# AI Infrastructure Market Landscape

## SPOTLIGHT REPORT

**May 2026**

## Report Overview

TBR's *AI Infrastructure Market Landscape* follows trends in the accelerated computing market. This includes tracking the performance of leading AI server and systems OEMs and following developments and the performance of key semiconductor manufacturers responsible for developing AI accelerators and supplying them to infrastructure OEMs and ODMs. The report includes an overview of vendors' revenue performance as well as key trends and strategies within each industry group, including developments between infrastructure OEMs, semiconductor providers and solution providers, and AI server and systems market share projections.

Publish date of latest edition: April 23, 2026

Click here to view a full list of the report's research topics and vendor coverage.

*"AI infrastructure demand is surging, but hyperscalers are reshaping the market by abstracting hardware, investing in custom ASICs, and shifting control up the stack, forcing OEMs and silicon vendors to rethink differentiation and long-term value capture."*

— Senior Analyst Ben Carbonneau

## Executive Summary Excerpt

### AI infrastructure demand continues to grow as new use cases emerge, though adoption is tempered by operational challenges, energy limitations and competitive disruption

**Key Market Growth Drivers**

- Large AI infrastructure investment announcements from service providers and sovereigns alike are becoming increasingly common, as are growing hyperscaler AI capex commitments.
- NVLink Fusion makes deploying NVIDIA GPUs more attractive to hyperscalers.
- The number of industry-specific AI solutions is rising.
- Pretrained models are becoming increasingly available.
- AMD solutions are gradually becoming more viable, increasing customer choice.
- Reasoning models — and agentic systems — require more GPU resources in the inferencing phase compared to traditional knowledge-based models.
- A growing number of GPU resources are becoming available through neoclouds.
- New hardware and software platforms and optimizations are reducing the cost per token of AI factories.
- A growing ecosystem of systems integrators and solution providers is architecting and deploying AI solutions.

**Key Market Growth Inhibitors**

- The macroeconomic and geopolitical environments are uncertain.
- Many AI use cases still lack a clear ROI.
- Hyperscalers are increasingly investing in servers and systems based on custom AI ASICs, negatively impacting merchant accelerator vendors.
- Data readiness and infrastructure complexity are major AI adoption inhibitors among enterprises.
- Energy constraints limit AI factory build-out.
- While AI compute demand is rising, so are tokens-per-watt outputs of next-generation AI servers and systems.
- The formation of a secondary merchant accelerator-based AI server and systems market caps OEMs' addressable market for new infrastructure sales but could drive adjacent services opportunities.
- Certain organizations have strict data privacy and security requirements.
- The semiconductor supply chain remains relatively fragile despite investments to enhance resiliency.

## Workload requirements drive hybrid AI deployments across cloud and on-premises environments, sustaining GPU demand despite rising ASIC adoption

### Workloads requirements drive AI deployment decisions

As AI infrastructure bifurcates between training and inference, enterprise and sovereign AI deployments are increasingly dictated by workload characteristics and economics. These factors drive a mix of cloud, on-premises and edge environments, with deployment decisions reflecting requirements for scale, efficiency, latency and data governance.


- High-volume, steady workloads drive cloud scale efficiency.
- Uniform, repeatable workloads enable ASIC-optimized inference.
- Intermittent and multitenant workloads favor cloud elasticity.
- Variable or heterogeneous workloads favor GPU-based infrastructure.
- Data gravity and sovereignty drive on-premises and edge deployment.
- Latency-sensitive workloads require proximity to compute.
- Deterministic performance needs favor a dedicated infrastructure.
- Data control and security requirements favor on-premises environments.

