LENOVO HYBRID AI ADVANTAGE™ WITH NVIDIA

Smarter technology for all

Lenovo

IGNITE YOUR ENTERPRISE INNOVATION

BUILD A CONNECTED AI FOUNDATION FROM WORKSTATION TO EDGE TO DATA CENTER

NVIDIA

ENTERPRISE AI IS READY FOR WHAT COMES NEXT

From AI investment to measurable impact 3

AI demand driving new infrastructure decisions 4

Put the right workload on the right compute 5

Connected systems for enterprise AI 6

Bring AI closer to people and data 7

Move faster from use case to working AI 8

Scale inference with better economics 9

Right-size your enterprise data center 10

Turn distributed data into trusted AI 11

Move faster with Lenovo AI Express 12

Keep software choice open 13

Build trust into every layer 14

Extend your AI capabilities with expertise 15

Hover over this icon throughout the eBook to learn more.



FROM AI INVESTMENT TO MEASURABLE IMPACT

AI IS NOW A CORE ENGINE FOR BUSINESS REINVENTION

AI is increasingly embedded across products, workflows, decisions, and customer experiences. The focus now is running AI efficiently in production: optimizing token economics, simplifying operations, and delivering value faster as adoption grows.

Enhancing, innovating, or reinventing the business with AI is the #1 global business priority for 2026.[1]

#1

96%

of organizations plan to increase AI investment over the next 12 months.[1]

THE OPPORTUNITY IS GETTING BIGGER

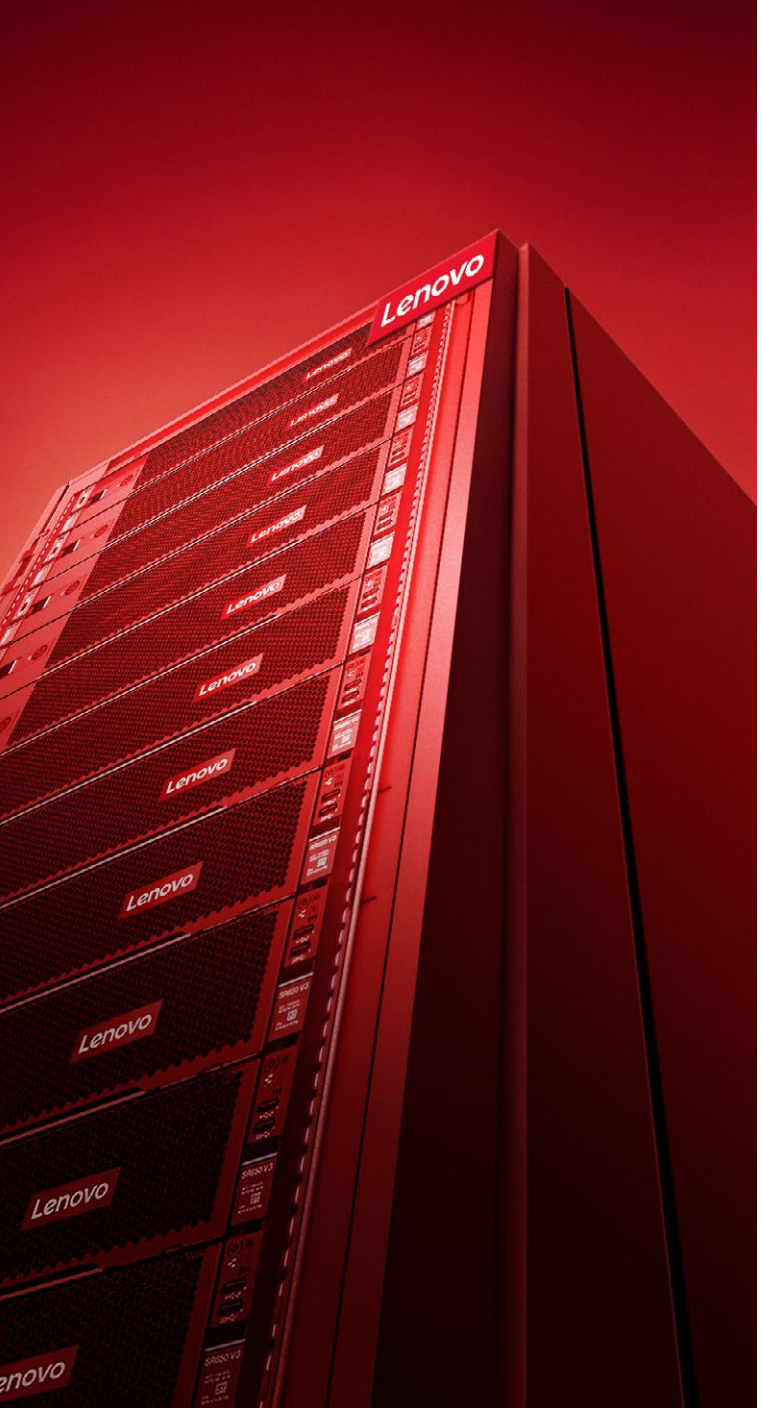
AI adoption is expanding across the business. Developers are building and customizing AI, while professionals are putting AI-enabled applications and agents to work across increasingly complex workflows. Customer service, marketing, operations, finance, and sales are all expected to see significant growth in AI implementation.[1]

