability of global enterprises to deploy low-latency inference services in critical Western markets, potentially impacting time-sensitive AI workloads like real-time customer service or fraud detection in those regions.

IBM

IBM, headquartered in Armonk, New York, U.S., provides a set of cloud AI infrastructure products mostly targeting large enterprises with hybrid and multicloud environments, and focusing on more highly regulated industries. Its focus is also on security and compliance, emphasizing flexible infrastructure deployment through the use of Red Hat's OpenShift and leveraging the watsonx platform.

Compute offerings include accelerators from NVIDIA, AMD and Intel. IBM also provides specialized AI hardware for their mainframe and midrange systems (IBM Z and Power systems) and offers a bare-metal service. IBM offers a wide range of storage offerings and Ethernet networking options. The watsonx platform enables model building, data lakes and governance. This platform can be deployed on top of OpenShift, which can be deployed in a hybrid or multicloud environment.

IBM is a Visionary in this Magic Quadrant, geared to enterprises that desire consistent software (OpenShift) that can be deployed across hybrid and multicloud environments. It offers consumption-based pricing.

Strengths

- **Broad compute portfolio for enterprise integration:** IBM's compute offering includes accelerators from NVIDIA, AMD and Intel, alongside specialized AI hardware for IBM Z and Power systems, enabling enterprises to protect and extend the value of their IT investments while tailoring the right compute for diverse and specialized AI workloads. This flexibility facilitates incremental AI adoption without requiring a complete overhaul of existing enterprise infrastructure.

- **Consistent hybrid and multicloud platform:** Red Hat OpenShift provides a consistent operating environment that supports AI workloads across public cloud, on-premises and multicloud locations, which is essential for enterprises that require a heterogeneous operational model to meet their hybrid AI infrastructure demand. This

reduces the friction associated with moving AI applications between environments and enables operational continuity, regardless of the deployment location.

- **Strong research capabilities:** IBM Research is well-funded and has a strong track record of driving innovation. This high volume of research underscores IBM's commitment to monetizing innovation, not only through client-facing capabilities, but also via robust research and intellectual property licensing.

Cautions

- **Limited appeal for frontier AI labs:** IBM Cloud lacks the necessary compute scale for training massive foundational models, hindering adoption by frontier AI research labs and model providers. This restricts the ability of IBM Cloud to improve its offering based on the lessons learned from training large models at scale, and forces customers to leverage IBM Cloud as their sole AI cloud for all workloads.
- **Heavy ecosystem dependency:** The value proposition is tightly coupled with adoption of the Red Hat OpenShift and watsonx ecosystem, which can result in vendor lock-in and a higher learning curve for development teams not already standardized on IBM's integrated software stack. Enterprises risk increased TCO and slower MLOps adoption if their development teams lack familiarity with the proprietary stack, constraining their ability to recruit from a broader talent pool.
- **Public cloud price/performance: T**he focus on hybrid deployment means the public cloud infrastructure as a service (IaaS) offering may not appear as competitive on price/performance or as broad in feature depth for purely cloud-native AI workloads, when compared to dedicated hyperscalers or specialized cloud AI vendors. Enterprises focused solely on public cloud deployments may experience higher running costs or diminished performance optimization for cloud-native AI applications, impacting budget predictability and efficiency.

Lambda

Lambda, headquartered in San Jose, California, U.S., provides on-demand and reserved access to cloud AI infrastructure centered around NVIDIA GPUs. Its primary service, Lambda GPU Cloud, provides high-performance infrastructure for training and inference requirements.

Compute resources are delivered from a wide variety of NVIDIA GPUs, often with large-scale multinode clusters for training massive foundational models. These compute nodes leverage InfiniBand for networking. A wide variety of storage services (excluding block storage) are also offered. Lambda also offers a managed Kubernetes service, along with a software development kit (SDK), Lambda Stack, that enables easy environmental setup. It also has a hosted private cloud offering.

Lambda is a Niche Player in this Magic Quadrant, specifically positioned for researchers and enterprises that need high-density, nonabstracted computing power. It uses a consumption-based pricing model.

Strengths

- **Cost-efficient model with zero egress fees:** Lambda offers consumption-based pricing that is often lower than other vendors in the market, and the zero egress fees reduce the cost barrier for organizations adopting multicloud strategies and lower the TCO for data-intensive AI workloads.
- **Prioritized access to next-generation hardware:** Lambda has an investment relationship with NVIDIA and has access to its newest GPUs, often before they are widely available to many other vendors in the market. This supply chain advantage is critical in a market defined by GPU scarcity, allowing enterprises to innovate with cutting-edge performance, lower overall compute costs, and reduce project delays caused by hardware constraints.
- **Developer-centric environment and managed services:** The platform provides a straightforward interface for skilled users to customize their access, complemented by an SDK (Lambda Stack) and a managed Kubernetes service. These capabilities reduce the specialized MLOps skills required to provision and manage training environments, enabling teams to onboard faster and focus primarily on model innovation.

Cautions

- **Narrow AI and PaaS ecosystem:** Lambda's portfolio remains fundamentally rooted in core IaaS compute, lacking the comprehensive suite of managed platform as a service (PaaS), serverless options and broader data services typically offered by hyperscalers. The absence of a robust, integrated AI ecosystem — including model marketplaces

and proprietary agentic AI platforms — limits Lambda's utility for enterprises seeking a single platform for the end-to-end AI life cycle.

