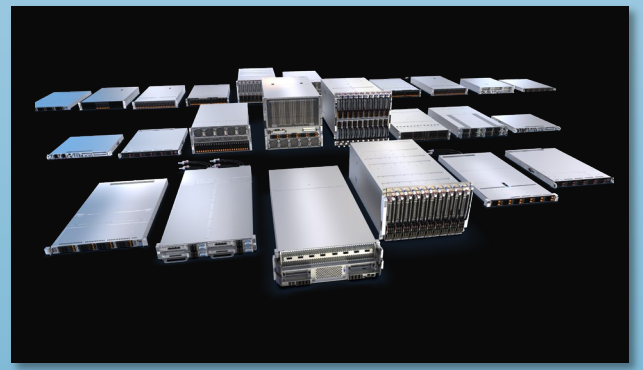
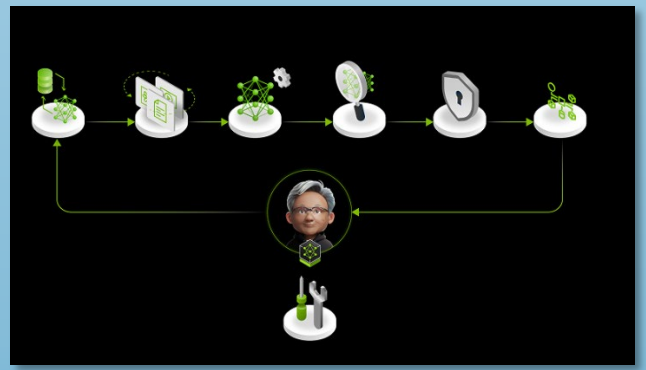
Selling Supermicro AI Factory and Enterprise AI Solutions

NVIDIA GTC Pre-GTC 2026 Edition — Channel Training by Philip Tamaki

The AI Opportunity in 2026



2026 is the year that AI Server TAM will overtake enterprise server TAM.



"Enterprise adoption of agents is skyrocketing." — Jensen Huang

The buildout of next-gen AI clouds continues to scale

Sources:
https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-fourth-quarter-and-fiscal-2026
https://www.aspeedtech.com/file/report/2024AnnualReport_en.pdf

philipt@supermicro.com

GTC 2026: Strategy Overview

Preface

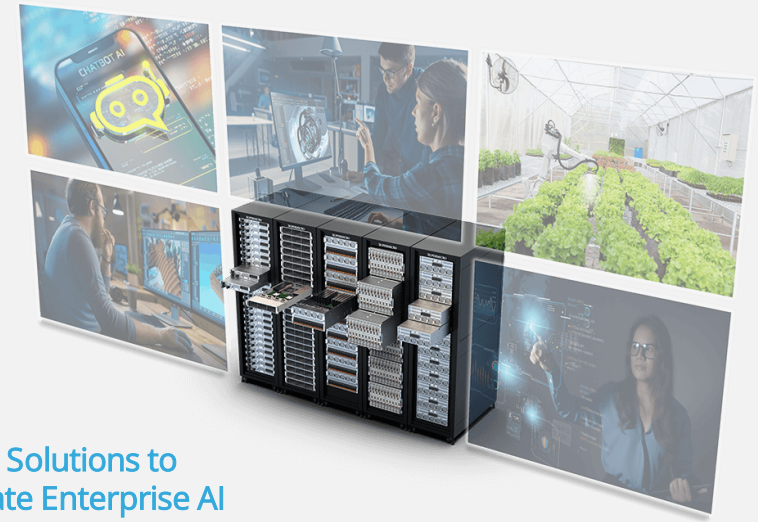
- GTC 2026 marks a shift from AI experimentation to enterprise-scale deployment
- Enterprises are no longer asking what is possible, they are asking how to deploy reliably
- Supermicro's role is to turn NVIDIA platforms into validated, deployable AI factories
- Our GTC presence focuses on how AI is built, integrated, and brought online as a **total solution**

Key Details on Supermicro's Exhibit at GTC26

- Date: March 16-19, 2026
- Location: San Jose McEnery Convention Center
- Sponsorship Level: Diamond
- Booth Size: 30'x30
- Booth Number: 1113

Promotional Focus:

