Datasheet

NVIDIA Vera Rubin

Seven new chips, one incredible AI supercomputer.

NVIDIA Vera Rubin

Built for the Age of Agentic AI

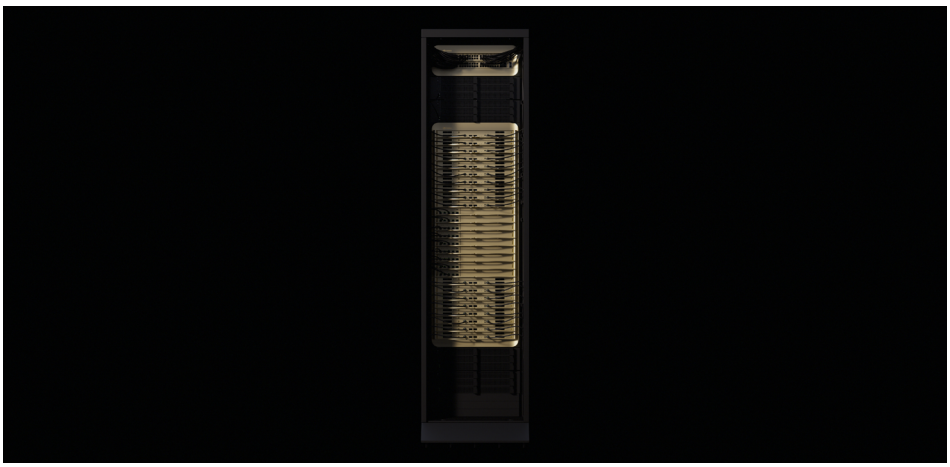
AI is evolving at a pace that is fundamentally transforming computing, turning data centers into AI factories where compute equals revenue. The rise of reasoning models and agentic workflows is driving an exponential surge in token demand, pushing infrastructure to its limits across performance, throughput, and cost efficiency. In this environment, the speed of intelligence and performance per watt determines profitability. The NVIDIA Vera Rubin platform is built through extreme co-design to deliver the highest performance, greatest efficiency, and lowest token cost for agentic workloads.

Key Offerings

- NVIDIA Vera Rubin NVL72
- NVIDIA HGX Vera Rubin NVL8
- NVIDIA HGX Rubin NVL8
- NVIDIA Vera Rubin NVL4

NVIDIA Vera Rubin NVL72

Building the Next Frontier of AI



NVIDIA Vera Rubin NVL72

Key Features

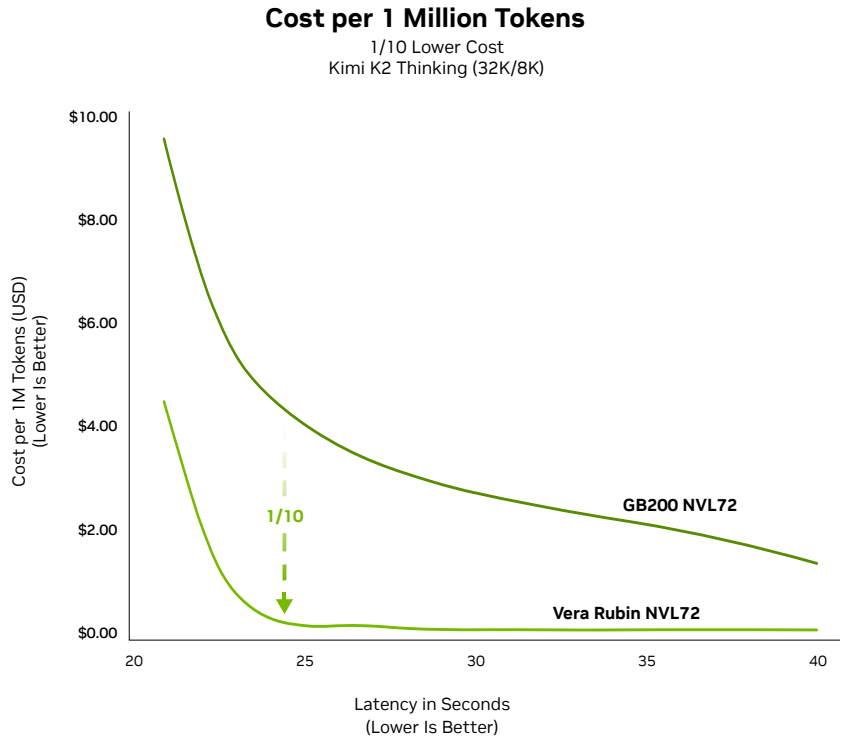
- 36 NVIDIA Vera CPUs
- 72 NVIDIA Rubin GPUs
- 54 TB of LPDDR5X memory
- 20.7 TB of HBM4
- 75 TB of fast-access memory
- NVLink domain: 260 Terabytes per second (TB/s) of low-latency GPU communication

NVIDIA Vera Rubin NVL72 unifies leading-edge technologies from NVIDIA—72 NVIDIA Rubin GPUs, 36 NVIDIA Vera CPUs, ConnectX-9 SuperNIC™, BlueField®-4 DPUs, and BlueField-4 STX storage processors. It scales up intelligence in a rack-scale platform with the NVIDIA NVLink™ 6 switch and scales out with NVIDIA Quantum-X800 InfiniBand and Spectrum-X™ Ethernet to power the AI industrial revolution at scale. When deployed with NVIDIA Groq 3 LPX racks, NVIDIA Vera Rubin NVL72 delivers a new class of inference performance for trillion-parameter models and million-token context.