- **Limited geographic presence:** The concentration of Lambda data centers primarily in the U.S. creates a hurdle for global enterprises, leading to potential latency issues for inference workloads outside of North America, and limiting the ability to meet strict digital sovereignty and compliance requirements in regions like Europe or Asia/Pacific. This geographical limitation necessitates an AI multivendor strategy for multinational organizations if they select Lambda.
- **Heavy training focus and lack of hybrid strategy:** As a vendor primarily focused on high-end training for researchers and adopting a bare-bones service approach, Lambda lacks a formal on-premises or robust edge strategy for AI inference. This is suboptimal for organizations moving toward a hybrid AI infrastructure model, where lighter inferencing must occur close to user activity or data location, forcing customers into fragmented deployment strategies and increasing their MLOps complexity.

Microsoft

Microsoft, headquartered in Redmond, Washington, U.S., offers a comprehensive suite of cloud AI infrastructure services as a part of Azure, intended to support training, tuning, and inference requirements. Key components include optimized compute services, specialized storage, and high-throughput, low-latency networking.

Microsoft offers a wide range of compute resources, both first- and third-party. It leverages InfiniBand networking for distributed training. Microsoft offers one of the broadest storage resource portfolios of anyone in this research. Microsoft also offers numerous platforms that enable user access and use. This includes Microsoft Foundry for building, customizing and deploying AI models. It also offers Azure AI Landing Zones, which provide preconfigured architectures that help the rapid setup of AI environments. Microsoft IQ is a shared intelligence layer connecting data, workflows and signals across organizations. This helps applications and agents to operate with better accuracy and relevance.

Microsoft is a Leader in this Magic Quadrant, with the ability to handle a wide range of AI use cases. Microsoft offers a flexible pricing model.

Strengths

- **AI solutions integration:** Microsoft excels in linking its cloud AI infrastructure with its other comprehensive cloud services. This deep integration simplifies development and deployment for organizations already invested in the broader Microsoft Azure and application ecosystem, accelerating time to market for integrated, AI-driven business solutions.
- **Generative AI ecosystem:** Microsoft's strategic partnerships enable it to access cutting-edge technology, which provides a unique value proposition, attracting AI innovation and offering customers the fastest route to deploying advanced GenAI capabilities. Enterprises can leverage leading AI technology immediately, minimizing the lag between AI innovation and production deployment.
- **Comprehensive managed services:** Services like Microsoft Foundry and Azure AI Landing Zones simplify the MLOps life cycle, providing preconfigured architectures for rapid environment setup. This reduces operational overhead and accelerates time to value by minimizing the specialized MLOps skills required, enabling development teams to focus on model innovation rather than infrastructure management.

Cautions

- **Widespread capacity constraints:** Like other hyperscale cloud providers, Microsoft has faced sustained demand pressure on AI infrastructure capacity, causing it to address this through continued infrastructure expansion and a multisource capacity strategy that combines first-party assets and ecosystem partners. Enterprises with demanding workloads (i.e., large-scale model training or high-throughput inference) should obtain explicit regional capacity commitments before committing to Azure.
- **Maturity of AI processors:** While Microsoft has introduced Maia processors to optimize inference efficiency, it currently plays a more limited role in model training compared to other top hyperscalers, which have longer-operated AI processors and adjacent AI supercomputing infrastructure that spans both training and inference. Enterprises must understand that Microsoft continues to rely heavily on third-party accelerators for the most compute-intensive training workloads, which may constrain long-term architectural differentiation and cost control at an extreme scale.
- **Complex pricing structure:** Similar to other hyperscalers, the expansive and layered Azure service catalog can create challenges around cost forecasting and governance for large-scale AI initiatives, potentially impacting the predictability of scaling

production workloads. Enterprises should carefully evaluate TCO for complex AI pipelines, and establish appropriate governance and optimization practices as deployments scale.

Nebius

Nebius, headquartered in Amsterdam, Netherlands, provides cloud AI infrastructure tailored for training and inference workloads. Nebius is a Reference Platform NVIDIA Cloud Partner that builds its own custom energy-efficient data center designs to provide highly performant infrastructure.

Nebius offers a suite of high-end NVIDIA GPUs enabled by high-performance, low-latency InfiniBand networking. They also offer high-performance object and file storage. Nebius also offers a set of platform services, including Managed Service for Kubernetes and Managed Service for Soperator, its fully managed Slurm-on-Kubernetes solution for high-performance AI training. It also offers an inference service (Nebius Token Factory) that allows users to deploy and run a wide set of models.

Nebius is a Visionary in this Magic Quadrant, aimed at AI-native startups, researchers, ISVs and forward-leading enterprises that need high-performance AI infrastructure. Nebius offers a comprehensive suite of pricing models, including consumption-based pricing.

Strengths

- **Early access to cutting-edge hardware:** As a Reference Platform NVIDIA Cloud Partner, Nebius provides access to the latest GPUs, often ahead of other vendors in this evaluation. This supply chain advantage is crucial in a market defined by GPU scarcity, which allows enterprises and researchers to accelerate innovation with the newest performance capabilities and avoid project delays caused by hardware constraints.
- **High-performance infrastructure:** By designing its own servers, racks and data centers, Nebius achieves superior performance, compared to many other providers in the evaluation. This specialized design, leveraging high-performance, low-latency InfiniBand networking, maximizes the efficiency of large-scale, distributed AI training and inference workloads, leading to accelerated time to market for production models.

- **Token-based SLA:** Nebius Token Factory — Nebius' serverless platform for inference at scale — offers a token-based SLA, specifying tokens per second and latency. Customers seeking a serverless platform for inference at scale can expect predictable results, backed by an SLA aligned with their business needs.

