Lenovo Hybrid AI Factory with NVIDIA

# LENOVO HYBRID AI FACTORY FOR AI ORCHESTRATION



## SCALE AI EFFICIENTLY WITH SHARED, OPTIMIZED INFRASTRUCTURE.

Turn your GPU infrastructure into one shared AI platform. Lenovo Hybrid AI Factory with NVIDIA® uses NVIDIA Run:ai to pool, share, and dynamically allocate GPU capacity across training, fine-tuning, and inference, so you get more from your GPUs as AI scales.

### HIGHER GPU UTILIZATION

Share GPUs across teams and workloads to keep capacity working.

### FLEXIBLE ISOLATION

Fractional sharing or hardware-isolated MIG, at matching throughput.[1]

### FASTER TIME TO INFERENCE

Go from model to live endpoint in a few guided steps.

## WHY AI FACTORY ORCHESTRATION?

### KEY CHALLENGES

GPU capacity has emerged as one of the most valuable and in-demand resources in the data center, giving organizations that optimize it a decisive competitive advantage. Demand keeps growing as more teams, models, and agentic workloads compete for the same GPUs. Keeping up means getting more from every GPU, instead of leaving capacity idle, reserved by one team, or unsafe to share.

AI practitioners wait for access, IT teams struggle to allocate capacity across projects, and leaders lack visibility into usage and cost. Enterprises need to share GPUs safely, without slowing teams down or compromising isolation where it matters.

AI inferencing is the **#1** AI workload enterprises want to optimize for better economics.[2]

### SOLUTION OVERVIEW

Lenovo Hybrid AI Factory with NVIDIA combines accelerated computing, networking, storage, AI software, models, and services in Lenovo validated configurations. NVIDIA Run:ai turns it into one shared GPU pool for every team.

The solution is built on AI servers validated by Lenovo to run NVIDIA Run:ai and NVIDIA AI Enterprise, backed by Lenovo services for deployment. GPUs are allocated on demand by policy and priority, raising utilization and improving cost per token, from a single server to multi-node clusters.

Data scientists launch jobs on demand, developers deploy live endpoints in a few steps, and IT sets team quotas with visibility across every cluster. Fractional GPU sharing starts and stops jobs without slowing down other workloads on the same GPU, while NVIDIA Multi-Instance GPU (MIG) adds hardware isolation at matching throughput.[1] You can choose isolation based on your security needs, not performance trade-offs.

# ORCHESTRATE EVERY AI WORKLOAD ON SHARED GPUS.

Lenovo gives you a practical, cost-effective path from shared GPUs to an enterprise AI factory. Start with Lenovo AI servers running NVIDIA Run:ai and NVIDIA AI Enterprise, then scale on the same Lenovo validated foundation.

## Lenovo Hybrid AI Factory with NVIDIA Run:ai

**AI WORKLOADS**

AGENTIC AI | DEVELOPMENT + EXPERIMENTS | MODEL TRAINING | MODEL SERVING

**NVIDIA AI Enterprise**

**MANAGE AND GOVERN**

**NVIDIA Run:ai control plane** One console for connected clusters

| Submit workloads | Pool and govern resources | Monitor cluster activity |
|---|---|---|
| Self-service access for AI teams | Projects, quotas, policies and priorities | Resource use, health and analytics |

Kubernetes-hosted control plane | Integrates with existing development and MLOps tools

HTTPS MANAGEMENT CONNECTIONS

**RUN AND SCALE**

**START AND DEPLOY** Enterprise inference and fine-tuning

Run:ai cluster: Priority scheduling | GPU allocation | Workload scaling

kubernetes | Ubuntu | Red Hat — Kubernetes with NVIDIA GPU Operator

**LENOVO AI SERVERS**



Lenovo ThinkSystem SR675 V3,
Up to 8 NVIDIA RTX PRO™ 6000 Blackwell Server Edition or NVIDIA H200 NVL GPUs

**SCALE** Large-scale training and inference

Run:ai cluster: Priority scheduling | GPU allocation | Workload scaling

kubernetes | Ubuntu | Red Hat — Kubernetes with NVIDIA GPU Operator

**LENOVO AI SERVERS**



Lenovo ThinkSystem SR680a V3,
NVIDIA HGX™ H200, 8 GPUs



Lenovo ThinkSystem SR780a V3,
NVIDIA HGX™ H200, 8 GPUs

Add GPU capacity or connect more clusters as needs grow, all managed from one control plane.

**ALLOCATE GPU CAPACITY**

GPU sharing at either scale

| Fractional GPU sharing (SOFTWARE) | MIG partitioning (HARDWARE) |
|---|---|
| Flexible memory allocations; GPU compute is shared. Fast workload start and stop without restarting neighbors. | Dedicated compute and memory per instance. For hardware isolation; plan maintenance for repartitioning. |

**AI FACTORY FOUNDATION**

| NETWORKING | STORAGE OPTIONS | VALIDATION, DELIVERY AND SERVICES |
|---|---|---|
| NVIDIA Spectrum-X Ethernet<br>NVIDIA BlueField-3 SuperNICs | Lenovo DM & DG Series<br>NVIDIA-Certified storage partners | Lenovo EveryScale and AI COE<br>Rack integration or standalone delivery<br>Advisory, deployment and managed services |

**VISIBILITY**

- Run:ai
- Prometheus + Grafana
- MLflow

## WHY LENOVO AND NVIDIA

More than 20 years of partnership, engineered into one Lenovo validated stack.

**674**
MLPerf™ world records across AI training and inference[3]

**12**
consecutive years ranked number one for server reliability[4]

**#1**
most secure and toughest to crack, ITIC 2025 server security survey[5]

**#5**
Gartner Supply Chain Top 25 for 2026, a Lenovo-owned supply chain[6]

**Leader**
in the Integrated AI Infrastructure Systems quadrant[7]

Lenovo | NVIDIA

Learn more — Lenovo Infrastructure with NVIDIA Run:ai Article

Learn more — Lenovo ThinkSystem servers with NVIDIA Run:ai Solution Brief

1 Lenovo validated testing: NVIDIA Run:ai fractional GPU sharing vs. MIG at matched scale (14 tenants), Lenovo ThinkSystem SR650 V3 with 2x NVIDIA H100 NVL
2 IDC CIO Playbook 2026
3 MLPerf benchmark results, Lenovo Press LP1145
4 ITIC Global Server Hardware and Server OS Reliability Study
5 ITIC 2025 Server Security Survey
6 Gartner Supply Chain Top 25 for 2026
7 2026 ISG Provider Lens AI-ready Infrastructure Solutions Report

© Lenovo 2026 All rights reserved. Lenovo, the Lenovo logo, Smarter Technology for All, ThinkCentre, ThinkPad, ThinkStation, and ThinkSystem are trademarks or registered trademarks of Lenovo. NVIDIA, the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Red Hat and the Red Hat logo are trademarks or registered trademarks of Red Hat, LLC or its subsidiaries in the U.S. and other countries. Ubuntu and Canonical are registered trademarks of Canonical Ltd. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. Kubernetes is a registered trademark of The Linux Foundation in the U.S. and other countries.