NVIDIA Vera Rubin NVL72 is built on the third-generation NVIDIA MGX™ NVL72 rack design, offering a seamless transition from prior generations. It delivers AI training with one-fourth the GPUs and AI inference at one-tenth the cost per million tokens versus NVIDIA GB200 NVL72. Featuring cable free modular tray designs and support from over 80 MGX ecosystem partners, the rack-scale AI supercomputer delivers world class performance with rapid deployment.

Driving Down Inference Costs

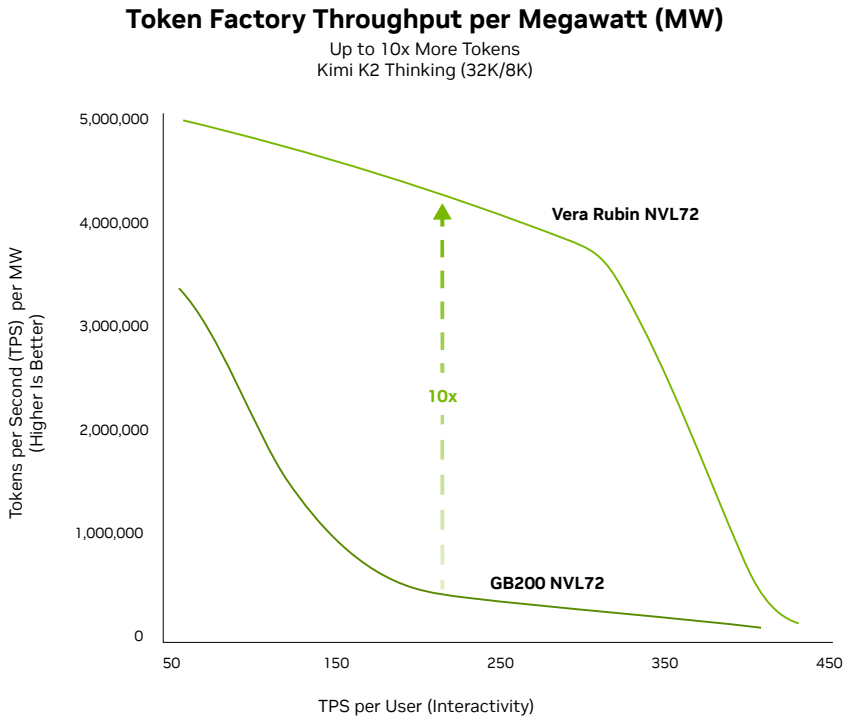
Deploying highly interactive, deep reasoning agentic AI requires predictable and manageable operational expenses. The NVIDIA Vera Rubin platform delivers one-tenth the cost per million tokens compared to NVIDIA GB200 NVL72. This breakthrough in cost efficiency enables enterprises to rapidly scale their most advanced AI deployments and achieve greater overall profitability.



LLM Inference performance subject to change. Cost per 1 million tokens based on Kimi K2-Thinking model using 32K/8K ISL/OSL comparing NVIDIA GB200 NVL72 and NVIDIA Vera Rubin NVL72.

Maximizing Token Factory Throughput

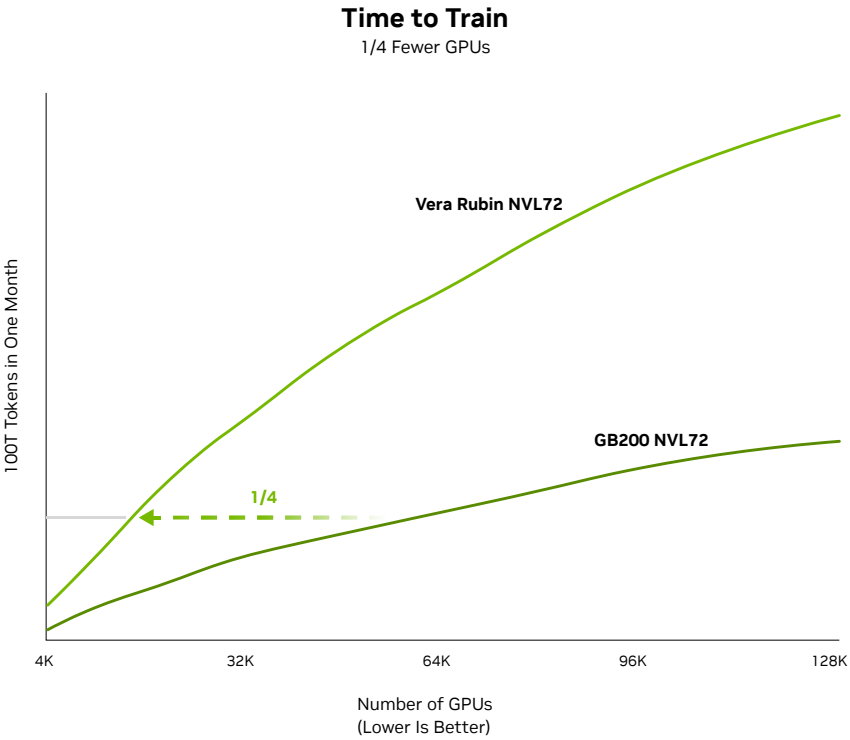
Scaling agentic and reasoning AI demands more intelligence from every megawatt of data center power. On Kimi K2-Thinking, the NVIDIA Vera Rubin NVL72 platform delivers up to 10x more tokens per megawatt than NVIDIA GB200 NVL72, expanding the throughput-versus-interactivity frontier that defines where modern workloads run. This leap in throughput efficiency allows enterprises to serve more users while delivering richer agentic experiences within the same power footprint.