Cautions

- **High customer concentration:** Nebius has fewer than 250 enterprise customers, with some large names, such as Microsoft and Meta. This customer concentration creates financial volatility and risk of service disruption or deprioritization if a major client shifts its strategy, posing a risk to the long-term stability and resource planning for other customers.
- **Limited hybrid strategy:** Nebius does not offer an on-premises solution, limiting deployment flexibility at a time when inference requirements necessitate a definitive hybrid focus. This is suboptimal for enterprises that require a seamless operational model for AI workloads spanning their data center and the cloud, increasing MLOps complexity.
- **Execution risk at scale:** The company is undergoing rapid expansion, which introduces risks in maintaining operational efficiency and reliability. This rapid, high-growth environment increases the potential for service disruptions, inconsistent performance and a lack of platform maturity, which could jeopardize mission-critical, production-level AI workloads.

Nscale

Nscale, headquartered in London, U.K., is a specialized cloud provider focused on delivering high-performance, sustainable, and sovereign AI cloud infrastructure. They support a range of AI workload requirements, including training, inference, and fine-tuning.

Nscale's core infrastructure provides bare-metal and virtualized NVIDIA GPU resources, based on off-the-shelf hardware primarily from Dell and Lenovo. It ensures low-latency networking through the use of InfiniBand and provides high-throughput storage via a strategic relationship with VAST Data. Management services include the Nscale Kubernetes Service, Managed Slurm scheduling, and a serverless inference service.

Nscale is a Niche Player in this Magic Quadrant, specifically appealing to enterprises that require customized, highly performant AI infrastructure that strongly emphasizes data sovereignty and sustainability. Nscale leverages both long-term and take-or-pay contracts for reserved infrastructure, and a consumption-based pricing model for on-demand AI services.

Strengths

- **Data center deployment speed:** Nscale utilizes prefabricated modular solutions, enabling Nscale to own and deploy integrated data center and GPU infrastructure with superior speed and cost-efficiency. This full-ownership model delivers more competitive unit economics compared to rivals that rely on third-party colocation leasing, which provides enterprises with a better cost-performance ratio.
- **Specialized, high-performance infrastructure:** The platform is optimized for demanding AI workloads, offering bare-metal/virtualized NVIDIA GPUs, low-latency InfiniBand networking, and high-throughput VAST Data storage. This high-performance stack is designed to be customized for customer requirements, accelerating time to completion for complex, resource-intensive training, inference and fine-tuning of projects.
- **Commitment to sustainability:** Nscale's data centers are strategically located to utilize renewable energy sources, with several sites using only renewable energy. The centers also use closed-loop direct liquid cooling and heat recovery. Enterprises seeking to meet strict internal ESG commitments can leverage this infrastructure to lower the carbon footprint of their compute-intensive AI training and inference workloads.

Cautions

- **Regional concentration:** Nscale's current infrastructure is primarily located in Europe and the U.S., with an expanding Asia/Pacific footprint. This limited geographic reach means global enterprises that require AI deployments to serve customers or process data in non-European or non-U.S. regions will face latency issues and may struggle to meet local compliance or data sovereignty requirements outside of Nscale's primary operational areas, necessitating a multivendor strategy.
- **New market entrant and lower platform maturity:** Nscale lacks the years-long maturity and comprehensive broad-spectrum service ecosystem of other providers in this

research. Enterprises may find that core infrastructure offerings are robust, but the absence of extensive managed PaaS, serverless options and integrated MLOps tools common to hyperscalers increases the operational burden, requiring high levels of specialized MLOps expertise to manage and secure the infrastructure.

- **Lack of hardware innovation:** Currently, Nscale relies entirely on off-the-shelf hardware from OEMs such as Dell and Lenovo, rather than engineering proprietary systems in-house; the vendor has not published research papers or filed patents within the last 24 months. Reliance on third-party hardware limits Nscale's ability to deeply optimize the full stack for massive scale, reducing its control over platform performance and hindering its capacity to drive down long-term cloud costs through specialized hardware engineering.

Oracle

Oracle Cloud Infrastructure (OCI), headquartered in Austin, Texas, U.S., delivers a high-performance AI infrastructure optimized for handling large-scale model training and inference, distinguished by its superior price/performance. OCI also provides a comprehensive range of distributed cloud offerings, catering to varied AI deployment needs.

OCI's foundational AI infrastructure includes bare-metal GPU instances powered by the latest NVIDIA and AMD GPUs, which can be deployed in OCI Superclusters for large-scale AI training and inference. These servers are integrated into OCI Superclusters, which are engineered for massive-scale AI training. The Superclusters utilize RDMA over Converged Ethernet (RoCE) or InfiniBand to ensure extremely low-latency communication crucial for large, distributed workloads. Oracle offers a wide range of high-performance storage options, including block, object and file storage, complemented by highly performant data services that specifically enable RAG use cases. Furthermore, Oracle's Acceleron SmartNIC technology improves AI workload efficiency by reducing infrastructure overhead in the host and optimizing the path between compute, storage and network services. At the services layer, OCI provides a managed service that offers access to pretrained foundation models, allows fine-tuning using proprietary data, and then provides ways to build, run and govern agentic applications.

Oracle is a Leader in this Magic Quadrant, asserting strong price/performance capabilities that effectively compete with other evaluated Leaders. Services are offered through a consumption-based pricing model.

