



SUPERMICRO AND DDN TOGETHER DEVELOP THE ENTERPRISE AI HYPERPOD



TABLE OF CONTENTS

Executive Summary	1
Solution Overview	2
Key Solution Components.....	2
DDN Enterprise AI HyperPOD Architecture	5
Summary	8
Further Information	8

Executive Summary

Supermicro, together with DDN, has created the Enterprise AI HyperPOD, which is built using Supermicro servers and accelerated by NVIDIA GPUs. The Enterprise AI HyperPOD is a purpose-built, turnkey solution for enterprise inferencing and Retrieval-Augmented Generation (RAG). In an era where speed and trust define AI outcomes, this collaboration between DDN, Supermicro, and NVIDIA delivers a validated platform that unifies compute, networking, and data intelligence. By

optimizing data paths with tiered caching and metadata acceleration, it minimizes storage-to-GPU overhead—reducing latency, sustaining high throughput, and achieving up to 99% GPU utilization. This results in more inferences per rack and per watt, with predictable linear scaling from pilot to production AI factories.

Enterprises often waste up to 60% of AI infrastructure capacity due to siloed data and underutilized GPUs. The DDN Enterprise AI HyperPOD addresses this by delivering breakthrough efficiency: 60% higher GPU utilization and up to 10x lower power costs than traditional architectures. It offers world-record density with up to 300 TB of capacity in just 1U and scales from 8 GPUs to more than 256 GPUs across tiered configurations.

Solution Overview

The DDN Enterprise AI HyperPOD provides production-grade inferencing and RAG as a pre-integrated system that moves data to GPUs with minimal friction. Preconfigured by Supermicro using the NVIDIA AI Data Platform (AIDP) architecture, hardware, and software, and DDN Infinia, it enables rack, stack, and run on day one. This flexible, scalable solution is optimized for specialized, information-intensive workloads, starting at a few hundred terabytes and scaling to multi-exabytes, including up to 100 PB in one rack.

Key benefits include:

- Performance: GPUs maintained at 90%+ utilization for sustained throughput.
- Latency: Sub-millisecond access for responsive inference.
- Density: Up to 100PB in one rack for compact AI factories.
- Efficiency: 10x power/cooling savings at scale compared to legacy builds.
- Availability: Five-9s design with multi-tenancy, encryption, and fault-domain protection.