At the same time, AI itself is evolving. AI agents are beginning to reason across multiple steps, interact with applications, and execute work with greater autonomy. AI for physical operations is creating new opportunities to connect intelligence with real-world environments. As these capabilities expand, organizations need proven practices, enterprise-ready AI infrastructure, trusted data, and a strong technology ecosystem to identify the highest-value use cases, deploy them efficiently, and manage them responsibly at scale.

For line-of-business leaders, decision-makers, and IT, the question becomes:

How do you deploy AI faster, operate more efficiently, improve token economics, and maintain the trust and control enterprise AI requires?





AI DEMAND DRIVING NEW INFRASTRUCTURE DECISIONS

Every new user, application, and AI agent adds demand for inference. Workloads can vary dramatically in model size, data sensitivity, latency requirements, utilization, and business criticality. And AI agents can generate repeated model calls as they reason, act, and interact across complex workflows. As those interactions multiply, token economics becomes part of the infrastructure equation. IT has to consider not only the value of each token, but how efficiently the environment can deliver scalable inference as demand grows.

That makes infrastructure decisions inseparable from AI strategy.



of organizations will leverage on-premises or edge deployments for AI workloads and applications as part of a hybrid deployment environment.[1]

MORE AI. MORE PLACES TO RUN IT.

Responsive, sensitive, or always-on workflows can benefit from dedicated compute close to users, applications, and data. Other workloads need shared capacity that can serve more users, teams, and applications. Cloud resources can complement those environments when organizations need burst capacity or additional elasticity.

The question is no longer simply: **Where does AI run?**

It is: **Where should each AI workload run?**

Performance, responsiveness, security, data control, capacity, and economics all factor into the answer. As requirements evolve, organizations need a consistent architecture that allows AI models and applications to scale from workstations to Lenovo Hybrid AI Factory with NVIDIA without requiring a new approach at every stage.

PUT THE RIGHT WORKLOAD ON THE RIGHT COMPUTE

ENTERPRISE AI HAS NO LIMITS — RUN IT ANYWHERE

The right environment depends on what the workload needs to accomplish.

AI AT WORKSPACES

Best when AI needs to run locally where work happens.

AI AT EDGE/BUSINESS LOCATIONS

Best when AI must operate where data is generated and decisions are made.

AI AT DISTRIBUTED DATA CENTERS

Best when workloads need broader access.

AI AT LARGE CENTRALIZED/ HPC DATA CENTERS

Best when demand and complexity grow.

CLOUD

Best when elasticity is the priority.

A hybrid AI approach gives IT the flexibility to place each workload where it delivers the optimal balance of performance, control, and economics.

CONNECTED SYSTEMS FOR ENTERPRISE AI

Built on more than 20 years of collaboration, Lenovo and NVIDIA help organizations accelerate AI adoption and innovate responsibly with enterprise-ready solutions spanning workstation, edge, and data center environments. Together, Lenovo AI Library solutions, validated Lenovo Hybrid AI Factory with NVIDIA reference architectures, end-to-end services, and a broad partner ecosystem provide a proven foundation for AI at scale.

Lenovo connects these capabilities across the hybrid AI environment to create a more integrated path from development to production and scale.