Strengths

- **Scalability and performance:** The OCI Supercluster architecture features ultra-low-latency networking and massive-scale interconnects, which enables highly efficient, large-scale, multinode GPU training and distributed model serving, is optimized for training foundation models and can handle the most demanding AI workloads. This significantly accelerates the development cycle for LLMs and complex AI algorithms, reducing time to market for critical business innovations and providing the necessary compute capacity for rapid iteration and deployment at massive-scale.

- **Flexible deployment options:** Oracle's comprehensive distributed cloud offerings, such as Dedicated Region, Cloud@Customer, and the Oracle Database@ family of multicloud services, allow enterprises to deploy AI infrastructure and services precisely at the location they desire, extending Oracle Cloud services outside of public cloud regions. Oracle addresses strict data sovereignty, regulatory compliance, and low-latency requirements by allowing data and compute to be localized, essential for global enterprises operating in regulated industries.

- **Enterprise integration:** OCI provides deep, native integration with core Oracle platforms, including Oracle Fusion Applications, AI Data Platform and the Autonomous Database, as well as easy connectivity to existing enterprise data stores. This streamlines the adoption of AI across the business by leveraging preexisting data and application ecosystems, which results in quicker implementation, reduced integration complexity, and immediate business value realized through automated and intelligent processes.

Cautions

- **Smaller ecosystem and community:** Compared to market leaders, OCI has a smaller ecosystem of prebuilt, third-party AI integrations, and a less extensive community resource base, which translates to fewer readily available, plug-and-play solutions and a more limited shared knowledge pool for troubleshooting. This caution may necessitate greater custom development for crucial integrations and increase reliance on internal teams or Oracle Consulting services for specialized support, potentially extending time to market for new AI solutions and increasing implementation complexity.

- **Less-mature AI support tooling:** While OCI's underlying AI infrastructure (such as the Supercluster) is robust, the supporting tooling for AI management, including

Kubernetes and observability, is considered less mature and feature-complete than competing offerings. Organizations seeking sophisticated capabilities may need to integrate and manage third-party tools or build custom solutions to achieve enterprise-grade orchestration, scheduling and observability, which introduces additional integration and maintenance complexity.

- **Support and documentation:** Gartner clients report that OCI's documentation can be challenging to navigate or fragmented when seeking guidance on complex AI infrastructure configurations; also, response times for general customer support inquiries can vary. This situation can create friction for development and operations teams, potentially leading to slower problem resolution, increased debugging time, and delays in project timelines when clear, self-service documentation or prompt support is critical.

OVHcloud

OVHcloud, headquartered in Roubaix, France, offers cloud AI infrastructure services designed to support the entire AI life cycle, from data experimentation to training to production-read inference, with a strong emphasis on data sovereignty, cost-efficiency and open-source technologies. OVH leverages a unit consumption pricing model.

OVHcloud's high-performance AI infrastructure includes high-end NVIDIA GPUs. Management services include a managed Kubernetes service, along with AI Training for training machine learning and deep learning models; these management services provide automatic scaling and optimization of underlying GPU resources. For demanding inference, OVHcloud also offers AI Endpoints, powered by SambaNova. OVH also leverages SambaNova for other AI services. Additionally, AI Deploy is designed to streamline model deployment, creating API endpoints for inference with automatic load balancing. AI Endpoints provide easy access to models, and AI Notebooks allow data scientists to work with models.

OVHcloud is a Challenger in this Magic Quadrant, primarily catering to European organizations with sovereignty ambitions. OVHcloud offers a consumption-based pricing model.

Strengths

- **No egress fee:** Unlike other providers in this evaluation, OVHcloud does not charge for cloud data transfer within its cloud environment nor for data egress. This no-egress-

cost policy makes OVHcloud more attractive for organizations that run their primary enterprise workloads with other providers.

- **In-house server manufacturing:** OVHcloud leverages more than 20 years of in-house server manufacturing and advanced water-cooling experience, reducing dependencies on third parties and enhancing carbon footprint and power usage effectiveness. Organizations seeking to meet strict internal ESG commitments can leverage this infrastructure to lower the carbon footprint of their compute-intensive AI training and inference workloads.

- **Business model:** OVHcloud has a business model that is externally profitable, with a primary focus on web hosting and bare-metal services; it had more than €1 billion of revenue in 2025, and a solid EBITDA margin of 40% (with IaaS/PaaS growing at 33% year over year). Customers looking for a nonhyperscale cloud provider for their AI workloads, and that also consider profitability a key factor, may wish to include OVHcloud in their shortlist.

Cautions

- **Limited ecosystem breadth:** Compared to major global hyperscalers, OVHcloud lacks the raw compute scale necessary for massive foundational model training and does not offer the same breadth of integrated PaaS, serverless options or extensive MLOps ecosystem services, potentially increasing the integration burden. Enterprises aiming to develop or fine-tune massive foundational models may face capacity constraints. They are also required to invest more heavily in custom integration work to connect OVHcloud's core infrastructure with necessary third-party MLOps and PaaS tools than they would with a global hyperscaler.

- **Limited networking services:** Networking options lack the breadth and depth of other vendors in the analysis. This deficiency in advanced networking capabilities can create performance bottlenecks and increase the complexity of distributing large-scale, high-throughput AI workloads across clusters, potentially slowing down training times and impacting the low-latency requirements of complex inference and agentic AI applications.

- **Limited operational management capabilities:** OVHcloud lacks the operational management capabilities that some of the vendors in this research offer, particularly cost allocation. Without native cloud financial management tooling, an enterprise is

left vulnerable to runaway costs and manual overhead, due to a lack of automated visibility and cost governance.