LLM Inference performance subject to change. Tokens per second per MW based on Kimi K2-Thinking model using 32K/8K ISL/OSL comparing NVIDIA GB200 NVL72 and NVIDIA Vera Rubin NVL72.

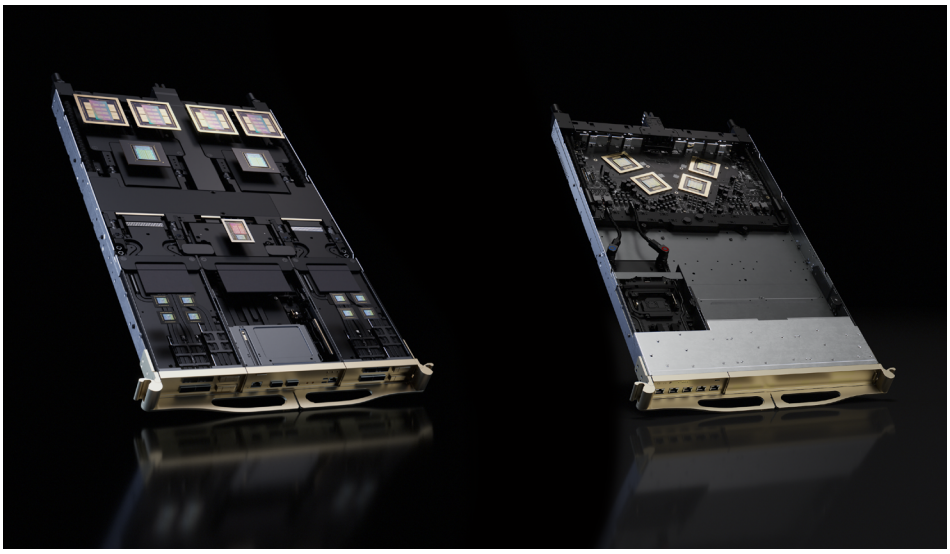
Built for Frontier MoE Pretraining

Pretraining frontier mixture-of-experts (MoE) models is the most compute-intensive workload in AI today, and it demands a system engineered for it. The NVIDIA Vera Rubin NVL72 platform trains these models using one-fourth as many GPUs as the NVIDIA GB200 NVL72. For organizations building AI factories, this leap in architectural efficiency translates into more research cycles per quarter, faster time to the next model generation, and greater output from every AI factory.



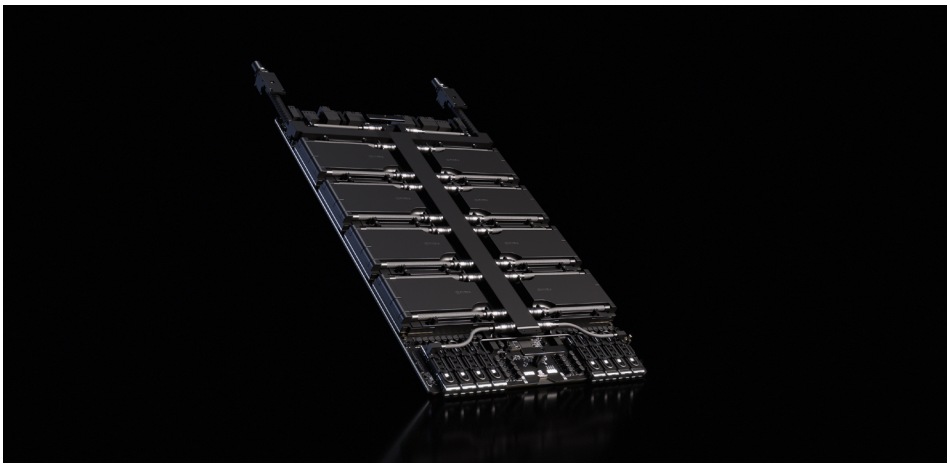
Projected performance subject to change. Number of GPUs based on a 10T MoE model trained on 100T tokens in a fixed time frame of 1 month comparing NVIDIA GB200 NVL72 and NVIDIA Vera Rubin NVL72.