AI Factory Total Solution
Development and Deployment



Tailored Solutions to
Accelerate Enterprise AI

philipt@supermicro.com

The Expanding AI Computing Requirements

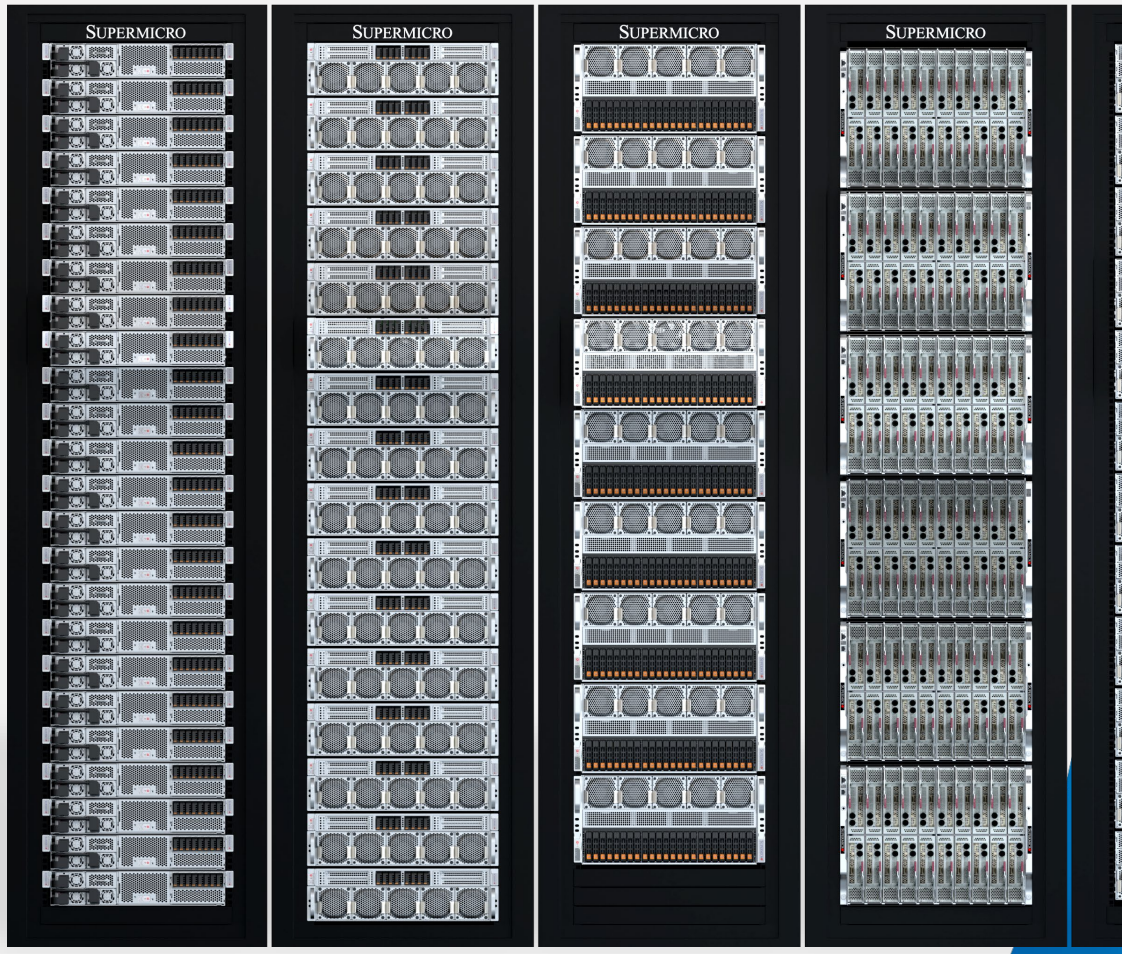
Increased AI Capabilities

Leads to

Greater Computing Requirements

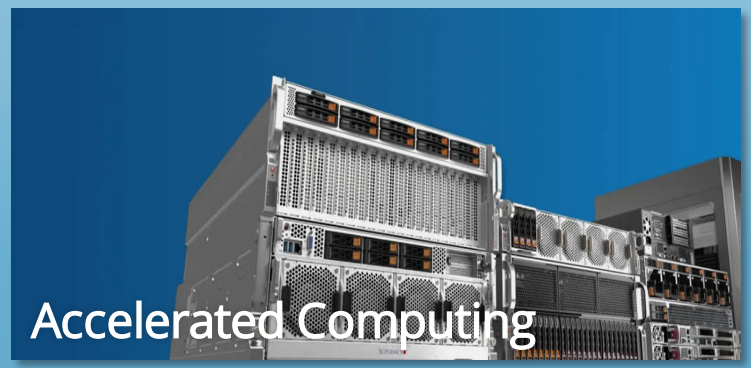
- Real-time Inference
- Agentic Execution
- Long-chain Reasoning
- Persistent Context

- All-to-all communication
- Memory bandwidth
- Memory capacity
- Context Memory Layer



philipt@supermicro.com

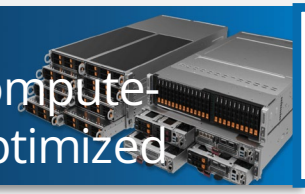
Traditional Data Center Computing Taxonomy



Traditional data center computing categories are labelled primarily by types of computing resources. Where does AI fit into this?

philipt@supermicro.com

How Solution Type Addresses Bottlenecks

 Compute-Optimized	**Addresses bottleneck:** CPU compute **Application Example:** CPU-bound workloads like Intensive algorithms (e.g., RAR, AES)
 Memory-Optimized	**Addresses bottleneck:** memory capacity, memory bandwidth **Application Example:** In-memory databases, big data analytics
 Storage-Optimized	**Addresses bottleneck:** Disk capacity, read/write speed, IOPS **Application Example:** Block storage, file storage, object storage, HPS, etc.
 HPC-Optimized	**Addresses bottleneck:** CPU Compute and networking **Application Example:** Computational Fluid Dynamics
 Accelerated Computing	**Addresses bottleneck:** GPU compute, GPU memory, networking **Application Example:** AI, ML, 3D, GPU-Accelerated Applications

Traditional data center computing categories cannot address the spectrum of modern AI workloads. That's where AI factories come in.

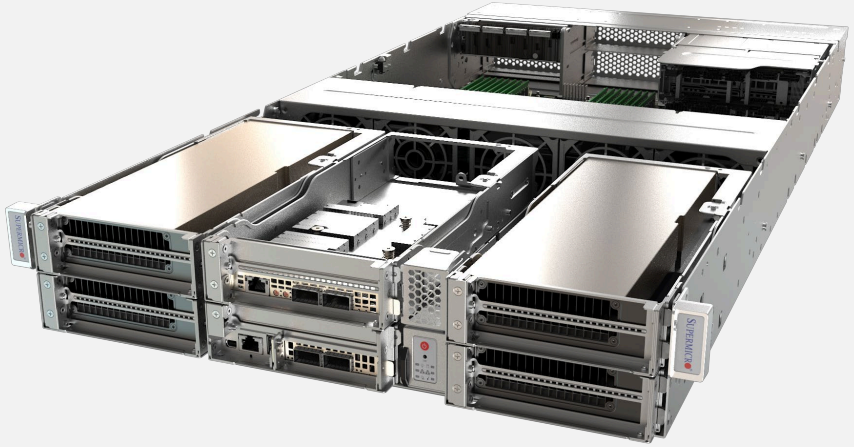
philipt@supermicro.com

Balancing Computing Resources

CPU-to-GPU Ratio

2:4

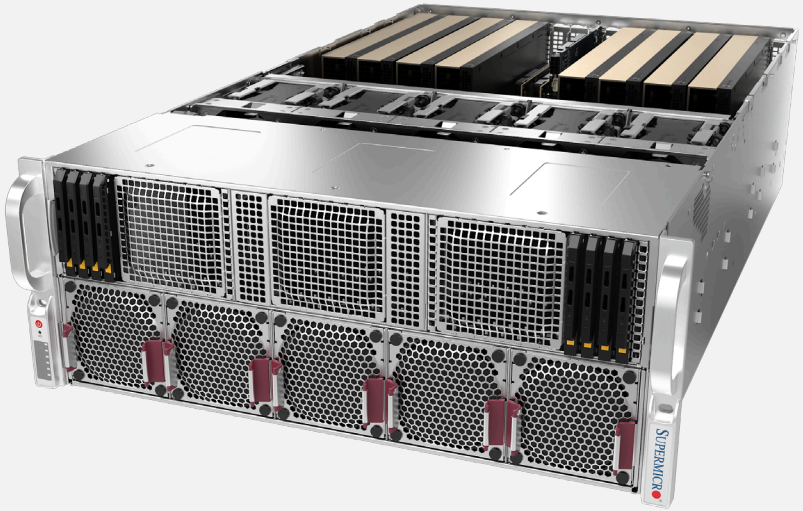
In a 2U Form Factor



CPU-to-GPU Ratio

2:8

In a 4U Form Factor



Would 2x 2U servers with 4x GPUs provide the same thing as 1x 4U server with 8x GPUs? Not quite.

