IDC MarketScape

IDC MarketScape: Worldwide Private AI Infrastructure Systems 2025 Vendor Assessment

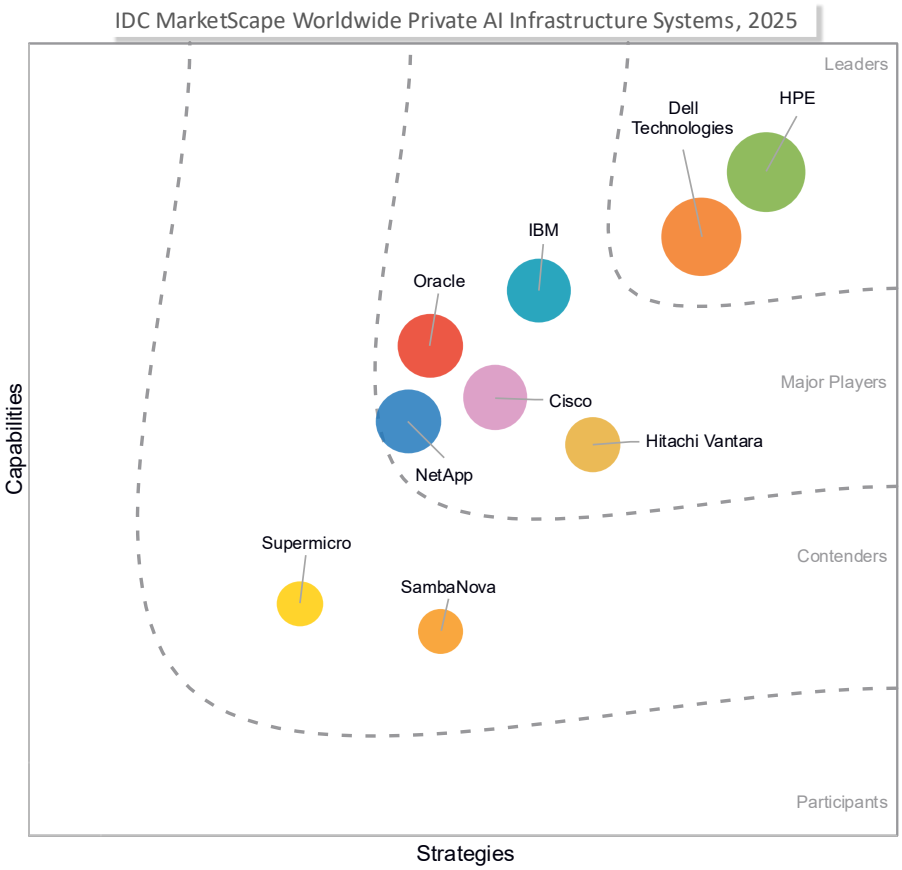
Mary Johnston Turner

THIS EXCERPT FEATURES HPE AS A LEADER

IDC MARKETSCAPE FIGURE

FIGURE 1

IDC MarketScape Worldwide Private AI Infrastructure Systems Vendor Assessment



Source: IDC, 2025

August 2025, IDC #US53002625e

Please see the Appendix for detailed methodology, market definition, and scoring criteria.

ABOUT THIS EXCERPT

The content for this excerpt was taken directly from IDC MarketScape: Worldwide Private AI Infrastructure Systems 2025 Vendor Assessment (Doc # US53002625).

IDC OPINION

As enterprise-scale organizations move AI proof-of-concept (POC) trials into production, many are finding that workload-specific cost, performance, and security considerations are creating requirements for on-premises AI-ready infrastructure deployment and operations. This is particularly true for organizations that are subject to data privacy and location regulations and compliance requirements as well as those that worry about the risk of confidential intellectual property and data seeping into public AI models.

Other organizations struggle to deliver ROI when using consumption-based public cloud infrastructure. Some organizations require local deployments for AI inference platforms used to support latency-sensitive workloads or very large data sets such as is often the case for digital twin use cases.

For all these reasons, many enterprises prefer to deploy at least some of their AI workloads in on-premises infrastructure. However, they often find that it can be challenging to accurately size, configure, integrate, scale, and operate GPU-based systems on their own. Agentic AI use cases are poised to further complicate these architectures with the introduction of multi-model use cases that depend on widely distributed inference, RAG, and reasoning engines.

For organizations that want to get started quickly with on-premises AI infrastructure, private AI infrastructure systems provide full-stack, standardized, turnkey solutions that can provide predictable costs, direct control over security and user access, modular scalability, ongoing access to the latest software tools and models, and a single-vendor support model.

As AI innovation accelerates, private AI infrastructure systems are emerging as important options for customers that want faster deployment of complete, optimized, fit-for-purpose stacks in dedicated on-premises or colocated facilities. IDC plans to track revenue associated with these systems via its Converged Infrastructure Tracker,

placing them within integrated platforms and systems, hyperconverged infrastructure, and/or composable infrastructure segments, depending on the system architecture.

