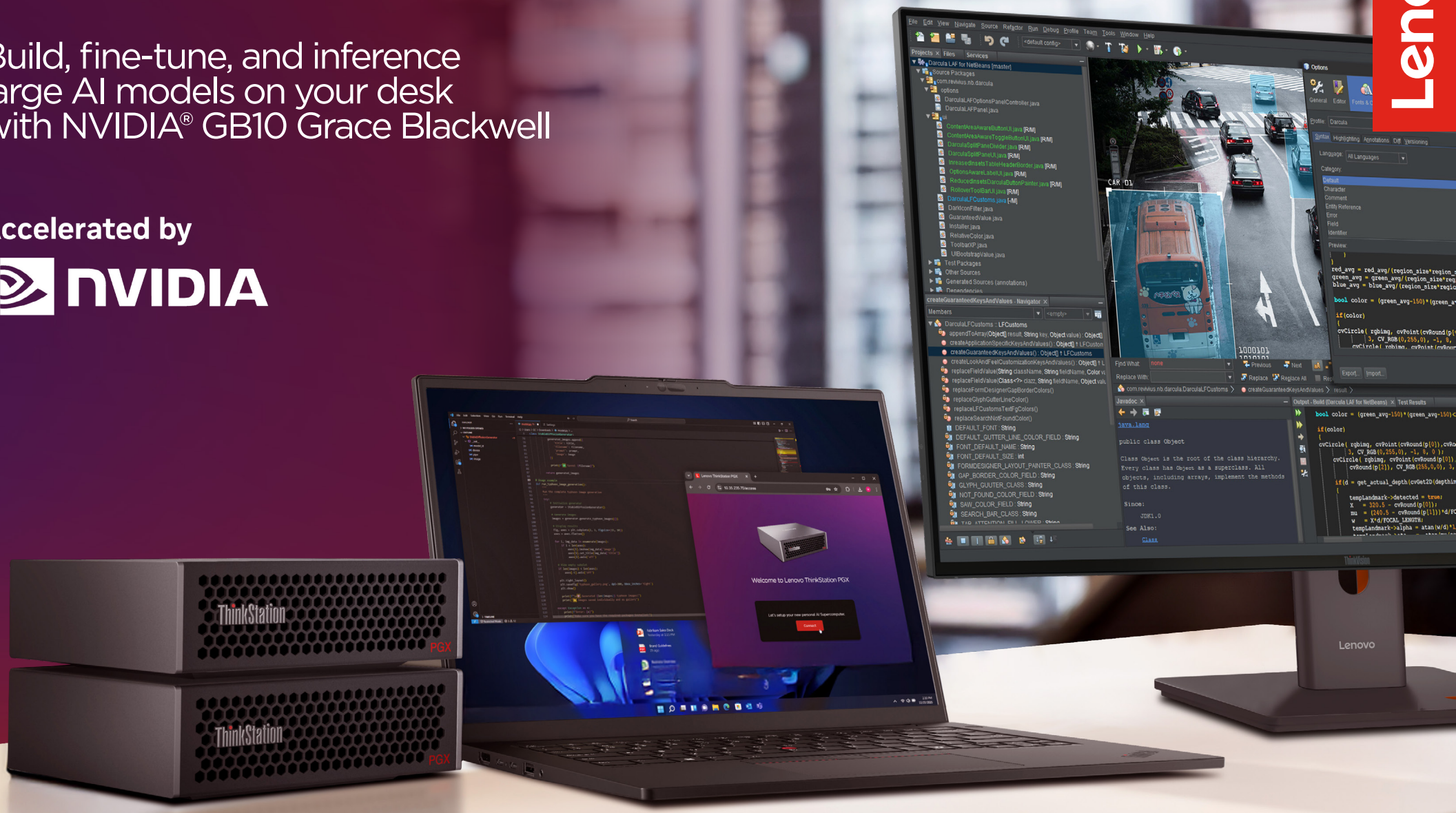
Lenovo ThinkStation PGX

Build, fine-tune, and inference large AI models on your desk with NVIDIA® GB10 Grace Blackwell