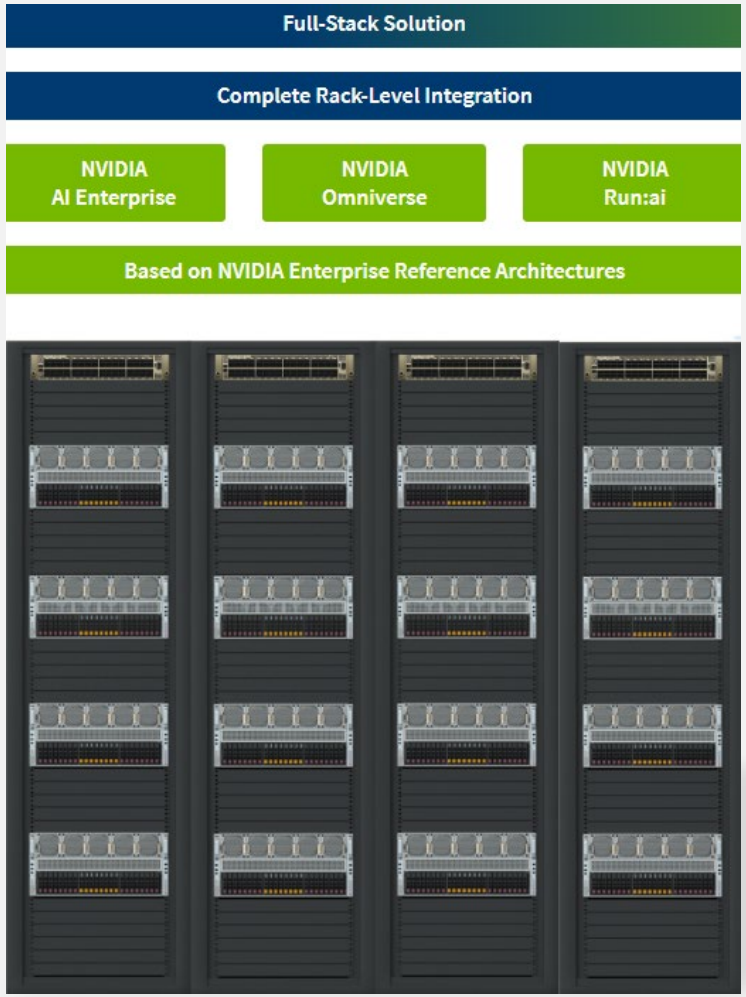
Other factors to consider: the number of CPUs (usually dual or single socket), system memory, GPU memory, number of GPUs, NICs, storage, and more. The ratio is critical since it established a pattern that gets magnified at data center scale.

philipt@supermicro.com

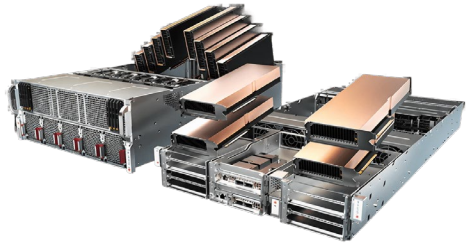
What is an AI Factory?

Supermicro NVIDIA AI Factory

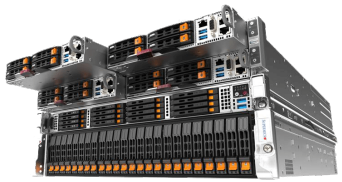
Turnkey solutions simplifying the deployment of enterprise AI at scale
Full-stack solutions including compute, software, networking, and storage.



Building Blocks (Compute and Storage)