### Enterprise AI adoption favors hybrid deployment, sustaining GPU demand

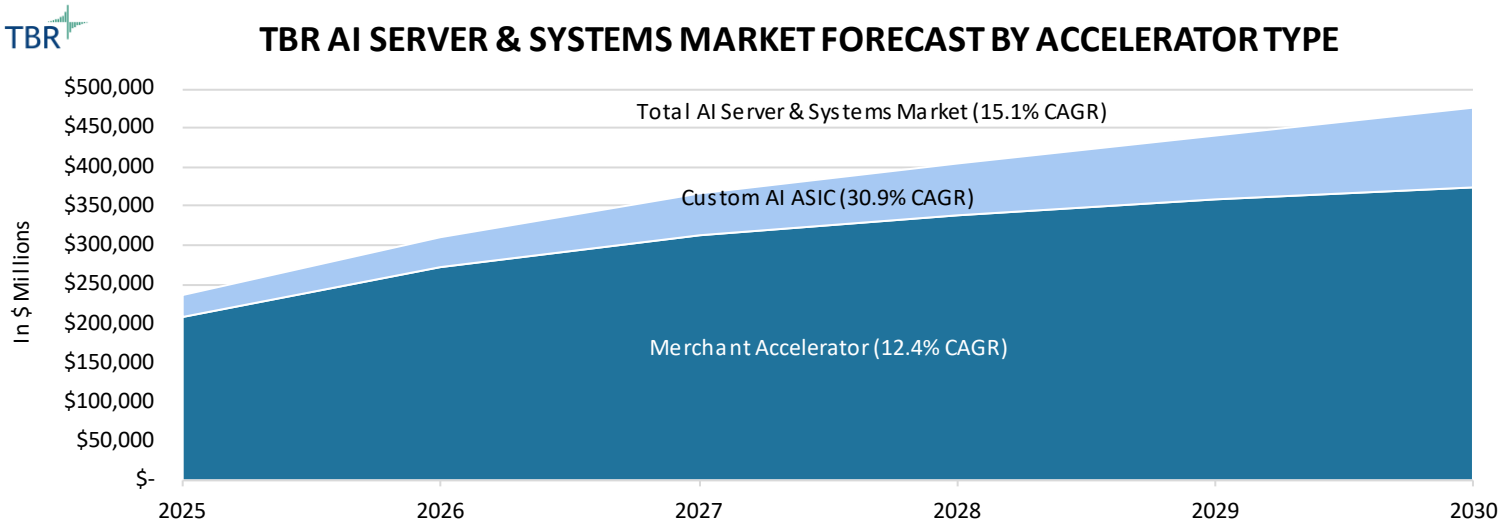
*Enterprise AI deployment ultimately converges on hybrid architectures, balancing cost, performance and control.*

- TBR believes enterprise inference deployment will be shaped by data governance, security and data gravity, driving adoption of hybrid cloud and on-premises environments. Real-time and latency-sensitive use cases, along with regulatory requirements, often favor on-premises and edge deployments, which are more likely to rely on merchant accelerator-based systems due to workload variability.
- Enterprise AI adoption will remain constrained by data readiness, talent availability and clarity of return on investment, resulting in uneven adoption across industries and concentration among large enterprises with well-defined use cases, particularly in regulated and edge-specific applications.
- Infrastructure constraints — including power availability, cooling requirements and rack density — will further influence deployment decisions, in some cases favoring cloud-based inference or less dense air-cooled systems despite latency and performance trade-offs.

## Custom AI ASIC adoption accelerates and diversifies the accelerator landscape as hyperscalers continue to anchor market growth

**TBR Projections**

- The total AI server market grew an estimated 46.9% year-to-year in 2025, driven primarily by large-scale AI factory build-outs.
- The total AI server market will grow at a 15.1% CAGR through 2030, driven by the ongoing initial AI infrastructure build-out through 2028 and a combination of service provider refresh and increasing sovereign and enterprise investments in the later part of the forecast period.
- TBR predicts custom AI ASIC-based systems growth will outpace merchant silicon-based systems through 2030.




SOURCE: TBR ESTIMATES

**Market Trends**

- Service providers continue to represent the lion's share of demand, with hyperscalers purchasing both node-level and system-level infrastructure based primarily on merchant accelerators. However, hyperscalers are increasingly investing in developing next-generation custom AI ASICs while funneling more resources into deploying systems based on these accelerators as certain inferencing workloads scale and become more stable.
- Aside from hyperscalers, the growing neocloud segment is a key driver of demand within the service provider market. Neoclouds, similar to model builders, require merchant accelerator infrastructure at scale to support increasing demand from AI labs and startups, as well as hyperscalers, which represent some of the largest neocloud clients. While neoclouds' value proposition originated in offering bare GPU instances, neoclouds such as CoreWeave are investing in moving up the value chain beyond orchestration to the platform layer. However, in contrast to hyperscalers, neoclouds are in closer alignment with NVIDIA because they are less actively investing in NVIDIA infrastructure abstraction with neoclouds, typically the customer, rather than the provider, owns their own stack, which is increasingly the case as hyperscalers invest in managed AI services.

## TBR predicts hyperscaler refresh cycles will drive a secondary AI infrastructure market

**Scenario Discussion: As build-out-driven demand transitions to replacement-driven demand, a secondary AI server and systems market will emerge, presenting new services opportunities and lowering the gate to enterprise on-premises and edge AI adoption**

**Key Takeaways**

- Inferencing workloads place increased emphasis on efficiency.
- In an energy-constrained market, generational tokens- per-watt improvements will dictate replacement cycle rates.
- TBR believes hyperscaler waterfalling will act as an initial buffer before the establishment of a material secondary merchant accelerator-based server and systems market.
- The secondary market will support both existing and new services opportunities for OEMs, solution providers and systems integrators while reducing the cost barriers of on-premises and edge AI adoption.

