ABI RESEARCH COMPETITIVE RANKING
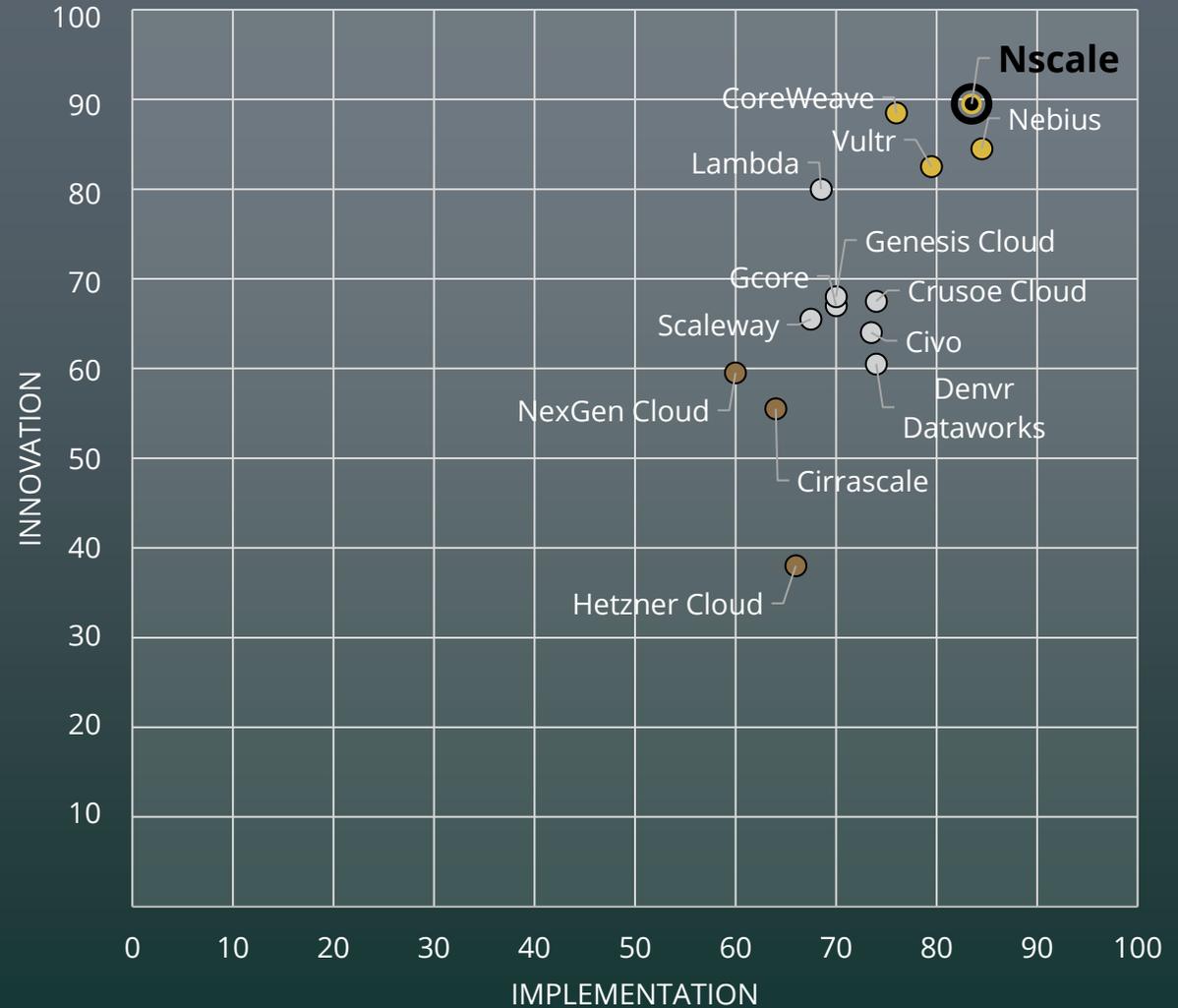
NEOCLOUD PROVIDERS

NSCALE

OVERALL LEADER
ABiresearch.

