A Blueprint for LLM Cluster Architecture

Scaling to the World's Largest Deployments

Agenda

Large Language Models New Computing Platform

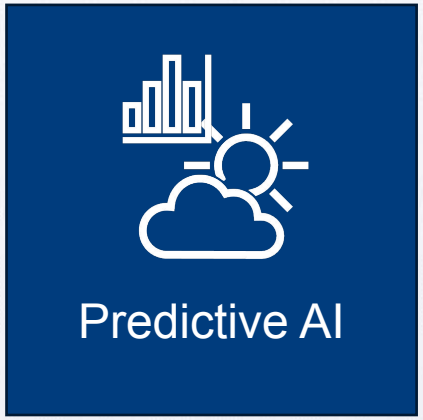
World’s Most Advanced AI Infrastructure Building Blocks

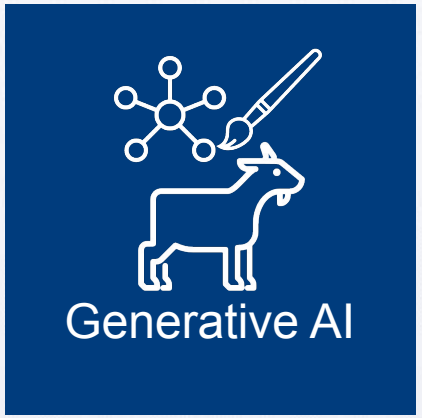
Rack as the New Unit of Compute

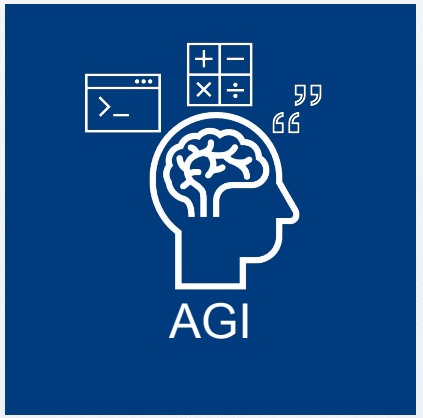
One of the world's largest AI Supercomputers

Supermicro’s complete plug-and-play Rack ‘n’ Roll AI cluster

Strategic Inflection Point



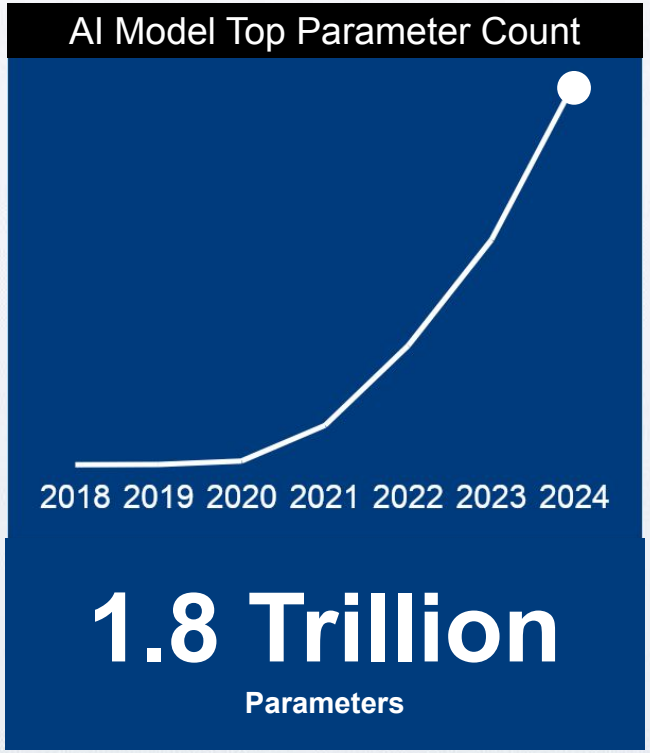




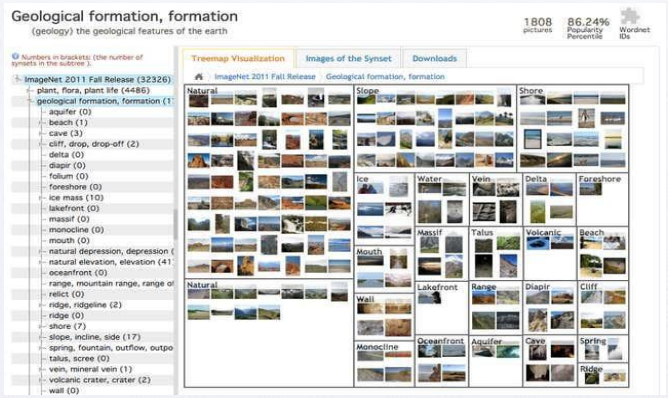
Why is this such a big deal?