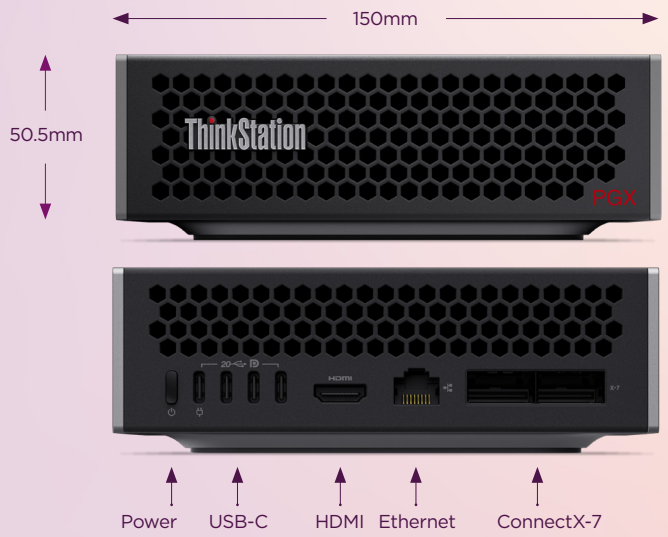
Accelerated by

NVIDIA



Executive summary

The **Lenovo ThinkStation PGX** is the first Lenovo workstation powered by the NVIDIA Grace Blackwell architecture—engineered to make AI development more accessible than ever. Built on the NVIDIA GB10 Grace Blackwell Superchip, it brings CPU, GPU, high-bandwidth unified memory, and networking together in a quiet, compact, and energy-efficient form factor. It provides a controlled local environment for prototyping, fine-tuning, and inference, and integrates seamlessly with existing workstation setups.



Lenovo

Key specifications:

Platform & Architecture

Architecture	NVIDIA Grace Blackwell
Processor	NVIDIA Grace 20-core Arm CPU (10x Cortex-X925 + 10x Cortex-A725)
GPU	Blackwell architecture - 5th-gen Tensor Cores - 4th-gen RT Cores - NVENC/NVDEC
AI Performance	1 PFLOP (FP4, sparsity)

Memory

Memory	128GB LPDDR5X 256-bit bus; 273 GB/s bandwidth

Storage

SSD	1TB or 4TB NVMe M.2 (self-encrypting)

Connectivity

Ports	(4) USB4 Type-C (20 Gb/s) (1) HDMI 2.1a
Networking	(1) RJ-45 10 GbE (2) ConnectX®-7 (PGX-to-PGX linking)
Wireless	Wi-Fi 7 Bluetooth 5.4 LE

Security

Storage	Self-encrypting NVMe (AES SED)
Platform	TPM 2.0 UEFI Secure Boot AMI Setup Password
NVIDIA	NVLink-C2C GPU-to-CPU interconnect NVIDIA FW Recovery

Power

Max power	240W

Physical

Dimensions	150 × 150 × 50.5mm (5.91 × 5.91 × 1.99in)
Volume	1.13 liters

Software

OS	NVIDIA DGX™ OS based on Ubuntu Linux
Stack & tools	NVIDIA AI software stack (incl. CUDA 13) GB10 Dashboard NVIDIA AI Workbench

The Problem: Why businesses need a local AI development platform

As AI adoption accelerates, organizations face mounting challenges that cloud-based development alone can't solve. A local AI development platform offers a strategic advantage by addressing key pain points:

1 Predictable costs

Cloud GPU pricing is volatile and rising, making budgeting for AI projects increasingly difficult.

2 Faster iteration, no waiting

Shared cloud clusters often suffer from capacity constraints and long queue times, slowing down experimentation and innovation.

3 Enhanced privacy & compliance

Sensitive assets like proprietary code, regulated datasets, legal documents, and internal communications, may be at risk of exposure in the cloud.

