Datasheet

NVIDIA DGX Rubin NVL8

Supercharged performance for agentic AI.

The Foundation for Intelligence at Scale

Agentic AI and reasoning models are redefining the limits of computation. As enterprises move from experimentation to production AI factories, workloads are demanding higher performance, longer context windows, and faster coordination between models and agents.

To support this shift, infrastructure must evolve beyond raw throughput. Training, post-training, and inference now depend on faster communication, lower latency, efficient memory movement, and optimized software that can keep systems productive at scale.

NVIDIA DGX™ Rubin NVL8 provides a blueprint for success in the age of agentic AI. Built on the NVIDIA Rubin architecture, DGX Rubin NVL8 is a turnkey AI infrastructure solution purpose-built to accelerate any AI workload and deliver intelligence at scale.

Powered by the NVIDIA Rubin Architecture

Powered by eight NVIDIA Rubin GPUs, two Intel® Xeon® CPUs, and sixth-generation NVIDIA NVLink, DGX Rubin NVL8 delivers 400 PFLOPS of NVFP4 inference performance and 280 PFLOPS of NVFP4 training performance. DGX Rubin NVL8 is designed for the world's most complex AI workloads, from large-scale model training to long-context reasoning and agentic inference. Equipped with 2.3 TB of total GPU memory and 176 TB/s of HBM bandwidth, data moves through the training and inference pipeline as quickly as possible.

Sixth-generation NVIDIA NVLink provides 28.8 TB/s of total bandwidth per system, enabling seamless peer-to-peer communication for massive model parallelism. With NVIDIA DGX SuperPOD, enterprises can scale from a single DGX Rubin NVL8 system to an AI factory blueprint without creating new communication bottlenecks.

Key Features

- Built on the NVIDIA Rubin architecture with eight NVIDIA Rubin GPUs and two Intel Xeon CPUs
- Sixth-generation NVIDIA NVLink™ connects GPUs for high-speed peer-to-peer communication
- Delivers 400 PFLOPS NVFP4 inference and 280 PFLOPS NVFP4 training performance
- Provides 2.3 TB of total GPU memory and 176 TB/s of HBM bandwidth
- Purpose-built for training, inference, post-training, and agentic AI workloads
- Designed as a 2RU, liquid-cooled system for next-generation data centers
- Includes NVIDIA Mission Control™ software to simplify AI factory operations
- Scales with NVIDIA DGX SuperPOD™ as a validated AI factory blueprint

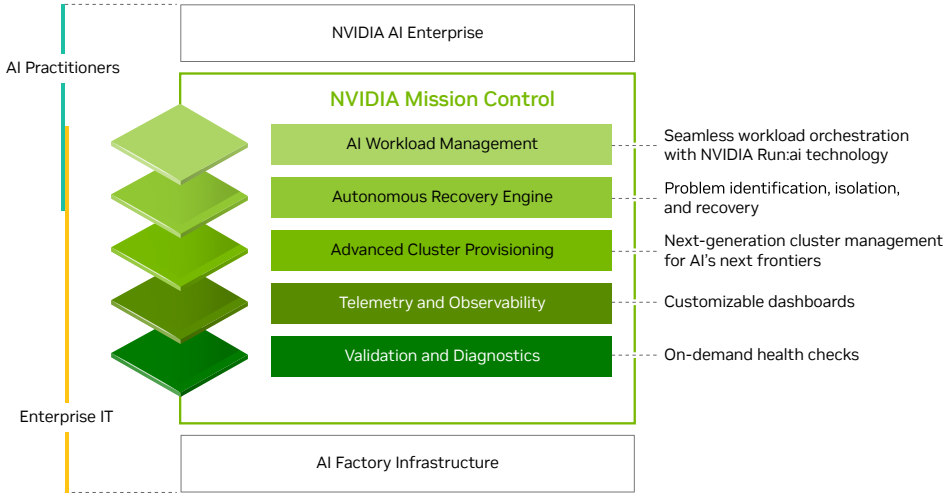
DGX Rubin NVL8 improves performance across the full AI lifecycle, including training, post-training, and inference. The NVIDIA Rubin architecture includes a specialized multi-agent engine for reasoning workflows and a dedicated reinforcement learning engine that optimizes memory movement in hardware.

These architectural advancements help enterprises ship better models faster and run reasoning models more efficiently. For teams managing trillions of tokens across production AI workloads, DGX Rubin NVL8 helps improve the token economics of continuous model improvement and deployment.

Run Models, Automate the Essentials With NVIDIA Mission Control

Deploying an AI factory is more than an infrastructure purchase. It is an ongoing operational commitment that requires visibility, automation, and resilience across workloads, infrastructure, and facilities. NVIDIA Mission Control accelerates every aspect of operations, from configuring DGX Rubin NVL8 to integrating with facilities to managing clusters and workloads, giving enterprises greater control over cooling and power events and redefining infrastructure resiliency.

By combining NVIDIA AI Enterprise, NVIDIA Mission Control, and NVIDIA DGX OS, enterprises can run AI with integrated software designed for production infrastructure. The software stack helps teams manage diverse training and inference workloads and leverage leading AI tools, improve utilization, recover from infrastructure events, and accelerate token production.



NVIDIA DGX software stack.

The Standard for Enterprise AI

DGX is built for the realities of enterprise AI deployment. It combines leadership-class computing, NVIDIA networking, integrated software, enterprise support, and NVIDIA expertise into a unified platform for production AI.

For enterprises building centralized AI factories, DGX Rubin NVL8 reduces the burden of infrastructure integration and provides a consistent foundation for AI innovation. With DGX SuperPOD and the NVIDIA partner ecosystem, organizations can deploy and scale AI infrastructure with a proven blueprint.

Technical Specifications

GPU	8x NVIDIA Rubin GPUs
Total GPU Memory	2.3 TB
GPU Memory Bandwidth	176 TB/s
Performance	NVFP4 inference: 400 PFLOPS NVFP4 training: 280 PFLOPS* FP8/FP6 training: 140 PFLOPS*
CPU	2x Intel Xeon 6776P processors
NVIDIA NVLink Bandwidth	28.8 TB/s total bandwidth
System Power Usage	~24 kW
Networking	8x OSFP ports serving 8x single-port NVIDIA® ConnectX®-9 VPI, up to 800 Gb/s NVIDIA InfiniBand and Ethernet 2x 400G QSP112 NVIDIA BlueField®-4 DPUs, up to 800 Gb/s NVIDIA InfiniBand and Ethernet
Management Network	1GbE RJ45 host baseboard management controller (BMC) 1GbE onboard NIC with RJ45
Storage	OS: 2x 2 TB NVMe M.2 Data storage: 8x 4 TB NVMe E1.S
Software	NVIDIA DGX OS Ubuntu Red Hat Enterprise Linux Rocky
Rack Units	2 RU
Enterprise Support	Three-year business standard hardware and software support.

*Dense specification.

PNY Technologies Europe
9 rue Joseph Cugnot
33708 Mérignac cedex | France
T +33 (0)5 40 240 240 | **pnypro@pny.eu**

Ready to Get Started?

To learn more about DGX Rubin NVL8, visit
www.pny.com/en-eu/professional/