That integrated approach helps organizations move from AI ideas and individual workflows to production AI at enterprise scale — improving deployment speed, performance, economics, and control across environments.

ACCELERATE AI VALUE WITH PROVEN USE CASES AND PRE-BUILT AGENTS

LENOVO AI LIBRARY

Cut the complexity. Start with ready-to-customize AI use case templates built on proven architectures, reusable assets, and Lenovo + NVIDIA technologies — helping teams move faster from pilot to deployment in production.

SCALE AI INFRASTRUCTURE WITH BETTER COST, PERFORMANCE, AND ENERGY EFFICIENCY

LENOVO HYBRID AI FACTORY WITH NVIDIA

Build and operate validated AI infrastructure from workstation to edge to data center, combining compute, networking, storage, software, open source models, enterprise applications, and services to improve inference performance, increase GPU utilization, and optimize AI economics.

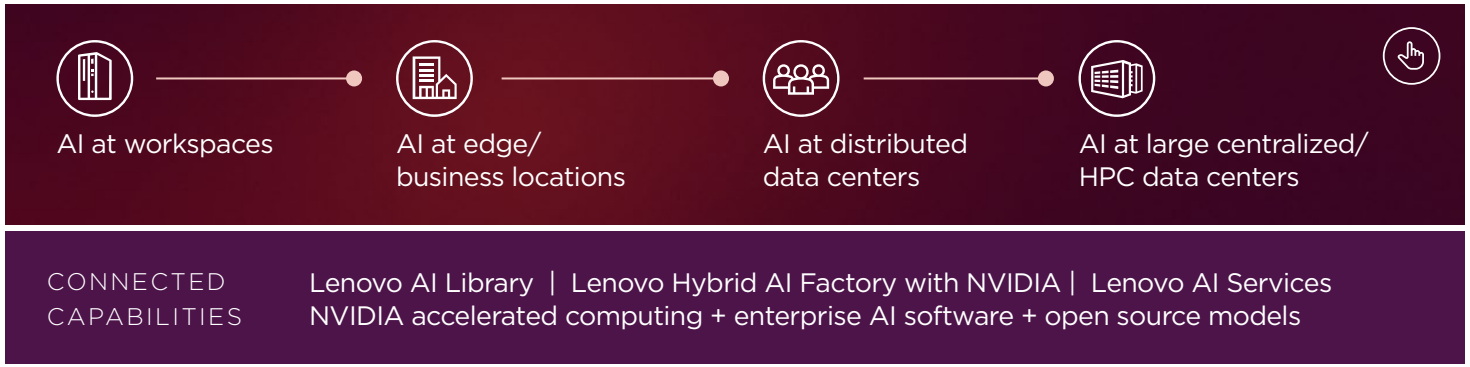
HARNESS AND MANAGE YOUR DATA EVERYWHERE WITH TRUST

LENOVO AI SERVICES + ECOSYSTEM

Create AI-ready data pipelines, modernize storage, and support low-latency retrieval while addressing security, governance, sovereignty, and compliance across the AI lifecycle.

FROM AI-ENABLED WORK TO ENTERPRISE-SCALE AI

This connected approach gives organizations greater flexibility to put each workload where it makes the most sense — while maintaining a more consistent strategy for performance, control, and scale.



BRING AI CLOSER TO PEOPLE AND DATA

AI innovation often begins close to the people developing, testing, and using it.

Lenovo Workstations with NVIDIA accelerated computing provide high-performance local environments for AI development, experimentation, inference, and demanding AI-enabled professional workflows.