Natural Language, Best programming Language
New Computing Platform LLM
New App Generative AI

Next Wave of AI Calls for Performance and Scalability



GPT-4



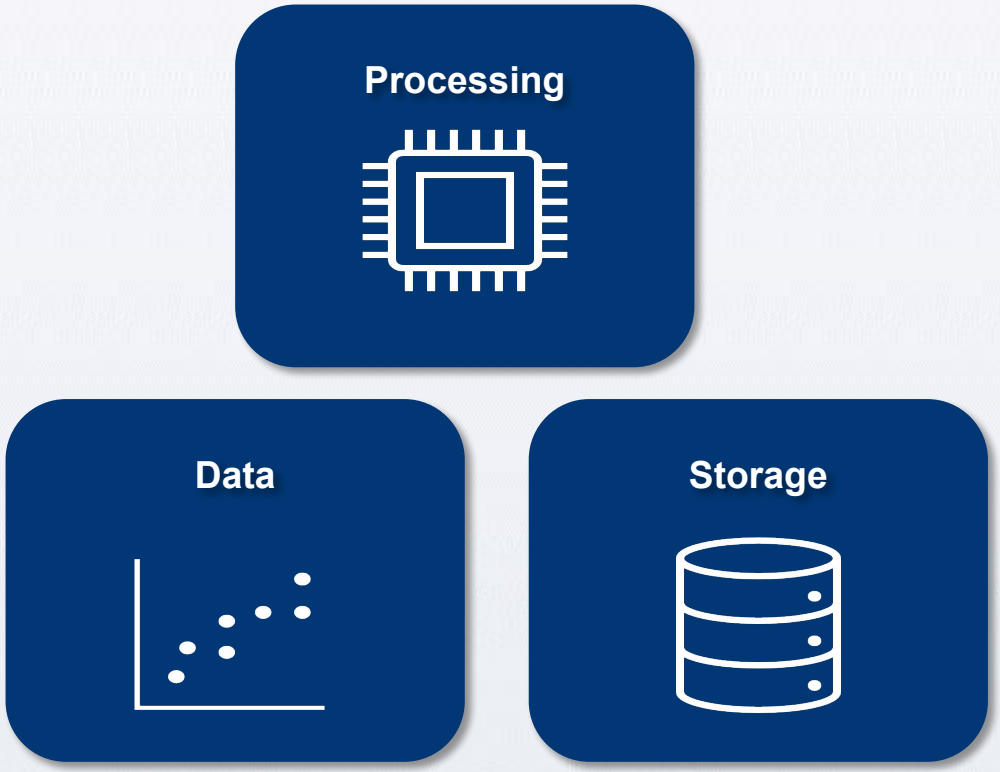
AlexNet

GPT-3

GPT-4

A massive business opportunity with continually increasing requirements: how do we meet this?

Modern Data Center Challenges



1. **Cater to disruptive growth of workloads that Industry has never witnessed before**
2. **Complexity of computation involved in addressing these workloads.**
3. **Calls for Full Stack Solutions**

AI Factories and AI Clouds by Supermicro and NVIDIA

AI Factories

- ☑️ Single or few workloads
- ☑️ Extremely large AI models

AI Cloud

- ☑️ Multi-tenant
- ☑️ Variety of workloads

8U 8-GPU System

GPU Server with NVIDIA HGX H100 8-GPU 80GB

Drop-in ready for NVIDIA H200 and NVIDIA B100

Proven Architecture for the Most Advanced AI Infrastructure

SYS-821GE-TNHR or AS -8125GS-TNHR

Key Features

- NVIDIA HGX H100 8-GPU with 80GB HBM3 memory per GPU
- 900GB/s GPU-GPU NVIDIA NVLink interconnect with 4x NVSwitch – 7x better performance than PCIe
- Dual 4th/5th Gen Intel Xeon® or AMD EPYC 9004 series processors
- Up to 32 DIMM slots: 8TB DDR5-5600
- PCIe 5.0 x16 1:1 networking slots for GPUs up to 400Gbps each supporting GPUDirect RDMA and Storage, and up to 8 U.2 NVMe drive bays
- Optimized thermal capacity and airflow to support CPUs up to 350W and GPUs up to 700W with air cooling
- Modular architecture for storage and I/O configuration flexibility

