HPE | NVIDIA

Solution brief

WORK AND COLLABORATE AT-SCALE

Operationalizing HPE AI Factory at-scale with visibility, control, and multi-tenancy

As AI adoption accelerates, model-building enterprises and AI service providers face a common challenge: moving from isolated AI pilots and discrete projects to productionizing AI throughout the organization. "Scaling AI Requires New Processes, Not Just New Tools," as outlined by the Boston Consulting Group[1]—it demands a unified approach that enables organizations to operationalize AI while seamlessly expanding operations from small-scale pilots to full enterprise-wide deployment. This includes supporting workloads ranging from tens to tens of thousands of GPUs while maintaining performance, efficiency, and manageability.

This means moving AI out of the lab and taking models beyond pilots or proofs of concept and embedding them into live workflows, systems, and decision-making processes—real business operations where AI can deliver measurable outcomes at enterprise scale. A critical capability is managing the full AI lifecycle end-to-end—from data ingestion and model training to tuning, inference, monitoring, and continuous retraining—while applying consistent governance, security, and compliance controls.

At-scale, operationalizing AI also demands infrastructure and orchestration that can support multiple workloads, teams, and environments efficiently and automatically. It transforms AI from isolated experiments into a repeatable, reliable enterprise capability—much like manufacturing operationalizes production lines.

When combined with centralized visibility and control as well as multi-tenancy support, this is precisely what HPE AI Factory at-scale, co-engineered with NVIDIA, is designed to deliver.

HPE and NVIDIA together

As part of the NVIDIA Computing by HPE portfolio, HPE AI Factory at-scale combines industry-leading HPE hardware and an intelligent control plane with NVIDIA accelerated computing, software, and networking. This enables organizations to build and operate AI at-scale by addressing three key foundational requirements:

1. Operationalizing AI across its lifecycle, turning AI from isolated experiments into reliable, repeatable, and business-critical systems that deliver ongoing value
2. Observability and control across hybrid environments
3. Secure, efficient multi-tenancy

Together, these capabilities can transform AI into a reliable, governed, and cost-effective enterprise-wide capability.

[1] "Scaling AI Requires New Processes, Not Just New Tools", Boston Consulting Group, January 20, 2026.

The challenges of scaling AI

Many organizations kick off their AI journey by focusing on specific, isolated use cases. While these initial successes showcase AI's potential, scaling it across the enterprise to achieve lasting business value proves to be a much greater challenge.

According to Zapier's "The state of RevOps automation report," only 1 in 5 companies have successfully adopted AI at an organizational level.[2]

Without a cohesive strategy, organizations often accumulate scattered infrastructure, duplicated effort, and disconnected tools. The result is a fragmented AI environment—models built in silos, overprovisioned infrastructure, inconsistent GPU utilization, and limited visibility across teams and environments—making it difficult to move from experimentation to impact.

To scale AI across the enterprise, organizations face the following operational challenges:

Challenge	Example / real-world scenarios
Scale	— A financial services firm struggles to train large AI models on decades of transaction data due to limited server capacity. — Security policies restrict data movement across departments and regions, delaying AI pipelines.
Complexity	— A retailer deploying AI for forecasting, recommendations, and inventory lacks shared expertise, slowing deployment. — Integrating AI solutions from multiple vendors causes compatibility issues due to lack of cross-system expertise.
Cost	— Maintaining AI infrastructure in-house requires hiring specialized engineers and data scientists. — Continuous monitoring, retraining, and maintenance of AI models create ongoing operational costs.
Data	— Global healthcare organizations need to manage and analyze data from diverse populations, regions, and languages to address public health challenges. — Inconsistent or incomplete data slows model training and reduces AI effectiveness.
Security	— A pharmaceutical company requires scalable infrastructure to manage and analyze data from diverse markets, regulatory environments, and patient populations to support global drug launches. — A defense contractor transmitting AI-generated simulations must safeguard against cyberattacks and insider threats.

Without addressing these challenges and adopting a scalable operating model, organizations cannot efficiently support multiple AI teams, business units, or customers—limiting innovation and reducing return on their AI investment.

[2] "The state of RevOps automation report," Zapier, accessed April 2026.

Overcoming the challenges: Three pillars for AI at-scale

1. Operationalizing AI at-scale across the entire AI lifecycle

How can organizations transform a fragmented landscape of disconnected AI efforts and pilot projects into a cohesive enterprise-wide ecosystem that drives value at-scale? The answer lies in operationalizing AI—enabling consistent, repeatable AI workflows from development through production.

HPE AI Factory at-scale, co-engineered with NVIDIA, seamlessly integrates with existing infrastructure and offers built-in orchestration, observability, and lifecycle automation, making it ideal for global enterprises and service providers delivering large-scale AI environments with multiple tenants. HPE AI Factory is built using NVIDIA-powered AI infrastructure as a foundational component and leverages NVIDIA accelerated computing, GPUs (such as NVIDIA® Blackwell series hardware), NVIDIA networking technologies, and the NVIDIA AI Enterprise software suite.