Third-Generation MGX Rack



The third-generation NVIDIA MGX rack architecture sets the production standard for AI infrastructure with engineering breakthroughs across mechanical, power, and cooling design. Built for the resiliency and energy efficiency AI factories demand, its single-wide, PCB-based design unlocks fully modular, cable-free, hose-free, and fanless compute and NVLink switch trays—paired with operational resiliency features that support swapping switch trays without interrupting rack operation. Dynamic power steering routes power between CPUs, GPUs, and NVLink switch trays based on real-time demand, while Intelligent Power Smoothing uses rack-level capacitors to cushion the steep load swings from AI workloads. And with 100% liquid cooling engineered for 45°C warm-water inlet temperatures, MGX drops seamlessly into existing liquid-cooled data centers and enables ambient-air, dry-cooler designs that reduce PUE and redirect power from cooling overhead to token generation.

NVIDIA HGX Rubin NVL8



Key Features

- 8 NVIDIA Rubin GPUs
- 2.3 TB of HBM4 memory
- 3.6 TB/s NVLink between GPUs via NVLink Switch chip

Supercharging AI and HPC for Every Data Center

The NVIDIA HGX™ platform brings together the full power of NVIDIA GPUs, NVIDIA Vera CPUs, NVIDIA NVLink, NVIDIA networking, and fully optimized AI and high-performance computing (HPC) software stacks to provide the highest application performance and drive the fastest time-to-insights for every data center.

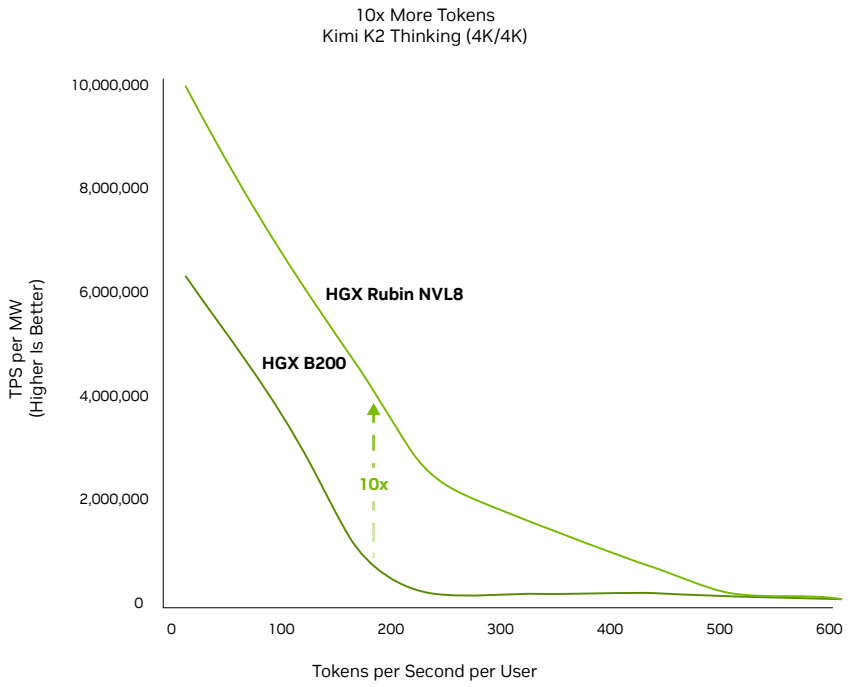
The NVIDIA HGX Rubin NVL8 integrates eight NVIDIA Rubin GPUs with sixth-generation high-speed NVLink interconnects, delivering up to 10x more token factory throughput versus HGX B200 and matching its training performance with 4x fewer GPUs. NVIDIA Rubin-based HGX systems are designed for the most demanding agentic AI, data analytics, and HPC workloads. NVIDIA HGX Rubin NVL8 can be paired with either NVIDIA Vera CPUs—configured as HGX Vera Rubin NVL8—or with x86-based CPU baseboards.

Accelerating the Next Generation of Agentic AI

Boost Token Factory Throughput With HGX Rubin NVL8

Serving agentic AI and reasoning models at scale demands extreme inference throughput. With architectural innovations including 400 PFLOPS of NVFP4 compute, 3x higher memory bandwidth at 176 TB/s, and 2x higher NVLink Switch bandwidth at 28.8 TB/s for high-throughput inter-GPU communication, HGX Rubin NVL8 delivers 10x higher token factory throughput than HGX B200. This leap in performance allows AI factories to serve more users, maximize token revenue, and lower cost per token.