IDC MARKETSCAPE VENDOR INCLUSION CRITERIA

This IDC MarketScape is focused on packaged, branded full-stack turnkey private AI infrastructure systems optimized for the needs of AI workloads that are deployed into dedicated on-premises and colocation facilities.

Private AI infrastructure systems are often focused on AI model training, fine-tuning, and/or inferencing use cases. They are offered as bundled, pre-integrated, and factory-certified builds that include computing, networking, storage, infrastructure software, security, and operational automation and support for AI model and agent libraries, blueprints, and developer tools. This assessment does not include AI PCs and workstations. Solutions defined only by reference architectures are also not included.

In determining whether vendors qualified for this analysis, IDC emphasized the following inclusion criteria:

- **Branded solution:** The vendor sells the entire system under a single brand, even if some of the components are sourced from one or more validated partners. The system must be advertised on the vendor's website and sold via predefined, preconfigured, modular packaging bundles, sometimes described as T-shirt sizes.
- **Full-stack infrastructure solution:** It includes full integrated, testing and validated AI-optimized compute, internal networking, and necessary management software as well as storage capabilities. The full system must be managed by a unified control plane for such activities as deployment, configuration, operations, scaling, load balancing, observability, security, and role-based access.
- **Enterprise-scale, turnkey solution**: It can be quickly deployed in self-owned and/or operated datacenters, colocation, hosting facilities, and/or edge environments and offers configurations that can be supported by power and cooling resources available in typical enterprise-scale datacenters. The vendor and/or the vendor's certified channel partners should be the primary sources of product support.
- **AI enabled:** The system includes access to a curated set of AI models, agents, development tools, and AI microservices to support AI model training, RAG, data management, and developer needs.
- **General availability:** The system was orderable and had general availability no later than January 1, 2025, and the vendor had announced plans for availability

in at least two major geographic regions for no later than the end of 2025. All vendors included in this IDC MarketScape were asked to provide IDC with reference customers running the system in production.

In many cases, these solutions are offered as part of a vendor's broader AI infrastructure portfolio which, in some cases, may also include highly customized solutions including cloud provider–scale systems. These types of customized solutions are not part of this IDC MarketScape analysis. This specific IDC MarketScape is focused on fully turnkey, certified, integrated solutions that offer enterprise-scale customers simplified options for deploying AI-ready infrastructure into on-premises and colocation environments.

ADVICE FOR TECHNOLOGY BUYERS

Enterprise technology buyers should structure their evaluation of private AI infrastructure systems by starting with an assessment of the workloads and data that will be running on the infrastructure and pair that assessment with consideration of how they may need to scale and update their AI-ready infrastructure and operating model over time. They also need to anticipate requirements to integrate on-premises infrastructure into their broader enterprise infrastructure environment including public cloud resources. Important evaluation and planning considerations include:

- **System sizing,** which is typically dependent on such use case–specific criteria including the size and complexity of the AI models being used as measured in terms of parameters and computational requirements (The size and complexity of the data sources also need to be factored into the sizing decision, particularly if the use case is focused on frequent model updates or extensive RAG for inference. The number of simultaneous users and numbers of tokens their usage will generate are also important factors. Many vendors provide sizing tools and use case guidelines or demonstration centers to help customers make these initial decisions.)
- **Infrastructure software** included as part of the full turnkey solution including AI software tools, libraries, and blueprints as well as infrastructure software capabilities such as operations control plans, system automation, and container or virtualization platforms included as part of the turnkey solution
- **System scalability and updates** as requirements evolve over time (Different systems will have different approaches in terms of using load balancers and expansion racks or other approaches to extending the initial system. As new generations of GPUs are introduced, it will also be important that first-generation systems are fully able to integrate with systems running next-generation GPUs.)

- **Time to results** based on system availability, shipping intervals, level of predelivery integration, and speed of activation once the system arrives onsite (Supply chain risks and contingencies should also be assessed.)
- **Facilities requirements** such as power, thermal, and cooling specifications (Many existing on-premises corporate datacenters may have power limitations or be unable to handle systems with liquid cooling. In some cases, customers are opting to deploy into colocation facilities that have been built specifically to support liquid cooling and the high levels of power consumption demanded by GPU-based systems.)
- **Vendor and partner support** including clear agreements about responsibilities for first, second, and third-line support (The majority of private AI infrastructure systems incorporate critical technologies from partners, so the availability of seamless multivendor support capabilities is important.)
- **Vertical industry-specific requirements** that can demand infrastructure ruggedization or form factor customization, particularly for industrial AI operations (Customers should evaluate any unique vertical industry requirements and ensure that the vendor or its channel partners can provide the required capabilities.)

