AIVRES

KRS8000V3

AI Rack Based on NVIDIA GB200 NVL72

NVIDIA

KRS8000V3 is an L11 AI rack based on NVIDIA GB200 NVL72, integrating 36 Grace CPUs and 72 Blackwell GPUs in a rack-scale, liquid-cooled architecture, achieving breakthrough performance in real-time trillion-parameter large language model (LLM) inference and training.

KRS8000V3 with GB200 NVL72 is poised to redefine performance benchmarks for AI, HPC, and data analytics, making it a pivotal component in next-generation computing infrastructure.

4X	**25X**	**30X**	**18X**
LLM Training	**Energy Efficiency**	**LLM Inference**	**Data Processing**
vs. H100	vs. H100	vs. Nvidia H100 Tensor Core GPU	vs. H100

KEY FEATURES

Blackwell Rack-Scale Architecture:

- Connects 72 Blackwell GPUs via NVIDIA® NVLink®.
- Delivers 130 TB/s of low-latency communication bandwidth.
- Acts as a single massive GPU for efficient processing.

Performance Enhancements:

- Achieves 30X faster real-time trillion-parameter LLM inference compared to previous generations.
- 4X faster training for large language models using FP8 precision.

Data Processing:

- Includes a hardware decompression engine supporting LZ4, Deflate, and Snappy formats.
- Provides up to 800 GB/s decompression throughput.
- Achieves 18X faster performance for database query benchmarks compared to traditional CPUs.

Memory and Bandwidth:

- Offers 8 TB/s high memory bandwidth.
- Grace CPU NVLink-C2C interconnect ensures high-speed data transfer.

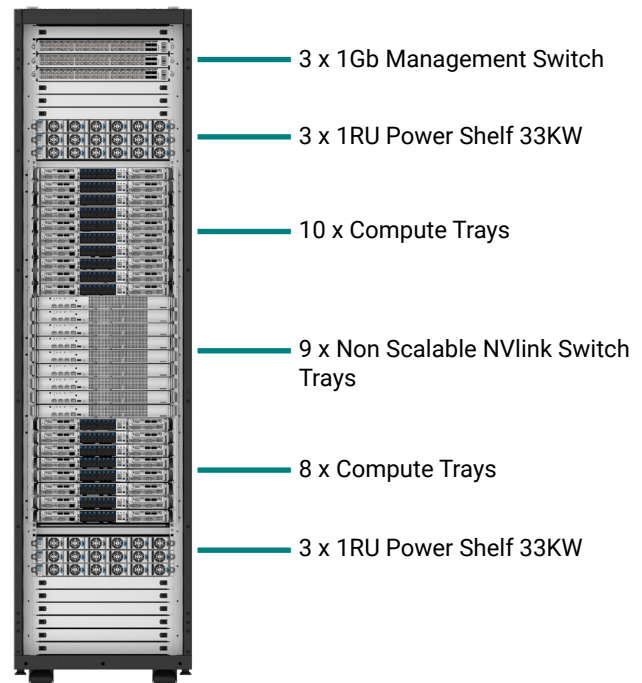## USE CASES

AI Training and Inference:

- Optimized for training large-scale models and performing real-time inference.
- Particularly effective for transformer-based models and other resource-intensive AI applications.

Big Data Analytics:

- Enhances the efficiency of big data processing pipelines.
- Reduces storage costs and processing times for large datasets.

Scientific and Engineering Simulations:

- Speeds up simulations in various domains, including fluid dynamics, circuit design, and more.

3 x 1Gb Management Switch

3 x 1RU Power Shelf 33KW

10 x Compute Trays

9 x Non Scalable NVlink Switch Trays

8 x Compute Trays

3 x 1RU Power Shelf 33KW

Technical Specifications

Model	NVIDIA® GB200 NVL72
Configuration	36 Grace CPU and 72 Blackwell GPUs
FP4 Tensor Core2	1,440 PFLOPS
FP8/16 Tensor Core2	720 PFLOPS
INT8 Tensor Core2	720 POPS
FP16/BF16 Tensor Core2	360 PFLOPS
TF32 Tensor Core	180 PFLOPS
FP32	6,480 TFLOPS
FP64	3,240 TFLOPS
FP64 Tensor Core	3,240 TFLOPS
GPU Memory | Bandwidth	Up to 13.5 TB HBM3e | 576 TB/s
NVLink Bandwidth	130TB/s
CPU Core Count	2,592 Arm® Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17 TB LPDDR5X | Up to 18.4 TB/s

Rack Specifications

Dimensions	600mm (23.6") W x 2236mm (84") H x 1068mm (42") L
NVL Config	72x 1
NV Switch tray	2x QM3
NVL Cartridge	4
Rack Type (Per Rack)	9x 1U NVlink Switch Trays 18x 1U Compute Trays 6x 1U Powershelf
N-S Networking	Support 2x FHFL PCIe 5.0 x16 (BF3 or NIC Card)
E-W Networking	2x Mezzanine card on board Support 4x HHHL PCIe 5.0 x16 with 400G bandwidth
Power-Shelf	6x 33kW
Busbar	1400A
Fan	CPU region: 8x 12V 4056 hot-swap fans with N+1 redundancy
Management	DC-SCM BMC management module

CPU Tray (MGX Base Tray)

	4x Blackwell GPUs + 2x Grace CPU
	1,728 GB Memory
	1U Liquid cooled
	18 per Rack
Storage	8x E1.S
Rear I/O	1x USB 3.0, 1x Mgmt I/O , 1x RJ45, 1x mini display port
M.2	1x Onboard NVMe / SATA M.2 (Optional)
OCP	Support 1x OCP 3.0 (Optional)

Switch Trays

2x NVLink X-800 Switch
14.4TB/s total Bandwidth
1U Liquid Cooled
Front IO: 2x RJ45, 1x USB, 1x UART
9 per Rack