NVIDIA RTX PRO™ Servers



Storage Servers

NVIDIA HGX Servers

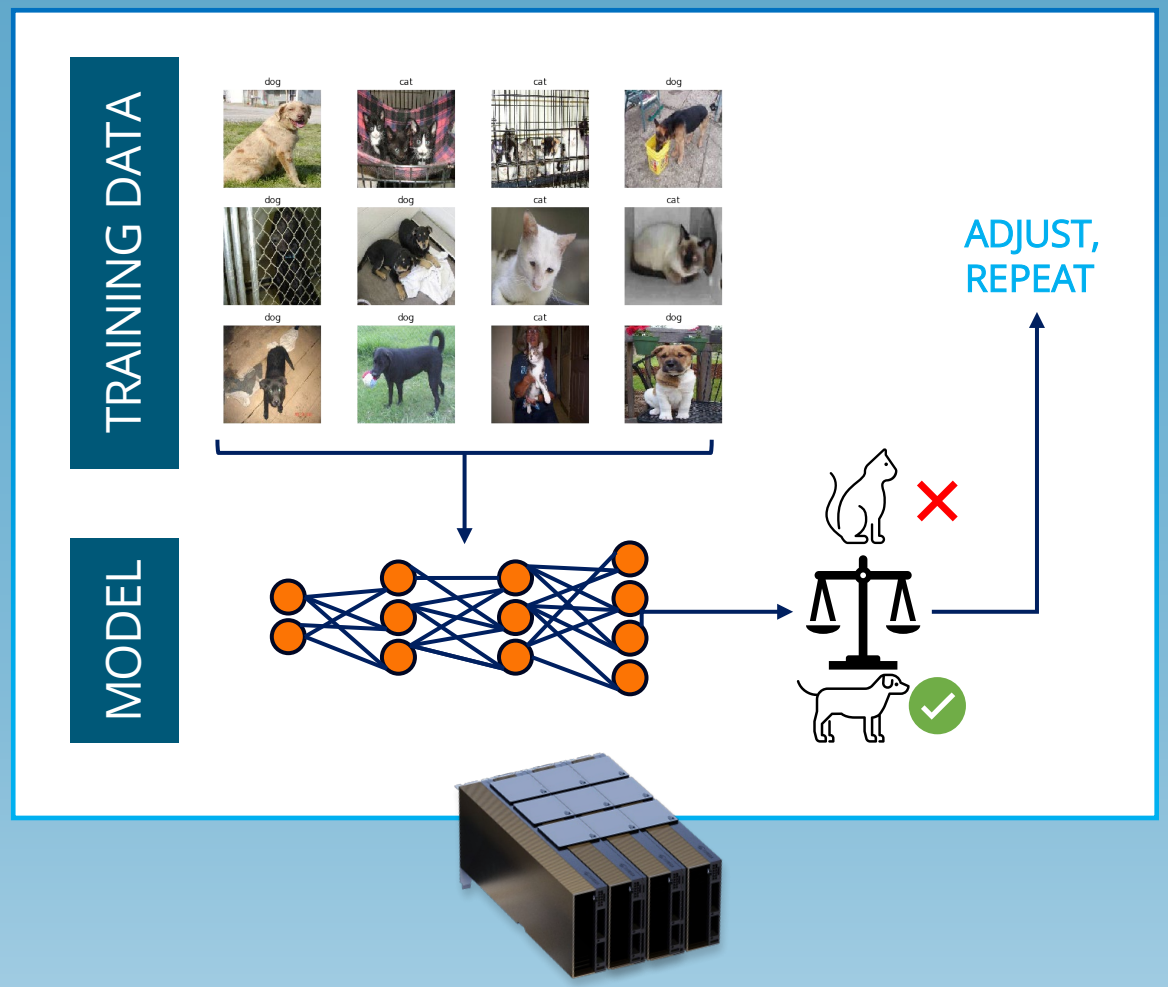
Integrated Rack Solutions

Unlike a data center that acts as a place to store data, AI factories are involved in actively producing tokens to run AI workloads of all types

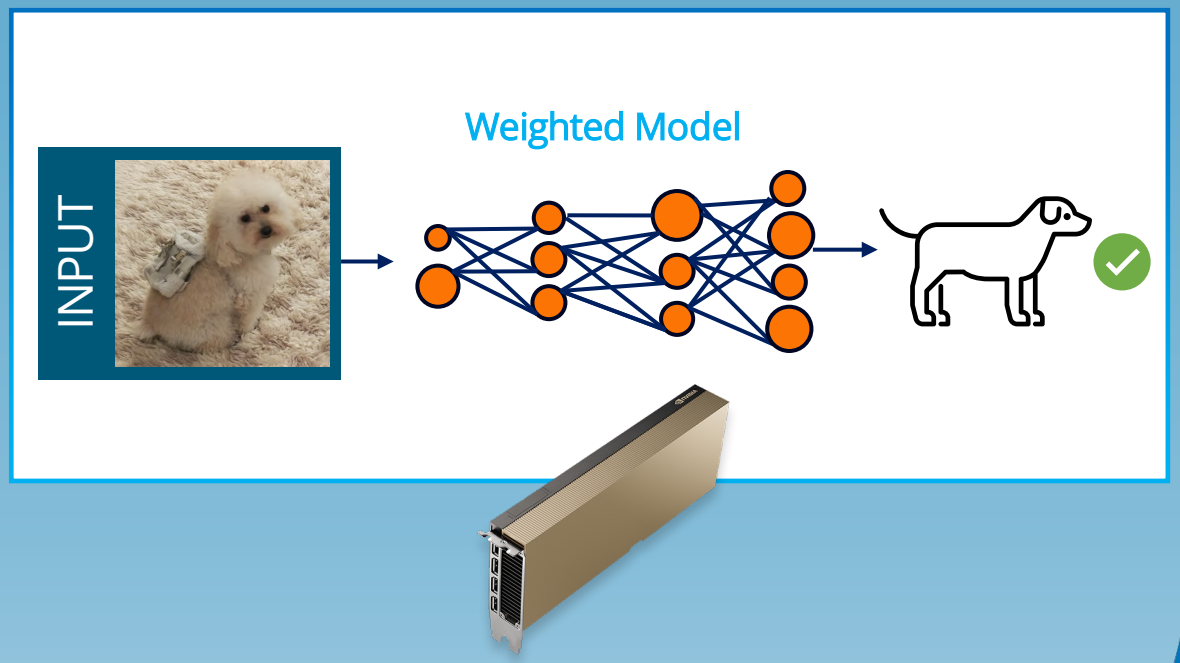
philipt@supermicro.com

AI Training vs. Inference GPU Memory Requirements

TRAINING



INFERENCE



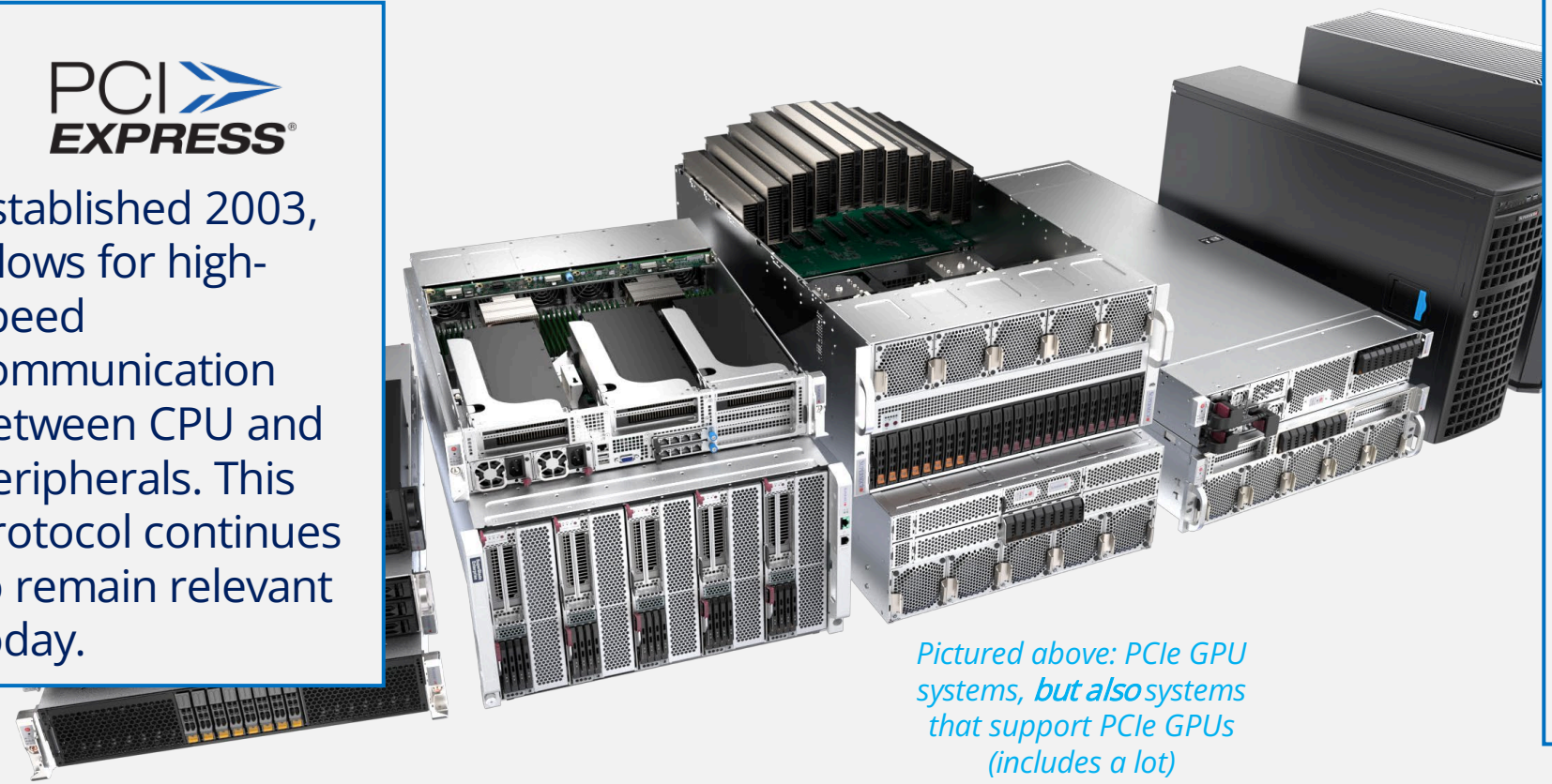
AI training can require 4x to 30x GPU memory capacity vs. AI inference