Scaleway

Scaleway, headquartered in Paris, France, is a provider of cloud AI infrastructure for training and inference, and is part of the iliad Group.

For compute, they offer a wide range of NVIDIA GPU instances and dedicated clusters for large-scale training, leveraging high-speed NVIDIA networking and NVLink. Data services include S3-compatible object storage, along with managed PostgreSQL and MySQL databases. They also leverage a partnership with VAST Data for high performance parallel storage. Scaleway's managed services simplify deployment, encompassing a managed inference platform, a model-as-a-service offering, and the managed Kubernetes service (Kapsule) for containerized AI applications.

Scaleway is a Niche Player in this Magic Quadrant, mostly suited for European small and midsize businesses. Scaleway uses a consumption-based pricing model.

Scaleway declined requests for supplemental information. Gartner's analysis is therefore based on other credible sources.

Strengths

- **Broad cloud portfolio:** Unlike other Niche Players in this evaluation, Scaleway offers a broader variety and depth of cloud services: Offerings include virtual machines and bare metal, PaaS (such as managed Kubernetes and DBaaS), Serverless, and even quantum computing. This comprehensive service range enables Scaleway to host a variety of workloads beyond just AI, which minimizes data transfer costs and latency between AI and non-AI components.
- **Sustainability:** Scaleway's data centers are highly energy-efficient and powered by low-carbon energy. Enterprises seeking to meet strict ESG commitments can leverage this infrastructure to lower the carbon footprint of their compute-intensive AI training and inference workloads.
- **Managed services:** Scaleway has a set of managed services that simplify the AI life cycle from training to inference, reducing the operation burden on users. By offering managed services, including Kapsule and an inference platform, Scaleway enables enterprises to accelerate the deployment of containerized AI applications and reduce

the need for specialized MLOps skills, allowing development teams to focus on model innovation.

Cautions

- **Regional footprint:** Scaleway offers only regions within Europe (France, Netherlands, Poland and Italy). This regional limitation means global enterprises or those with significant operations and data processing needs outside of Europe will be unable to use Scaleway for AI workloads that require localized low-latency services in other major regions (e.g., North America, Asia/Pacific), necessitating a multivendor strategy.
- **Smaller ecosystem:** Scaleway has fewer native AI services, integrations and PaaS offerings, compared to many hyperscalers in the evaluation. The limited ecosystem means enterprises must invest more effort and resources in custom integration and middleware development to connect Scaleway's infrastructure with existing enterprise toolchains, CI/CD pipelines and MLOps platforms, increasing development costs and operational complexity.
- **Billing complexity:** Although transparent in base pricing, total costs can become complex due to separate billing for storage, bandwidth and instances. Complex billing structures create governance and forecasting challenges for finance and procurement teams, making it difficult to accurately predict the TCO for large-scale AI projects, and potentially leading to budget overruns if usage across different service components is not meticulously tracked.

Tencent Cloud

Tencent Cloud, headquartered in Shenzhen, China, offers a comprehensive suite of cloud AI infrastructure centered around its machine learning platform (TI-ONE) that supports data labeling, model training, evaluation and deployment. Tencent provides a wide range of pricing options for its cloud AI infrastructure, including consumption-, seat- and value-based, as well as revenue sharing.

Compute resources offered include computing clusters that utilize compute nodes and high-speed RDMA technology that provides low-latency, high bandwidth communication for training. Tencent Cloud also provides a set of proprietary AI chips (e.g., Zixiao for inference, Xuanling for network optimization). A high-performance file system

(GooseFS/CFS Turbo) designed for challenging data scenarios is offered, along with a multimodal data lake (TCLake) for data ingestion and processing.

Tencent Cloud is a Challenger in this Magic Quadrant, with a focus on optimized, vertically aligned AI infrastructure.

Strengths

- **Proven scale via Tencent's first-party AI workloads:** Tencent Cloud's AI infrastructure is validated by large-scale first-party workloads from Tencent Group, including social platforms, real-time advertising, gaming, and content services. These internally operated, latency-sensitive and revenue-critical AI workloads provide continuous validation and optimization feedback, resulting in higher system maturity.
- **Integrated ML platform for the full life cycle:** The machine learning platform (TI-ONE) provides a comprehensive, vertically integrated suite. This vertically integrated approach simplifies the MLOps life cycle, reducing operational overhead and accelerating time to value by enabling development teams to focus on model innovation, rather than integrating disparate tools.
- **Integrated high-performance AI platform:** Tencent Cloud's platform combines proprietary high-performance AI chips, advanced computing clusters with high-speed RDMA technology, and Cloud Object Storage (COS), which integrates high performance file caching and acceleration through GooseFS/CFS Turbo. It also includes a high-performance file system and multimodal data lake (TCLake), ensuring low-latency, high-bandwidth communication and seamless management of petabytes of complex, multimodal data. As a result, enterprises can accelerate large-scale AI model training and inference, optimize project completion times, and efficiently support mission-critical AI workloads with enhanced throughput and scalability.

Cautions

- **Limited global footprint:** Tencent Cloud has a strong regional execution in Asia, but a smaller global footprint compared to many others in the evaluations. This regional concentration restricts the ability of multinational enterprises to deploy low-latency inference services in critical Western markets and may pose challenges in meeting global data sovereignty and compliance requirements, necessitating a multivendor strategy for global operations.

