Datasheet

NVIDIA DGX Vera Rubin NVL72

The foundation for gigascale AI training and inference.



Scaling Intelligence in the Age of Agentic AI

Generative AI, reasoning models, and emerging agentic AI systems are reshaping enterprise AI strategies. As organizations move from model experimentation to production AI factories, they must support the convergence of pretraining, post-training, and inference-time scaling laws.

Modern data centers face performance and efficiency bottlenecks when trying to support this new class of AI. The immense size and complexity of frontier models demand high-performance distributed computing, advanced networking, massive memory pools, and operational software that can keep infrastructure resilient and productive.

NVIDIA DGX™ Vera Rubin NVL72 is a turnkey, rack-scale AI infrastructure solution purpose-built for gigascale AI training and inference. Built on the NVIDIA Vera Rubin platform, DGX Vera Rubin NVL72 unifies 72 NVIDIA Rubin GPUs and 36 NVIDIA Vera CPUs through sixth-generation NVIDIA NVLink™ into a massive shared-memory fabric.

With a 100% liquid-cooled rack-scale design and an integrated NVIDIA software stack, DGX Vera Rubin NVL72 helps enterprises break through performance and energy bottlenecks without taking on the complexity of building a full hardware and software platform themselves.

Accelerated Training and Inference for Any AI Model

DGX Vera Rubin NVL72 is engineered as a universal AI engine for the full spectrum of modern workloads. It accelerates large language models, mixture-of-experts models, high-fidelity video generation, graph neural networks, physical AI simulations, and other demanding AI pipelines that require scale-up and scale-out performance.

Sixth-generation NVIDIA NVLink connects 72 Rubin GPUs into a single NVLink domain, enabling high-speed GPU-to-GPU communication across the rack. When deployed with NVIDIA DGX SuperPOD, DGX Vera Rubin NVL72 provides a validated blueprint for scaling AI factories across multiple racks.

Key Features

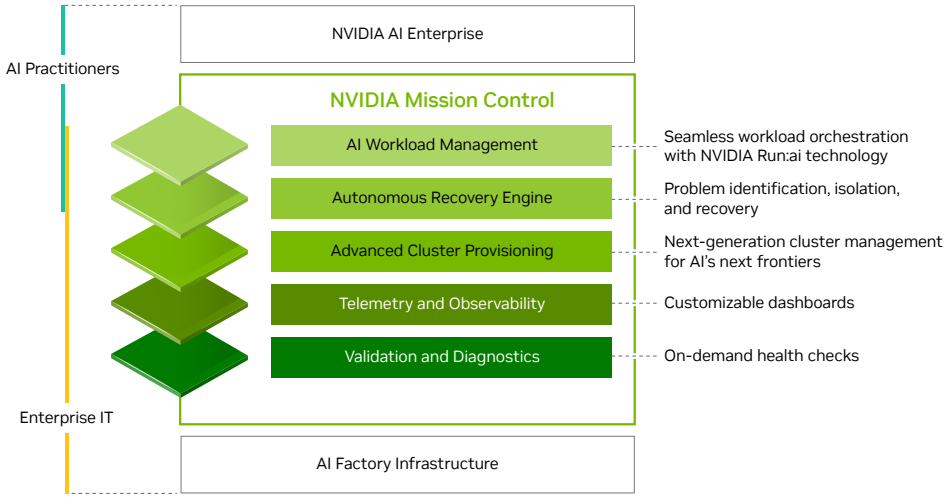
- Built on the NVIDIA Vera Rubin platform with NVIDIA Rubin GPUs and NVIDIA Vera CPUs.
- 72 NVIDIA Rubin GPUs connected as one massive GPU with sixth-generation NVIDIA NVLink.
- Delivers up to 3,600 PFLOPS NVFP4 inference and 2,520 PFLOPS NVFP4 training performance.
- Provides 20.7 TB of GPU memory for large-scale reasoning, simulation, and AI workloads.
- 100% liquid-cooled, rack-scale architecture for power-constrained data centers.
- Scales with NVIDIA DGX SuperPOD™ as a validated blueprint for gigascale AI factories.
- Includes NVIDIA Mission Control™, NVIDIA AI Enterprise, and NVIDIA DGX OS software.

The system delivers up to 3,600 PFLOPS of NVFP4 inference performance and 2,520 PFLOPS of NVFP4 training performance. With 20.7 TB of GPU memory and a high-bandwidth NVLink fabric, DGX Vera Rubin NVL72 enables acceleration of the world's largest models.

Run Models, Automate the Essentials

Deploying an AI factory is more than an infrastructure purchase. It's 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 Vera Rubin NVL72 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.

Designed for Enterprise AI Factories

DGX Vera Rubin NVL72 extends the proven NVIDIA DGX platform for the next generation of enterprise AI. It combines leadership-class compute, NVIDIA networking, integrated software, enterprise support, and the DGX ecosystem into a unified platform that helps organizations move faster from AI strategy to production value.

For enterprises building centralized AI factories, DGX reduces the burden of infrastructure integration and provides a consistent foundation for AI innovation. With DGX SuperPOD, organizations can deploy and scale AI infrastructure with a proven blueprint rather than assembling a custom stack from the ground up.

Technical Specifications

	DGX Vera Rubin NVL72
GPU and CPU	72x NVIDIA Rubin GPUs and 36x NVIDIA Vera CPUs
Total GPU Memory | Bandwidth	20.7 TB HBM4 | 1,400 TB/s
Performance	NVFP4 Inference: 3,600 PFLOPS* NVFP4 Training: 2,520 PFLOPS** FP8/FP6 Training: 1,260 PFLOPS**
NVIDIA NVLink Switch Systems	9x L1 NVIDIA NVLink Switches
NVIDIA NVLink Bandwidth	216 TB/s
Networking	144x OSFP single-port NVIDIA ConnectX™-9 VPI with 800 Gb/s NVIDIA InfiniBand and Ethernet 18x dual-port NVIDIA BlueField™-4 VPI with 400 Gb/s NVIDIA InfiniBand and Ethernet
Management Network	Host baseboard management controlled (BMC) with RJ45
Storage	E1.S
Software	NVIDIA Mission Control NVIDIA AI Enterprise NVIDIA DGX OS Supports Ubuntu
Rack Units	48 RU
Enterprise Support	Three-year enterprise business-standard support for hardware and software

* Sparse specification.

** Dense specification.

Ready to Get Started?

To learn more about DGX Vera Rubin NVL72, visit nvidia.com/dgx-vera-rubin-nvl72

© 2026 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, DGX, DGX SuperPOD, Mission Control, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 5857622. OCT26

NVIDIA