philipt@supermicro.com

The PCIe GPU Platform

PCI EXPRESS®

Established 2003, allows for high-speed communication between CPU and peripherals. This protocol continues to remain relevant today.



*Pictured above: PCIe GPU systems, **but also** systems that support PCIe GPUs (includes a lot)*

Supermicro PCIe GPU Systems

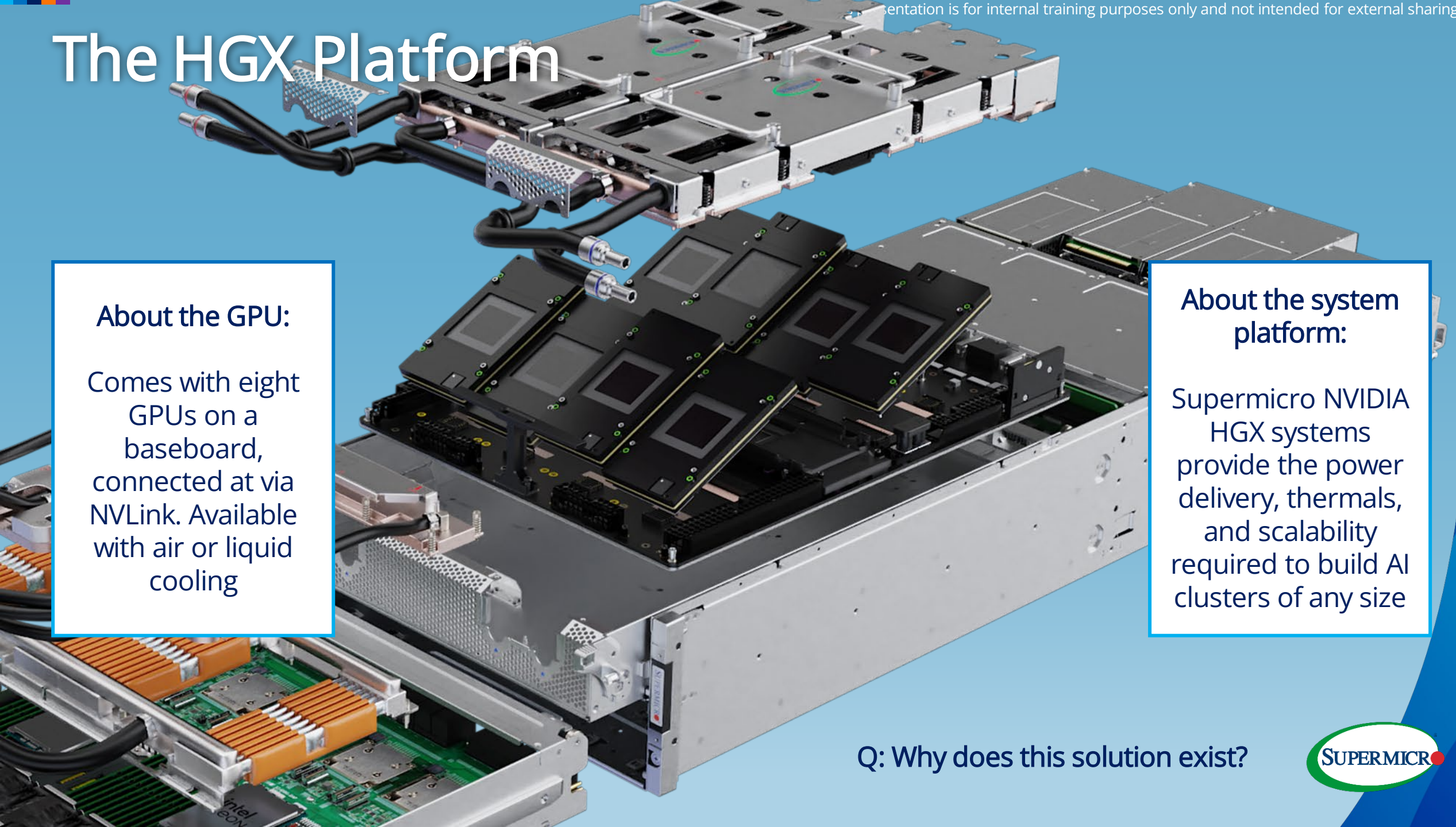
Unlike consumer desktop PCs, a single GPU server can host 8 to 10 double-width PCIe GPUs. These systems offer the flexibility to use a variety of PCIe-based accelerators/GPUs. Some popular choices for AI today include RTX PRO 6000 Blackwell Server Edition and H200 NVL.

PCIe-based GPUs are a cost-effective choice for AI workloads not limited by GPU memory capacity, such as AI inference.

SUPERMICRO

philipt@supermicro.com

The HGX Platform



About the GPU:

Comes with eight GPUs on a baseboard, connected at via NVLink. Available with air or liquid cooling

About the system platform:

Supermicro NVIDIA HGX systems provide the power delivery, thermals, and scalability required to build AI clusters of any size

Q: Why does this solution exist?

