Datasheet

NVIDIA DGX Station

The ultimate deskside AI supercomputer.

Data Center-Class AI at Your Deskside

NVIDIA DGX Station™ is the ultimate deskside AI supercomputer built for data scientists, AI researchers, and developers. Powered by the NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip, DGX Station delivers data center-level performance to AI teams, alongside massive memory, high-speed NVIDIA ConnectX™ networking, and seven-way NVIDIA Multi-Instance GPU (MIG) capabilities to supercharge local AI development, fine-tuning, agentic AI, and inference workloads. Run and develop with large AI models and autonomous long-running AI agents locally, providing enhanced data security and continual iteration without requiring cloud resources.

- The NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip uses NVIDIA NVLink™-C2C to interconnect an NVIDIA Blackwell Ultra GPU and Grace CPU.
- 748 GB of coherent memory enables the system to tackle the largest AI workloads locally.
- Up to 20 petaFLOPS of compute bring immense speed to AI training and inferencing tasks.
- MIG can be used to partition the GPU cores of DGX Station to up to seven users, empowering AI teams with a powerful workgroup cluster.
- Preinstalled Ubuntu with NVIDIA AI Developer Tools enables seamless scaling of AI workloads from your deskside to the data center and acceleration of NVIDIA CUDA-X AI™ libraries.

Built for AI Acceleration

DGX Station fuses the power of the GB300 Grace Blackwell Ultra Superchip with a hyperoptimized, preconfigured NVIDIA AI software stack. The software stack includes Ubuntu with NVIDIA AI Developer Tools, simplifying AI deployment from deskside to data center, and preinstalled CUDA-X™ libraries, providing developers a full-stack solution for AI workloads, including fine-tuning, inference, and data science. With the NVIDIA AI software stack, developers can easily work on AI models locally and deploy to the cloud or data center using the same tools, libraries, frameworks, and pretrained models. Enterprises can fast-track AI deployment

Key Features

NVIDIA GB300 Superchip

- NVIDIA Blackwell Ultra GPU, Grace CPU
- Up to 20 petaFLOPS of AI compute
- 748 GB of coherent memory