Importantly, the HPE AI Factory offering spans the full AI lifecycle—model building, training, fine-tuning, inference, and continuous monitoring—across the entire IT landscape. Plus, unlike bespoke, rigid platforms, HPE AI Factory at-scale is:

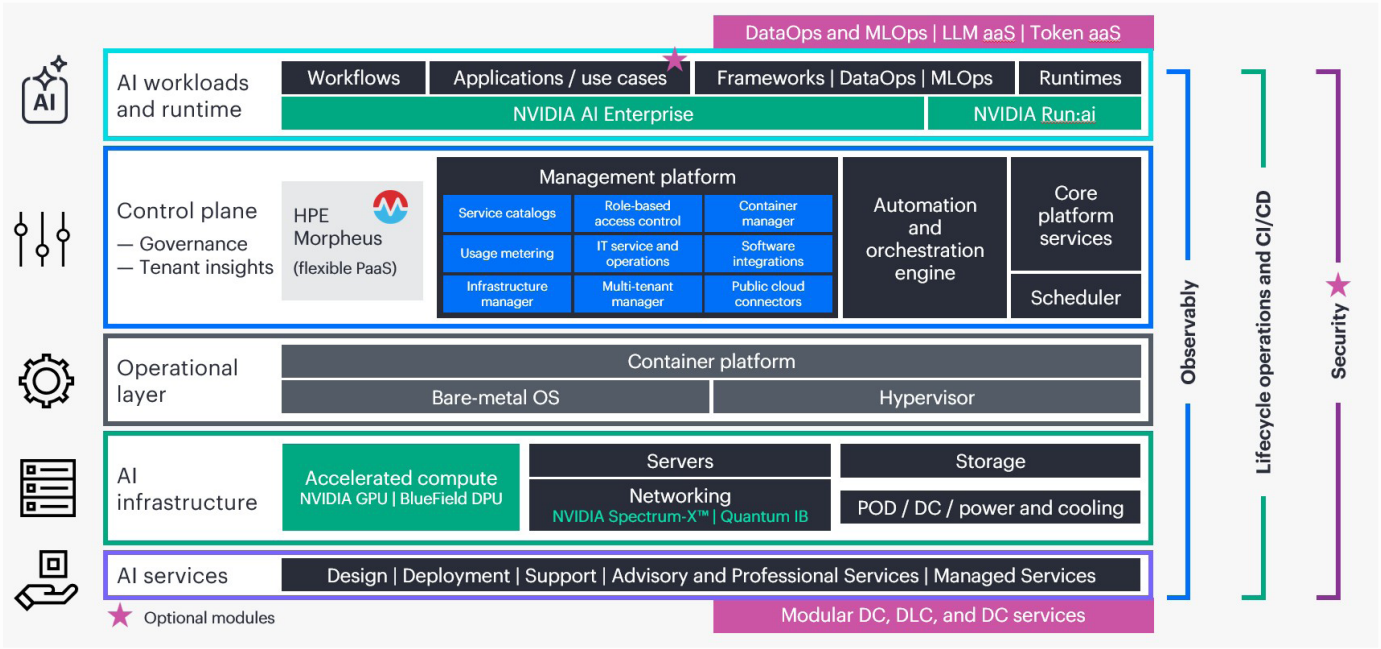
— **Customizable and composed** of standardized and validated hardware, software, and services building blocks
— **Designed for scale** with cluster performance guarantees and a multi-tenant GPU-cloud-like experience
— **Able to support** a variety of AI and high-performance computing (HPC) workloads including model pre-training, fine-tuning, inference, and mixed simulations
— **Ready for direct liquid cooling (DLC)**—to support the newest NVIDIA accelerators. Additionally, it offers performance-optimized data center (POD) support, enabling the fastest deployment to new locations

Once implemented, operationalizing AI at-scale becomes much easier due to centralized workflows, reusable assets, and a unified operating model environment. In addition, integrated orchestration, automation, and lifecycle management reduce manual effort, while built-in governance and security ensure AI workloads meet enterprise and regulatory requirements. This provides organizations seeking to operationalize AI at-scale with faster time to production, reduced operational overhead, and scalable AI deployments aligned with business goals.

2. Observability and control

As AI environments become distributed, multi-user visibility and centralized control are critical. HPE AI Factory at-scale provides a unified view across decentralized infrastructure, enabling teams to monitor performance, optimize workload placement, and manage GPU utilization efficiently.

HPE AI Factory at-scale provides centralized control and observability, including the ability to monitor cluster usage and analyze consumption, as well as dynamic GPU-instance provisioning along with automated resource monitoring and billing, It is the only factory that incorporates a control plane to both observe and manage the entire system.