4 Reduced latency & fewer dependencies

Remote development introduces latency and reliance on external networks and services, which can hinder productivity.

5 Seamless deployment

Mixed toolchains between development and deployment environments often lead to compatibility issues, bugs, and delays.

6 Office ready, deploy anywhere

Traditional AI hardware is bulky and power hungry. Enterprises are finding they need a device that can be deployed in diverse environments, from labs to remote offices, without specialized infrastructure.

The Solution: Lenovo ThinkStation PGX

The ThinkStation PGX is a compact, personal AI developer platform that complements ThinkStation and ThinkPad workstation workflows. It provides a local sandbox for prototyping, fine-tuning, and inference.

With a ThinkStation PGX you can:

- Run experiments locally with affordable, always-available local compute to reduce cloud dependence and make budgets more predictable.
- Avoid queuing for shared clusters and iterate faster with always available local compute.
- Keep sensitive data on premises to support governance, compliance, and IP protection.
- Shorten development loops by eliminating remote latency and external dependencies.
- Align development with production using a Linux native environment and the NVIDIA AI software stack, including CUDA 13, GB10 Dashboard, NVIDIA AI Workbench, and familiar tools like PyTorch and Jupyter Notebooks.
- Fit a complete AI development workspace on your desk thanks to a 1.13 liters footprint.
- Manage power and operating costs with 240W power consumption.
- Scale as your needs grow. A single PGX supports models up to 200B parameters. Users can link two PGX units via integrated NVIDIA ConnectX-7 ports to handle larger models.

The ideal local AI development experience

Thoughtfully engineered for ease of use and seamless integration, the ThinkStation PGX delivers a ready-to-go experience that helps teams start building from day one, without the complexity or overhead of traditional AI hardware.

Here's how ThinkStation PGX empowers your team to develop locally with confidence:

- Ubuntu Linux Pro with NVIDIA Base OS for Linux-native workflows.
- NVIDIA AI software stack including CUDA 13, GB10 Dashboard, and NVIDIA AI Workbench, all preinstalled.
- Ready for popular frameworks and tools like PyTorch and Jupyter Notebooks.
- Remote access from your existing ThinkPad or ThinkStation workstations.
- Windows developer support via Secure Shell (SSH) or NVIDIA AI Workbench's remote code execution features.
- Day-one productivity with minimal setup required.

NVIDIA jupyter Linux ubuntu PyTorch

Architecture

NVIDIA GB10
Grace Blackwell Superchip

128GB
Unified System Memory
(273 GB/s bandwidth)

1 PFLOP
for AI acceleration

5th Generation Tensor Cores
4th Generation RT Cores

NVENC | NVDEC
engines for efficient video processing.

Security features
NVMe self encrypting
TPM 2.0 UEFI Secure boot
NVLink C2C GPU-to-CPU interconnect

Built for the demands of AI development

Prototyping & fine-tuning

Rapidly adapt LLMs and generative models locally. Up to ~200B parameters on one PGX. Larger with two PGX devices clustered.

Inference serving

Serve trained models for development and pre-production testing.

Data science acceleration (RAPIDS)

Turbocharge ETL, feature engineering, and classical ML with end-to-end GPU pipelines.

Robotics (Isaac | Metropolis)

Build, simulate, and validate perception and control apps for robots.

Agents & enterprise apps experimentation

Safely explore agent frameworks and internal AI apps in an on-prem, sandboxed environment.

Edge

Package, optimize, and validate models for on-premises/IoT devices to deliver secure, low-latency on-device inference.

Who benefits from the ThinkStation PGX

- AI developers
- Data scientists
- Researchers
- Students
- IT and DevOps (on premises)

Deployment modes and integration

ThinkStation PGX integrates seamlessly into your workflow—whether in a dedicated standalone mode, a companion mode, or cluster mode for high-capacity workloads. Built on the same Linux-native NVIDIA AI software stack, PGX enables smooth transition from local development to data center or cloud.

