2U Server with Dual Intel® Xeon® 6 Processors and 3 x NVIDIA RTX PRO™ 4500 SE

SR710-3

Senao SR710-3 is a 2U GPU server powered by Intel® Xeon® 6 processors. Designed for storage-intensive AI workloads, it supports 16 x U.2 hot-swappable NVMe SSDs and 3 x NVIDIA RTX PRO™ 4500 SE GPU cards, an ideal platform for a compact AI inference host.

GPUs

3 x NVIDIA RTX PRO™ 4500 SE

Inference— FP16:up to 28B/ FP8:upto 56B / NVFP4: up to 112B
Fine-Tuning — LoRA: up to 29B / QLoRA: up to 72B / Full-Param: up to 5B

Performance

Dual Xeon® 6 SP

Dual Intel® Xeon® 6 with P-core/E-core architecture handles AI workloads.

Memory

16 channels DDR5

Up to 32 x DDR5 ECC RDIMM.

Storage

NVMe SSD and RAID

Up to16 x 2.5-inch U.2 NVMe with Intel® VROC premium support.

Management

Modular OCP

DC-MHS design keeps infrastructure serviceable and upgradable.

Scalability

CXL 2.0 Ready

CXL 2.0 support enables future-ready memory and device expansion.

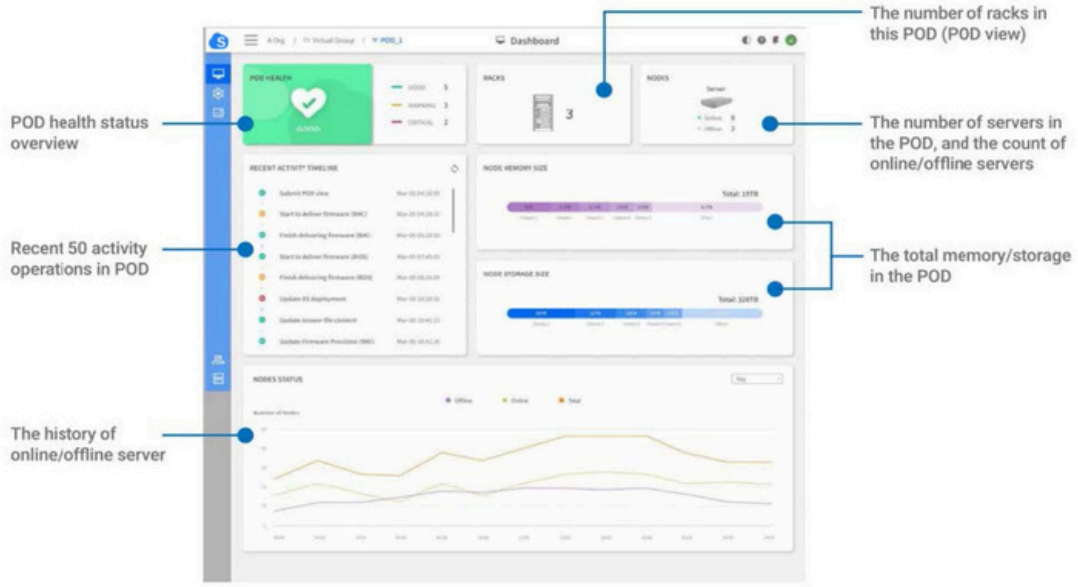
P owe r

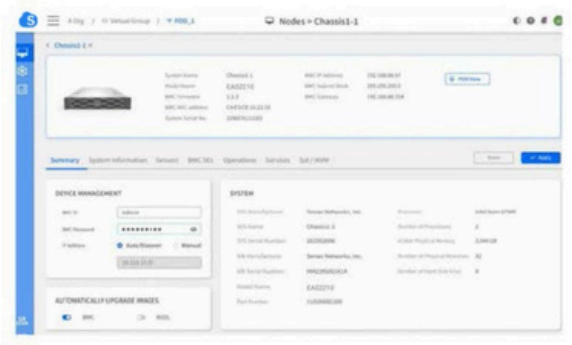
Redundant Power

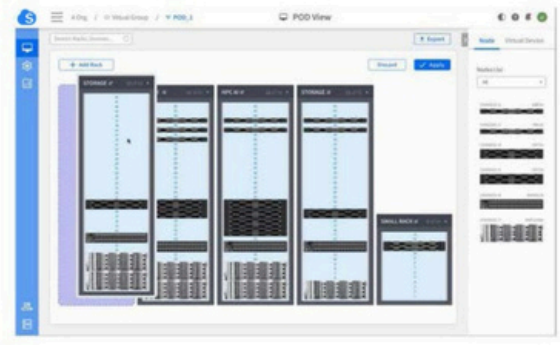
1+1 redundant 2200W CRPS
Designed for 80 PLUS Titanium Efficiency

Server Centralized Management

The server management is the cornerstone of simplified, efficient management tool for your Senao GPU and Rack servers. This powerful, license-free platform provides comprehensive out-of-band management capabilities, empowering IT teams with unparalleled control and visibility over their entire server fleet from a single interface.







Simplified Firmware Updates

Push unified firmware updates to multiple servers simultaneously, ensuring consistency and reducing maintenance overhead.

Visual POD View

Gain a physical-to-logical view of server groupings with intuitive POD visualization, enabling more efficient resource mapping, configuration, and infrastructure monitoring.

Out-of-Band Control

Manage servers even if the operating system is unresponsive or powered off, ensuring continuous uptime and remote troubleshooting.

Remote Diagnostics & Troubleshooting

Gain real-time insights into server health, performance metrics, and system logs, enabling proactive issue identification and rapid resolution from anywhere.

Unified Device Management

Centralized control over all Senao servers, eliminating the need for individual server access.