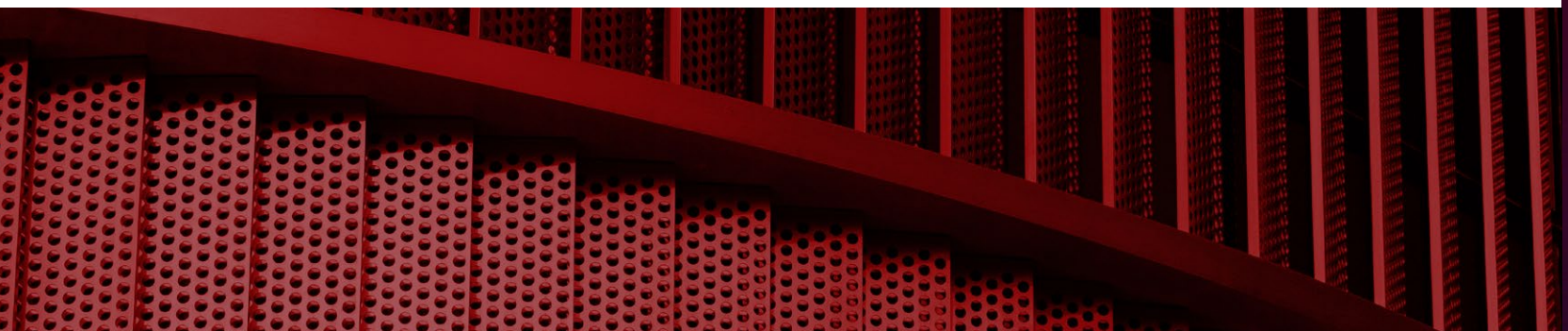
That gives teams another option beyond relying exclusively on shared cloud or data center GPU resources.

Lenovo ThinkStation® PGX, powered by the NVIDIA GB10 Grace Blackwell Superchip, provides a purpose-built AI environment for AI developers to build, test, and run AI workloads locally.

ACCELERATE AI DEPLOYMENT WITH NVIDIA NIM™ MICROSERVICES

Pre-integrated NVIDIA NIM™ microservices on Lenovo ThinkStation systems powered by NVIDIA RTX PRO™ 6000 Blackwell Workstation Edition GPUs and NVIDIA RTX PRO™ 4500 Blackwell Workstation Edition GPUs give developers a faster path to deploying optimized AI models locally — helping reduce the time and complexity of building AI capabilities from scratch.

NVIDIA NIM microservices provide optimized, containerized inference capabilities that can help teams move models into applications faster. NVIDIA AI Blueprints complement them with proven recipes and workflow guidance, supporting a more consistent path from workstation-based experimentation to deployment on Lenovo Hybrid AI Factory validated infrastructure.



THE ADVANTAGE OF AI WORKLOADS ON WORKSTATIONS

Keeping the right workloads closer to users and data can help:

- **Accelerate iteration —** Reduce wait time for shared resources.
- **Improve responsiveness —** Support low-latency local inference.
- **Maintain greater control —** Keep sensitive data, models, and intellectual property closer to the organization.
- **Reduce unnecessary data movement —** Process data where it is created or used when appropriate.
- **Use shared resources more efficiently —** Run appropriate development and inference workloads locally, reserving centralized accelerated capacity for workloads that require greater scale.

And local AI on workstations does not end when development does.

Once solutions move into production, workstations can continue supporting **AI-driven workflows** for engineers, designers, data scientists, creators, and other professionals.

MOVE FASTER FROM AI VISION TO BUSINESS OUTCOMES

A MORE REPEATABLE PATH TO AI VALUE

The Lenovo AI Library helps organizations accelerate AI value with ready-to-customize agents and use case templates rather than building each new solution from the ground up.

Lenovo AI Library use cases combine validated architectures, reusable assets, and technologies and services from Lenovo, NVIDIA, and select Lenovo AI Innovator software partners to create a more repeatable path from business need to deployment. **NVIDIA Blueprints** provide proven recipes and workflow guidance to accelerate agent development, while **NVIDIA NIM™ microservices** help simplify and speed model deployment.

Each Lenovo AI Library use case is validated with governance controls for data access, security, compliance, and lifecycle management — helping reduce the complexity and risk of moving AI into production.

For agentic AI use cases, the **Lenovo xIQ Agent Platform** provides tools to create, deploy, monitor, and manage enterprise-ready AI agents with built-in governance and hybrid deployment flexibility.

PUT AI TO WORK ACROSS THE ENTERPRISE

Lenovo AI Library use cases span functions and industries, helping organizations address opportunities such as:

- **Lenovo Knowledge Super Agent**
- **Lenovo AI Computer Vision**
- **Lenovo Edge AI Orchestration**
- **Lenovo Super Agent for Retail**
- **Lenovo AI Robotic Inspection**

Organizations can also draw on Lenovo-validated AI Innovators and NVIDIA ecosystem partners for specialized technologies and expertise that can extend AI deployments.