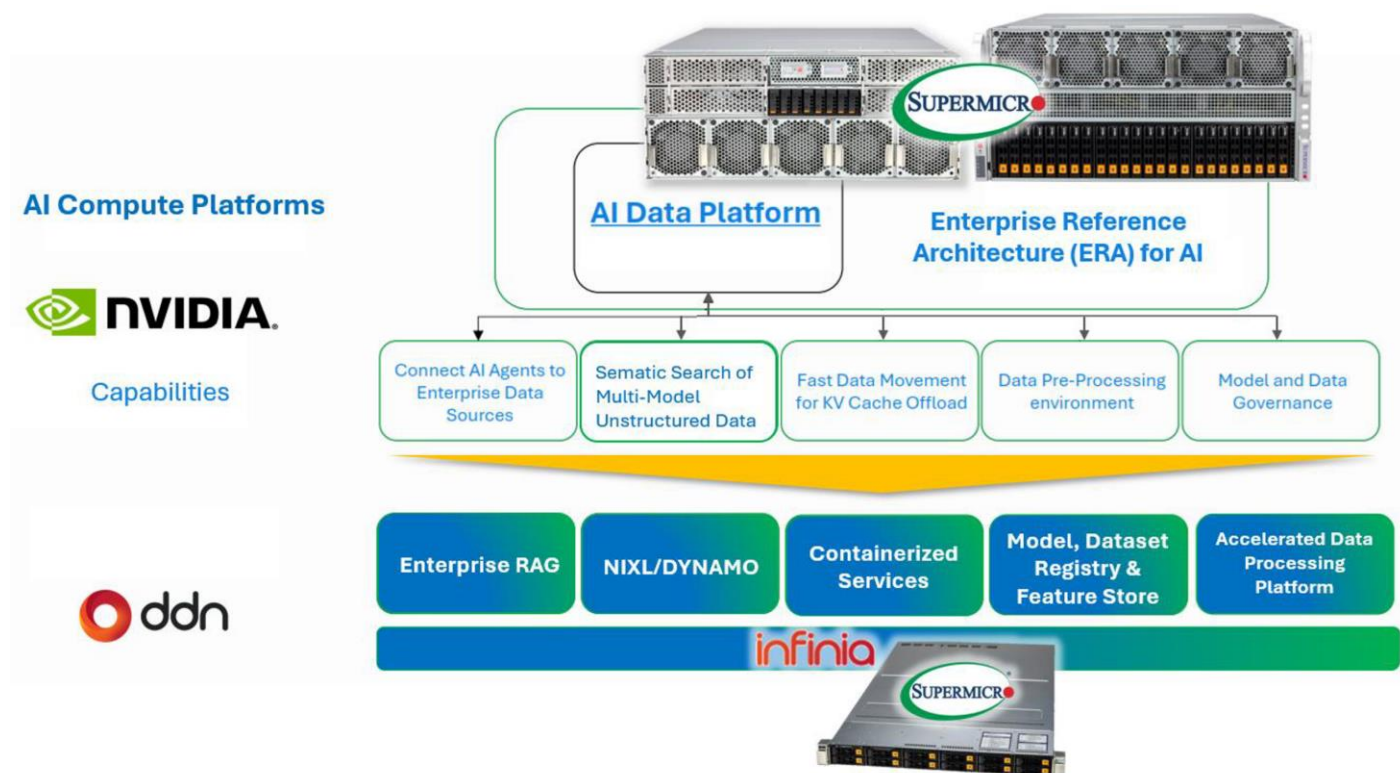




Figure 1 - DDN Enterprise AI HyperPOD, Built on Supermicro, Accelerated by NVIDIA

Key Solution Components

Supermicro supplies dense, power-efficient GPU systems tuned for AI throughput, delivering compact footprints and Day-1 consistency.

Supermicro GPU Platform Hardware for Accelerated Infrastructure

5U 10 PCIe GPU Server – SYS-522GA-NRT	4U 8 PCIe MGX GPU Server – SYS-422GL-NR
	
<p>Product Specifications</p> <ul style="list-style-type: none"> Up to 8 NVIDIA RTX PRO™ 6000 Blackwell Server Edition cards Dual Intel® Xeon® 6 Scalable series processors p-cores 24 DDR5 DIMMs supporting up to 6TB or DDR5 MRDIMM supporting up to 6TB Up to 24 U.2/U.3 NVMe and 2 M.2 NVMe 13 PCIe 5.0 slots and 1 AIOM/OCF 3.0 slot 6 redundant (N+N) 2700W Titanium level power supplies 	<p>Product Specifications</p> <ul style="list-style-type: none"> Up to 8 NVIDIA RTX PRO™ 6000 Blackwell Server Edition cards Dual Intel® Xeon® 6 Scalable series processors p-cores 24 DDR5 DIMMs supporting up to 6TB or DDR5 MRDIMM supporting up to 6TB Up to 8 front hot-swap E1.S NVMe drive bays 13 PCIe 5.0 slots; 8 double wide + 4 FHHL + 1 LP 4 redundant (2+2) 3200W Titanium level power supplies

DDN Infinia Storage System Running on Supermicro Systems

DDN Infinia is a next-generation, metadata-smart, software-defined Data Intelligence Platform designed for the performance, scalability, and efficiency demands of AI. It unifies structured and unstructured datasets across multi-cloud, on-prem, and edge environments with native metadata intelligence and extreme low-latency access. DDN Infinia supports sub-millisecond latency, faster metadata-driven AI pipelines, and seamless integration with NVIDIA-powered AI stacks—delivering 10x higher efficiencies than traditional file systems.

Built on DDN Infinia 2.3 with Dynamo metadata acceleration, it provides 100x faster queries and multi-tenant QoS for sovereign AI and NCP workloads. Infinia solves four key enterprise AI challenges:

- **Reduces Complexity:**
 - Unified AI Data Fabric: Unify multimodal data across core, cloud, and edge using smart metadata.
 - Seamless Integration: Native multi-protocol support (S3, CSI, etc.) avoids reformatting and reconfiguration delays.