4U 8-GPU Liquid-Cooled System

NEW

GPU Server with NVIDIA HGX H100 8-GPU 80GB

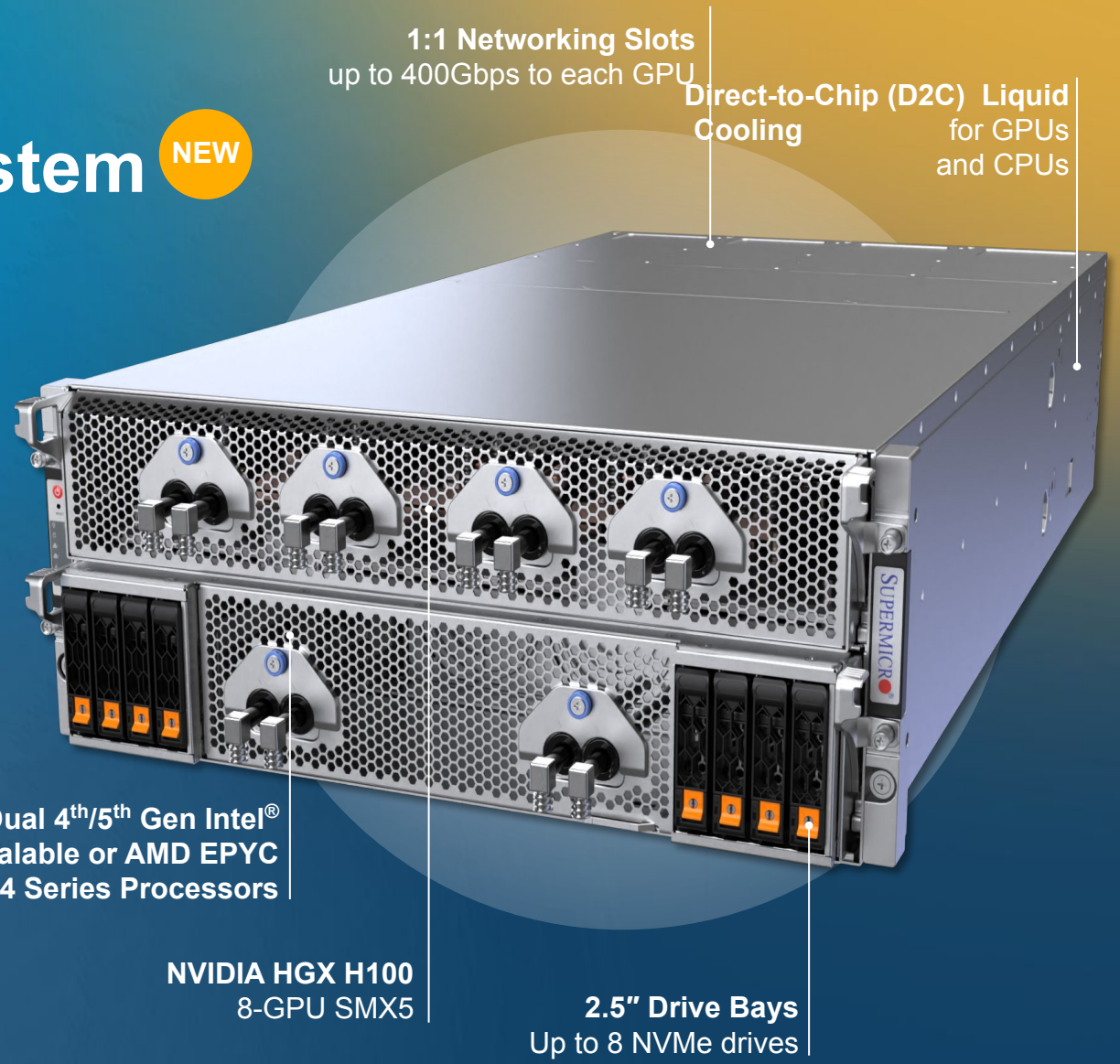
Drop-in ready for NVIDIA H200 and NVIDIA B200