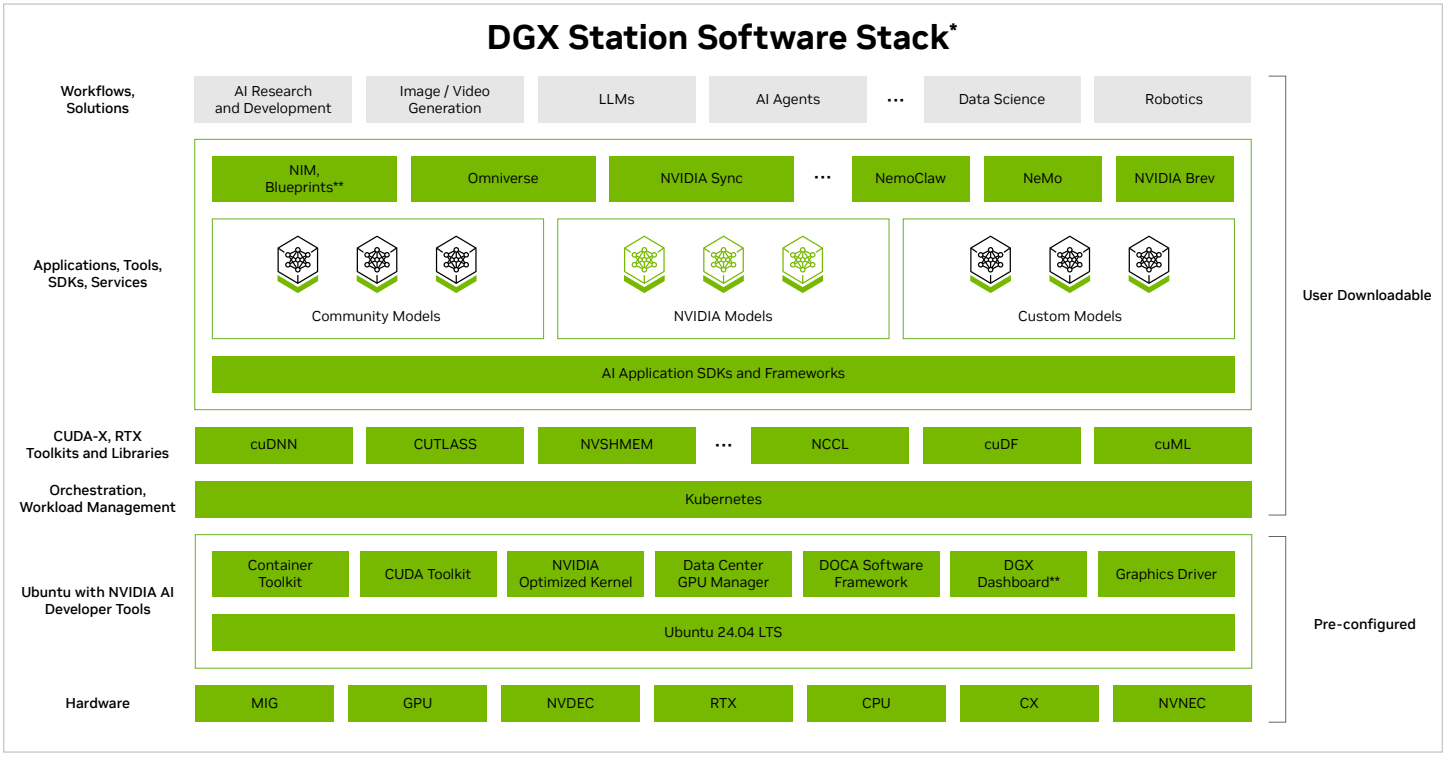
Ready-to-Run AI Software

- Ubuntu with NVIDIA AI Developer Tools
- NVIDIA AI Enterprise
- NVIDIA NIM™ microservices
- NVIDIA CUDA-X libraries

Deskside Server for AI Teams

- Partition compute and memory for up to seven instances with NVIDIA MIG
- Link up to two DGX Stations with the NVIDIA ConnectX-8 SuperNIC™

further with an NVIDIA AI Enterprise software license, which provides a cloud-native suite of software tools, libraries, and frameworks, simplifying the development, deployment, and scaling of AI applications.



* Software stack components subject to change.
** Feature support planned in future update.

In addition to support for familiar AI models and tools, enterprises deploying autonomous agents across proprietary workflows need governance, isolation, and control. With NVIDIA DGX Station and NemoClaw™, an open source stack that simplifies running OpenClaw always-on assistants more safely with a single command, developers have the ultimate platform for building and running AI agents locally.

Powerful Supercomputer for AI Teams

NVIDIA DGX Station can serve as either a personal supercomputer for one user to run advanced AI models with local data or as a centralized compute node for multiple team members to fine-tune and run IP-specific models on demand. DGX Station supports NVIDIA MIG technology, providing AI teams with a powerful local system for distributed compute workloads. MIG partitions the GPU memory and compute capabilities of DGX Station into as many as seven instances for multiple users, each fully isolated with its own high-bandwidth memory, cache, and compute cores. Performance can be scaled further by connecting up to two DGX Stations via NVIDIA ConnectX networking—doubling the memory and compute capacity to over 1.4 TB of coherent memory and up to 40 petaFLOPS of AI compute. DGX Station provides a unified architecture from desktop to data center. If even more compute capability is needed, projects can seamlessly be sent up to the data center without impacting productivity or reworking project compatibility.

DGX Station Technical Specifications

Technical Specifications	
Architecture	NVIDIA Grace Blackwell
NVIDIA GPU	1x NVIDIA Blackwell Ultra
NVIDIA CPU	1x NVIDIA Grace™-72 Core Neoverse V2
GPU Memory	252 GB of HBM3e | 7.1 TB/s
CPU Memory	496 GB of LPDDR5X | 396 GB/s
NVLink-C2C	900 GB/s
Networking | Peak Bandwidth	NVIDIA ConnectX-8 SuperNIC | up to 800 GB/s | Ethernet
Supported OS	Ubuntu with NVIDIA AI Developer Tools
MIG	Up to 7x instances
FP4 Tensor Core	20 | 15[3] PFLOPS
FP8 / FP6 Tensor Core	10 PFLOPS
Int8 Tensor Core	330 TOPS
FP16/BF16 Tensor Core	5 PFLOPS
TF32 Tensor Core	2.5 PFLOPS
FP32	80 TFLOPS
FP64 / FP64 Tensor Core	1.3 TFLOPS
Supported RTX PRO GPU[4]	NVIDIA RTX PRO™ 6000 Workstation Edition, RTX PRO 6000 Blackwell Max-Q Workstation Edition, RTX PRO 4000 Blackwell SFF Edition, RTX PRO 2000 Blackwell
Total System Power	1600 W
Decoders	7 NVDEC 7 nvJPEG
Storage	4x M.2 Gen 5 Slots
Ethernet	2x QSFP112 (400 Gb/s per port) 1x RJ45 10GbE 1x RJ45 1GbE (BMC for System Management)
PCIe Slots	1x PCIe Gen 5 x16 2x PCIe Gen 5 x16 (x8 electrical)
Additional Ports and Connectors[4]	Front IO: 2x USB Type C (USB3.1), 2x USB Type A (USB3.1), Audio Rear IO: 4x USB Type A, USB Micro-B (BMC) mDP[5] (BMC), Audio

Ready to Get Started?

To learn more about DGX Station, visit:
nvidia.com/dgx-station-gb300

1. Peak rates are based on GPU boost clock.
2. All Tensor Core specifications are with sparsity unless otherwise noted.
3. Without sparsity.
4. Support may vary by system. Check with your OEM to confirm specific support and configurations.
5. Mini DisplayPort is intended only for system management.

NVIDIA