- Due to their immense size and workload share, hyperscalers are poised to remain the dominant industry group driving the AI server and systems market for the foreseeable future. While increasing AI compute demand will remain the core driver of hyperscalers' investments, today's build-out-driven demand will increasingly shift toward replacement-cycle demand as the market becomes energy-constrained.
- At scale, inference prioritizes efficiency above all else, and the rate at which hyperscalers refresh AI servers and systems will be closely correlated to the rate of tokens-per-watt improvements in future-generation silicon — both custom AI ASICs and merchant accelerators — as well as the servers and systems leveraging them.
- However, TBR expects hyperscalers to waterfall prior-generation infrastructure, putting next-generation systems — delivering higher tokens-per-watt efficiency — in place to support the most demanding workloads while transitioning previous-generation systems to relatively less-intensive workloads, thereby optimizing inference costs. However, as this cycle persists, at some point previous-generation systems will be phased out of hyperscaler environments. TBR believes the current rate of generational tokens-per-watt efficiency gains supports the thesis that merchant accelerator-based servers and systems will be phased out of hyperscaler environments before they burn out, while custom AI ASIC-based servers and systems will be sweated out. As a result, there will be residual value in merchant accelerator-based servers and systems, and given the scale of hyperscalers' initial deployments, the initial wave of these servers and systems being phased out of hyperscaler environments will coincide with what TBR predicts will be the birth of a material secondary market.
- Secondary rack-scale integrated systems are likely to be repurposed by neoclouds and other more advanced customers that have the workload capacity and existing data center environments necessary to house and run the equipment, with excess systems potentially being farmed for components that will find their way into refurbished servers. Secondary node-level servers are more likely to end up in enterprise on-premises and edge environments due to their lower cost and operational demands.
- The implications of hyperscaler waterfalling and the eventual establishment of a material secondary merchant accelerator- based server and systems market create new opportunities for OEMs, solutions providers and systems integrators to capitalize on secondary system refurbishment, certification and redeployment, while also supporting the AI solution architecting and services initiatives that are already being established to support the first wave of enterprise on-premises and edge AI adoption.

## Supporting Research

TBR's *AI Infrastructure Market Landscape* follows trends in the accelerated computing market. This includes tracking the performance of leading AI server and systems OEMs and following developments and the performance of key semiconductor manufacturers responsible for developing AI accelerators and supplying them to infrastructure OEMs and ODMs. Access the full *AI Infrastructure Market Landscape* and all of the supporting research below with a 60-day free trial of TBR Insight Center™.

### Vendor Analysis

Deep-dive analysis of a single vendor across corporate strategies, tactics, SWOT analysis, financials, go-to-market strategies and resource strategies

- Dell Technologies
- Hewlett Packard Enterprise
- Lenovo Group
- NVIDIA

### Benchmarks

Comparison of vendor performance in a market, including analysis on vendor strategies, financial performance, go-to-market and resource management

- IT Infrastructure

### Market Forecasts

Analysis of market opportunity and contain current market sizing and five-year forecasts, including analysis on growth drivers, top trends and leading market players

- IT Infrastructure

### Market Landscapes

Analysis of an emerging or disruptive market segment or technology, including insight into how vendors and customers address the emerging technology as well as market sizing, vendor positioning, strategies, acquisitions, alliances and customer adoption trends

- AI Infrastructure
- Infrastructure Services

### Customer Research

A view of end-customer adoption within a market, including information on adoption time frames and preferences, spending/budgeting, vendor preference and selection criteria, and satisfaction analysis

- Infrastructure Strategy

## Table of Contents and Vendor Coverage

Publish date of latest edition: April 23, 2026



**Vendor Coverage**

**Server vendors**
- Dell Technologies (Dell)
- Hewlett Packard Enterprise (HPE)
- Lenovo
- Supermicro

**Semiconductor companies**
- AMD
- NVIDIA

**Coverage-adjacent players**
- Amazon
- Broadcom
- Google
- Intel
- Meta
- Microsoft



*Technology Business Research, Inc. is a leading independent market, competitive and ecosystem intelligence firm specializing in the business and financial analyses of hardware, software, professional services, and telecom vendors and operators. Serving a global clientele, TBR provides timely and actionable market research and business intelligence in formats that are tailored to clients' needs. Our analysts are available to address client-specific issues further or information needs on an inquiry or proprietary consulting basis.*