Double Density for the Most Advanced AI Infrastructure

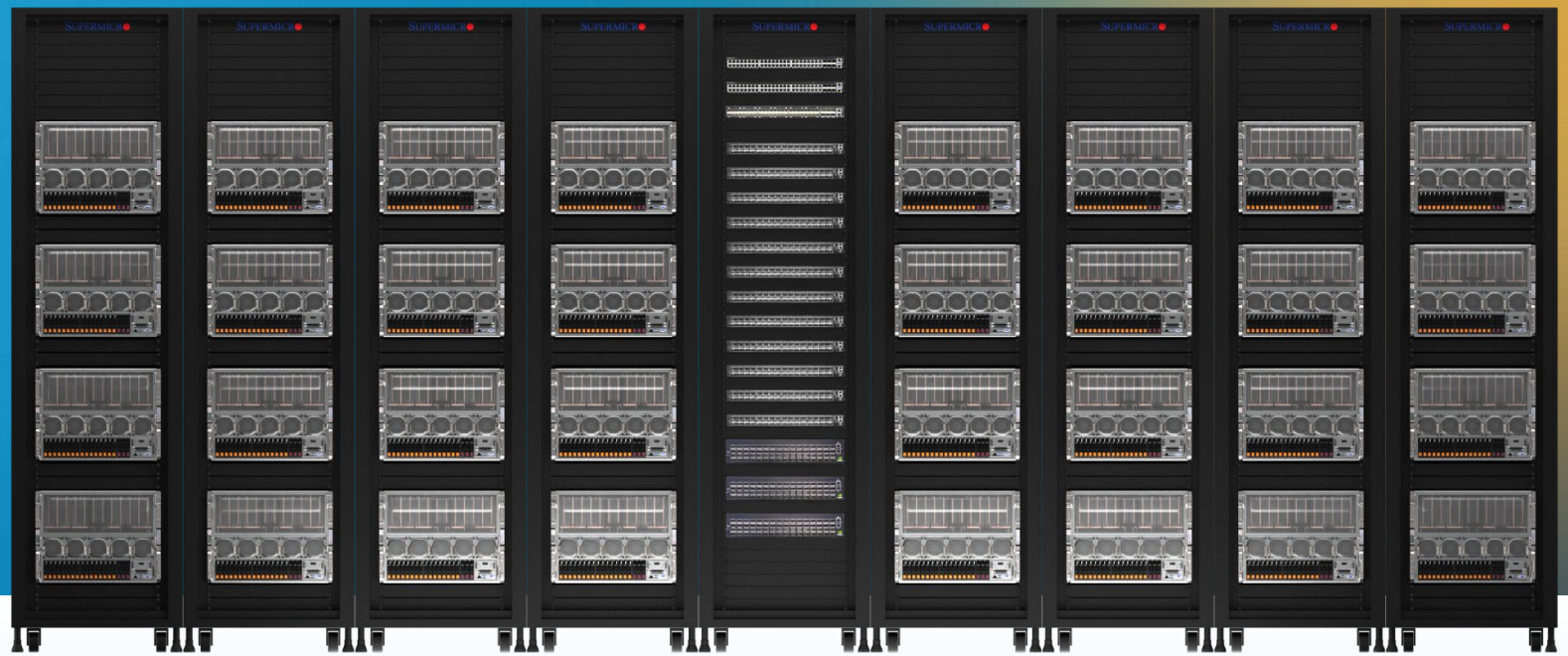
SYS-421GE-TNHR2-LCC or AS -4125GS-TNHR2-LCC

Key Features

- Highest density and efficiency with D2C liquid cooling for both GPUs and CPUs to optimize performance and energy cost
- NVIDIA HGX H100 8-GPU with 80GB HBM3 memory per GPU
- 900GB/s GPU-GPU NVLink interconnect with 4x NVSwitch – 7x better performance than PCIe
- Dual 4th/5th Gen Intel® or AMD EPYC 9004 series processors
- Up to 32 DIMM slots: 8TB DDR5-5600
- PCIe 5.0 x16 1:1 networking slots for GPUs up to 400Gbps each supporting GPUDirect RDMA and Storage, and up to 8 U.2 NVMe drive bays



RACK is the New Unit of Compute!