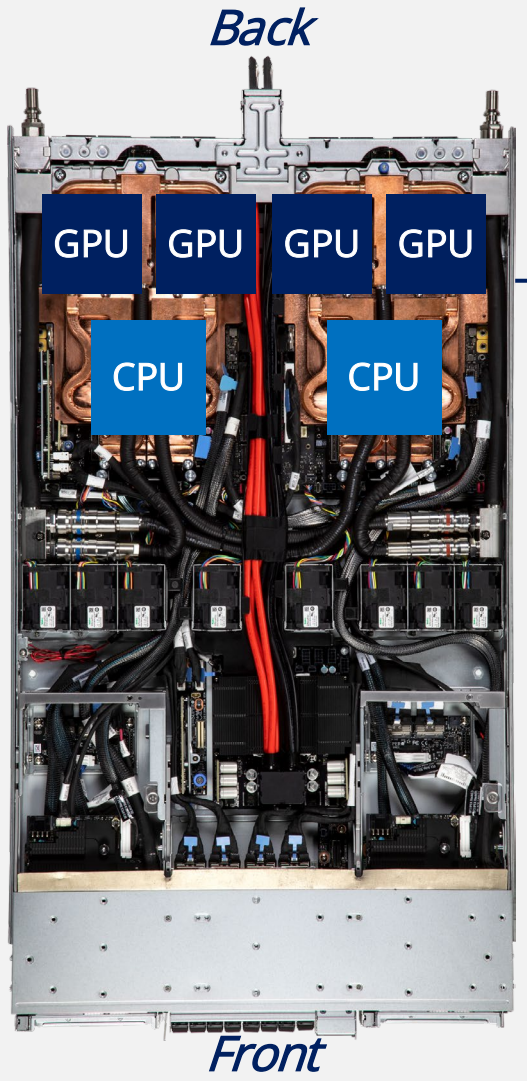
The NVL72 Platform

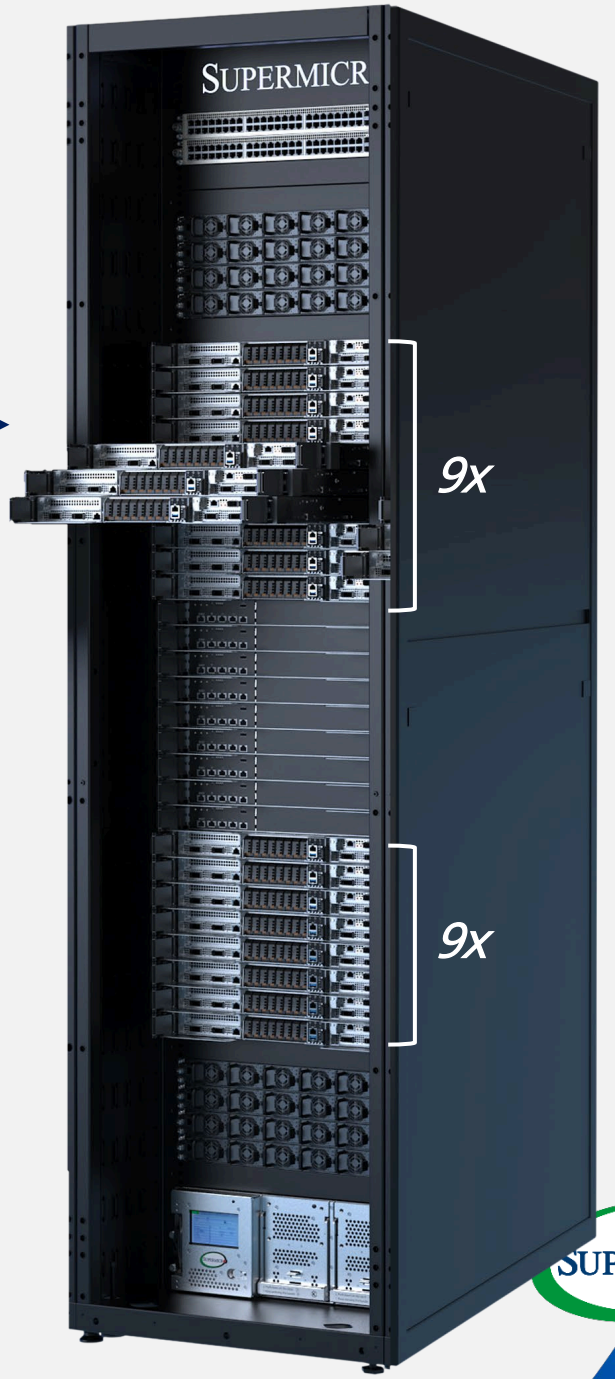
"NVL72" Solutions

Includes NVIDIA rack-scale solutions including GB200 NVL72, GB300 NVL72, and Vera Rubin NVL72.

NVL72 refers to **72 GPUs** interconnected using a high-speed NVLink fabric, all residing in a single rack enclosure.

It's designed with a specific BOM and rack elevation, treating the entire rack like a single powerful system.

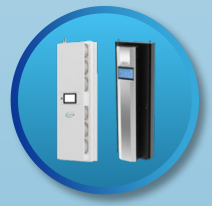


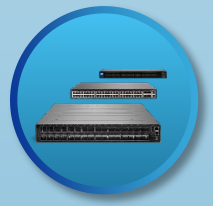


Q: Once again, consider why does this solution exist?

philipt@supermicro.com

Supermicro Data Center Building Block Solutions


Chilled Door


Switches


Power Shelf

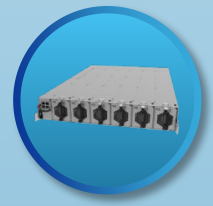

In row DLC


Storage


Cooling Tower


Dry Tower


In-Rack DLC


Battery Back Up

Supermicro's DCBBS provides broad coverage of critical components required for AI factory buildout.

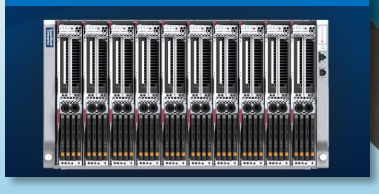
philipt@supermicro.com

GTC 2026 Booth Overview

Enterprise AI Inference

Liquid & Air-Cooled AI Training

SuperBlade®

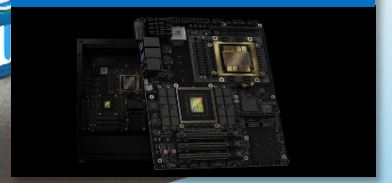


Edge AI and Micro Data Center

Storage Solutions for AI

Full-Stack Software for AI

Local AI Development



Rack-scale solutions: VR NVL72 and HGX

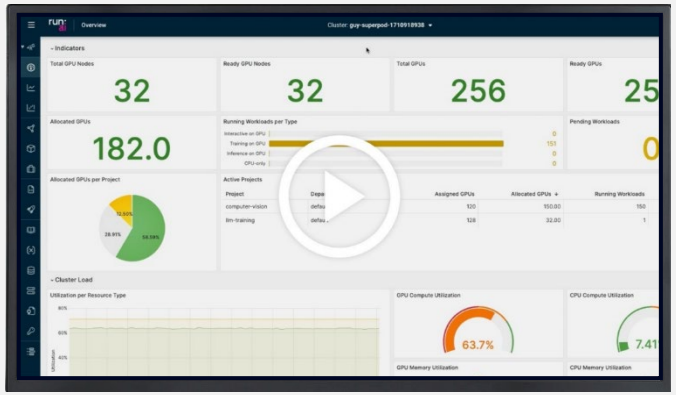
Supermicro at GTC26
Built to Accelerate AI