TOP INNOVATOR
ABiresearch.

TOP IMPLEMENTER
ABiresearch.

OVERALL: 86.6 | INNOVATION: 89.5 | IMPLEMENTATION: 83.5 | RANK: 1

INNOVATION

INNOVATION
SCORE: 89.5

Nscale scores 9 for its hardware and supply chain strategy. The company announced an expanded agreement with Microsoft to deploy ~200,000 NVIDIA GB300 GPUs across Europe and the United States, described as one of the largest AI infrastructure contracts signed. Alongside its US$1.1 billion Series B funding, the vertically-integrated model and long-term NVIDIA/Microsoft contracts ensure predictable large-scale GPU supply and rapid global deployment. The sole weakness in this otherwise excellent strategy is the reliance on a single hardware vendor (NVIDIA H100, H200, GB200/GB300 NVL72, and A100).

In terms of maximimum distributed cluster scale, Nscale scores 10 as the company has several proof points of large deployments. Most importantly, Nscale provides *Stargate Norway* (together with Aker & OpenAI), an AI Gigafactory that targets 100,000 NVIDIA GPUs by the end of 2026 (renewable-powered, with ambitions to scale significantly beyond the first phase). Nscale also supports multi-thousand GPU workloads using Slurm-on-Kubernetes (SLONK) and is building similarly large, high-density clusters at its roughly 240 Megawatt (MW) Texas AI campus.

Nscale scores 8 for its AI optimization stack maturity, as it sells a productized AI platform surface, including training clusters, inference service, serverless inference, fine-tuning, and a Marketplace with a "model library" and "tools/resources" for development and deployment. Proprietary runtimes remain less visible, but integrated GEMM tuning and quantization within its orchestration layer strengthen performance.

For adjacent cloud, data & AI services, Nscale scores 8 as it provides not only GPU nodes, but also managed training and inference capabilities, including fine-tuning and serverless inference. The platform integrates Kubernetes, Slurm, modular MLOps tooling, observability, and development workflows end-to-end, while the portfolio of legacy services is very limited.

Nscale scores 9 for ISV partner & co-development innovation, as it has a large network of partners from different backgrounds: Microsoft (anchor offtaker), OpenAI (Stargate Norway), Dell (server OEM partner for NVIDIA GB300), Nokia (networking), and VAST Data (storage). Infrastructure partners across the Nordics and Southern Europe add further depth for high-density, renewable-powered sites.

Nscale scores 9 for interconnect bandwidth & topology, as it emphasizes "fabric optimized for collectives," "non-blocking design," and "high bandwidth/low latency." It combines InfiniBand, high-bandwidth Ethernet with Remote Direct Memory Access (RDMA), and NVLink/NVSwitch for NVL72 systems, supported by joint innovation with Nokia across switching and optical layers.

IMPLEMENTATION

NSCALE

IMPLEMENTATION
SCORE: 83.5

TOP IMPLEMENTER

ABIresearch.

Nscale scores 9 for GPU availability, as the company's commitment to deploy approximately 200,000 NVIDIA GB300s (across Europe and the United States) illustrates strong infrastructure availability—primarily driven by large-scale deployments in Texas (United States), Norway, and Portugal. Queue times remain undisclosed, though most on-demand jobs typically start within minutes, with reservation options for large training workloads.

Nscale scores 8 for price performance, as the company monetizes on GPU-hours and token pricing. This gives potential customers and Large Language Model (LLM) builders more predictability. The lowest published tier on that page includes examples like US$0.01 input/US$0.03 output per 1 million tokens (Qwen), while other models are priced higher (e.g., US$0.10/US$0.40 for some OpenAI entries).

For SLAs & reliability track record, Nscale scores 7, as it targets enterprise-grade reliability (hyperscale campuses, sovereign positioning, major counterparties). While no binding SLAs have been published, the presence of multi-year hyperscale delivery contracts also implies operational rigor. Modular data center redundancy, monitoring, and automated recovery underpin operational resilience.