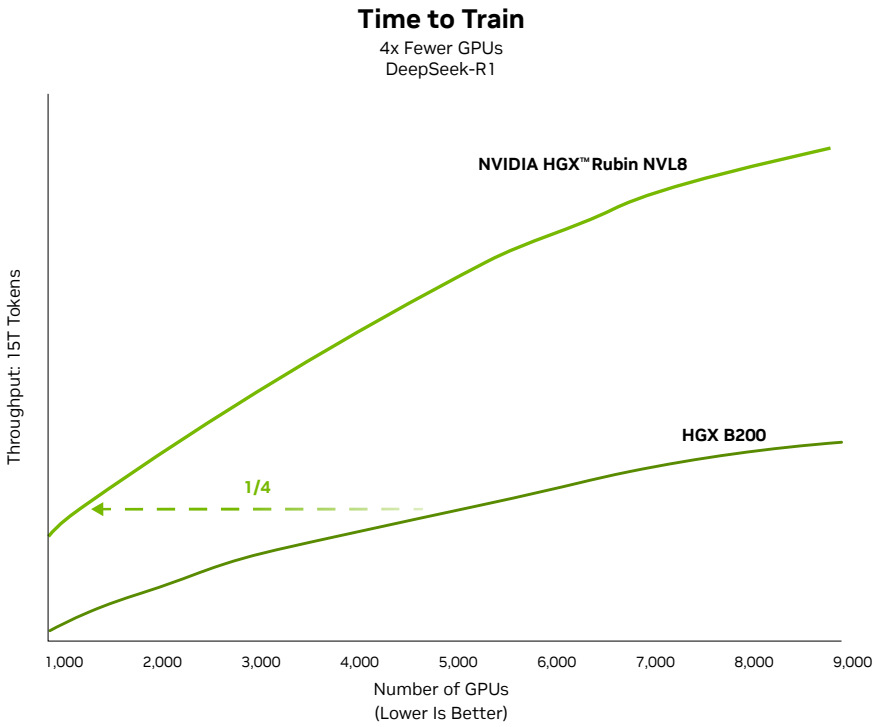
Token Factory Throughput With HGX Rubin NVL8



Projected performance subject to change. Kimi K2-Thinking model with FTL<=500ms, ISL=4K, OSL=4K. HGX Rubin NVL8 with Sparse NVFP4, HGX B200 with Dense NVFP4.

Train Next-Generation AI Models With 4x Fewer GPUs

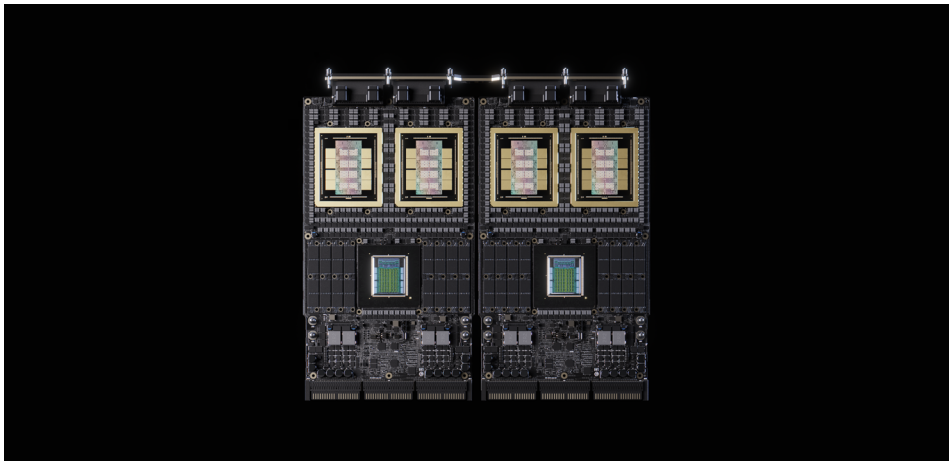
HGX Rubin NVL8 brings breakthrough MoE pretraining to the 8 GPU server form factor, training next-generation agentic AI models with 4x fewer GPUs, enabled by architectural innovations including 4x more NVFP4 training FLOPS, 1.6x more high-speed HBM memory capacity, and 2x more NVLink bandwidth versus HGX B200. This leap in training efficiency enables organizations to train more models within the same infrastructure footprint, lower the cost of model development, and maximize the return on AI infrastructure investment.



Projected performance subject to change. Number of GPUs based on DeepSeek-R1 pretrained on 15T tokens with 4K sequence length.

NVIDIA Vera Rubin NVL4

Accelerating Automated Scientific Discovery and Agentic AI



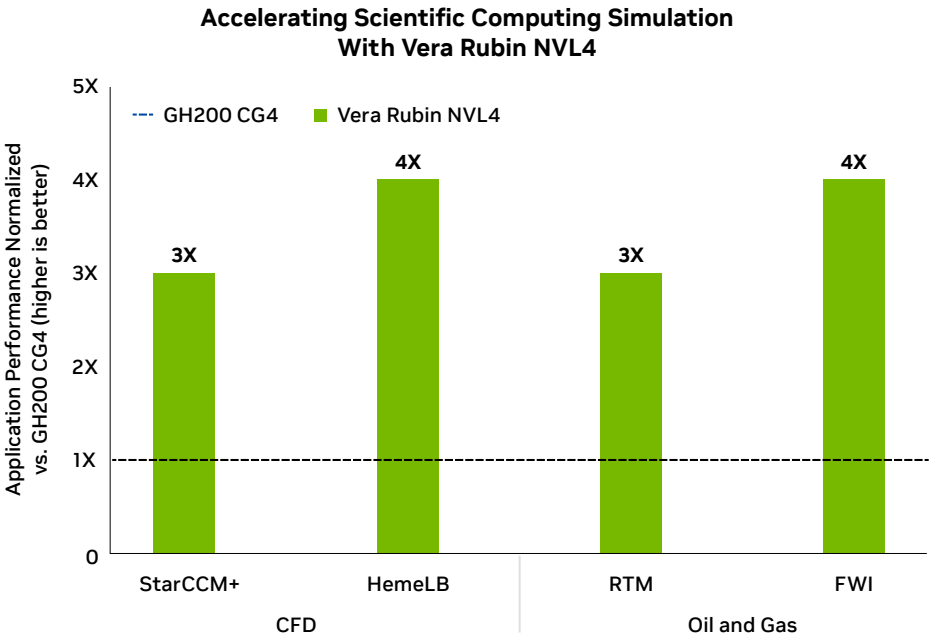
NVIDIA Vera Rubin NVL4 delivers revolutionary performance through four NVIDIA Rubin GPUs interconnected by a second-generation NVLink bridge running sixth-generation NVIDIA NVLink, paired with two NVIDIA Vera CPUs over NVLink-C2C. Compatible with liquid-cooled NVIDIA MGX™ modular servers, it delivers up to 4x the performance for scientific computing simulation, 6x for AI-for-Science training, and 8x for AI-for-Science inference versus NVIDIA Grace Hopper.