- One Platform for AI: Centralize tools for AI Data Analytics, Data Preparation, Model Loading, and Inference, including NVIDIA AI Enterprise, Trino, Spark, and TensorFlow.
- **Accelerates Innovation:**
 - GenAI & LLMs: Reduce training and deployment time for faster insights.
 - GPU Optimization: Maximize GPU utilization up to 99% by minimizing data movement.
 - Real-Time AI: Support RAG-enabled indexing and ultra-low sub-millisecond latency for instant inferencing responses.
- **Lowers Costs:**
 - 10x Data Reduction: Metadata-driven intelligence minimizes data movement and egress costs in cloud environments.
 - Power Savings: Reduce power & cooling costs up to 70% at massive scale.
 - Optimal Resource Allocation: Meet varying demands of GPUs, efficiency, and performance.
- **Provides Proven Reliability & Security:**
 - Validated with NVIDIA: Scale seamlessly from terabytes to exabytes with proven reliability, stability, and availability.
 - Security-Driven: Built-in secure multi-tenancy, encryption, fault-domain-aware erasure coding, and data protection.
 - 100% Software Defined: Optimized for any NVMe flash (TLC, QLC, or PLC) for cost optimization.

Optimized System for Infinia Storage Solution

Supermicro Hyper A+ Server AS -1115HS-TNR

Per System:

- Network Port Options: 2x 400GbE OSFP or 4x 200GbE QSFP112
- Drive Slots: 12x Gen5 NVMe U.2 Slots
- Drive Options: 30TB/60TB/120TB TLC/QLC NVMe Drives



Figure 2 - Supermicro AS -1115HS-TNR

NVIDIA Enterprise AI Software

NVIDIA's AI Data Platform (AIDP) is a proven reference design for enterprise AI, implemented with DDN Infinia software and Supermicro systems. It includes:

- Production AI software with NVIDIA AI Enterprise, including NIM for endpoints and NeMo Retriever for RAG.
- Accelerated networking with NVIDIA® Spectrum-X™ Ethernet and BlueField®-3 DPUs for low-latency data paths and offload.
- Fast data-to-compute pipelines to keep apps responsive and GPUs saturated.
- Agentic and RAG-ready foundations for applications that reason, retrieve, and act.

DDN Enterprise AI HyperPOD Architecture

Supermicro and DDN deliver enterprise-ready architectures for AI training and inferencing. The DDN Enterprise AI HyperPOD includes Supermicro GPU computing systems, all-flash storage, ultra-high-speed networking, support systems, and operational management. Enterprise AI Reference Architectures scale from 8 GPUs to more than 256 GPUs with tiered configurations (XS, Small, Medium) for workloads from inference to large-scale training.

Hybrid Cloud Integration enables seamless extension to popular clouds for secure hybrid deployments.

