DATASHEET

KRS8000V3

AI Rack Based on NVIDIA GB200 NVL72

KRS8000V3 is an L11 AI rack based on NVIDIA GB200 NVL72, integrating 36 NVIDIA Grace™ CPUs and 72 NVIDIA 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.

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

KEY FEATURES

NVIDIA 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.

Advanced 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.
- NVIDIA Grace CPU NVLink®-C2C interconnect ensures high-speed data transfer.

USE CASES

AI Training and Inference:

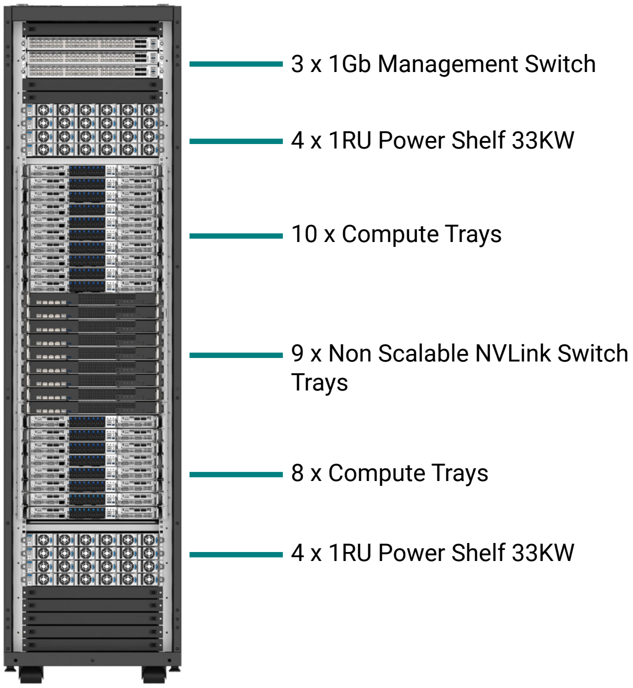
- Optimized for training large-scale models and performing real-time inference.
- Particularly effective for transformer-based models and other resource-intensive AI applications.

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.



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 13.39 TB HBM3e | 576 TB/s
NVLink Bandwidth	130TB/s
CPU Core Count	2,592 Arm® Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17.28 TB LPDDR5X | Up to 18.4 TB/s

Rack Specifications

Dimensions	600mm (23.6") W x 2236mm (88") H x 1200mm (47.2") L
Weight	1,553.27 kg (3,434.37 lbs)
NVL Config	72x 1
NV OOB Switch	Option 1: 3 x SN2201 DC Option 2: 4 x SN2201 DC
NVL Cartridge	4
Rack Type (Per Rack)	9x 1U NVLink Switch Trays 18x 1U Compute Trays 8x 1U Power Shelves
Power-Shelf	8x 33kW
Busbar	1,400A
Rack Manifold	Option 1: 44RU, BF Option 2: 44RU, TF
CDU	Option 1: L2L In-Row Option 2: L2L In-Rack Option 3: L2A Sidecar

Compute Tray

CPU/GPU	2x Grace CPUs + 4x Blackwell GPUs
Cooling	1U liquid cooled
Storage	8x E1.S NVMe SSDs
M.2	1x Onboard NVMe / SATA M.2
Front I/O	1 x USB 3.0, 1x Mgmt I/O, 1x RJ45, 1x Mini Display Port
N-S Networking	2x FHFL PCIe 5.0 x16 (BF3 or NIC Card)
E-W Networking	2x Mezzanine card on board 4x HHHL PCIe 5.0 x16 with 400G bandwidth
Fan	CPU region: 8x 12V 4056 hot-swap fans with N+1 redundancy
Management	DC-SCM BMC management module
TPM	Supports TPM 2.0

Switch Tray

Type	2x NVLink X-800
Bandwidth	14.4TB/s
Cooling	1U liquid cooled
Front I/O	2x RJ45, 1x USB, 1x UART