- **Lack of visibility on Tencent proprietary AI chips:** Tencent Cloud has not participated in publicly available hardware performance benchmarks. While Tencent's AI infrastructure supports heterogeneous hardware configurations, customers leveraging its proprietary AI chips should conduct comprehensive performance and capability evaluations prior to production deployment.
- **Lack of access to latest GPUs:** Geopolitical issues have limited Tencent's access to the latest GPUs from NVIDIA. This supply chain constraint can disrupt large-scale training projects and foundational model development, forcing enterprises to rely on older or less performant hardware, which slows down innovation and increases the time to market for cutting-edge AI capabilities.

Vultr

Vultr, headquartered in West Palm Beach, Florida, U.S., offers a comprehensive suite of AI cloud infrastructure as a part of its overall cloud services. Its focus is on cost-optimized infrastructure served across 33 worldwide locations. It offers consumption-based pricing, along with revenue sharing for some of its partners.

Vultr offers compute resources from both NVIDIA and AMD (a strategic relationship), along with single-tenant bare-metal services. It also offers a wide range of storage services and high-performance networking. Preconfigured operating system images are available. Vultr Kubernetes Engine (VKE) provides a managed orchestration service for deploying and scaling AI workloads across GPUs. A serverless inference service is also offered.

Vultr is a Challenger in this Magic Quadrant and will appeal to enterprises that want cost-optimized cloud AI infrastructure along with other cloud-native services, which are offered in many worldwide locations. Vultr offers a flexible pricing model.

Strengths

- **Cost-efficient and transparent pricing:** Vultr offers competitive, consumption-based pricing that is often lower than hyperscalers, along with a transparent cost model and zero data egress fees. This allows enterprises, particularly those with cost-optimization mandates, to achieve a lower TCO for data-intensive AI workloads; it also facilitates multicloud strategies without prohibitive network charges.

- **Broad GPU and bare-metal diversity:** Vultr provides a more diverse compute portfolio (with NVIDIA and AMD) than many other vendors in the evaluation. This hardware diversity allows enterprises to optimize the price/performance ratio for specific AI workloads and reduces the risk of vendor lock-in associated with a single GPU ecosystem.
- **Extensive global infrastructure:** With data center regions served across many worldwide locations, Vultr offers a geographic footprint that reduces latency for inference workloads deployed closer to global end users, and helps enterprises address regional data residency requirements.

Cautions

- **Limited AI ecosystem and managed services:** Vultr's ecosystem of integrated and management AI services is less-extensive than those offered by hyperscalers. This requires enterprises to invest more effort in custom integration and MLOps tools to build end-to-end AI pipelines, which increases operational overhead and development complexity.
- **Support response and operational maturity:** As a specialized, growing cloud provider, Vultr may face execution risk, and some users have indicated that customer support times for urgent, complex issues can occasionally be slow. This lack of established operational scale compared to hyperscalers can pose a risk to mission-critical, production-level AI workloads where fast resolution times are essential for maintaining SLAs.
- **Nascent hybrid and on-premises strategy:** Vultr does not offer a standard kit for enterprises looking to deploy an on-premises private cloud offering, despite offering hybrid capabilities. This limits deployment flexibility for organizations moving toward a hybrid AI infrastructure model that requires a seamless, unified operational model for both public cloud and data center workloads.

Inclusion and Exclusion Criteria

In addition to Gartner client relevance, as determined by analyst expertise and opinion, providers need to meet the following criteria to qualify for inclusion:

- Mandatory Features contained in the Market Definition

- Operational on at least two continents
- At least 50 paying customers
- At least $50 million in revenue (USD constant currency)

Evaluation Criteria

Ability to Execute

We assessed vendors' Ability to Execute in this market by using the following criteria:

Product or Service: This criterion looks at the core cloud AI infrastructure services that the vendor offers to the market in terms of breadth, depth and quality of features. Consideration is given to a vendor's ability to deliver the comprehensive set of cloud AI infrastructure that enable training, inference and agentic AI requirements. Weight is also given to a vendor's particular capabilities in adjacent technical areas.

Overall Viability: This criterion includes an assessment of the organization's overall financial health, as well as the financial and practical success of its cloud AI infrastructure business unit. Considerations include a track record of growth, commitment to this market, and stability.

Sales Execution/Pricing: This criterion assesses the vendor's capabilities in all presales activities and the structure that supports them. This includes deal management, pricing and negotiation, presales support, and the overall effectiveness of the sales channel. Consideration is given to the depth and quality of the vendor's sales force, as well as its pricing and discounting models. Weight is also given to how well a vendor adapts its selling to specific geographies and industries.

Market Responsiveness/Record: This criterion looks at a vendor's ability to successfully respond and change direction based on the evolving needs of the market. Considerations include response to competitors, ability to perceive and adapt to changing customer needs, and pace of introduction of new services, features and programs.

Marketing Execution: This criterion looks at the quality and effectiveness of programs that deliver the vendor's message in order to influence the market, promote the brand,

increase awareness of products and establish a positive identification in the minds of customers. Considerations include a vendor's ability to demonstrate thought leadership, as well as the ability to convey truthful yet compelling market messages to each target buyer, region and industry.

Customer Experience: This criterion covers the vendor programs that enable customers to achieve anticipated results with their products. This includes presales interactions, customer enablement and implementation assistance, and ongoing technical and account support. Consideration is also given to the quality of a vendor's partner programs and partner involvement processes, as well as delivery of positive customer experiences in all target geographies and industries.