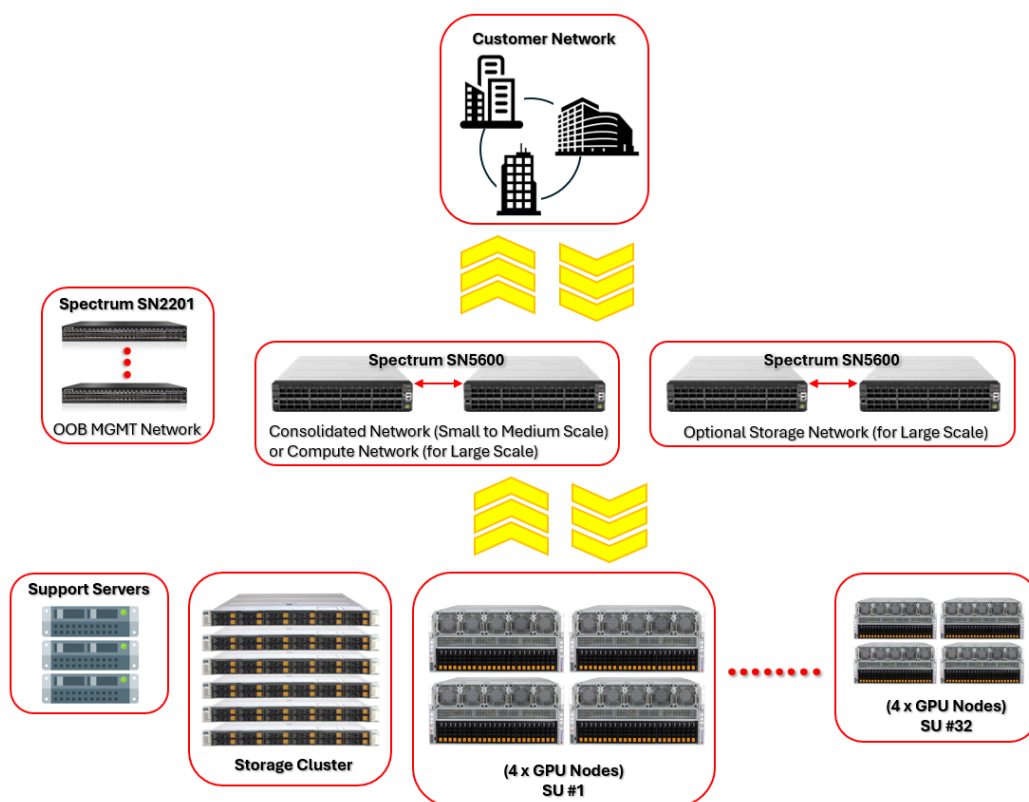


Figure 3 - DDN Enterprise AI HyperPOD Solution Architecture

Data Platform Scale	XS	S	M	L
Total GPUs - NVIDIA RTX PRO 6000	4	32	64	256
NVMe SSD Capacity (TB)	7.68	15.36	30.72	122.88
DDN Infinia Storage Capacity (PB)	0.5+	1+	3+	12+
Quantity of Storage Servers	6	6	6	6
Storage Node Type	AS -1115HS-TNR	AS -1115HS-TNR	ASG-1115S-NE316R	ASG-1115S-NE316R

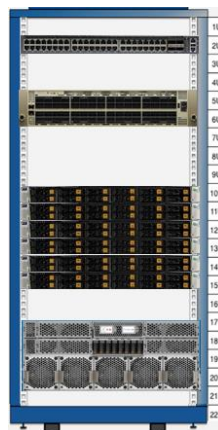
Table 2 – Sizing of AI Data Platform Reference Designs

AI workloads optimized for this system:

- Real-time inference APIs for apps, agents, and tools.
- RAG over enterprise content (NeMo Retriever).
- Analytics for AI data prep and model loading.
- Agentic workflows needing low-latency retrieval with high GPU throughput.

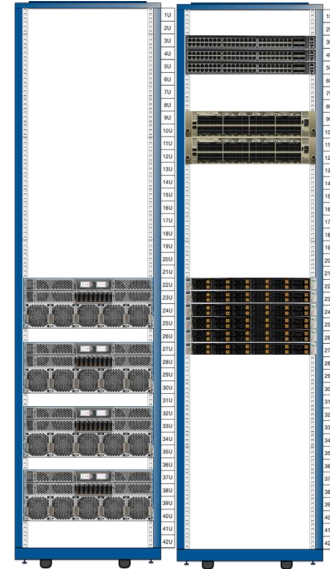
AI Data Platform – Extra Small (XS)

- 1x 4U MGX GPU Server SYS-422GL-NR
 - NVIDIA 2-4-3-200 ERA
 - 4x NVIDIA RTX PRO 6000 GPU
 - 2x BF3 B3140H (400GbE, E-W)
 - 1x BF3 B3220 (Dual 200GbE, N-S)
- 0.5PB Infinia (Raw Capacity):
 - 6x AS -1115HS-TNR
 - 2x 7.68TB NVMe drives
- Network:
 - 200GbE (QSFP112) Consolidated Network
 - 1x NVIDIA SN5610 (128x OSFP 400GbE)
 - 1GbE Management Network
 - 1x NVIDIA SN2201
- Power Budget:
 - 12,100W