Up to 24x faster to production

Across multiple deployments, organizations reached production up to 24x faster than with custom-built approaches. Lenovo AI Library solutions can support production-ready agentic AI deployment in as little as one week.[2]

THE PATH FROM IDEA TO DEPLOYMENT

1. **IDENTIFY**
   Choose a business problem with measurable value.
2. **CUSTOMIZE**
   Adapt the agent, model, or workflow to business requirements.
3. **VALIDATE**
   Test performance, data access, and governance.
4. **DEPLOY**
   Place the workload on the right compute.
5. **OPTIMIZE**
   Manage performance, economics, and adoption over time.

SCALE INFERENCE WITH BETTER ECONOMICS — FROM THE WORKSTATION TO ENTERPRISE INFRASTRUCTURE

MAKE EVERY TOKEN COUNT

As AI use expands across workstations and enterprise infrastructure, the infrastructure equation extends beyond token economics alone. Organizations need to consider the value and cost of each token, how efficiently and securely the environment can deliver scalable inference, and how enterprise data remains protected and controlled as usage grows.

AI agents raise the stakes. A single business task can trigger multiple model interactions as an agent reasons, acts, and interacts with applications — multiplying token consumption and increasing the importance of security and data control across everyday workflows.

Improving token economics means optimizing and right-sizing the environment that produces those tokens — including deciding which workloads should run locally and which should scale through shared infrastructure — not simply finding the lowest model price.

WHAT DRIVES TOKEN ECONOMICS?

Token economics depends on more than model pricing. It reflects how effectively a right-sized AI environment turns compute, energy, time, and infrastructure investment into useful inference.

- **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.

FULL AI STACK AS A SYSTEM

Lenovo Hybrid AI Factory with NVIDIA brings together the infrastructure elements that determine how effectively and securely AI operates:

- **Compute**
- **Networking**
- **Storage and data**
- **Software and models**
- **Management and orchestration**
- **Power and cooling**

Across these layers, Lenovo AI Services can help organizations plan, integrate, deploy, operate, and optimize the environment as requirements evolve.

PROVEN EFFICIENCY FOR PRODUCTION AI

The goal is not simply more compute. It is **more productive compute.**

Up to 8x lower cost per token[3]

Up to 40% energy savings compared with equivalent air-cooled systems[4]

RIGHT-SIZE YOUR ENTERPRISE DATA CENTER

CHOOSE VALIDATED INFRASTRUCTURE BASED ON WORKLOAD DEMANDS, DEPLOYMENT PRIORITIES, AND YOUR EXISTING IT ENVIRONMENT

Lenovo Hybrid AI Factory with NVIDIA provides NVIDIA-certified systems built on Lenovo ThinkSystem accelerated servers, with Lenovo Neptune® liquid and air cooling, plus Lenovo ThinkAgile hyperconverged infrastructure (HCI) solutions with partner options. Together, these options help organizations match infrastructure to AI requirements, accelerate time to value, and retain flexibility as model and workload needs evolve — without designing every environment from scratch.



AI AT EDGE/BUSINESS LOCATIONS

More compact enterprise configurations can provide a dedicated AI factory for focused, smaller size inference and AI development requirements and pilots.

TYPICAL FIT

- Focused production inference
- Departmental AI
- Smaller shared workloads
- AI development and testing
- Organizations establishing dedicated AI capacity



AI AT DISTRIBUTED DATA CENTERS

As AI serves more users, agents, and applications, organizations can scale up the infrastructure designed for greater AI factory capacity and inference.

TYPICAL FIT

- Larger production inference environments
- Shared enterprise AI
- Higher concurrent demand
- More complex agents and applications
- Greater performance and networking requirements



AI AT LARGE CENTRALIZED/ HPC DATA CENTERS

Larger scale-out configurations support organizations with the most intensive AI factory, including more advanced model and enterprise-scale workloads.

TYPICAL FIT

- Large-scale inference
- Advanced model customization
- Training and fine-tuning
- High-performance AI environments
- More intensive networking requirements

TURN DISTRIBUTED DATA INTO TRUSTED AI

AI IS ONLY AS USEFUL AS THE DATA IT CAN REACH

Data may live across devices, workstations, enterprise storage, data centers, and cloud environments — and each workload can have different requirements for access, latency, security, and sovereignty.

