R945 4U All-round AI Server

The R945 is a high-end 4U rack-mounted server powered by 4th/5th generation Intel® Xeon® Scalable processors. It supports up to 8 double-slot, full-height, and full-length professional GPU AI accelerator cards, effectively meeting the demands of ultra-high-density AI computing applications. With its superior hardware configuration and optimized design, it is highly suitable for large-scale cluster deployments, delivering reliable and efficient performance for intensive AI workloads.



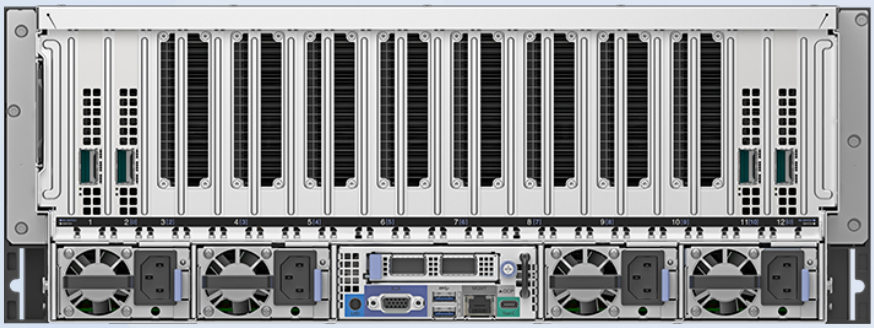
Extremely Cost-Effective AI Computing Server

Specifications

Model	R945
Processor	Supports 2*4th/5th Gen Intel Xeon® Scalable Processors (TDP 350W)
Memory	Up to 32*DDR5 slots, up to 5600MHz
RAID	Supports BCM 95 series RAID cards.
Storage	Up to 24*3.5" hot-swappable drives (16*NVMe compatible) Internal: 2*SATA/PCIe M.2
Network Interfaces	Network card: 1*OCP 3.0 with NCSI Optional OCP 3.0 interfaces: 2*GE, 2*10GE, 2*25GE, 2*100GE, 2*200GE
GPU	Supports up to 8*double-slot GPU

Additional Interface	Rear: 1*RJ-45 management port, 2*USB 3.0, 1*VGA Front:2*USB 3.0 + VGA
Power Supply	Supports 2700W CRPS high-efficiency Platinum PSU(hot-swappable, 2+2 redundant) Voltage: 100-240V AC (refer to nameplate)
Fan	8*hot-swappable fans, supports N+1 redundant
Management	Integrated BMC with 1Gbps RJ-45 port Supports Redfish, SNMP, IPMI 2.0, HTML5/iKVM remote control BMC dual-flash backup
OS Support	Windows Server, RHEL, SLES, CentOS, Ubuntu, VMware ESXi, etc.
Operating Temperature	5°C ~ 35°C

Product View



8GPU (8*double-slot) riser model

Product Features

Ultimate Performance and Flexible Scalability

Powered by 4th/5th generation Intel® Xeon® Scalable processors, equipped to support up to 8*double-slot, full-height, and full-length professional GPU AI accelerator cards, it combines powerful hardware configuration with flexible computing matching, tailored specifically to meet the demands of ultra-high-density AI computing applications.

Streamlined Operations and Intelligent Management

It features streamlined operations and intelligent management capabilities to reduce operational complexity and improve maintenance efficiency. The out-of-band visualization management enables remote and precise location tracking of physical devices, further optimizing operational efficiency. It supports automatic crash logging, with online viewing and download capabilities for crash logs, enabling precise fault diagnosis and reducing troubleshooting time. Additionally, it provides real-time health monitoring and reporting for critical components, delivering comprehensive intelligent services throughout the entire product lifecycle.

Extremely Quiet and Energy-Efficient

It features extremely quiet operation and excellent energy efficiency, balancing high performance with low energy consumption and low noise pollution. It supports high-performance online inference with 8*GPUs, realizing seamless interconnection of 8*professional-grade GPU cards to ensure stable and efficient data transmission. Meanwhile, it optimizes GPU Direct efficiency specifically for deep learning training scenarios, effectively reducing data transmission latency and enhancing the overall training efficiency of deep learning models.