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 parameters[1] locally, helping keep data secure and costs predictable[2].

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 billion[1] 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 parameter[1] 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 billion[1] 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 parameter[1] 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[2]
Processor family[1]	NVIDIA® Grace Blackwell Ultra Desktop Superchip
Available Processors[1]	NVIDIA® GB300 Grace Blackwell Ultra Desktop (NVIDIA® Grace-72 Core Neoverse V2 CPU) with NVIDIA® Blackwell Ultra GPU (252 GB HBM3e)
Form factor	Tower
Maximum memory	496 GB LPDD5X 6400 MT/s SOCAMM Memory bandwidth up to 396 GB/s
Internal storage	2 TB HP Z Turbo Drive PCIe® M.2 SSD[7] 4 TB HP Z Turbo Drive PCIe® M.2 SSD[7] 2 TB Z Turbo PCIe NVMe™ M.2 SED NMIC SSD[7]
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 GbE); 1 RJ-45 1 Gb Ethernet (BMC for system management only); 1 micro USB (BMC for system management only); 1 microphone audio connector; 1 Mini DisplayPort™ 2.0 (BMC systems management only); 2 QSFP112 (400Gb per port); 1 audio-in; 1 audio line-out ; 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	44.6 x 22.15 x 66.7 cm; (Standard tower orientation, ships with rack rails for rack configuration availability out of the box.); 80.65 x 58.9 x 43.5 cm (Package)
Weight	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.
[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.

Specification Footnotes

[1] Multi-core is designed to improve performance of certain software products. Not all customers or software applications will necessarily benefit from use of this technology. Performance and clock frequency will vary depending on application workload and your hardware and software configurations.
[2] Not all features are available in all editions or versions of Ubuntu. Systems may require upgraded and/or separately purchased hardware, drivers, software or BIOS to take full advantage of Ubuntu functionality. Ubuntu may be automatically updated. ISP fees may apply and additional requirements may apply over time for updates.
[3] Wireless access point and Internet service required and sold separately. Availability of public wireless access points limited. Wi-Fi 7 (802.11BE) functionality requires compatible Windows 11 24H2 OS, compatible processor, and separately purchased Wi-Fi 7 router to support backward compatibility with prior 802.11 specs. Available in countries where Wi-Fi 7 is supported.
[7] 2TB or 4TB storage configuration must be selected at time of purchase.

© Copyright 2026 HP Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

July 2026