In terms of security, compliance & governance depth, Nscale scores 9. The narrative is built around sovereign AI infrastructure and renewable-powered sites. Controls align with the General Data Protection Regulation (GDPR), International Organization for Standardization (ISO) 27001, and SOC 2 principles, embedded directly into its vertically integrated stack.

For the breadth of its ecosystem partnership network, Nscale scores 9, as it is unusually strong for its age and maturity within the industry (Microsoft, OpenAI, Aker, Dell, Nokia, and VAST Data + NVIDIA). Engineering and site-delivery partners in the Nordics and Southern Europe further expand its operational ecosystem for renewable-powered, high-density builds.

For cloud & platform integration depth, Nscale scores 9, as it has a coherent AI platform surface area (including libraries, training/fine-tuning/inference clusters and GPU nodes). The platform integrates Kubernetes, Slurm, and MLOps workflows into a unified control plane with development tools and lifecycle management baked into the same vertically-integrated stack.

CRITERIA & METHODOLOGY

VENDOR MATRIX

Methodology: After individual scores are established for innovation and implementation, an overall company score is established using the Root Mean Square (RMS) method:

$$Score = \sqrt{\frac{innovation^2 + implementation^2}{2}}$$

The resulting overall scores are then ranked and used for percentile comparisons.

The RMS method, in comparison with a straight summation or average of individual innovation and implementation values, rewards companies for standout performances.

For example, using this method, a company with an innovation score of nine and an implementation score of one would score considerably higher than a company with a score of five in both areas, despite the mean score being the same. ABI Research believes that this is appropriate as the goal of these matrices is to highlight those companies that stand out from the others.

RANKING CRITERIA

Leader: A company that receives a score of **75 or above** for its overall ranking.

Mainstream: A company that receives scores **between 60 and 75** for its overall ranking.

Follower: A company that receives a score of **60 or below** for its overall ranking.

Innovation Leader: A company that receives a score of **75 or above** for its innovation ranking.

Implementation Leader: A company that receives a score of **75 or above** for its implementation ranking.

INNOVATION CRITERIA

Hardware & Supply Chain Strategy: This criterion assesses the relationship between neocloud provider and accelerator manufacturers, diversified multi-vendor sourcing, and long-term supply agreements. It also considers the provider's data center approach (owned facilities versus colocation, regional coverage, and availability of sovereign cloud options) along with the transparency of its supply chain governance. A strong strategy ensures that enterprises have predictable access to compute, reduced geopolitical and vendor risk, and the sovereignty assurances required for regulated AI workloads.

Max Distributed Cluster Scale: This criterion evaluates the maximum number of GPUs that can be reliably used in a single distributed training job. Large-scale AI research, frontier model development, and advanced pretraining workflows require clusters spanning hundreds or thousands of GPUs. Providers with higher-scale limits enable enterprises to train larger models, reduce training time, and take on more ambitious AI initiatives. Scale is a foundational differentiator: without it, enterprises are effectively capped in how far their AI roadmaps can go.

AI Optimization Stack Maturity: This criterion assesses the sophistication of the provider's AI software ecosystem. A mature optimization stack can deliver 20% to 40% better throughput without changing hardware, improving both performance and cost-efficiency. Enterprises rely on this layer to stabilize long-running jobs, simplify scaling, and produce reliable results. Providers with strong AI optimization capabilities empower customers to deploy AI faster, cheaper, and with greater reliability.

Adjacent Cloud, Data & AI Services: This criterion measures innovation in complementary platform services that support end-to-end AI workflows. Enterprises increasingly depend on integrated platforms, rather than standalone compute. Providers that innovate across adjacent layers dramatically reduce integration overhead and accelerate time-to-value. Strong adjacent services enable organizations to build unified data-to-AI pipelines, streamline deployment, and operationalize models across the business.