As AI grows, data architecture becomes part of AI architecture to support:

PERFORMANCE

AI agents and inference depend on rapid access to context, enterprise knowledge, and model data. Storage and data architecture can directly affect responsiveness.

CONTROL

Keeping selected data and workloads within enterprise-controlled environments can help organizations maintain greater control over sensitive information and intellectual property.

SOVEREIGNTY

Workload placement and data architecture can help support requirements for where data is stored and processed.

SCALE

As AI serves more users, models, and applications, data infrastructure must support more users, models, and interactions without becoming a bottleneck.

MAKE DATA AI-READY

Lenovo enterprise storage, data management solutions, and services help organizations create AI-ready data foundations that support:

DATA PIPELINES

Move and prepare data for AI applications.

MODERN STORAGE

Support growing AI data requirements.

LOW-LATENCY RETRIEVAL

Give models and agents faster access to relevant information.

DISTRIBUTED DATA

Bring compute closer to data where appropriate rather than moving everything to one location.

MOVE FASTER WITH LENOVO AI EXPRESS

Lenovo AI Express with NVIDIA helps you accelerate your path to the Lenovo Hybrid AI Factory with NVIDIA in weeks with an order-to-ship goal of 15 business days for eligible configurations.*

REDUCE RISK AND PROTECT YOUR TIME

Validated, ready-to-ship AI factory configurations in three easy quick-start sizes combine Lenovo's validated systems with NVIDIA accelerated computing, networking and AI software sized to your workloads.

MOVE FAST FROM DECISION TO DEPLOYMENT

Curated services give you a stronger starting point and greater confidence for day two and beyond.

SCALE WITHOUT STARTING OVER

Lenovo Hybrid AI Advantage™ with NVIDIA helps you build on what's already working, to turn early AI momentum into broader business impact.

SMALL	MEDIUM	LARGE
FOCUSED AI INFERENCING	HIGHER-THROUGHPUT INFERENCING & AGENTIC AI	FULL-SCALE GENERATIVE AI MODELS
		
NVIDIA AI ENTERPRISE SOFTWARE		
Lenovo ThinkSystem SR650a V4 powered by 2x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs	**Lenovo ThinkSystem SR675 V3** powered by 8x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs	**Lenovo ThinkSystem SR680 V4** powered by NVIDIA HGX B300
NVIDIA ConnectX-7, NVIDIA Bluefield-3, and NVIDIA Spectrum-X Ethernet networking	NVIDIA ConnectX-8, NVIDIA BlueField-3, and NVIDIA Spectrum-X Ethernet networking	NVIDIA ConnectX-8, NVIDIA BlueField-3, and NVIDIA Spectrum-X Ethernet infrastructure
15 business days order to ship goal*	**20 business days** order to ship goal*	**25 business days** order to ship goal*

* For eligible Lenovo AI Express configurations and selected parts in non-restrictive markets. Clock starts after completion of order validation, payment/credit clearance, End User Certification (EUC), and applicable due diligence/export approvals.

KEEP SOFTWARE CHOICE OPEN

Red Hat | Canonical | SDI partners

Lenovo Hybrid AI Factory combines the NVIDIA AI Enterprise software platform with a choice of enterprise platform and orchestration options. Organizations can select Red Hat OpenShift, Canonical Ubuntu and Kubernetes, Nutanix, VMware, Microsoft, or other validated SDI partner solutions based on their existing IT environment, skills, and operational requirements.

LENOVO HYBRID AI FACTORY WITH NVIDIA

Flexible software stack. Open choice. Enterprise ready.

AI APPLICATIONS AND WORKLOADS	Agents · RAG · Inferencing · Model customization · Enterprise AI workloads
AI SOFTWARE FOUNDATION	NVIDIA AI Enterprise · NIM Microservices · Frameworks & Tools · Optimized Models
PLATFORM & ORCHESTRATION OPTIONS	Red Hat OpenShift · Canonical Ubuntu & Kubernetes · NUTANIX · vmware by Broadcom · Microsoft · Other validated options
MANAGEMENT & OPERATIONS	Infrastructure management · Orchestration · Monitoring · Security · Lifecycle services
LENOVO HYBRID AI INFRASTRUCTURE	Compute & GPUs · Networking · Storage · Cooling

BUILD TRUST INTO EVERY LAYER

SUCCESSFUL AI AT SCALE REQUIRES MORE THAN PERFORMANCE GAINS

Organizations also need consistent ways to protect data, manage access, govern AI behavior, and meet changing regulatory requirements across every environment where AI operates.