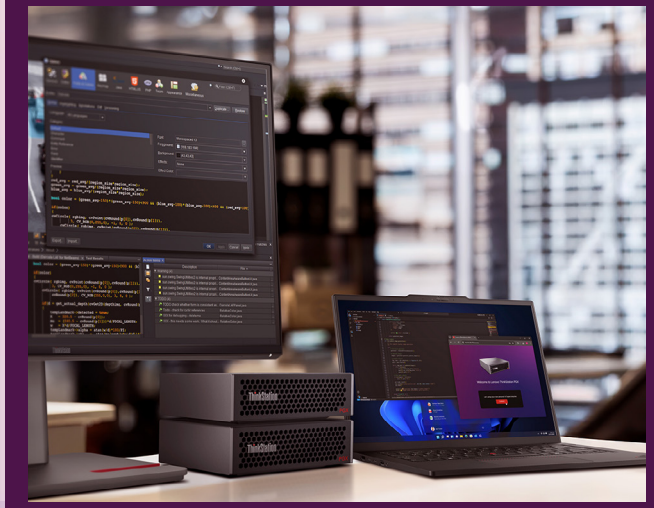
Standalone mode

Connect the ThinkStation PGX to a monitor, keyboard, and mouse and use it like a desktop system for development.

Companion mode

Use the ThinkStation PGX as a companion to your ThinkPad or ThinkStation, enabling:

- Faster results by offloading GPU-intensive workloads.
- A consistently responsive primary workstation for interactive work.
- The ability to tackle bigger models and run concurrent jobs locally.
- Pair locally or work remotely from your ThinkPad or ThinkStation.



Cluster mode

Link two PGX units via integrated ConnectX-7 ports to support larger model capacities. No external switch is required.

Accelerate ROI with smarter AI development

Cost and agility (TCO)

- Reduce cloud GPU spending during development by running key workloads locally.
- Keep costs predictable with Lenovo Device as a Service (DaaS).
- Accelerate development cycles by avoiding shared-cluster queues and minimizing delays.
- Maintain sensitive data on-premises to support privacy, compliance, and IP protection.

ROI levers

- Increase utilization of local compute resources across projects.
- Deliver faster results with shorter, more reliable dev cycles.
- Reduce expenses from data transfer and storage overhead during experimentation.

How the ThinkStation PGX fits into scaling

- The ThinkStation PGX is an ideal entry point for local AI development, designed for 1–2 nodes per developer or team.
- For larger workloads that require higher throughput, multi-GPU parallelism, or greater local VRAM capacity, Lenovo's ThinkStation P5, P7, P8, and PX systems provide the next step up.