HPE AI Factory at-scale provides:

— Centralized control of decentralized AI infrastructure

— Optimized workload placement, GPU usage, and multi-tenant management

— Built-in automation and observability with HPE Morpheus and HPE OpsRamp

— Flexibility across data centers, edge sites, and public clouds

— Integration with existing DevOps and IT tools

With built-in observability and automation, organizations can streamline operations and gain proactive insight into performance, cost, and reliability—reducing risk and improving outcomes.

3. Multi-tenancy

Multi-tenancy enables multiple teams, departments, or customers to share the same AI infrastructure securely. Workloads remain isolated, while shared GPU pools maximize utilization and reduce costs. Tenant-level visibility enables organizations to track usage, manage priorities, and allocate costs accurately.

HPE AI Factory at-scale is an integrated platform that provides a multi-tenant, GPU-cloud-like experience for various AI workloads. It offers complete multi-tenant enablement including GPU tenants, automation, and self-service for tenants with automated resource monitoring and billing, as well as provisioning of value-added services to individual tenants.

HPE AI Factory at-scale, co-engineered with NVIDIA:

— **Supports multi-tenancy**, shared GPU pools, and workload isolation, including hybrid AI factories

— **Meets the unique needs** of a specific client and enables multiple users and departments to share the same infrastructure while maintaining isolation and security

— **Optimizes performance and cost optimization** by identifying the specific needs and usage among individual tenants

This approach provides organizations with secure, scalable AI for multiple users as well as innovation at-scale without duplicating environments or compromising security. HPE AI Factory offers full control over sensitive data and the ability to meet product sovereignty compliance and solution compliance support. In addition, organizations can improve infrastructure efficiency as well as cost transparency and optimization.

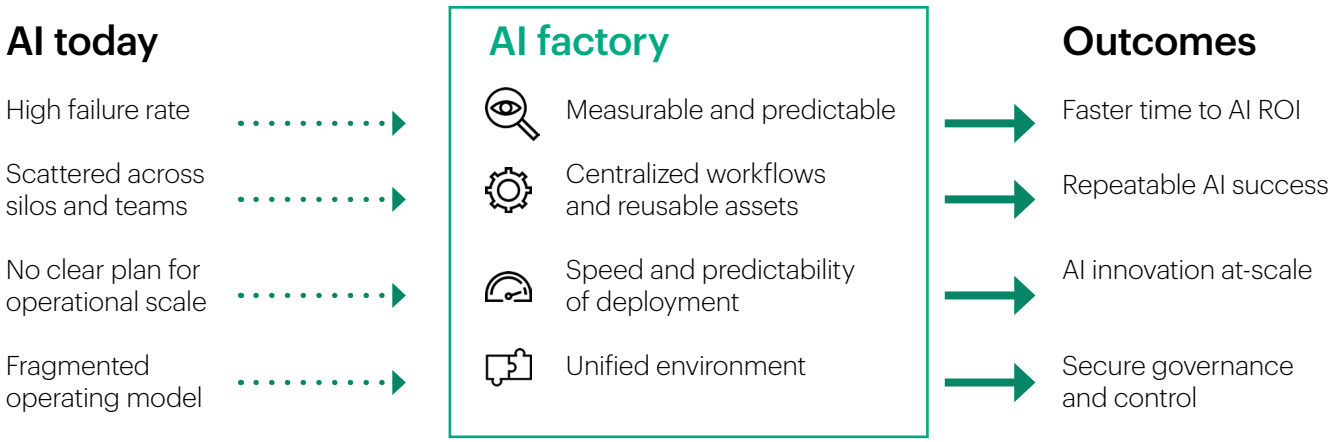
Achieve scalable AI operations across your organization

Scaling AI across an organization requires more than just adding infrastructure—it demands a cohesive strategy that simplifies complexity, optimizes resources, and delivers real business impact. Whether building large-scale AI factories or managing highly complex workloads, HPE AI Factory at-scale is designed to operationalize AI across its entire lifecycle.

Beyond technology, HPE brings deep expertise and comprehensive services to every engagement. Together, HPE and NVIDIA Services teams partner closely with organizations from initial strategy and planning through deployment and ongoing operations. Whether supporting solution design, financing, education, or long-term optimization, HPE and NVIDIA help accelerate implementation and ensure AI factories can evolve as business needs change.

AI factories provide a blueprint for AI success

From pilots to profits and purpose-driven outcomes



For model-building enterprises and AI service providers, multi-tenancy, observability, and control are essential to operationalizing AI at-scale. AI is no longer a single team's pilot project or experiment—it's a shared, business-critical capability that must support many users, workloads, and priorities simultaneously.

By combining operationalized AI with observability, control, and multi-tenancy, organizations can confidently scale AI across the enterprise. The result is a production-ready AI foundation that improves efficiency, reduces costs, strengthens governance, and accelerates innovation.

Learn more at

hpe.com/us/en/ai-factory

hpe.com/ai

Visit HPE.com

Take the next step

Ready to get started? Explore purchasing options or engage with HPE experts to determine the best solution for your business needs.

Chat now

a00156329ENW

HEWLETT PACKARD ENTERPRISE

hpe.com