SuperCluster Scalable Unit:

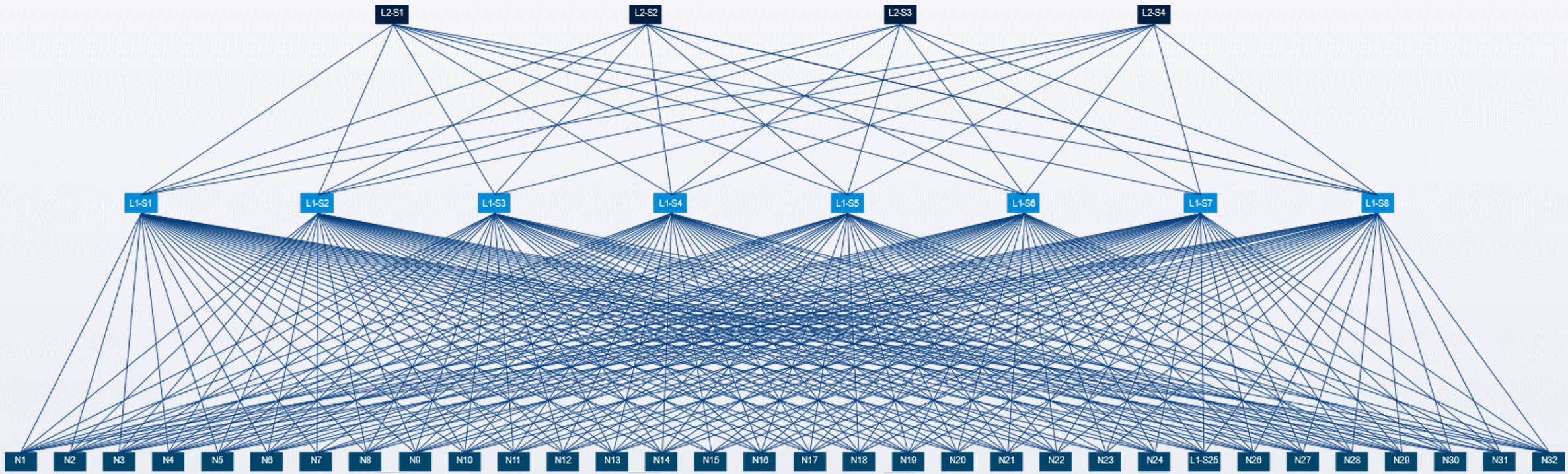
- 9 Rack Design
- 32x NVIDIA HGX H100
- 1+ EFLOPS AI
- 256 NVIDIA H100 GPUs per SU

- 20TB HBM3 GPU Memory
- 102.4Tbps Network B/W Non-blocking
- NVIDIA Quantum-2 InfiniBand

- Software: NVIDIA AI Enterprise, SLURM, Kubernetes
- Flexible 3rd Party SW Integration

- Flexible Storage Options
- Scalable to thousands of units (32 node SU)

