Brief

ENGINEERING THE FUTURE OF AI AT SCALE

Aging systems can't support today's AI demands. Here's how organizations can build scalable, composable infrastructure for what's next.

Meeting the demands of next-generation AI isn't just about adding more hardware. It's about engineering a cohesive, scalable foundation that delivers agility, performance, and control across every layer.

For customers implementing AI on a massive scale or with complex workload needs, Hewlett Packard Enterprise delivers customized solutions based on proven architectures, deployed through our expert services.

Yesterday's data centers aren't a good fit for AI

Yesterday's data centers weren't built for AI. They lack the orchestration tools to coordinate complex workloads, the observability to monitor performance in real time, and the tenancy controls needed to support multiple users securely and efficiently. As a result, expensive GPUs sit idle, energy is wasted, and deployments drag.

Modern AI workloads, especially GPU-intensive ones, demand more than manual provisioning and static environments. They require dynamic scheduling, airtight isolation among tenants, and the ability to move workloads seamlessly across locations.

For service providers, the old model leaves revenue on the table. For enterprises, it slows innovation at the exact moment speed matters most.

What IT leaders need: Connect and control multitenant AI

To support modern AI at scale, organizations need infrastructure that's purpose-built for multi-tenancy, an increasingly common architecture that allows multiple workloads or clients to share these costly computing resources, thus maximizing their utilization and minimizing chargebacks. That means centralized orchestration, fine-grained access control, tenant isolation, and end-to-end observability — all working across multiple sites and clouds.

"You want to make sure the right resources are used at the right time by the right people," says Piyush Shukla, director of AI product marketing at HPE. "You don't want GPUs sitting idle, so you need strong observability, and that comes from multitenancy, where you can see who's using what and where."

It's not enough to make infrastructure smarter. It has to be operable, too. Teams need to deploy, monitor, and manage AI workloads with the same ease they expect from public cloud, but with the performance, compliance, control, and data sovereignty of on-premises infrastructure.

It ultimately comes down to a balancing act that delivers cloud-like self-service while maintaining tight governance. Without that foundation, AI deployments can either stall or spiral out of control.

AI factory at scale: A new model for modern AI

Modern AI infrastructure needs to function like a factory — modular, efficient, and built to grow. That's the idea behind AI factory at scale: a layered, composable architecture designed to support high performance AI across tenants and locations.

At the foundation are high-density GPU servers like HPE Cray XD, paired with liquid cooling and high-speed interconnects to handle power and thermal demands. Above that sits the virtualization and container layer, combining VMs and Kubernetes to support diverse AI workloads.

The control plane brings in essential services through a data software stack that starts with the NVIDIA® AI Enterprise software platform. This provides usage metering, role-based access control, tenant isolation, and a self-service catalog. Automation handles lifecycle ops, observability, and policy enforcement. And at the top are your AI workloads — large language models, retrieval-augmented generation, and agentic systems — all tuned and ready to deploy.

The result is a modular platform that grows with your needs but without costly overhauls.

Designing for performance at scale

Running AI at scale starts with making the right infrastructure choices. That means evaluating each site's power, cooling, and rack density, especially when deciding between air- and liquid-cooled systems. Liquid cooling enables higher-density deployments, making it ideal for data centers with limited space. High-speed networking is another must. Without fast interconnects, even the most powerful GPUs can become bottlenecked.

Placement matters, too. GPUs should be sized for the workload and deployed close to the data to reduce latency. Data gravity is real, and moving large datasets across regions adds both delay and cost.

Finally, there's the issue of observability. As Shukla puts it, "Normally you take four or five years to depreciate an IT asset. GPU depreciation can be accelerated by technology advancement, usage patterns, high wear and tear, and demand in secondary markets. So, you need to maximize your usage of that resource very quickly. That's why observability becomes critical."

These decisions aren't just technical or operational. They're financial. Every watt consumed, rack deployed, and workload placed has a direct impact on utilization and return on investment. The right infrastructure supports AI and accelerates its value.

Where cloud experience meets control

The control plane is where cloud convenience meets enterprise control. With built-in metering, chargeback, and a self-service catalog, it reduces the burden on IT teams while giving users what they need fast.

Tenant isolation is enforced at both the network and RBAC levels, keeping environments secure and separated without extra overhead. That makes it possible to offer true AI-as-a-service operations, whether you're serving internal business units or external customers.

"Through the control plane, tenants can access open-source software, monitor their own billing and metering, and get the resources they need — all from a single dashboard," says Shukla. "If you want more GPUs, it's a click, not a ticket. That kind of flexibility removes the complexity from getting work done."

Granular cost controls also mean IT can track usage down to the GPU and bill accordingly. That means budget clarity for enterprises and new revenue streams for service providers.

Why modernized AI infrastructure pays off

Efficiency is only part of the story. A modernized AI data center helps teams get more done. Developers can spin up what they need from curated, pre-tested stacks, reducing time to deployment from weeks to hours. Hardware gets used more effectively, driving better ROI. And modular scaling keeps costs aligned with demand.

Compliance stays intact, too. Sensitive data remains on-prem or in-country, aligned with local regulations and enterprise policies. For service providers, the opportunity is even bigger. Instead of reselling generic compute, they can package and sell advanced AI services.

That shift — from static infrastructure to dynamic, service-driven platforms — is what separates the AI leaders from the laggards.

A practical path to modernization

Modernizing for AI doesn’t mean starting from scratch. Begin with an AI readiness assessment: evaluate power, cooling, governance, and in-house expertise. From there, stand up a sandbox, which is a pilot stack to test real workloads and validate ROI.

HPE can help design composable solutions that integrate with what you already have, reducing risk and accelerating time to value. And as demand grows, you can scale in place by adding tenants, nodes, and workloads without tearing out what works.

Don’t wait for the “perfect” setup. Aim for something you can observe, operate, and profit from, then build on it. That’s how real progress starts.

From foundation to future

AI is evolving fast, and scale is no longer optional. With an AI factory at scale approach, IT leaders can engineer infrastructure that matches the ambition of modern AI: agile, observable, and built for growth.

Learn more at

HPE.com/ai/insights

Visit HPE.com

Chat now

a50013339ENW

HEWLETT PACKARD ENTERPRISE

hpe.com