- **Supermicro's GTC26 exhibit** focuses on how AI total solutions are built, integrated, and brought online
- **Showcases Enterprise AI factories**, from edge and inference systems to rack-scale training platforms
- **Demonstrates validated, deployable building blocks** based on NVIDIA platforms
- **Connects compute, storage, networking, and software** into end-to-end, production-ready pipelines

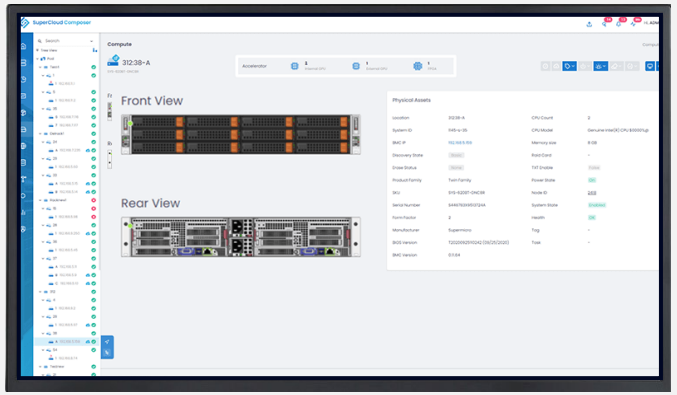
philipt@supermicro.com

Live Demos

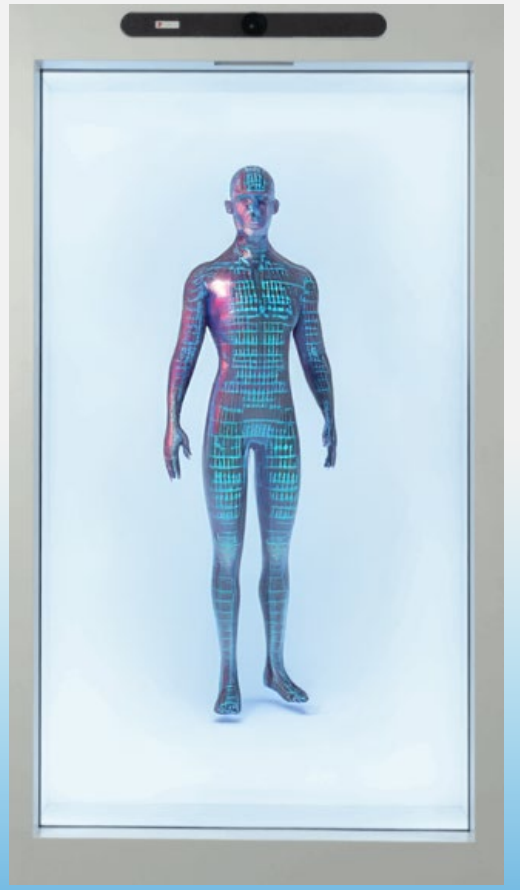
NVIDIA Run:ai



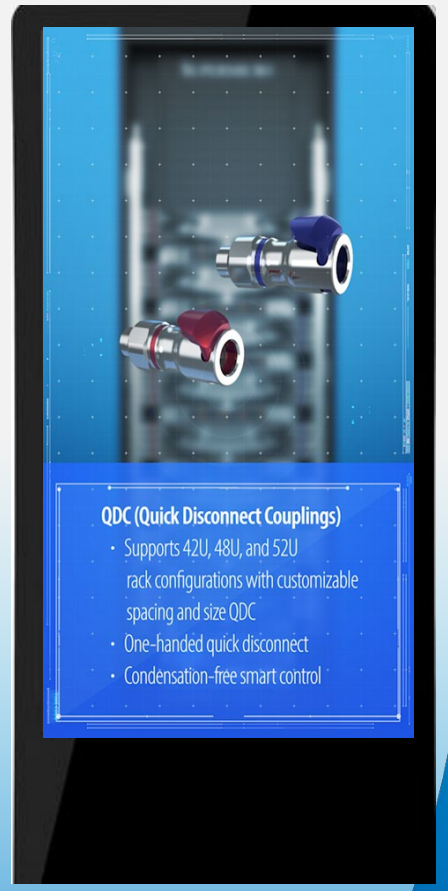
Supermicro Management Suite



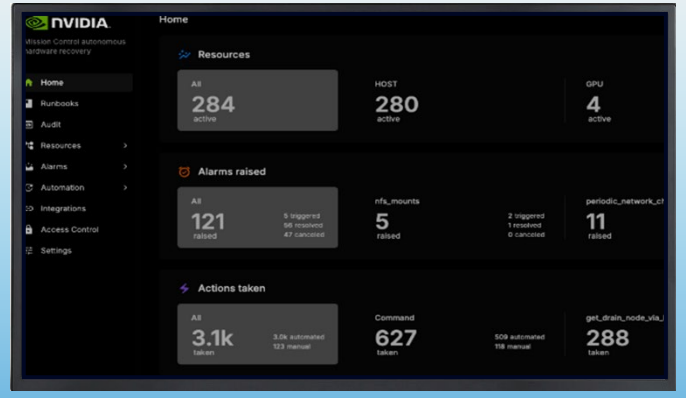
Hologram Concierge



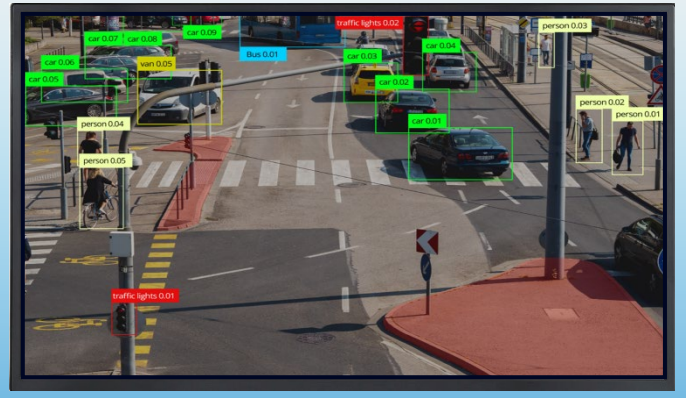
Touchscreen



NVIDIA Mission Control



Edge AI Machine Vision

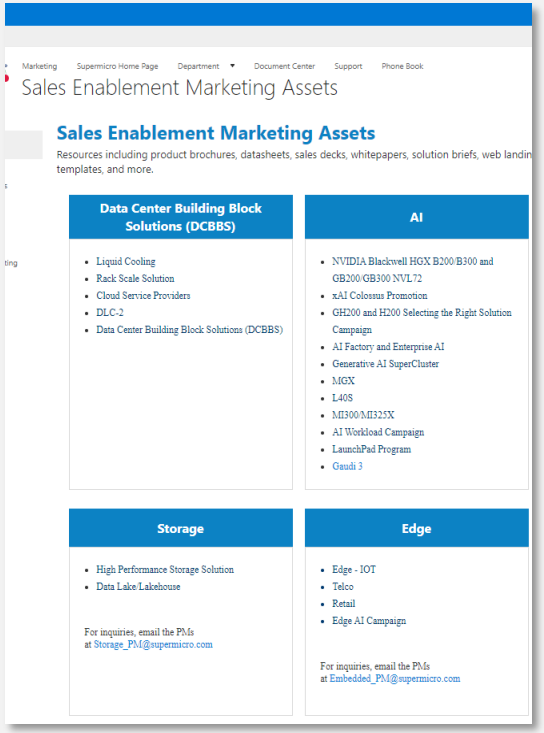


Live demos in booth combine both hardware and software to demonstrate real workloads and applications, ranging from an RTX PRO 6000 BSE-powered hologram, to NVIDIA's infrastructure and orchestration software.

philipt@supermicro.com

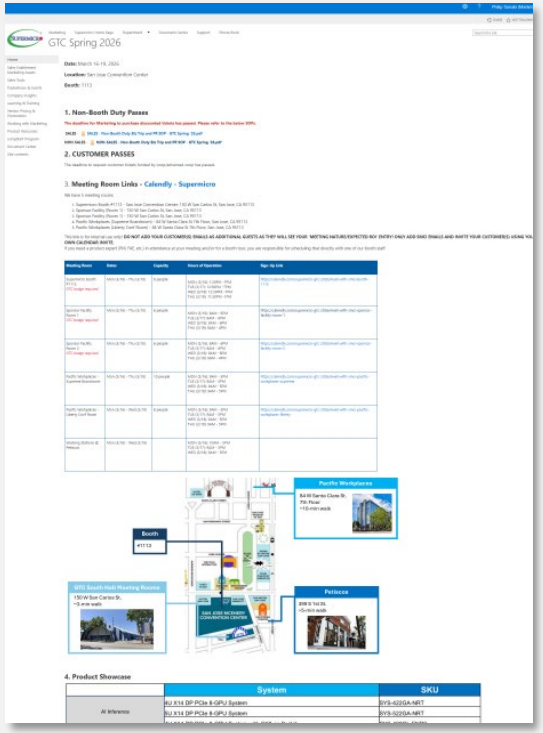
Resources (Top 3)

Sales Enablement Marketing Assets