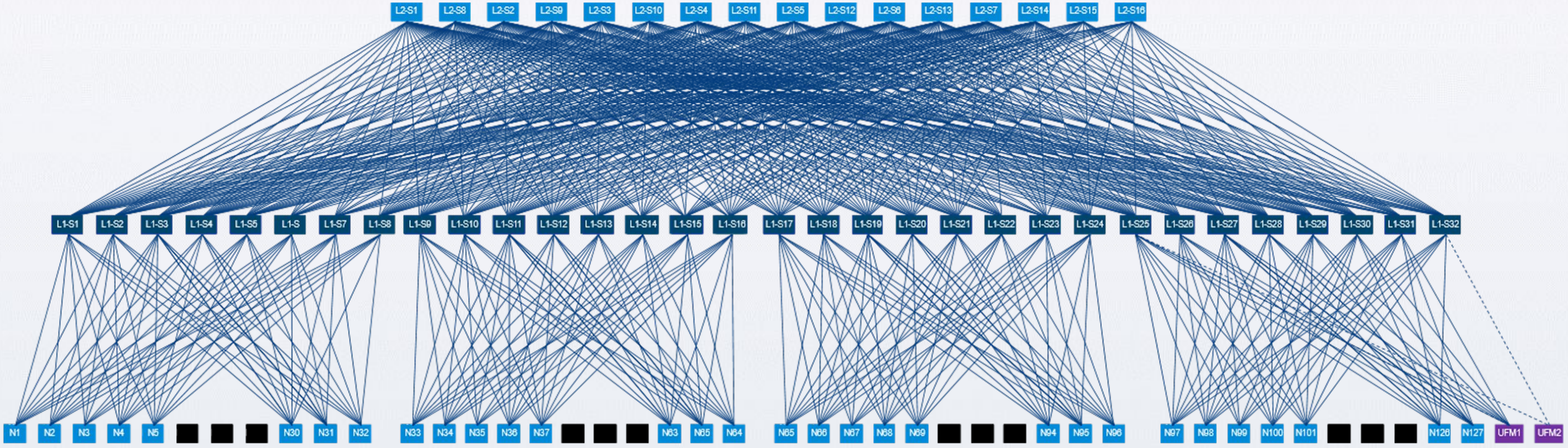
32-Node (SU) Rail- Optimized Network Topology



Networking Details

- 12x NVIDIA QM9700 64-Port InfiniBand NDR, 32 OSFP ports switches
  - Back-end compute non-blocking network
  - 8x Leaf switches (L1)
  - 4x Spine switches (L2)
- Each NVIDIA HGX H100 node has 8x back-end NVIDIA ConnectX-7 adapters, creates 1 leaf switch group / SU

127-Node SuperCluster (Rail- Optimized Network Topology)



