

Supermicro AI Platform for Inferencing



The The AI Data Platform solution with Supermicro and VAST Data, built on the NVIDIA AI Data Platform reference design, provides a turnkey AI storage and workflow solution for enterprise AI deployment. This solution integrates All-Flash storage arrays, high performance servers, GPU accelerators with VAST's AI OS which provides universal storage and integrated vector database, event streaming services, and low-code data pipeline builder for orchestrating AI workloads.

The Supermicro, NVIDIA and VAST solution eliminates fragmentation with unified infrastructure purpose-built for

simplified enterprise AI deployments. This portfolio spans storage with Supermicro's CloudDC A+ servers powering the EBox nodes (using the AMD EPYC™ 9005/9004 processor with up to 192 cores and 3TB DDR5 memory for VAST's all-flash DASE architecture) to GPU inference using Supermicro's compact 2U GPUs servers with up to four NVIDIA RTX PRO™ 6000 Blackwell Server Edition or NVIDIA H200 GPUs for on-premise data center and edge deployments.

Together with VAST AI OS managing storage and GPU inference as one system, including InsightEngine, built on the NVIDIA AI Data platform reference design for automated RAG pipelines, Data Engine for event-driven processing, and native Database for SQL and vector workloads with unified governance, Supermicro infrastructure eliminates waste from fragmented solutions. Continuous context refresh keeps data current at AI factory speed. Native vector database propagates access controls from source files through embeddings to inference outputs. Whether deploying federal GPU clusters, hyperscale AI cloud operations, or foundation model training, Supermicro and VAST delivers validated configurations powered by NVIDIA where performance stays predictable, rework stays controlled, and intelligence compounds through traceability that turns production speed into business value.

From Fragmentation to Integration The Unified Enterprise AI Stack

Enterprise IT leaders spend months validating disparate solutions across dozens of vendors databases, orchestration platforms, and AI engines—resulting in slow adoption, high costs, and security risks. The Supermicro and VAST unified platform eliminates this complexity.

Supermicro Integrated AI Solution with VAST InsightEngine

Total AI Software Stack

Leverage the integrated hardware and software stack powered by NVIDIA and VAST:

VAST Accelerated Core

- Scalable, Event-Driven GPU Compute for AI pipelines
- Real-Time RAG and Multimodal workflows
- VAST Container Orchestration (RKE2) for NVIDIA NIM microservices

VAST Compute Core

- Transactional, Analytical & Vector Database Services
- VAST Event Broker For Distribution Event Streaming

VAST Capacity

- Globally Distributed DataStore & KV Cache Services
- Universal Storage (NFS, S3, SMB, NVMe/ITCP)

Platforms | EBox (see page 2) + CNode-X (see page 2)



Supermicro AI- Ready Infrastructure

Network & Fabrics

- 2x NVIDIA Spectrum™ SN5600 Switches (800GbE)

Accelerated Compute Nodes

- 2x Supermicro "CNode-X"
- 2x NVIDIA RTX PRO™ 6000 GPUs per node
- 2x NVIDIA® ConnectX®-7 400G

Storage and Data Platform

- Built on NVIDIA AI Data Platform reference design
- VAST InsightEngine on Supermicro (EBox)
- Unified datastore & vector intelligence
- 2x NVIDIA® ConnectX®-7 400G

Built for Scalability and Simplicity

- Scales from single rack to multi rack
- Unified design simplifies deployment and management

Unified AI Platform Solution

Intelligence Layer (NVIDIA)

AI-Ready Data

with NVIDIA AI Data Platform reference design

Key Technologies

NVIDIA RTX Pro Server, NVIDIA Spectrum-X™ networking, NVIDIA CUDA®/NVIDIA AI Enterprise software and Cuda-X Libraries

AI Workloads

NVIDIA AI Enterprise, including NVIDIA NeMo Retriever, NIM inference microservices, and cuVS

Data Platform (VAST Data)

