AI FACTORY

BUILD AI INFRASTRUCTURE FOR PRODUCTION REALITY

OPTIMIZE PERFORMANCE, EFFICIENCY, AND ECONOMICS FOR PRODUCTION AI

AI infrastructure requirements change when applications and agents move into everyday business operations.

Inference becomes continuous. Agents generate repeated model interactions as they reason, act, and connect across workflows. More users and applications increase concurrency and token consumption. Data has to move quickly enough to keep accelerated compute productive. And IT has to manage performance, energy use, capacity, security, and cost as demand grows.

That makes production AI an infrastructure-wide challenge.

Lenovo Hybrid AI Factory with NVIDIA is Lenovo's systems approach to AI where enterprises can transform their businesses with their choice of AI model, with full control of their own data and through token efficient, validated Lenovo hybrid infrastructure with NVIDIA — compute, networking, storage, software, and expertise. The AI factory can be deployed locally from employee workstations, stores, factory floors to enterprise, and gigawatt cloud data centers.

Up to 8x lower cost per token

Lenovo Hybrid AI Factory with NVIDIA can deliver up to 8x lower cost per token versus cloud IaaS and up to 18x versus AI APIs, with break-even in under four months.[1]

Up to 40% energy savings compared with equivalent air-cooled systems

Lenovo Neptune® liquid cooling can help reduce the energy demands of high-density AI infrastructure.[3]

NVIDIA

Smarter technology for all

Lenovo

OPTIMIZE THE AI FACTORY AS A SYSTEM

EVERY LAYER AFFECTS THE PERFORMANCE AND ECONOMICS OF PRODUCTION AI

Accelerated compute drives AI performance, but GPUs depend on a larger system. Networking moves traffic, storage supplies data, applications provide context, software and orchestration allocate resources, and power and cooling determine how much capacity can run efficiently.

Lenovo Hybrid AI Factory with NVIDIA integrates these elements around production AI requirements, reducing integration complexity and helping IT operate them as one coordinated system.

WHAT MAKES AN AI FACTORY PRODUCTIVE?

Productive AI depends on how well every layer of the environment works together.

AI MODELS

Choose open or proprietary models, from specialized to large language models, based on performance, accuracy, privacy, and cost.

COMPUTE

Match **NVIDIA accelerated computing,** memory, and capacity to workloads with Lenovo ThinkSystem®, ThinkAgile®, and ThinkStation®.

NETWORKING

Use NVIDIA Spectrum-X™ Ethernet networking, purpose-built for AI factories, to reduce bottlenecks in multi-node training and high-concurrency inference and improve GPU utilization, throughput, and latency.

STORAGE AND DATA

Give models and agents fast access to enterprise data and context with Lenovo Data Storage solutions.

AI SOFTWARE

Use **NVIDIA AI Enterprise** and enterprise software options to support AI development, deployment, monitoring, and security for production operations.

MANAGEMENT AND ORCHESTRATION

Use NVIDIA Run:ai to allocate GPU resources, manage queues and priorities, and improve cluster utilization — increasing throughput and optimizing cost per token.

POWER AND COOLING

Support high-density accelerated infrastructure while managing energy and thermal demands with Lenovo Neptune® liquid cooling and air-cooling options.

SECURITY

Protect data and systems with Lenovo enterprise security capabilities, a Lenovo-owned global supply chain, and on-premises options for sensitive data.

Lenovo Hybrid AI Services can help plan, integrate, deploy, operate, and optimize the environment. Lenovo TruScale provides flexible consumption, while partner services can extend implementation and operational support.

MAKE EVERY TOKEN COUNT

As AI use expands, organizations need to consider token value and speed, compute utilization, energy use, and how those factors affect cost at scale.[2]

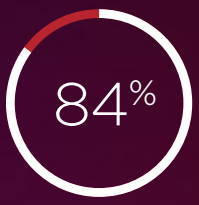
- **Value per token —** how effectively AI output contributes to the intended business outcome
- **Time to first token —** how quickly useful inference begins
- **Token throughput —** how much inference the environment can deliver
- **GPU utilization —** how effectively accelerated resources are put to work
- **Cost per token —** the infrastructure cost of producing inference at scale
- **Tokens per watt —** how efficiently infrastructure turns energy into AI output

Lenovo-validated infrastructure and NVIDIA Run:ai help optimize these measures as AI demand scales, maximizing AI value while optimizing the resources required to produce it.

BUILD AROUND THE WORKLOAD — AND THE WAY YOU OPERATE

MATCH INFRASTRUCTURE TO AI DEMAND WITHOUT CREATING UNNECESSARY COMPLEXITY

Enterprise AI requirements can range from focused inference and development to shared production AI, advanced model customization, and highly intensive training or large-scale inference.



of organizations expect to use on-premises and/or edge deployments as part of a hybrid AI environment.[4]

Lenovo Hybrid AI Factory with NVIDIA provides validated infrastructure options across workstation, edge, shared enterprise, and data center environments — designed for different workload demands and levels of scale. Organizations can align compute, networking, storage, software, power, and cooling to the workload, then add accelerated capacity as requirements grow.

