



## High Performance Storage for AI Applications

### Create an Effective and Responsive Complete Solution for Demanding AI Environments

The recent explosion in AI for everything from large language models to recommender systems is pushing demand for increases in GPU performance in order to maximize the value and efficiency of GPU servers. A complete solution which includes the right combination of CPUs, GPUs, tiered storage and networking will ensure optimal performance to meet users' specific application requirements.

One of the biggest challenges facing businesses looking to capitalize on the growth of AI, is finding a storage solution that won't become the bottleneck in their high performance GPU cluster. High throughput, low latency storage is vital to feed massive amounts of data to train models and perform complex simulations and analysis, reducing AI model training and inference times, as well as TCO.

Choosing a High Performance Storage Solution for AI requires an understanding of the following:

1. How much storage do I need?
  - a. 2 to 4 Bytes Per Parameter in a Large Language Model
2. What are the options for object storage?
  - a. Single/dual node for redundancy? Understand the application requirements.
  - b. What capacity do I need for warm storage with 3.5" top-loading servers?
3. How much fast flash do I need?
  - a. 1U or 2U All Flash featuring EDSFF storage devices
4. What about a Hybrid System?
  - a. How much hot and how much warm storage do I need?
5. Will workloads be executed?
  - a. Networking requirements: 100G/200G/400G Ethernet/InfiniBand

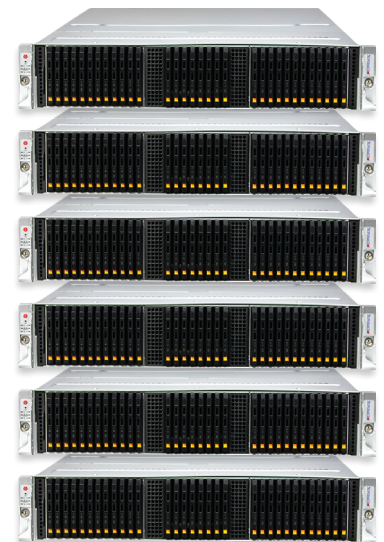


### AI Rack Scale Storage Solutions from Supermicro

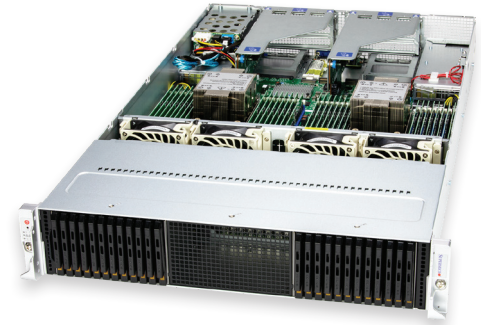
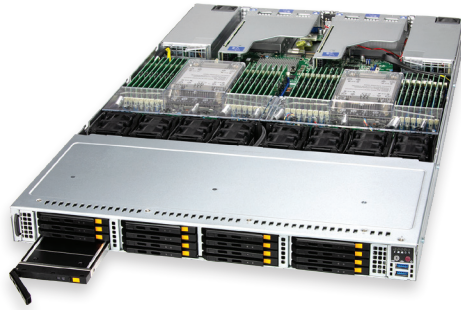
Storage for high end AI environments where very large (>1 Trillion parameters) or multiple training scenarios execute at the same time, require solutions designed at the rack level.



Benefits of Supermicro EDSFF E3.S storage solutions

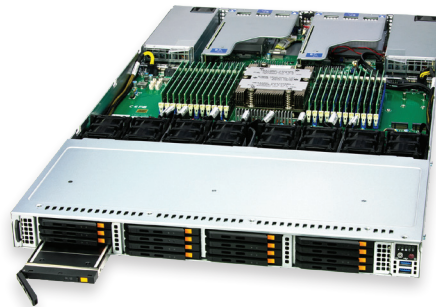
- Balanced architecture to reduce latency
- PCIe 5.0 x16 rear I/O for connection to GPU servers via NVMe-oF
- Lower TCO due to increased density and efficiency
- E3.S is optimized for PCIe 5.0 compatibility and performance
- Enhanced thermal performance of the E3.S form factor
- Increased storage density compared to U.2
- Increased number of devices per server





Supernicro solutions for high performance storage for AI environments:



MODEL	SSG-121E-NE316R	SSG-221E-NE324R
Media Type	E3.S	E3.S
CPU	2x 4th Gen Intel® Xeon® Scalable Processors	2x 4th Gen Intel® Xeon® Scalable Processors
Number Storage Drives	16 	24 
Height (U)	1U	2U



MODEL	ASG-1115S-NE316R	ASG-2115S-NE332R
Media Type	E3.S	E3.S
CPU	1x 4th Gen AMD EPYC™	1x 4th Gen AMD EPYC™
Number Storage Drives	16 	32 
Height (U)	1U	2U

Supernicro High Performance Storage Partners:



Go to <https://www.supernicro.com/en/products/storage> or scan the QR code to visit the Supernicro Storage for AI web page:

