

HP ZGX Fury G1n AI Station

On-prem AI, built for scale. No usage-based token costs.

Quickly scale and streamline AI development, inference, and agentic AI workflows with groundbreaking on-prem compute and a flexible software stack. The HP ZGX Fury AI Station hosts large-scale models up to 1 trillion parameters¹ locally, helping keep data secure and costs predictable².

Accelerate local AI development

Power AI development with the HP ZGX Fury, featuring the NVIDIA® GB300 Grace Blackwell Ultra Desktop Superchip with 748GB of unified memory. Speed fine-tuning iteration of 100 billion¹ class AI models with local control and no ongoing cloud charges.

Production-grade inference. Predictable costs.²

Bring enterprise-grade inference on-prem and leave token fees behind. Deploy HP ZGX Fury as a shared AI system to host up to 1T parameter¹ AI models, proprietary apps, and AI agents locally, with high-throughput, low-latency performance for teams.

Open, flexible tools. Fast setup. No lock-in.

Use built-in NVIDIA AI software stack and streamline device discovery, pairing, and installation of open-source tools with the HP ZGX Toolkit.³ Pull and deploy Hugging Face models in minutes with the pre-installed HP Z Runtime CLI app.



*Product image may differ from actual product

Sustainability in action

Shareable AI compute. Office-deployable design.

Run inference for multiple users simultaneously or give developer teams a shared platform to experiment and fine-tune in parallel on a single HP ZGX Fury. Its rack-mountable, office-deployable design plugs into standard enterprise environments.



HP ZGX Fury G1n AI Station

Featuring

Accelerate local AI development

Power AI development with the HP ZGX Fury, featuring the NVIDIA® GB300 Grace Blackwell Ultra Desktop Superchip with 748GB of unified memory. Speed fine-tuning iteration of 100 billion¹ class AI models with local control and no ongoing cloud charges.

Production-grade inference. Predictable costs ².

Bring enterprise-grade inference on-prem and leave token fees behind. Deploy HP ZGX Fury as a shared AI system to host up to 1T parameter¹ AI models, proprietary apps, and AI agents locally, with high-throughput, low-latency performance for teams.

Open, flexible tools. Fast setup. No lock-in.

Use built-in NVIDIA AI software stack and streamline device discovery, pairing, and installation of open-source tools with the HP ZGX Toolkit.³ Pull and deploy Hugging Face models in minutes with the pre-installed HP Z Runtime CLI app.

Shareable AI compute. Office-deployable design.

Run inference for multiple users simultaneously or give developer teams a shared platform to experiment and fine-tune in parallel on a single HP ZGX Fury. Its rack-mountable, office-deployable design plugs into standard enterprise environments.

NVIDIA® DGX™ OS and AI Software Stack

Hit the ground prototyping and inferencing large-scale models with the familiar NVIDIA® DGXTM OS. Easily configure NVIDIA AI Enterprise or open ecosystems with the HP ZGX Fury Developer Software Stack, purpose-built for modern AI workflows.

NVIDIA® GB300 Grace Blackwell Ultra Superchip

The NVIDIA® GB300 Grace Blackwell Ultra Superchip delivers incredible on-prem performance, with up to 20 petaFLOPS of FP4 AI performance and 748 GB of unified memory for developing and running frontier AI workloads.

Unified Memory

Effortlessly develop and fine-tune up to 100 billion class or inference up to trillion parameter class AI models with 748 GB of large, unified memory.¹

Fast and Secure Storage

Choose between 2 or 4 TB of NVMe M.2 self-encrypted storage to work efficiently with large files and keep more of your data local and secure.⁴

Office-Deployable Design

Desk, data closet, or data center – enable secure on-prem AI compute how you need it. The office-deployable, rack-mountable design plugs into standard enterprise environments.

HP ZGX Toolkit for AI

Get open-source frameworks, MLflow tracking, and Ollama testing. With instant discovery, sync, and export, train locally, track results, and deploy with cloud, data center, or workstation.⁵

HP Z Runtime CLI app

Go from model to inference in minutes with the pre-installed HP Z Runtime, giving AI teams a fast, streamlined path to pull, serve, and manage locally.⁶

Graphical AI support

Offload video output to an optional NVIDIA RTX™ PRO GPU, while the GB300 GPU stays locked in on heavy AI workloads.