Scalable Deployment

Designed to manage a growing number of servers, providing seamless scalability as your data center expands.

Technical Specifications

GPU Server	
Model Name	2U GPU Server
Model Number	SR710-3
Form Factor	2U Rackmount
Processor	Intel® Xeon® 6 (6500/6700 Series)
GPU	3 x NVIDIA RTX PRO™ 4500 SE
Memory	32 x DIMM slots DDR5 ECC RDIMM 6400 MT/s 1DPC 5200 MT/s 2DPC
Drive Bays	16 x 2.5-inch NVMe PCIe Gen5 Hot-Swap drive Intel® VROC Premium
System Cooling	Air Cooling
Available Expansion Slot	1 x PCIe Gen5 x 8 HHHL 1 x PCIe Gen5 x 16 FHHL 2 x OCP3.0 Gen5 x 8
I/O Port	Front: 2 x USB3.0 Type A Rear: 1 x USB3.0 Type A 1 x Mini Display 1 x USB Type A Console 1 x MGNT RJ45 GE
Power Supply	1+1 CRPS 2200W 80 PLUS Titanium Level
Security	TPM 2.0 PFR 2.0 Intrusion Detection
Chassis	438 x 790 x 88 mm Weight: Approx. 25 kg
Environment	Operating Temperature: 10–35°C Storage Temperature: -20–60°C Relative Humidity: 10–90% RH, non-condensing

Technical Specifications

GPU Server	
Model Name	2U GPU Server
Model Number	SR710-3
System Management	IPMI 2.0 Management Redfish Management HTTPS Web Management IPv4/ IPv6 Address Dashboard Hardware Sensor Monitor KVM console Serial over LAN Virtual Media Firmware Upgrade(BMC/BIOS) Event Log & Alert (SMTP/ SNMP/Syslog) LDAP/ Windows AD/ Radius User Account Management System Firewall (IP Access Control) Hardware Root of trust (RoT) Backup & Restore Configuration Protocol support: SNMP, SMTP, DHCP, NTP, DNS, Dynamic DNS Intel Platform Feature: Platform Firmware Resilience (PFR), Autonomous Crash Dump (ACD)
Central Management	Server Management

High-Performance AI & Graphics Acceleration

GPU Support: 3 x NVIDIA RTX PRO™ 4500 SE

LLM Workload Capacity:
Inference — FP16: up to 28B / FP8: up to 56B / NVFP4: up to 112B
Fine-Tuning — LoRA: up to 29B / QLoRA: up to 72B / Full-Param: up to 5B

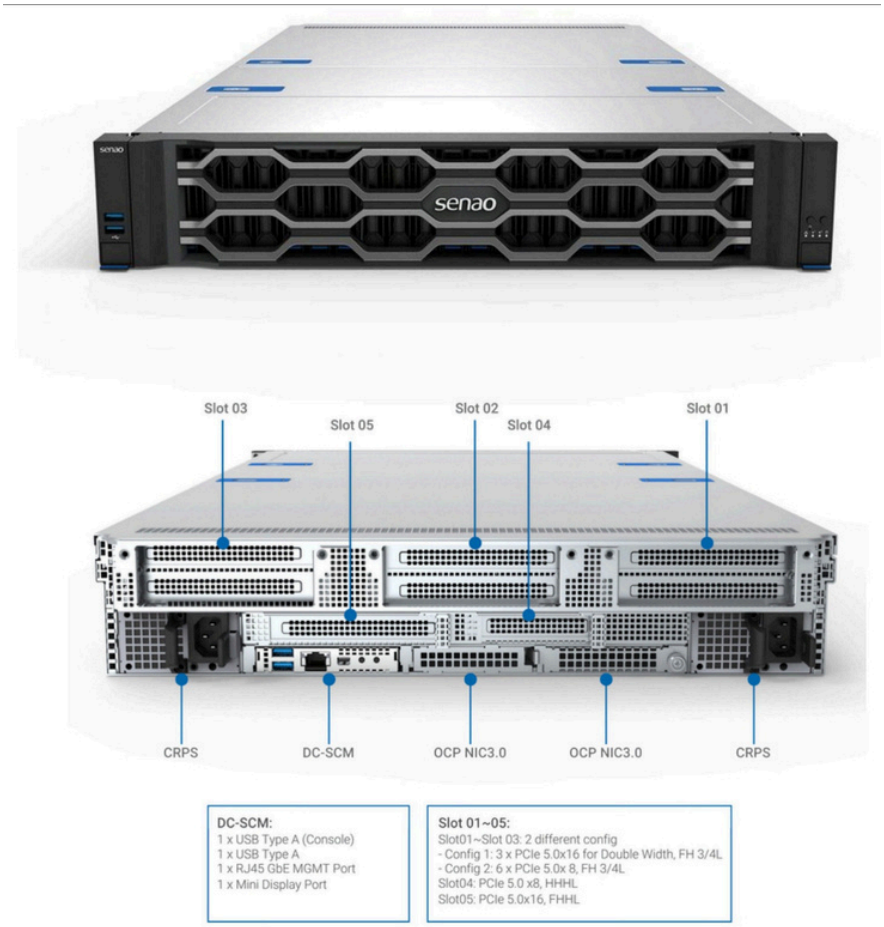
Target Workloads:

Parallel Small-Model Hosting
Concurrently serves several 7B–13B models for separate applications — customer-service chatbots, document summarization, text classification — or a single 28B-class model at moderate concurrency. Also works well as a shared development, test, and POC environment for teams evaluating multiple models at once.

System Memory Size: 192 GB

Hardware Overview

SR710-3 (2U GPU Server)



Senao Networks

Product specifications, configurations, certifications, and availability are subject to change without notice and may vary by region and customer program. Final commitments are governed by the applicable quotation, statement of work, or executed agreement.