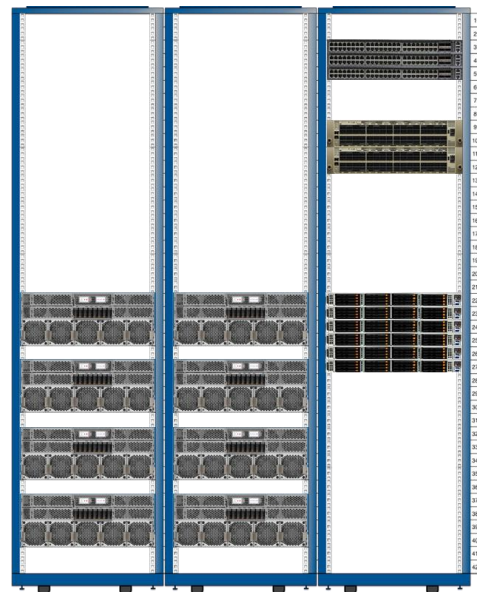
AI Data Platform - Small (S)

- 4x 4U MGX GPU Server SYS-422GL-NR
 - NVIDIA 2-8-5-200 ERA
 - 32x RTX PRO 6000 GPU
 - 16x BF3 B3140H (400GbE, E-W)
 - 4x BF3 B3220 (Dual 200GbE, N-S)
- 1PB Infinia (Raw Capacity):
 - 6x AS -1115HS-TNR
 - 2x 15.36TB NVMe drives
- Network:
 - 200GbE (QSFP112) Consolidated Network
2x NVIDIA SN5610 (128x OSFP 400GbE)
 - 1GbE Management Network
2x NVIDIA SN2201
- Power Budget:
 - GPU Rack: 33,400W
 - Storage / Network Rack: 8,080W



AI Data Platform - Medium (M)

- 8x 4U MGX GPU Server
 - NVIDIA 2-8-5-200 ERA
 - 64x RTX PRO 6000 GPU
 - 32x BF3 B3140H (400GbE, E-W)
 - 8x BF3 B3220 (Dual 200GbE, N-S)
- 3PB Infinia (Raw Capacity):
 - 6x ASG-1115S-NE316R
 - 96x 30.72TB NVMe drives
- Network:
 - 200GbE (QSFP112) Consolidated Network
2x NVIDIA SN5610 (128x OSFP 400GbE)
 - 1GbE Management Network
3x NVIDIA SN2201
- Power Budget:
 - GPU Rack: 33,400W
 - Storage / Network Rack: 8,860W



AI Data Platform – Large (L)

- 32x 4U MGX GPU Server
 - NVIDIA 2-8-5-200 ERA
 - 256x RTX PRO 6000 GPU
 - 128x BF3 B3140H (400GbE, E-W)
 - 32x BF3 B3220 (Dual 200GbE, N-S)
- 12PB Infinia (Raw Capacity):
 - 6x ASG-1115S-NE316R
 - 96x 122.88TB NVMe drives
- Network:
 - 200GbE (QSFP112) Consolidated Network
 - 4x NVIDIA SN5610 (128x OSFP 400GbE)
 - 1GbE Management Network
 - 4x NVIDIA SN2201
- Power Budget:
 - GPU Rack: 33,400W
 - Storage / Network Rack: 10,920W



Summary

Supermicro's professional solutions team, collaborating closely with DDN and NVIDIA, delivers tailored project planning, pre-sales and post-sales consulting, technical support, and other integrated one-stop services. This enables accelerated solution deployment with minimal time and labor costs, achieves Day 2 operations, and focuses on product and service development without any concerns.

Further Information:

Supermicro

<https://www.supermicro.com/en/solutions/ddn>

<https://www.supermicro.com/en/accelerators/nvidia/pcie-gpu>

DDN

<https://www.ddn.com/products/ddn-enterprise-ai-hyperpod/>

<https://www.ddn.com/products/infinia>

<https://www.ddn.com/products/data-intelligence-platform/>

NVIDIA

<https://www.nvidia.com/en-us/data-center/ai-data-platform>

<https://www.nvidia.com/en-us/technologies/enterprise-reference-architecture/>

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, AI, High-Performance Storage, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Data Center Building Block Solutions® (DCBBS) approach allows us to provide a broad range of SKUs and enables us to build and deliver application-optimized solutions in all scales based upon your requirements.

For more information - www.supermicro.com

DDN

DDN is a company of pioneers in high-performance data storage and management, dedicated to delivering innovative solutions that empower organizations across the globe. Our commitment to excellence, coupled with our cutting-edge technology, enables us to drive performance, scalability, and reliability for our clients. Discover how our expertise and passion for data are transforming industries and shaping the future.

For more information - www.ddn.com