Gartner

Gartner Insights

Market Guide for Hybrid AI Infrastructure

Dennis Smith, Chandra Mukhyala

7 May 2025

Gartner

Market Guide for Hybrid AI Infrastructure

7 May 2025 - ID G00818310 - 13 min read

By: Dennis Smith, Chandra Mukhyala

Initiatives: Technologies and Markets; Reinvent I&O as an Enabler of AI Value

Emerging AI requirements drive demand for solutions enabling AI workloads to be deployed within hybrid environments. This requires infrastructure that can handle performance demands introduced by AI workloads. This research helps I&O leaders navigate this dynamic market.

Overview

Key Findings

- AI, specifically GenAI, is transformational — particularly industry-specific models — and is becoming a top priority for all enterprises. This requires infrastructure and IT operations (I&O) teams to plan for the impact on their infrastructure strategies.
- Data is a key component of AI deployments, and since most data is still and often must stay within the confines of enterprise locations, I&O teams must prepare for infrastructure that enables the necessary processing and operationalizing of data.
- Many vendors are converging around this domain with varied compute, storage and networking offerings optimized for AI. I&O needs a structured plan to navigate the market.

Recommendations

- Make a clear identification of what use cases must have private AI infrastructure and what can be deployed on cloud provider PaaS to be certain of the scope of your use cases and capacity needs before investing.
- Differentiate between the infrastructure needed for training and fine-tuning from the infrastructure needed for inference and monitoring of deployed models as they will differ because inference can often be done on less expensive infrastructure located closer to users.
- Favor fully managed stack offerings that are conformant to industry standard reference architectures for self-managed private data center infrastructure (DCI).

Strategic Planning Assumption

By 2028, more than 20% of enterprises will run AI workloads (training and/or inference) locally in their data centers, an increase from fewer than 2% as of early 2025.

Market Definition

Gartner defines hybrid AI infrastructure as offerings that address the need to enable AI and machine learning (ML) workloads across enterprise data centers, colocation and the edge, and public cloud. The offerings include a combination of compute, storage and networking components, along with the requisite enablement tooling, middleware and libraries. Infrastructure and operations (I&O) leaders can use these solutions to deploy AI in the most strategic location for their needs, whether on-premises, at the edge or in the cloud.

Hybrid AI infrastructure addresses enterprise needs for deployment of inference and possibly training to enterprise data centers. The solutions provide foundational capabilities for companies unable or unwilling to solely leverage public cloud AI infrastructure (e.g., due to digital sovereignty or costs involved with moving data from an existing location).

Mandatory Features

Hybrid AI infrastructure must have infrastructure and/or infrastructure-related tooling in support of the building and/or operation of traditional AI and generative AI (GenAI) workloads. This includes the enablement of:

- Automated aspects of building, deploying and maintaining AI/ML models

- Data processing and preparation of data for model training and inference
- Packaging, deploying and managing the excursion of AI/ML training and model inference, as well as fine-tuning the operation
- Integration of AI/ML libraries (e.g., PyTorch, Keras, CUDA)
- Tools enabling the deployment and operation of AI/ML workloads (e.g., InstructLab)
- Security, including encryption, access controls and compliance

Specific capabilities include the support of:

- High computational demands
- High-throughput, low-latency and high-capacity storage to support the large datasets and high-speed data transfers necessary to keep graphics processing units (GPUs) engaged with large quantities of data
- 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])
- The ability to support workloads on-premises, at the edge and within a public cloud hyperscaler

Common Features

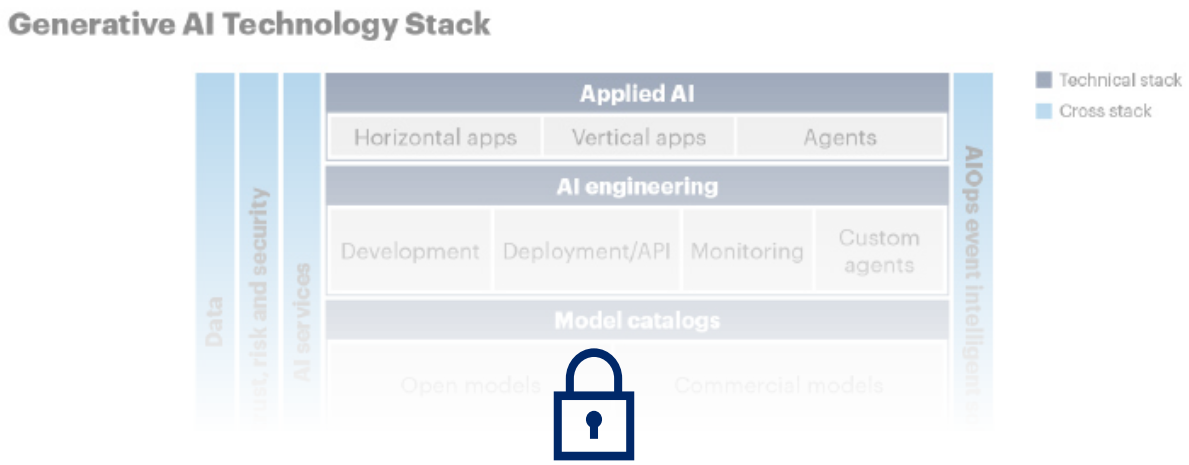
- A combination of compute, storage and networking infrastructure in support of AI workloads, which can be installed in enterprise data centers, colocation centers, at the edge and within the public cloud.
- Software that enables the operation of the underlying infrastructure that supports AI workloads. This includes some combination of automated provisioning, optimizing, observing and/or securing capabilities for the underlying infrastructure.
- Global data management that can scalably process and store high volumes of data across hybrid multicloud environments.

- Dedicated networks for training (e.g., a supercomputing network with lossless architecture) and inference (lower cost but low latency).
- Infrastructure-enabling frameworks (e.g., PyTorch, TensorFlow, JAX) and programming languages that enable workload processing.

Market Description

Enterprises have leveraged variations of AI (from traditional AI to recently released GenAI technologies) for many years, often leveraging more traditional infrastructure. GenAI and new AI use cases have placed additional demands on enterprise infrastructure due to the need to rapidly process a massive amount of data. Figure 1 depicts a GenAI technology stack.

Figure 1: GenAI Technology Stack



Actionable, objective insights

Explore these additional complimentary resources and tools for digital workplace leaders:

Conferences

Cloud Conferences and Events 2026

Join our cloud conferences to prepare for the future of infrastructure and operations.

View Calendar

Roadmap

3 Key Trends Shaping the I&O Landscape in 2026

Accelerate value, cut risk and build future-ready infrastructure and IT operations with strategic priorities.

Download Now

Webinar

2026 Technology Adoption Roadmap for Infrastructure and Operations

Explore how I&O leaders can benchmark technology adoption, assess risk and navigate AI-driven change in 2026.

Watch Now

Tool

Gartner Heads of Infrastructure and IT Operations Effectiveness Diagnostic

Benchmark your performance and develop key actions that differentiate highly effective infrastructure and IT operations leaders.

Learn More

Already a client? Get access to even more resources in your client portal.
Log In ↗

Connect with us

Get actionable, objective business and technology insights to deliver on your mission-critical priorities. Our expert guidance and tools enable faster, smarter decisions and stronger performance. Contact us to become a client:

U.S.: 1 855 322 5484
International: +44 (0) 3300 296 946

Become a Client

Learn more about Gartner for Digital Workplace Leaders
gartner.com/en/infrastructure-and-it-operations

Stay connected to the latest insights

Attend a Gartner conference
View Conference

Gartner®