Only 27% of organizations have comprehensive AI governance frameworks today.[1]

TRUST HAS TO TRAVEL WITH THE WORKLOAD

Whether AI runs on a workstation, in shared enterprise infrastructure, or in the cloud, organizations need to address:

SECURITY

Protect AI infrastructure, applications, and data.

PRIVACY

Control access to sensitive information.

GOVERNANCE

Establish policies for how AI systems, data, and users interact.

COMPLIANCE

Support changing regulatory and industry requirements.

PROTECTION

Maintain appropriate control over where data and workloads reside.

LIFECYCLE MANAGEMENT

Continue monitoring and managing AI after deployment.

WHY LENOVO AND NVIDIA

TRUSTED INFRASTRUCTURE

Lenovo is recognized as a Leader in the Integrated AI Infrastructure Systems quadrant of the 2026 ISG Provider Lens® global AI-ready Infrastructure Solutions report.[5] Lenovo also delivers market-leading reliability — #1 in x86 server reliability for 12 consecutive years.[6]

PROTECTED ACROSS THE ENVIRONMENT

Lenovo ThinkShield® platform helps protect people, devices, and data across the AI environment.

TRUSTED SUPPLY CHAIN

Lenovo's global supply chain is ranked #5 in Gartner's 2026 Global Supply Chain Top 25[7] and spans PC through data center infrastructure, helping provide greater visibility, consistency, and control across the AI lifecycle.

PROTECTED INFERENCE

NVIDIA Confidential Computing creates a hardware-enforced boundary for inference, protecting model weights, prompts, and enterprise data across CPU and GPU memory — even from infrastructure operators. On supported GPUs, remote attestation verifies data remains private.

GOVERNED AI

Lenovo AI Services help organizations address privacy, governance, and compliance controls across data, infrastructure, models, and users.

AI should move faster without leaving trust behind.

EXPERTISE THAT EXTENDS YOUR AI CAPABILITIES

ENTERPRISE AI BRINGS TOGETHER INFRASTRUCTURE, DATA, SOFTWARE, MODELS, APPLICATIONS, GOVERNANCE, AND ORGANIZATIONAL CHANGE

Lenovo AI Services provides expertise across the AI lifecycle to help organizations move from strategy to implementation and ongoing operations.

ADVISORY

Identify high-value use cases, assess readiness, and define a practical AI roadmap.

IMPLEMENTATION

Design, integrate, and deploy AI environments based on business and technical requirements.

MANAGED SERVICES

Operate, monitor, and optimize AI infrastructure as workloads and requirements evolve.

AI ADOPTION AND CHANGE MANAGEMENT

Help prepare employees and processes for new ways of working while supporting responsible AI practices and governance.