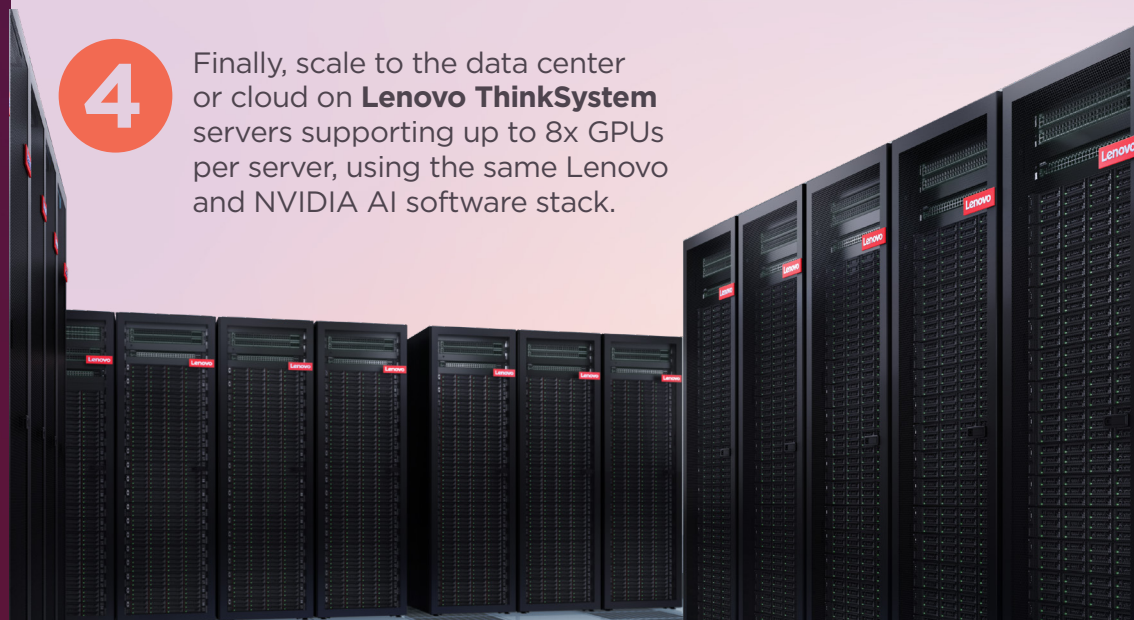
Scaling up the Lenovo ecosystem

1. Start with a **ThinkStation PGX** for on-desk AI development and local inference.

2. Expand to the ThinkStation P Series, with the **ThinkStation P3 Ultra** or **P3 Tower** utilizing 1x GPU, for more headroom.

3. Scale further to the **ThinkStation P5**, **P7**, **P8**, or **PX** with support for up to 4x GPUs for increased performance with larger models.

4. Finally, scale to the data center or cloud on **Lenovo ThinkSystem** servers supporting up to 8x GPUs per server, using the same Lenovo and NVIDIA AI software stack.



Why Lenovo

Lenovo stands apart by delivering a complete, enterprise-ready AI development experience. Choosing Lenovo means more than just selecting a device, it's about partnering with a trusted global leader in innovation, support, and end-to-end solutions.

Strategic partnership with NVIDIA

Lenovo works hand-in-hand with NVIDIA across devices, infrastructure, and services. This close collaboration ensures seamless integration, optimized performance, and faster access to the latest AI technologies.

Global reach, local support

With Lenovo's robust global supply chain and partner ecosystem, customers benefit from reliable delivery, localized support, and scalable services, wherever they operate.

Comprehensive solution blueprints

With Lenovo AI Developer, your enterprise can take advantage of fully customized, end-to-end solutions tailored for their specific goals. This includes:

- Step-by-step guidance for NVIDIA AI Workbench and NVIDIA AI Enterprise.
- Access to the latest open-source AI frameworks, pre-trained models, hardware sizing recommendations, and technical documentation.

End-to-end services & programs

Lenovo Hybrid AI Advantage with NVIDIA

Accelerate agentic AI adoption with full-stack, validated solutions and a hybrid infrastructure that supports scalable deployment across devices, edge, and cloud, boosting productivity and delivering real business impact.