https://portal.supermicro.com/sites/Marketing/SitePages/Sales%20Enablement%20Marketing%20Assets.aspx

GTC Spring 2026 Portal Page



https://portal.supermicro.com/sites/Marketing/Pages/GTC2026.aspx

Supermicro.com NVIDIA Page

supermicro.com/en/accelerators/nvidia

philipt@supermicro.com

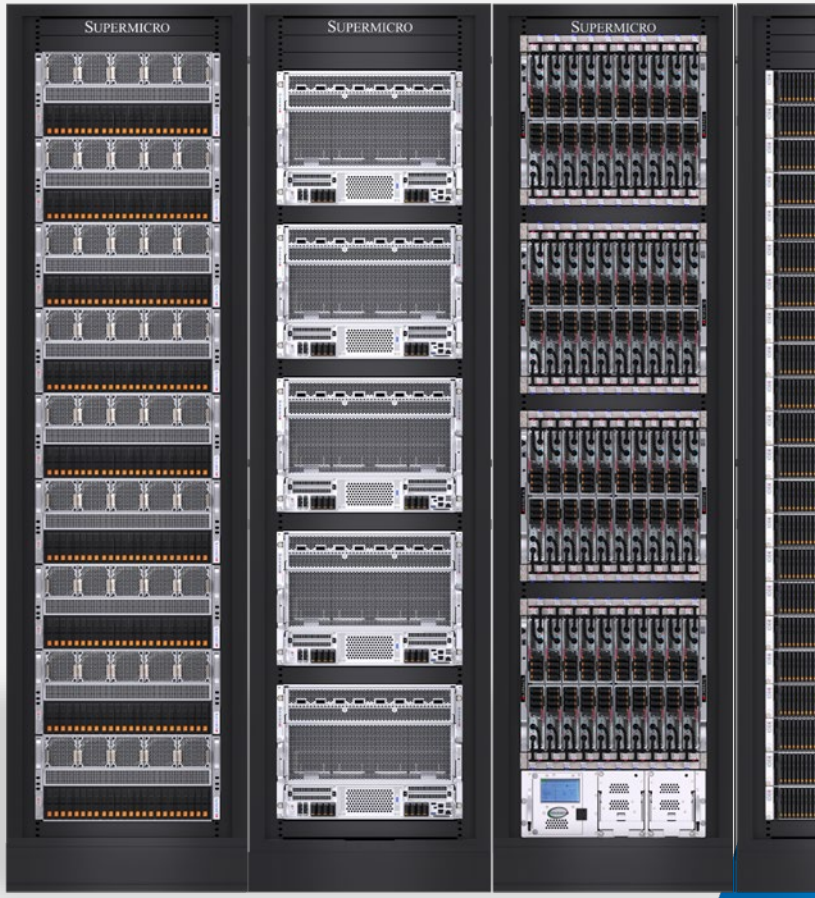
Next Week's Channel Training

Time: March 11th, 11:00am - 12:00pm

Title: The Power of Choice: Scaling Enterprise AI with the Industry's Widest Portfolio

Abstract: As AI shifts from massive training labs to distributed enterprise environments, a "one-size-fits-all" approach no longer works. This presentation showcases how Supermicro's Building Block Solutions provide partners with the industry's most expansive portfolio of NVIDIA-Certified Systems, including systems optimized for the new RTX Blackwell GPUs. We will explore how this modularity allows resellers to right-size infrastructure for any environment from compact edge nodes to liquid-cooled racks ensuring your customers achieve the fastest Time-to-Online in the market

Sign-up link:



philipt@supermicro.com

Thank you!

philipt@supermicro.com