NVIDIA ConnectX®-8 SuperNIC™ Networking

Work with even larger AI models locally. With up to 800 Gb/s, the NVIDIA ConnectX®-8 SuperNIC™ delivers extremely fast, efficient network connectivity and supports linking two ZGX Fury devices to scale model capacity and performance.^{6,7}



HP ZGX Fury G1n AI Station

Technical specifications

Available Operating Systems	Ubuntu 24.04 LTS with NVIDIA AI Developer Tools
Processor family	NVIDIA® Grace Blackwell Ultra Desktop Superchip
Available Processors	NVIDIA® GB300 Grace Blackwell Ultra Desktop (NVIDIA® Grace-72 Core Neoverse V2 CPU) with NVIDIA® Blackwell Ultra GPU (252 GB HBM3e)
Product colour	Black
Form factor	Liquid cooled rackable tower
Maximum memory	496 GB LPDDR5X 6400 MT/s SOCAMM Maximum of 748 GB coherent memory, including 496 GB LPDDR5X CPU memory, 396 GB/s (SOCAMM) and 252GB HBM3e GPU memory, 7.1 TB/s.
Internal storage	2 TB HP Z Turbo Drive PCIe® M.2 SSD 4 TB HP Z Turbo Drive PCIe® M.2 SSD 2 TB Z Turbo PCIe NVMe™ M.2 SED NMIC SSD All configurations include two M.2 SSDs for OS (RAID 1 mirrored pair) and two M.2 SSDs for data (RAID 0). Available capacity will vary by configuration.
Additional storage	Four M.2 storage slots: Two PCIe Gen5 x4 M.2 devices for OS (RAID 1 mirrored) and two PCIe Gen6 x4 for data (RAID 0 striped). Drive types and capacities vary by configuration.
Available Graphics	Integrated: NVIDIA® Blackwell Ultra GPU Discrete: NVIDIA RTX PRO™ 4000 Blackwell Generation (24 GB GDDR7 dedicated); NVIDIA RTX PRO™ 2000 Blackwell Generation (16 GB GDDR7 dedicated)
Expansion slots	3 PCIe Slots: 1 PCIe Gen5 x16; 2 PCIe Gen5 x16 (x8 electrical)
Ports and connectors	Rear: 4 USB Type-A 10Gbps signaling rate; 1 microphone; 1 RJ-45 1 Gb Ethernet (BMC for system management only); 1 micro USB (BMC for system management only); 1 Mini DisplayPort™ 2.0 (BMC systems management only); 2 QSFP112 (400Gb per port); 1 audio-in; 1 audio line-out; 1 RJ-45 (10 GbE) ; Front: 2 USB Type-C® 10Gbps signaling rate; 2 USB Type-A 5Gbps signaling rate; 1 universal audio jack
Communications	NVIDIA® ConnectX-8, 2x 400Gb QSFP112 ports (supports optical and DAC connectivity)
Software	HP Z Runtime; NVIDIA AI Developer Tools; NVIDIA approved partner-developed BIOS; NVIDIA approved partner-developed BMC Firmware
Security management	Trusted Platform Module (TPM) 2.0; UEFI version 2.9; Kensington Cable Lock; UEFI Secure Boot



HP ZGX Fury G1n AI Station

Technical specifications

Management features	Out of band with NVIDIA integrated BMC; Redfish API support
Power	1600 W Titanium
Dimensions	17.56 x 8.72 x 26.26 in; 44.6 x 22.15 x 66.7 cm; (Standard tower orientation, ships with rack rails for rack configuration availability out of the box.); 31.75 x 23.19 x 17.125 in 80.65 x 58.9 x 43.5 cm (Package)
Weight	Starting at 90 lb; Starting at 41 kg; (Exact weights depend upon configuration (System weight only))
Certification and compliance	TAA compliant



HP ZGX Fury G1n AI Station

Footnotes

Messaging Footnotes

- ⁴ 2TB or 4TB storage configuration must be selected at the time of purchase. OS drives are RAID 1, mirrored. Available capacity is half of total drive capacity.
- ⁵ The HP ZGX Toolkit is provided free of charge. Use requires a client device running Windows 11 or Ubuntu 24.04 (or later) with Visual Studio Code installed and host device ZGX Fury. The client device must be x86-based, but otherwise there are no restrictions regarding hardware specifications or device manufacturer. Availability may vary by region and is subject to applicable local laws, regulations, and restrictions.
- ⁶ Speed varies on both model size and gated models.
- ⁷ Requires compatible QSFP cable. Sold separately.
- ¹ Model harness quantizes at FP4.
- ² AI stations enable more predictable costs by eliminating usage-based pricing (such as token fees), reducing reliance on variable cloud infrastructure, and shifting AI workloads to a fixed, on-prem devices.
- ³ The HP ZGX Toolkit is provided free of charge. Use requires a client device running Windows 11 or Ubuntu 24.04 (or later) with Visual Studio Code installed and host device ZGX Fury. The client device must be x86-based, but otherwise there are no restrictions regarding hardware specifications or device manufacturer. Availability may vary by region and is subject to applicable local laws, regulations, and restrictions.