Involvement in AI Partner & ISV Co-Development Innovation: This criterion captures how actively the provider collaborates with enterprise Software-as-a-Service (SaaS) vendors, Independent Software Vendors (ISVs), and industry partners to develop new AI capabilities, vertical solutions, and reference architectures. Co-development accelerates the creation of AI-powered business applications. Providers with high involvement in partner innovation help enterprises deploy AI where it matters most: in real workflows across finance, healthcare, manufacturing, and retail. This ecosystem-level innovation drives faster adoption and unlocks value beyond raw compute.

Interconnect Bandwidth & Topology: This criterion evaluates the quality of the provider's high-performance network fabric (based on InfiniBand or advanced Ethernet, for example) and the effective bandwidth and latency delivered per node. Strong interconnect design is critical for multi-GPU and multi-node training, where communication overhead can dominate performance. High bandwidth and low latency ensure efficient scaling, faster training times, and more predictable job stability. For enterprises, a robust interconnect topology directly reduces cost, accelerates large-scale AI development, and enables more complex model architectures to run reliably in production.

IMPLEMENTATION CRITERIA

GPU Availability & Queue Times: This criterion measures how consistently a provider can deliver access to high-demand GPUs, which directly affects whether enterprises can start, continue, or scale AI workloads without disruption. High availability and low queue times are essential for predictable development cycles, maintaining training momentum, and supporting time-sensitive product roadmaps. When GPUs are scarce or wait times are long, enterprise AI teams face delays, cost overruns, and missed opportunities; reliable access makes the difference between experimentation and production-grade AI.

Price–Performance Score: This criterion evaluates the true economic efficiency of running AI workloads, factoring in both the cost of GPU hours and the real throughput delivered (training speed, inference efficiency, and tuning performance). Enterprises increasingly assess value in terms of *output per dollar*, not raw infrastructure cost. A high price–performance score enables organizations to scale models more affordably, deploy more use cases, and justify AI investments internally. It has a direct impact on Total Cost of Ownership (TCO), Return on Investment (ROI), and competitiveness.

Service-Level Agreement (SLA) & Reliability Track Record: This criterion measures the stability and predictability of the GPU-as-a-Service (GPUaaS) environment, including uptime guarantees, historical reliability, job interruption rates, and multi-zone robustness. Enterprise AI workloads cannot tolerate instability or repeated restarts. Strong SLAs and operational resilience reduce production risk, protect timelines, and prevent budget overruns tied to failed or delayed jobs. This directly underpins trust in AI pipelines and the ability to support mission-critical AI applications.

Security, Compliance & Governance Depth: This criterion assesses the provider's ability to meet enterprise-grade security and regulatory requirements through certifications, data isolation, encryption, key management, identity integration, and governance controls. Enterprises need strong guarantees around data handling, workload isolation, and compliance frameworks. A mature security and governance posture enables organizations to deploy AI at scale without violating internal policies or external regulations, unlocking adoption in high-sensitivity use cases.

Enterprise SaaS & ISV Ecosystem Breadth: This criterion evaluates how well the GPUaaS platform integrates with the broader enterprise software ecosystem, including Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Human Resources Information System (HRIS), security tools, analytics platforms, data governance suites, and AI-specific ISVs. Real enterprise AI success depends on embedding models into business workflows, so deep SaaS/ISV integration dramatically accelerates deployment and reduces the need for custom engineering. Providers with strong ecosystems enable enterprises to operationalize AI faster, more consistently, and across more business functions.

Cloud & Platform Integration Depth: This criterion measures how seamlessly the GPUaaS offering ties into the provider's own cloud and platform services, including storage, networking, identity, Development Operations (DevOps)/Machine Learning Operations (MLOps) tooling, data platforms, and AI utilities. Integrated platforms reduce friction, enable end-to-end AI pipelines, and simplify infrastructure management. Mature integration ensures that enterprises can move from data ingestion → model training → deployment → monitoring without building custom plumbing. This lowers operational overhead, accelerates innovation, and improves the reliability of AI in production.

Published February 26, 2026