Operations: This criterion looks at the ability of the vendor to meet its operational responsibilities. Factors include the quality of its organizational structure, its technical and commercial operational processes, its platform resilience, and its ability to meet SLAs. This assessment also includes an evaluation of vendors' outages, security vulnerabilities and capacity shortage events. Consideration is also given to the quality and consistency of customer-facing interfaces and interactions. Lastly, weight is given to the degree to which a vendor offers options for customers to easily and reliably automate their own operational activities within the vendor's environment.

Ability to Execute Evaluation Criteria

Evaluation Criteria	*Weighting*
Product or Service	High
Overall Viability	High
Sales Execution/Pricing	Medium
Market Responsiveness/Record	Medium
Marketing Execution	Low

Evaluation Criteria	*Weighting*
Customer Experience	High
Operations	Medium

Source Gartner (July 2026)

Completeness of Vision

We assessed vendors' Completeness of Vision in this market by using the following criteria:

Market Understanding: This criterion assesses a vendor's ability to understand custome needs and translate them into products and services. Consideration is given to the abili of a vendor to understand enterprise requirements in each of five areas: efficiency and scalability, speed and agility, location flexibility, technical innovation, and digital transformation.

Marketing Strategy: This criterion looks for clear, differentiated messaging consistently communicated internally and externalized through social media, advertising, customer programs and positioning statements. Consideration is given to how well a vendor markets to different types of customers. Weight is also given to the maturity of a vendor's approach to marketing in regions outside its home country.

Sales Strategy: This criterion considers whether the vendor has a sound strategy for selling that uses the appropriate networks. Consideration is given to a vendor's strategies for selling to business leaders and IT leaders, for selling internationally, for selling into specific vertical industries, and for selling with and through partners.

Offering (Product) Strategy: This criterion evaluates whether a vendor's approach to product and service development and delivery emphasizes market differentiation, functionality, methodology, and features that cover current and future market requirements. Consideration is given to how well-articulated a vendor's product strategy is.

Business Model: This criterion looks at the design, logic and execution of the vendor's business proposition to achieve continued success. Consideration is given to how well a

vendor articulates its value proposition as a provider of cloud AI infrastructure services, as well as its value as a strategic IT partner.

Vertical/Industry Strategy: This criterion looks at a vendor's strategy to direct resources, skills, products and product integrations to meet the specific needs of individual market segments, including verticals. Consideration is given to the vendor's solution strategy and roadmap, as well as its partner ecosystem. Some weight is given to a vendor's breadth of coverage across major industry verticals, such as manufacturing, healthcare, telecom, banking and financial services, pharma and life sciences, retail, and insurance.

Innovation: This criterion looks at how a vendor applies its resources, expertise and partnerships in a coordinated way to lead and differentiate in the market. Consideration is given to a vendor's track record of innovation in the cloud AI infrastructure area. Weight is also given to business innovations, such as new approaches to market making, product licensing and customer digital transformation.

Geographic Strategy: This criterion looks at a vendor's strategy to direct resources, skills and offerings to meet the specific needs of geographies outside its home or native region. Consideration is given to a vendor's strategy for establishing global sales and delivery capabilities, and whether it makes available a complete product offering in regional markets and distributed locations. Weight is also given to how well a vendor responds to the shifting geopolitical landscape and regulatory requirements of the markets it sells into.

Completeness of Vision Evaluation Criteria

Evaluation Criteria	*Weighting*
Market Understanding	Medium
Marketing Strategy	Low
Sales Strategy	Low
Offering (Product) Strategy	High

Evaluation Criteria	*Weighting*
Business Model	Medium
Vertical/Industry Strategy	Low
Innovation	High

Source Gartner (July 2026)

Quadrant Descriptions

Leaders

Leaders distinguish themselves by offering a set of services suitable for strategic adoption and having an ambitious and well-aligned roadmap. They can serve a broad range of use cases, although they do not excel in all areas, may not necessarily be the best providers for a specific need and may not serve some use cases at all. Leaders in this market have appreciable market share and many referenceable customers.

Challengers

Challengers are well-positioned to serve some current market needs. They deliver a good service that is targeted at a particular set of use cases, and they have a track record of successful delivery. However, they are not adapting to market challenges quickly enough or do not have a broad enough scope of ambition.

Visionaries

Visionaries have an ambitious vision of the future and are making significant investments in the development of unique technologies. Their services are still emerging, and they have many capabilities in development that are not yet generally available. Although they may have many customers, they might not yet serve a broad range of use cases well, or may have a limited geographic scope.

Niche Players

Niche Players may be excellent providers for particular use cases or in regions in which they operate, but they should ultimately be viewed as specialist providers. They often do not serve a broad range of use cases well or have a broadly ambitious roadmap. Some may have solid leadership positions in markets adjacent to this market, but are limited in their cloud AI infrastructure capabilities.

Context

AI requirements, particularly GenAI, necessitate highly performant infrastructure that traditional computing, storage and networking cannot fully support. This emergent requirement also includes upper-layer services designed to optimize the underlying infrastructure, a capability often overlooked in traditional setups. Practical applications of AI are broadly categorized by their purpose, such as:

- Automating tasks
- Enhancing customer service with chatbots
- Analyzing data for insights
- Generating content, like text or images
- Personalizing recommendations
- Supporting medical diagnosis through image analysis
- Detecting fraud in finance
- Optimizing supply chains
- Enabling new forms of creativity and research

These applications span numerous industries. To fully understand these uses, you must understand the infrastructure activities associated with training and inferencing, along with the impact of agentic AI.