FLEXIBLE CONSUMPTION

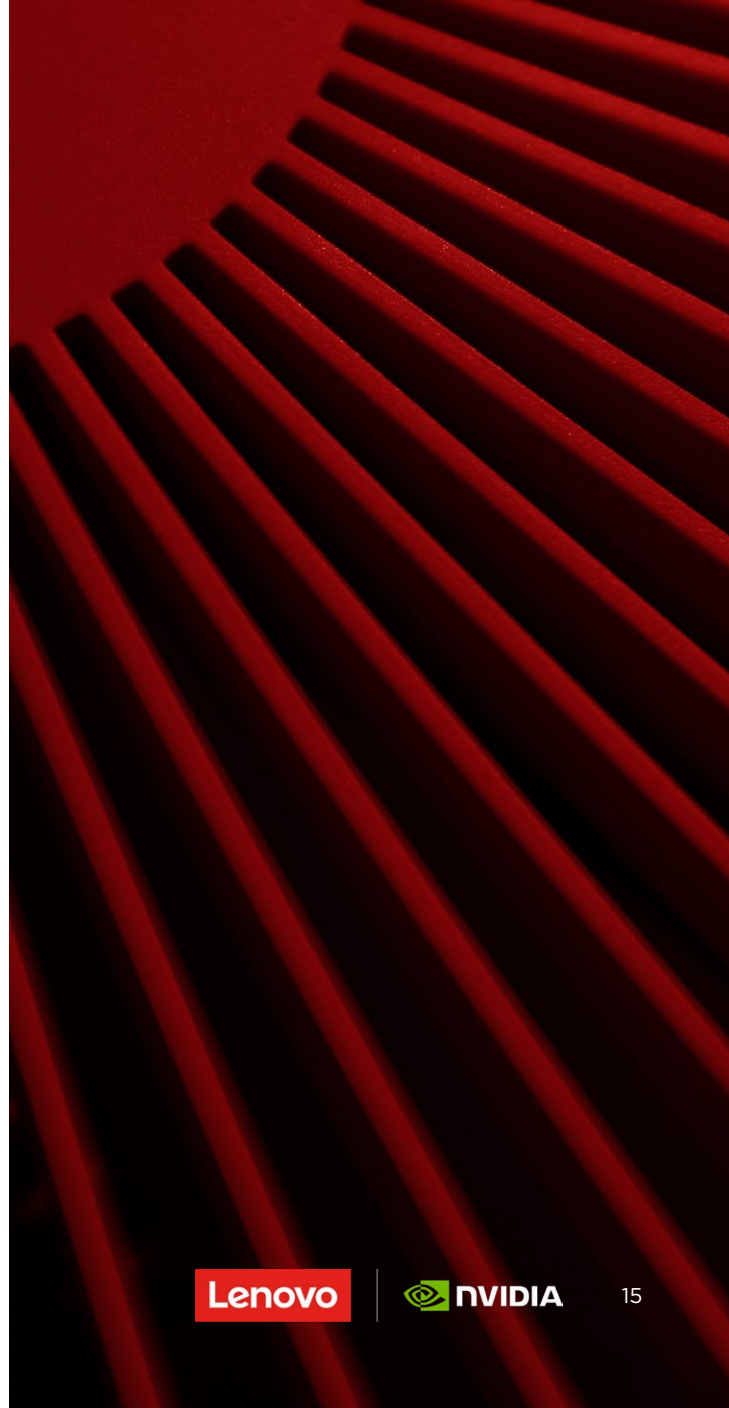
Lenovo TruScale provides flexible consumption options that can help shift AI infrastructure from large, unpredictable capital investments toward more predictable operating costs as requirements change.

LENOVO AND NVIDIA: A CONNECTED PATH FORWARD

Built on more than 20 years of collaboration, Lenovo and NVIDIA combine workstations, enterprise infrastructure, data and storage capabilities, software, services, and a global partner ecosystem with NVIDIA accelerated computing and AI technologies.

Together, Lenovo and NVIDIA help organizations:

- Start with proven use cases
- Place AI where it works best
- Scale inference more efficiently
- Protect data and maintain control
- Adapt as AI requirements change



IGNITE YOUR ENTERPRISE INNOVATION

AI will not live in one application, one infrastructure platform, or one location. It will move across people, workflows, and environments.

Lenovo Hybrid AI Advantage™ with NVIDIA provides connected systems for that evolution — from AI development and AI-driven work to shared inference and enterprise-scale infrastructure.

Move proven AI use cases into production faster.
Put the right workload on the right compute.
Scale performance, economics, and trust together.

READY TO BUILD YOUR ENTERPRISE AI ROADMAP?

Connect with Lenovo to determine where your AI workloads should run and how to scale them with the right balance of performance, economics, and control.

Book an AI Discover workshop today



Smarter technology for all

Lenovo

Sources

1 Lenovo, "CIO Playbook 2026: The Race for Enterprise AI," January 2026
2 Lenovo, "Lenovo Enables One-Week Deployment of Production-Ready Agentic AI to Transform Enterprise Workflows," Lenovo StoryHub, May 2026
3 Lenovo, "On-Premise vs. Cloud: Generative AI Total Cost of Ownership (2026 Edition)," July 2026
4 Lenovo, "ThinkSystem SC750 V4 Neptune Datasheet," June 2026
5 Information Services Group, "ISG Provider Lens®: AI-ready Infrastructure Solutions," July 2026
6 ITIC, "ITIC 2025 Global Server Hardware, Server OS Reliability Report," December 2025
7 Gartner, "Global Supply Chain Top 25," June 2026

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

NVIDIA