Key Features

- Two NVIDIA Vera CPUs
- Four NVIDIA Rubin GPUs
- 2.5 TB of fast-access memory
- 4x higher performance for scientific computing simulation

Advancing Scientific Breakthroughs

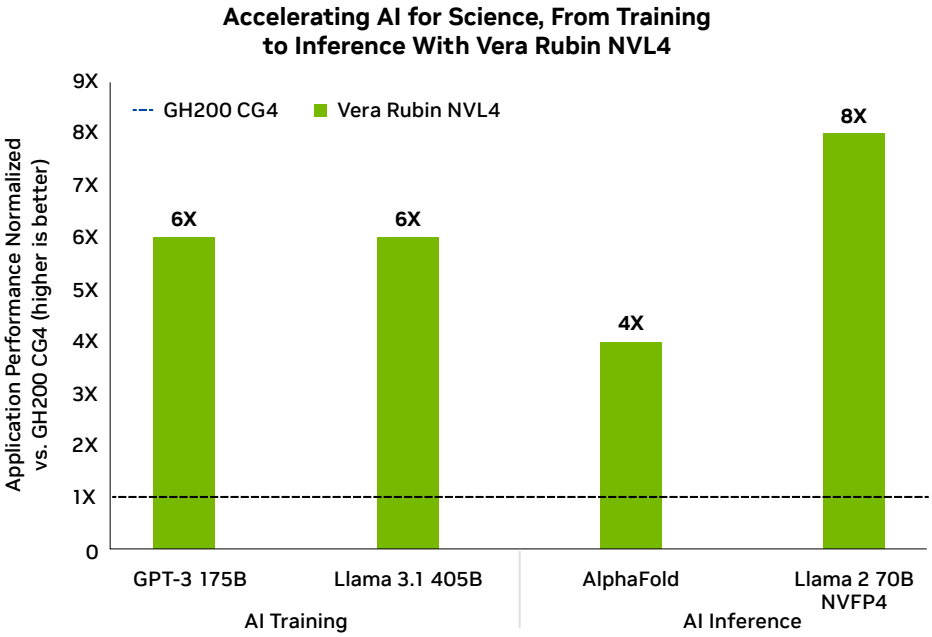
The next era of scientific computing is defined by problems too large and too complex to solve at the pace discovery demands. The NVIDIA Vera Rubin NVL4 is built for it, with 2.5 TB of fast-access memory in a single compute node, engineered to accelerate scientific computing simulations. Delivering up to 4x the application performance of NVIDIA Grace Hopper across computational fluid dynamics and oil and gas workloads, it gives researchers the headroom to model larger, simulate faster, and accelerate the pace of scientific discovery.



Projected performance on Vera Rubin NVL4. Subject to change. Vera Rubin NVL4 vs GH200 CG4 (4x Grace Hopper Superchips). GH200 CG4: Network at 200Gbps/GPU, Power at 700 Watts/GPU. Vera Rubin NVL4: Network at 800Gbps/GPU, Power at 1800 Watts/GPU. Star-CCM+ is a commercial computational fluid dynamics and Multiphysics simulation tool. HemeLB is an HPC biomedical simulation code for blood-flow / hemodynamics modeling. RTM: Reverse Time Migration and FWI: Full Wave Inversion. RTM solves the acoustic or elastic wave equation forward and backward in time to create an image of underground structures. FWI iteratively updates a subsurface model, usually velocity such that seismic waveforms match observed data.

AI for Science - From Training to Deployment

NVIDIA Vera Rubin NVL4 is built for the AI-for-Science workflows that drive modern scientific discovery, including fine-tuning domain-specific models, generating and using synthetic data, accelerating simulations, and running high-throughput inference across large experimental spaces. With 2.5 TB of fast-access memory in a single compute node, NVIDIA Vera Rubin NVL4 delivers up to 6x faster training on models like GPT3-175B and Llama 3.1 405B and up to 8x faster inference on Llama 2 70B versus Grace Hopper, giving researchers the power to build bigger models and the speed to put them to work.