Private AI infrastructure system tech buyers should recognize that their organization's AI strategies will continue to evolve, as will the enabling infrastructure technologies. Agentic AI is rapidly emerging to support more dynamic and computationally intensive use cases while GPU innovation continues at a rapid pace. Tech buyers need to ensure that private AI infrastructure systems are deployed and maintained in ways that will enable ongoing refresh and integration and be ready to support new use cases and AI technologies as they emerge. Conversations with potential vendors should balance immediate considerations related to cost, performance, and security, with longer-term road maps for expansion, tech debt avoidance, and optimizing operations over time.

VENDOR SUMMARY PROFILES

This section briefly explains IDC's key observations resulting in a vendor's position in the IDC MarketScape. While every vendor is evaluated against each of the criteria outlined in the Appendix, the description here provides a summary of each vendor's strengths and challenges.

HPE

HPE is positioned as a Leader in this 2025 IDC MarketScape for worldwide private AI infrastructure systems.

The HPE Private Cloud AI solution was introduced in June 2024. At the time, it was one of the first fully turnkey private AI infrastructure systems focused on enterprise-scale inference use cases, model fine-tuning, and on-premises deployments. A second generation of the portfolio was introduced in June 2025 to add configurations supporting the latest generation of NVIDIA GPUs and more advanced large language models (LLMs) and agentic AI use cases.

The system is anchored by NVIDIA GPUs as well as software and hardware codeveloped as part of the ongoing NVIDIA AI Computing by HPE initiative. It is engineered to leverage the full HPE infrastructure portfolio including ProLiant Gen12 servers and HPE Alletra disaggregated file storage with object included. Networking is provided by NVIDIA Spectrum Ethernet networking switches. HPE also provides AI systems based on AMD and Intel processors, but those are not currently offered as HPE Private Cloud AI turnkey solutions.

The management control plane for full-stack ITOps functionality is powered by the HPE GreenLake hybrid cloud operations platform. This control plane unifies infrastructure observability across the full stack, including infrastructure runtimes for VMs, bare metal hosts, and Kubernetes clusters. It also supports consistent set up and control of user roles and permissions and zero trust security. Future editions are expected to take advantage of the recently announced GreenLake Intelligence agentic AIOps framework. The HPE GreenLake management control plane is typically enabled as a GreenLake Cloud service but is also available for fully air-gapped deployment.

All HPE Private Cloud AI deployments come pre-integrated with the HPE AI Essentials Software platform that provides seamless low-code, wizard-based developer access to open source software, HPE and NVIDIA tools, models, and frameworks from a persona-based model catalog. NVIDIA NIMs inference microservices and ongoing support for NVIDIA blueprints are included, as is support for Hugging Face, Langflow, and other pre-validated solutions from HPE ISV partners.

This updated portfolio announced in June 2025 features the following use case–defined options:

- **Developer system:** 2 NVIDIA H100 NVL GPUs, 32TB integrated storage, customer networking up to 2.2kW
- **Small:** 8 NVIDIA RTX Pro 6000 GPUs, 109TB HPE Alletra storage, 400GbE NVIDIA networking up to 12kW a rack
- **Medium sized:** 8 NVIDIA H200 GPUs, 217TB HPE Alletra storage, 400GbE NVIDIA networking up to 13kW a rack
- **Large:** 16 NVIDIA H200 GPUs, 217TB HPE Alletra storage, 400GbE NVIDIA networking up to 17.4kW a rack

The small, medium-sized, and large configurations can be horizontally scaled by adding expansion racks of up to 16 H200 GPUs per rack and connecting them to the top of rack networking switch in the primary rack.

With the introduction of an updated portfolio running the next generation of GPUs, HPE has taken a number of steps to ensure customer investment protection over time. One-click software updates are automated and coordinated across the full stack, and workloads running on first-generation processors have full interoperability with workloads deployed on the second generation. In addition, the company has implemented a federated architecture across expansion racks so multiple GPU types can run in the same environment and act as a shared resource pool.

Delivery ranges from two weeks for the developer system to several months for larger systems. HPE Private Cloud AI systems can be purchased as capex, subscription, or via the HPE GreenLake as-a-service consumption model. HPE has partnerships with colocation providers such as Equinix to allow customers to explore and experience Private Cloud AI first hand for POCs/pilots/demos.

Regional and worldwide HPE distributors and resellers have access to turnkey systems. GSI partners such as Deloitte, Accenture, Infosys, and Wipro are often involved in customer AI engagements, implementation, and support. HPE also provides free presale discovery sessions as well as paid transformation workshops and deployment services.

The Private Cloud AI system is fully racked, networked, and tested by HPE prior to being shipped to the customer or colocation site. Once the rack is connected to power and external networking, the system is generally up and running and ready for customers to deploy workloads in a few hours or days. HPE is the primary support provider.