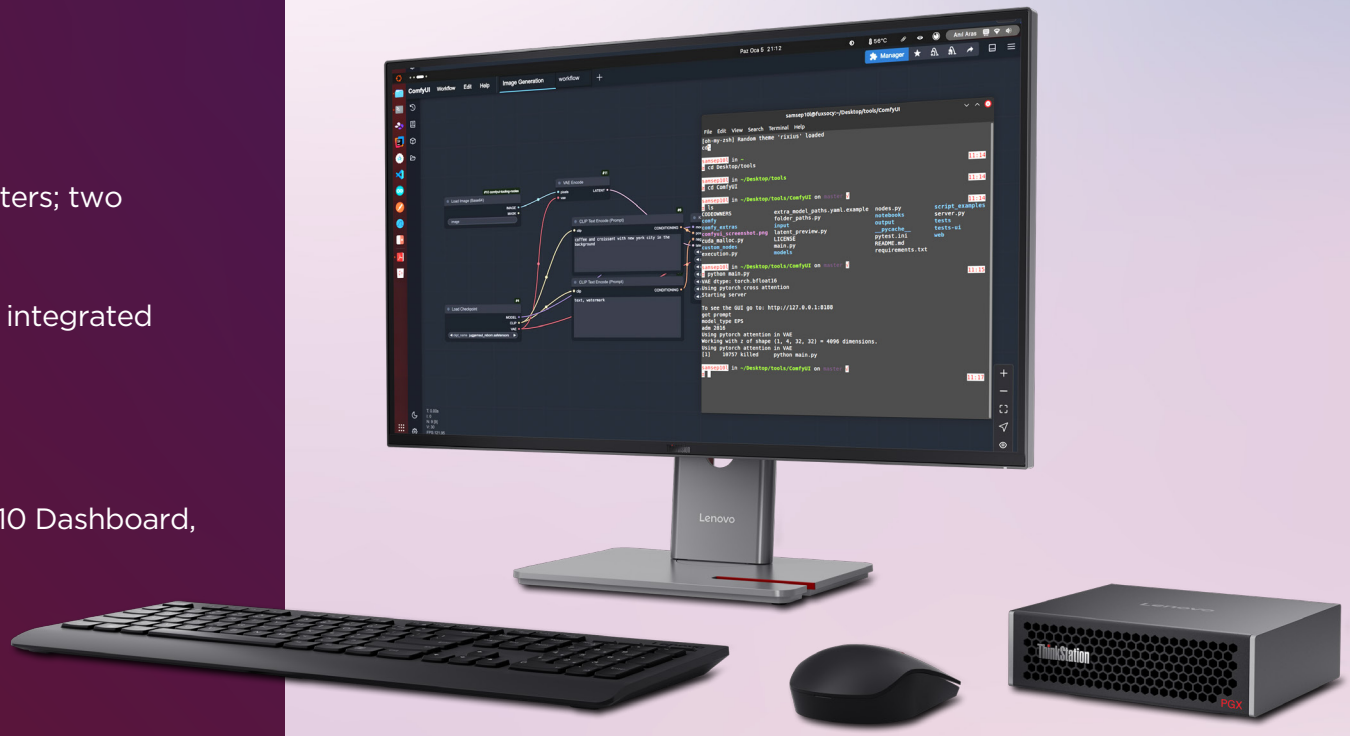
Lenovo TruScale Device as a Service (DaaS)

Simplify device deployment and lifecycle management with a flexible model that combines hardware, software, training, and support, all under one subscription.

FAQ and how to get started

Frequently Asked Questions

- What model sizes can I run locally?
  A single PGX supports models up to 200B parameters; two linked PGX units up to 405B parameters.
- Can I cluster multiple units?
  Yes, two ThinkStation PGX nodes can be linked via integrated ConnectX-7.
- What operating system does it run?
  NVIDIA DGX OS based on Ubuntu Linux.
- What other software comes preinstalled?
  NVIDIA AI software stack (including CUDA 13), GB10 Dashboard, and NVIDIA AI Workbench.
- What are the key I/O and connectivity options?
  (4) USB4 Type-C (20 Gb/s), HDMI 2.1a,
  10 GbE LAN; Wi-Fi 7 and Bluetooth 5.4 LE.
- What is the power consumption?
  240 W.
- What are the size and weight characteristics?
  1.13 liters volume; 150 × 150 × 50.5 mm; starting at 1.2 kg.



October 2025 ©Lenovo. All rights reserved.

Lenovo is not responsible for photographic or typographic errors. Lenovo makes no representation or warranty regarding third-party products or services. LENOVO, ThinkStation, ThinkPad, ThinkSystem and Lenovo Hybrid AI Advantage are trademarks of Lenovo. NVIDIA, ConnectX-7, and NVIDIA DGX are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Microsoft and Windows are registered trademarks of Microsoft Corporation. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries. Ubuntu is a registered trademark of Canonical Ltd.