AIVRES

Accelerated Solutions for AI

Industry-Leading AI Technology
on Cutting-Edge GPU Platforms

Aivres and NVIDIA are working together to develop solutions to help businesses harness the power of AI. Through integration and innovation with NVIDIA's leading-edge AI technology, Aivres server platforms deliver breakthrough performance, efficiency, and flexibility to tackle the most challenging AI workloads in the industry today – large models, machine learning, generative AI, and beyond.

AI Platforms for Every Scenario

Broad product portfolio for inference, training, large language models, Gen-AI, digital twins, real-time graphics.

Supporting Workloads at Any Scale

Flexible AI building blocks to rack-scale supercomputing clusters, for enterprise AI to the largest training models.

Extreme GPU Density Platforms

Up to 8 of the latest NVIDIA GPUs in 6U with advanced interconnect for breakthrough speed and computing.

Industry-Leading Performance

Best-in-class for training and inference performance benchmarks.

Rack-Scale Solution Based on NVIDIA GB200 NVL72

Exascale Rack for Trillion-Parameter LLM Inference

KRS8000

- NVIDIA Blackwell rack-scale architecture connects 72 Blackwell GPUs via NVIDIA® NVLink™
- 130 TB/s of low-latency communication bandwidth
- Acts as a single massive GPU for efficient processing
- 4X faster training for large language models using FP8 precision
- Up to 800 GB/s decompression throughput
- Achieves 18X faster performance for database query benchmarks vs traditional CPUs
- 8 TB/s high memory bandwidth
- NVIDIA Grace™ CPU NVLink®-C2C interconnect ensures high-speed data transfer

NVIDIA HGX™ H200 8-GPU Server

6U Modular AI Pod for Extra-Large Scale AI Training Models

KR6288

- Delivers 32 PFLOPS industry-leading AI performance
- 6U modular form factor for data center deployment
- Unified GPU modules for heterogeneous computing, deploy and scale as needed
- Lossless scalability with PCIe 5.0 slots for NDR 400Gb/s InfiniBand
- 300TB massive local storage with 24x 2.5-inch SSDs, up to 16x NVMe

PCIe GPU Servers

Multi-Config 4U for Gen-AI, Training & Inference

KR4268

- Supports NVIDIA L40S, NVIDIA H100 NVL
- Supports up to 10 double-width PCIe GPUs for performance up to 15 PFLOPS
- PCIe 5.0 architecture supporting E3.S
- CXL1.1 supports storage-level memory expansion
- Flexible topologies and configurations to support various applications

2U Flexible Mainstream Platform for Enterprise AI

KR2280

- Supports NVIDIA L40S, NVIDIA H100 NVL
- Supports 4 double-width or 8 single-width PCIe GPUs
- CXL1.1 supports storage-level memory expansion
- Versatile 2.5", 3.5", E3.S all-flash storage options
- Cold-plate liquid cooling for enhanced energy-efficiency

NVIDIA H200 NVL Tensor Core GPU Server

KR6268

- Supports 8x NVIDIA H200 NVL GPUs via PCIe 5.0
- 6U modular form factor for data center deployment
- Unified GPU modules for heterogeneous computing, deploy and scale as needed
- Up to 24x 2.5-inch SSDs, 16x E3.S
- Delivers high performance for enterprise AI

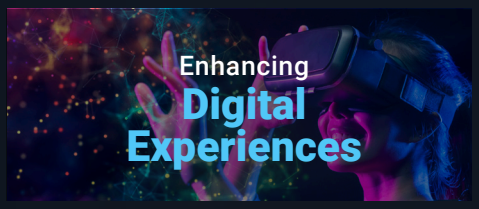
NVIDIA HGX™ B200 8-GPU Server

10U High-Performance AI Training Server

KR9288

- Supports NVIDIA HGX™ B200 with Blackwell 8-GPU and NVLink™ interconnect
- Modular design for easy deployment and maintenance
- Supports up to 12x full-height expansion cards
- Supports cold-plate liquid cooling on 2x NVSwitch, 2x CPU

Empowering New Possibilities with Aivres AI



Sustainability through Enhanced Cooling

Aivres integrates a combination of liquid cooling methods to maximize the energy efficiency of our highest performance density AI solutions and platforms so you can supercharge your AI workloads while maintaining a low PUE.

Server level cold-plate cooling

Direct-to-chip thermal transfer on heat components helps maintain optimal temperatures in high-heat workload scenarios.

Rack-scale air-liquid hybrid cooling

Versatile solution with combined efficiency of both mediums that is easy to implement into existing air-cooled infrastructure.

aivres.com/artificial-intelligence/nvidia
linkedin.com/company/aivres

AIVRES