Projected performance on Vera Rubin NVL4. Subject to change. Vera Rubin NVL4 compared to GH200 CG4 (4x Grace Hopper Superchips). Assumes CG4 network at 200G/GPU and Vera Rubin NVL4 network at 800G/GPU. Power assumptions CG4 = 700 W/GPU; Vera Rubin NVL4 = 1800 W/GPU. Precision info – GPT3175B: FP8, Llama 3.1 405B: FP8, Alphafold: BF16, Llama2 70B: NVFP4, ISL/OSL – 1024/1536 with in-flight batching, chunk size 1024.

Technical Specifications[1]

System Specifications	NVIDIA Vera Rubin NVL72	NVIDIA Vera Rubin Superchip	NVIDIA Vera Rubin NVL4
Configuration	72 NVIDIA Rubin GPUs | 36 NVIDIA Vera CPUs	2 NVIDIA Rubin GPUs | 1 NVIDIA Vera CPU	4 NVIDIA Rubin GPUs | 2 NVIDIA Vera CPUs
NVFP4 Inference	3,600 PFLOPS	100 PFLOPS	200 PFLOPS
NVFP4 Training[2]	2,520 PFLOPS	70 PFLOPS	140 PFLOPS
FP8/FP6 Training[2]	1,260 PFLOPS	35 PFLOPS	70 PFLOPS
INT8[2]	18 POPS	500 TOPS	1 POPS
FP16/BF16[2]	288 PFLOPS	8 PFLOPS	16 PFLOPS
TF32[2]	144 PFLOPS	4 PFLOPS	8 PFLOPS
FP32	9,360 TFLOPS	260 TFLOPS	520 TFLOPS
FP64	2,400 TFLOPS	67 TFLOPS	133 TFLOPS
FP32 SGEMM[3]	28,800 TFLOPS	800 TFLOPS	1.6 PFLOPS
FP64 DGEMM[3]	14,400 TFLOPS	400 TFLOPS	800 TFLOPS
GPU Memory | Bandwidth	20.7 TB HBM4 | 1,580 TB/s	576 GB HBM4 | 44 TB/s	1.1 TB HBM4 | 88 TB/s
NVIDIA NVLink	Sixth Generation		
NVLink Bandwidth	260 TB/s (NVLink 6 Switch Bandwidth)	7.2 TB/s	14.4 TB/s
NVLink-C2C Bandwidth	65 TB/s	1.8 TB/s	1.8 TB/s
CPU Core Count	3,168 custom NVIDIA Olympus cores (Arm®compatible)	88 custom NVIDIA Olympus cores (Arm® compatible)	176 custom NVIDIA Olympus cores (Arm® compatible)
CPU Memory	54 TB LPDDR5X	1.5 TB LPDDR5X	1.5 TB LPDDR5X
Networking Bandwidth (Scale Out)	28.8 TB/s	0.8 TB/s	0.4 TB/s
Total NVIDIA + HBM4 Chips	1,296	30	60

Technical Specifications[1]

System Specifications	NVIDIA HGX Vera Rubin NVL8	NVIDIA HGX Rubin NVL8
Configuration	8x NVIDIA Rubin SXM with Single-Socket Vera CPU	8x NVIDIA Rubin SXM
CPU | Core Count	NVIDIA Vera CPU | 88 Custom NVIDIA Olympus Cores (Arm® compatible) with Spatial Multithreading (SMT)	x86 CPU[4]
CPU Memory | Bandwidth	1.5TB LPDDR5X | 1.2 TB/s	x86 CPU[4]
NVFP4 Inference	400 PFLOPS	
NVFP4 Training[2]	280 PFLOPS	
FP8/FP6 Training[2]	140 PFLOPS	
INT8[2]	2 POPS	
FP16/BF16[2]	32 PFLOPS	
TF32[2]	16 PFLOPS	
FP32	1,040 TFLOPS	

FP64	265 TFLOPS
FP32 SGEMM[3]	3,200 TFLOPS
FP64 DGEMM[3]	1,600 TFLOPS
GPU Memory | Bandwidth	2.3 TB HBM4 | 176 TB/s
NVLink Switch Bandwidth	28.8 TB/s
NVIDIA NVLink	Sixth Generation
Networking Bandwidth	1.6 TB/s

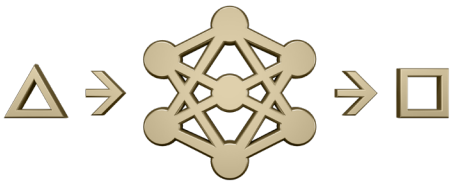
Individual GPU Specifications	**NVIDIA Rubin GPU**
NVFP4 Inference	50 PFLOPS
NVFP4 Training[2]	35 PFLOPS
FP8/FP6 Training[2]	17.5 PFLOPS
INT8[2]	250 TOPS
FP16/BF16[2]	4 PFLOPS
TF32[2]	2 PFLOPS
FP32	130 TFLOPS
FP64	33 TFLOPS
FP32 SGEMM[3]	400 TFLOPS
FP64 DGEMM[3]	200 TFLOPS
NVLink Bandwidth	3.6 TB/s
NVIDIA NVLink	Sixth Generation
GPU Memory | Bandwidth	288 GB HBM4 | 22 TB/s

1. Preliminary information. All values are up to and subject to change. NVFP4 Inference specification is sparse.
2. Dense specification.
3. Peak performance using NVIDIA Tensor Core-based emulation algorithms.
4. CPU and memory specs are defined by OEM offerings.