AI training is the process of feeding an AI model with large datasets to learn patterns and make predictions. Training is the most infrastructure-intensive activity, demanding highly performant compute, storage and networking, which must be orchestrated and optimized efficiently for the activity.

AI inference is the operational phase where a model applies learned knowledge to new data to generate outputs or decisions. While less resource-intensive than training, inference is often distributed and requires efficient data ingestion, highly elastic compute that scales for real-time demand, and complex stateful management to handle long-term memory for grounding RAG and agentic AI use cases.

Agentic AI refers to autonomous AI systems that plan and execute multistep tasks with minimal human intervention. Agentic AI requires a highly dynamic and elastic infrastructure, with resilience and low latency being critical factors, as this type of AI simulates real human interactions and activities.

The cloud AI infrastructure market is characterized by logistical challenges and long lead times for purchasing and deploying infrastructure privately, as the demand for high-performance processors continues to exceed the supply. This has led to a surge in activity for specialty cloud providers that can meet the growing demand, particularly as leading frontier AI labs consume a large share of the critical GPU components.

Most enterprises turn to cloud AI infrastructure providers because they lack the internal skills to build and maintain private GPU environments, and worry about the high cost of purchasing, configuring and maintaining the expensive infrastructure required to scale for peak training loads.

Additionally, AI will prompt organizations to reassess their multicloud strategies. The importance of data location will drive infrastructure expansion beyond the public cloud, resulting in hybrid AI infrastructure deployments. This means training and heavy inferencing will likely occur in the public cloud, while lighter inferencing and servicing will occur where the data is located (on-premises, regional and edge environments).

Finally, organizations will require support for AI operations that meet data sovereignty requirements, necessitating controls to ensure data remains compliant within specific operational jurisdictions. Furthermore, addressing security and compliance will be paramount for all AI deployments that rely on organizational data.

Market Overview

Market Segmentation and Landscape

The global market for cloud AI infrastructure is segmented into two primary vendor categories:

- **Hyperscale vendors (e.g., AWS, Google Cloud, Microsoft Azure):** These providers offer comprehensive capabilities beyond GPU-as-a-service (GPUaaS), including model catalogs/marketplaces with first- and third-party generative AI (GenAI) models, AI engineering platforms, and AI-infused applications. A key strategic endeavor for hyperscalers is investing in their own proprietary AI-optimized hardware.
- **Specialized cloud providers:** This category includes startups focused on scalable, high-performance and cost-effective infrastructure for AI/ML workloads. Other providers started as bare-metal IaaS and expanded their offerings to GPU instances, clusters and AI frameworks. Some focus on sovereignty and/or niche areas, such as AI at the edge or partnering with specific jurisdictions.

Current demand is for high-performance model training, low-cost inference, and large-scale data processing for AI/ML use cases. Enterprise adoption is driven by several critical challenges:

- **Lack of internal expertise:** Most enterprises lack the internal skills to build and maintain private GPU-based AI environments for training.
- **Cost and scaling risk:** Enterprises worry about the cost of purchasing, configuring and maintaining expensive AI infrastructure that must be scaled to meet peak training load capacities.
- **Infrastructure barriers:** The availability of GPUs in some geographies and the electrical power required to run them often act as barriers to building a private AI infrastructure.
- **System integrator requirements:** Solution providers, including system integrators, require AI-optimized infrastructure to support complex, distributed AI solutions for their clients, enabling enterprises to purchase preconfigured capacity and pay only for what they use.

Supply Chain Constraints and Market Dynamics

The market is currently defined by a supply-demand imbalance:

- **GPU scarcity:** Logistical challenges and long lead times for hardware acquisition persist as demand for high-performance processors significantly exceeds supply. Gartner believes this demand will continue to exceed supply for the foreseeable future.
- **Market concentration:** Leading frontier AI labs are entering into multibillion-dollar agreements with GPU manufacturers (e.g., NVIDIA and AMD), which crowds out smaller entities and private users, and impacts hyperscalers. This competition is resulting in a surge of activity for specialty cloud providers that can meet growing demand.
- **Future inference strain:** While the focus has been on the highest-performing processors for training, logistical challenges will also arise for the lower-performing processors needed for inference use cases, as those deployments grow and exceed training capacity needs.
- **Energy demands:** AI infrastructure will create energy demands that current sources cannot meet, necessitating new energy sources and efficient data center designs. This will also impact energy strategy in metropolitan areas, where demand for AI-enabling data centers might conflict with residential needs.

Evolving Deployment Models and Strategic Imperatives

The market is shifting from a public-cloud-only model to more distributed architectures:

- **Hybrid AI infrastructure:** The importance of data location will drive AI infrastructure expansion beyond the public cloud to on-premises, regional and edge environments, culminating in hybrid AI deployments. Training and heavy inferencing will likely occur in the public cloud, while lighter inferencing and servicing will occur where the data is located. This hybrid deployment is often necessitated by the differing requirements (optimization and latency) of inference and agentic AI, compared to training.
- **Data sovereignty and compliance:** Organizations will require support for AI operations that meet data sovereignty requirements, demanding AI infrastructure deployment in specific operational jurisdictions with controls to ensure data remains compliant. Addressing security and compliance is paramount for all AI deployments relying on organizational data.

- **Multicloud reassessment:** AI will prompt organizations to rethink their multicloud strategies, as it will become common to utilize technology in one cloud while deploying services in another where the data may be located.

⊕ Evaluation Criteria Definitions