Strengths

HPE Private Cloud AI is tightly integrated and fully turnkey. It requires little onsite integration to get the system up and running. Integrated, full-stack automation, observability, and security functions are enabled via the HPE GreenLake hybrid cloud operations platform. The recently announced GreenLake Intelligence agentic operations capabilities may add additional efficiencies over time.

The system's disaggregated storage architecture allows customers to scale compute and storage independently. For customers that want a tightly integrated stack that includes robust storage and management integrations, the HPE Private Cloud AI reduces the complexity related to ensuring data pipelines and AI workloads are efficiently integrated.

HPE's long history of working with high-performance computing and liquid cooling systems is also valuable for customers that face power or thermal constraints and are looking for partners that can help them optimize datacenter facilities for AI.

Challenges

HPE has opted to prioritize its own technologies and those provided by or co-engineered with NVIDIA. HPE storage and HPE VM runtime and Kubernetes cluster management software are the system standards. These turnkey solutions do not include certified testing, integration, or support for third-party storage systems or Kubernetes platforms. For customers that are aiming to reuse existing non-HPE storage or licenses for VM runtimes and Kubernetes platforms, these HPE-specific design choices may be challenging to overcome. Customers can always opt to work with HPE using reference architectures but will then lose out on the fast deployment and tight integration and operational automations provided by the turnkey system.

Customers with plans for rapid expansion may want to dig deeper into HPE's current scalability constraints that are limited to four expansion racks for the turnkey solution portfolio.

APPENDIX

Reading an IDC MarketScape Graph

For the purposes of this analysis, IDC divided potential key measures for success into two primary categories: capabilities and strategies.

Positioning on the y-axis reflects the vendor's current capabilities and menu of services and how well aligned the vendor is to customer needs. The capabilities category focuses on the capabilities of the company and product today, here and now. Under this category, IDC analysts will look at how well a vendor is building/delivering capabilities that enable it to execute its chosen strategy in the market.

Positioning on the x-axis, or strategies axis, indicates how well the vendor's future strategy aligns with what customers will require in three to five years. The strategies category focuses on high-level decisions and underlying assumptions about offerings, customer segments, and business and go-to-market plans for the next three to five years.

The size of the individual vendor markers in the IDC MarketScape represents IDC's estimate of the vendor's current number of private AI infrastructure system

deployments running in production, recognizing that this is an emerging solution area where initial ordering and shipping for most vendors started less than 18 months ago.

IDC MarketScape Methodology

IDC MarketScape criteria selection, weightings, and vendor scores represent well-researched IDC judgment about the market and specific vendors. IDC analysts tailor the range of standard characteristics by which vendors are measured through structured discussions, surveys, and interviews with market leaders, participants, and end users. Market weightings are based on user interviews, buyer surveys, and the input of IDC experts in each market. IDC analysts base individual vendor scores, and ultimately vendor positions on the IDC MarketScape, on detailed surveys and interviews with the vendors, publicly available information, and end-user experiences in an effort to provide an accurate and consistent assessment of each vendor's characteristics, behavior, and capability.

Market Definition

Private AI infrastructure systems are emerging as important options for vendors to offer standardized, certified, turnkey offerings for complete, optimized, fit-for-purpose AI infrastructure stacks in on-premises or colocated facilities. IDC plans to track revenue associated with these systems via its Converged Infrastructure Tracker, placing them within integrated platforms and systems, hyperconverged infrastructure, and/or composable infrastructure segments, depending on the system architecture.

LEARN MORE

Related Research

- *Preparing Enterprise AI-Ready Infrastructure for Agentic-Driven Disruption: Impacts and Opportunities 2025–2027* (IDC #US53299325, April 2025)
- *Planning for GenAI Inferencing Impact on Infrastructure Investment Decisions* (IDC #US51994524, March 2025)
- *AI-Ready Infrastructure Tech Buyer Adoption Trends Research Highlights Demand for Scalable Infrastructure Platforms and Alignment with Business ROI* (IDC #US52101825, December 2024)
- *Harnessing Hybrid Infrastructure to Fuel AI Business at Scale: A C-Suite Playbook* (IDC #US52101325, August 2024)

Synopsis

This IDC study represents a vendor assessment of full-stack, turnkey private AI infrastructure systems using the IDC MarketScape model.

"Full-stack, turnkey private AI infrastructure systems offer enterprise customers a fast track on-ramp to sizing and deploying on-premises AI computing and storage resources while maintaining predictable costs and performance and ensuring that private and confidential data remain secure," explains Mary Johnston Turner, research vice president, Digital Infrastructure Strategies at IDC. "Vendors that offer fully integrated and validated solutions simplify AI-ready infrastructure configuration, deployment, and operations while allowing customers to focus on time to results and maximizing the business value of AI infrastructure investments."

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com