Unified Platform

Combines datastore + database + AI functions in one platform

InsightEngine

Real-time RAG and multimodal workflows optimized for NVIDIA NIM automation

Capabilities

Event-driven agentic computing stack with real-time queries, analytics, and governance at scale

Infrastructure & Integration (Supermicro)

End-to-End Provider

Single partner: compute, storage, networking, integration

Validated Architectures

Rack-scale designs, liquid cooling, deployment services

Reference Architecture

Component	Configuration
Network & Fabrics	2x NVIDIA Spectrum™ SN5600 Switches (800GbE)
AI Compute Nodes	2x Supermicro 2U PCIe GPU servers for the CNode-X with 2x NVIDIA RTX PRO™ 6000 Blackwell Server Edition per node, NVIDIA BlueField™ equipped
Storage & Data Platform	VAST InsightEngine on Supermicro EBox, unified datastore & vector intelligence
Scalability	Scales from single rack to multi-rack cluster

Key Benefits

Benefit	Value
Pre-validated Integration	2x NVIDIA Spectrum™ SN5600 Switches (800GbE)
Unified Stack	Eliminates data movement penalties, unbeatable global data reduction, simplified AI development.
Turnkey Infrastructure	Lowers operational complexity and total cost of ownership

VAST EBox: Data Services Infrastructure



AS-1116CS-TN CloudDC 1U AMD EPYC™ Storage Node

The Supermicro AS-1116CS-TN is optimized as a VAST Data EBox storage node, delivering the NVMe density, memory bandwidth, and throughput required for VAST's all-flash DASE architecture. Built on OCP DC-MHS standards for hyperscale compatibility.

Ideal For: VAST DataStore nodes, HPC storage, cloud infrastructure, virtualization

EBox Specifications

Form Factor

1U Rackmount server with 12x SSDs OCP DC-MHS compliant

Processor

Single AMD EPYC™ 9555P (64 cores, 350W TDP)

Memory

12x 32G DDR5 DIMM, up to 384GB @6400MT/s

Storage

12 hot-plug NVMe (3x SCM, 9x QLC)

Management Network

1x AIOM slot with 1GbE x 4 + 1GbE BMC

Network

2x NVIDIA® ConnectX®-7 (200G x 2 or 400G x 1, Ethernet or IB)

Management

DC-SCM with OpenBMC

Learn More:

Contact your Supermicro representative for validated VAST configurations.

supermicro.com | vastdata.com | vastdata.com/forward

VAST CNode-X: Agentic Services Infrastructure



SYS-212GB-FNR 2U Multi-GPU Compute Server

The Supermicro SYS-212GB series delivers high-density GPU compute for AI training, inference, and edge deployments. Supports up to 4x high-power GPUs (NVIDIA H200, RTX PRO™ 6000) in a compact 2U form factor, ideal for VAST InsightEngine RAG workloads.

Ideal For: LLM training, AI inference, RAG with InsightEngine, Edge AI (Telecom, Healthcare, Retail)

CNode-X Specifications (SYS-212GB-FNR)

Form Factor

2U Rackmount (X14SBGM + CSE-GP201TS)

Processor

Intel® Xeon® 6774P, 64 cores, 2.5GHz, 350W TDP

Memory

512GB DDR5 RDIMM 6400 ECC (8x 64GB MEM-DR564L-SL03-ER64)

GPU

2x NVIDIA RTX PRO™ 6000 Blackwell Server Edition, 96GB GDDR7, PCIe 5.0

Storage

2x Micron 7450 PRO 960GB NVMe E1.S 15mm

Networking

2x NVIDIA® ConnectX®-7 400GbE NDR QSFP112 + 4-port 1GbE Intel i350

VROC

AOC-VROCSTNM0D (Standard)

Security

TPM 2.0 (AOM-TPM-9672V-P)

Management

DataCenter Management Package (SFT-DCMS-SINGLE)