For organizations that want a faster, simpler path to production AI, **Lenovo AI Express** provides a differentiated way to accelerate access to Lenovo Hybrid AI Factory. By streamlining the purchasing process and reducing acquisitioncomplexity and uncertainty, it helps organizations move more quickly from planning to deployable AI capacity — with a clear path to expand as requirements grow.

SMALL — FOCUSED AI INFERENCING

Support iterative workstation-based AI development, from prototyping and model customization to testing, validation, and dedicated inference — providing accelerated local performance before scaling to broader enterprise AI environments.

MEDIUM — HIGHER-THROUGHPUT INFERENCING & AGENTIC AI

Extend validated models, agents, and applications into shared infrastructure that can serve more users, teams, and workloads with greater accelerated capacity — helping support broader adoption while avoiding siloed AI investments.

LARGE — FULL-SCALE GENERATIVE AI MODELS

Run demanding AI workloads reliably at scale — including high-volume inference, large shared models, advanced customization, fine-tuning, training, and other intensive workloads — with the performance, networking, storage, power, and cooling those environments require.

Cloud resources can complement these environments when organizations need burst capacity, additional elasticity, or access to cloud-based AI services and frontier models.

VALIDATION CAN SHORTEN THE PATH TO OPERATIONS

AI infrastructure can span hardware, networking, operating systems, container platforms, workload management, AI software, storage, and lifecycle services. Lenovo validates configurations and interoperability across those layers to help reduce the effort required to design and integrate each environment independently.

Lenovo EveryScale can pre-integrate, cable, load, and test AI infrastructure at the rack level before delivery, helping minimize installation time in the data center.

KEEP SOFTWARE CHOICE OPEN

Lenovo Hybrid AI Factory with NVIDIA supports NVIDIA AI Enterprise, including NVIDIA Run:ai for GPU resource orchestration, alongside enterprise software options such as Red Hat and Canonical. This helps IT align AI infrastructure with existing skills, platforms, and operational strategies.

PRODUCTION AI NEEDS A CONNECTED FOUNDATION

INFRASTRUCTURE PERFORMANCE IS ONE PART OF THE ENTERPRISE AI EQUATION

An efficient AI factory still depends on the use cases it supports, the data it can reach, the controls surrounding it, and the expertise required to keep it operating as demand changes.

Lenovo Hybrid AI Advantage™ with NVIDIA connects those requirements in one enterprise AI approach — helping organizations move from AI opportunity to production with a foundation designed for performance, economics, control, and scale.



ACCELERATE AI INNOVATION — LENOVO AI LIBRARY

Use ready-to-customize agents and use cases to turn business opportunities into working AI.



OPTIMIZE AI INFRASTRUCTURE — LENOVO HYBRID AI FACTORY WITH NVIDIA

Bring together validated compute, networking, storage, software, orchestration, and services to operate AI with better performance and economics.



SECURE AND TRUST ENTERPRISE AI —SECURITY, CONTROL, AND PROTECTION

Protect data and AI assets while supporting governance, compliance, sovereignty, and control across the AI lifecycle.

TURN ACCELERATED INFRASTRUCTURE INTO PRODUCTIVE AI CAPACITY

Lenovo and NVIDIA combine validated enterprise AI infrastructure, accelerated computing and AI software, advanced cooling, open software choice, and lifecycle expertise to help organizations deploy AI with less complexity and continuously improve performance and economics as workloads grow. Lenovo is also recognized as a Leader in AI-ready Infrastructure Solutions,[5] reinforcing the enterprise foundation behind Lenovo Hybrid AI Factory with NVIDIA.

READY TO BUILD FOR PRODUCTION AI?

Connect with Lenovo to align workloads, infrastructure, software, data, and services to your AI priorities.

Book an AI Discover workshop today

Explore Lenovo's hybrid AI infrastructure solutions

Sources

1 Lenovo, "On-Premise vs Cloud: Generative AI Total Cost of Ownership (2026 Edition)," Lenovo Press, July 2026
2 NVIDIA, "Rethinking AI TCO: Why Cost per Token Is the Only Metric That Matters," NVIDIA Blog, April 2026
3 Lenovo, "Lenovo ThinkSystem SC750 V4 Neptune Datasheet," Lenovo Press, June 2026
4 Lenovo, "CIO Playbook 2026: The Race for Enterprise AI," Lenovo, January 2026
5 Information Services Group, "ISG Provider Lens®: AI-ready Infrastructure Solutions," Information Services Group, July 2026

© Lenovo 2026. All rights reserved. v1.00 September 2026.

NVIDIA

Smarter technology for all

Lenovo