Explore the Technological Breakthroughs of NVIDIA Vera Rubin

Transformer Engine

The Rubin GPU features a new Transformer Engine (TE) with hardware-accelerated adaptive compression to boost NVFP4 performance while preserving accuracy. This enables up to 50 petaFLOPS of NVFP4 inference. Fully compatible with NVIDIA Blackwell, the Transformer Engine ensures seamless upgrades, so previously optimized codes transition effortlessly to the Vera Rubin platform.



Third-Generation Confidential Computing

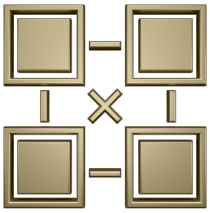
The third generation of NVIDIA Confidential Computing expands security to full-rack scale with NVIDIA Vera Rubin NVL72. This platform creates a unified, trusted execution environment across all 36 NVIDIA Vera CPUs, 72 NVIDIA Rubin GPUs, and the NVIDIA NVLink™ fabric that seamlessly connects them. The platform maintains data security across CPU, GPU, and NVLink domains. With attestation services for cryptographic proof of compliance, it combines massive scale with uncompromised protection, all to protect the world's largest proprietary models, training data, and inference workloads

Learn More About NVIDIA Confidential Computing

Sixth-Generation NVLink and NVLink Switch

The sixth-generation NVLink delivers a major leap for NVIDIA's high-speed GPU interconnect fabric that unifies 72 NVIDIA Rubin GPUs into a single performance domain. Doubling NVIDIA Blackwell's performance, the Rubin GPU delivers 3.6 terabytes per second (TB/s) of bandwidth per GPU and 260 TB/s of connectivity with low latency to facilitate faster communication. Combined with NVIDIA® Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ that reduces network congestion by up to 50 percent for collective operations, this next-generation interconnect accelerates training and inference for the world's largest models, at scale and without compromise.

Learn More About NVIDIA NVLink and NVLink Switch



Second-Generation Reliability, Availability, and Serviceability (RAS) Engine

The NVIDIA Vera Rubin platform delivers rack-scale resiliency with advanced reliability features. NVIDIA Rubin GPUs feature a dedicated second-generation RAS engine for proactive maintenance and real-time health checks without downtime. NVIDIA Vera CPUs add enhanced serviceability with small-outline compression-attached memory modules (SOCAMM) LPDDR5X and in-system tests for the CPU cores. The rack introduces modular, cable-free tray designs for 18x faster assembly and serviceability versus NVIDIA Blackwell, combined with intelligent resiliency and software-defined NVLink routing, which ensures continuous operation and reduces maintenance overhead.

NVIDIA Vera CPU

The NVIDIA Vera CPU is engineered for data movement and agentic reasoning across accelerated systems, with full confidential computing support. It pairs seamlessly with NVIDIA GPUs or operates independently for analytics, cloud, orchestration, storage, and high-performance computing (HPC) workloads. Vera combines 88 NVIDIA-designed cores, up to 1.2 TB/s of LPDDR5X memory bandwidth, and NVIDIA Scalable Coherency Fabric to deliver predictable, energy-efficient performance for data- and memory-intensive workloads with full Arm® compatibility. Integrated NVIDIA NVLink-C2C connectivity enables high-bandwidth, coherent CPU–GPU memory access to maximize system utilization and efficiency.

Learn More About NVIDIA Vera

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™ provides an optional software stack for accelerating every aspect of operations, from configuring NVIDIA Vera Rubin systems to integrating with facilities to managing clusters and workloads, giving enterprises greater control over cooling and power events and redefining infrastructure resiliency.

AI-Ready Enterprise Platform

NVIDIA AI Enterprise brings together microservices, frameworks, and libraries for AI development, along with advanced GPU orchestration and infrastructure management, into a fully supported, production-ready, commercial software suite. Deploy leading open source tools and AI models with confidence, improve productivity and time to value, and run AI workloads at scale with optimized resource utilization.

Part of NVIDIA AI Enterprise, NVIDIA NIM™ is a set of easy-to-use microservices that accelerate the deployment of foundation models on any cloud or data center and help keep your data secure. Enterprises that run their businesses on AI rely on the security, support, manageability, and stability provided by NVIDIA AI Enterprise to ensure a smooth transition from pilot to production.

Together with the NVIDIA Vera Rubin platform, NVIDIA AI Enterprise not only simplifies the building of an AI-ready platform but also accelerates time to value.

Learn about AI workload workflows with NVIDIA AI Enterprise via NVIDIA Launchpad's hands-on labs.

Ready to Get Started?

To learn more about NVIDIA Vera Rubin, visit nvidia.com/vera-rubin/