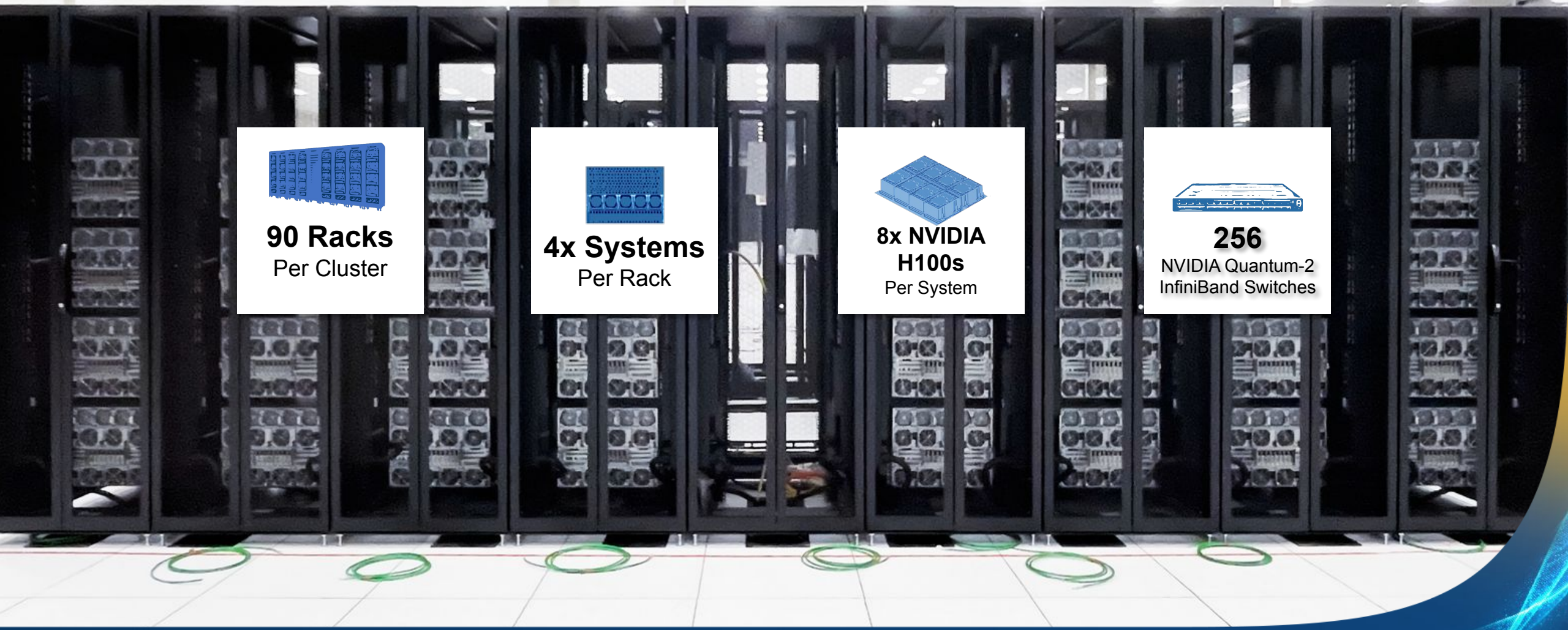
Networking Details

- 48x NVIDIA Quantum-2 QM9700 64-Port 400Gb/s Switches (32 OSFP ports)
  - Back-end compute non-blocking network
  - 32x Leaf switches (I-I)
  - 16x Spine switches (L2)
- Each NVIDIA HGX H100 node has 8x back-end NVIDIA ConnectX-7 adapters to create 1 leaf switch group
  - 4x switch groups/ scalable unit (SU) makes a Super Cluster
- 1x NVIDIA HGX H100 node removed for UFM SDN appliances
- Max cluster size up to 255-nodes with UFM

Case Study: Rack Scale SuperCluster Deployment

Customer: GPU Cloud Service Provider

Case Study: Rack Scale SuperCluster Deployment



90 Racks
Per Cluster

4x Systems
Per Rack

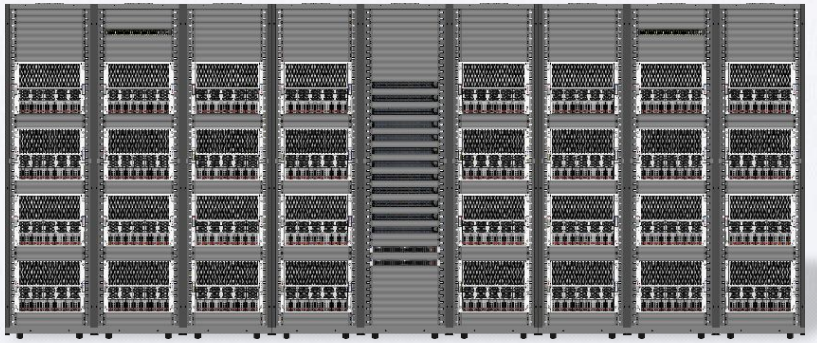
8x NVIDIA H100s
Per System

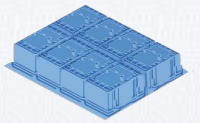
256
NVIDIA Quantum-2 InfiniBand Switches

Case Study: Rack Scale SuperCluster Deployment

102.4Tbps
InfiniBand NDR

9-Rack
Design



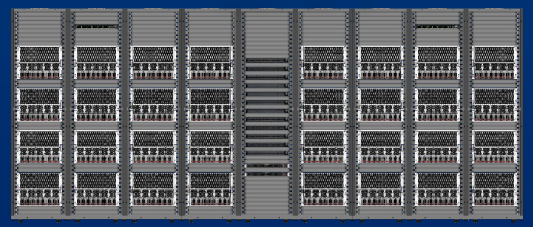


256 GPUs
NVIDIA H100

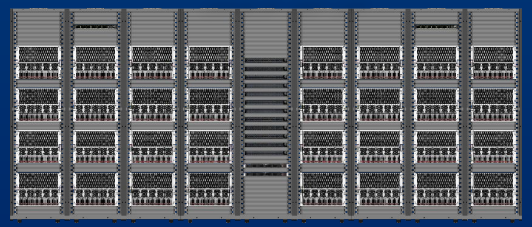
20TB HBM3
GPU Memory

3064 NVIDIA H100 GPUs

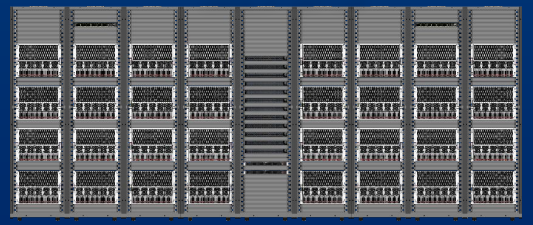
SU # 1



SU # 2

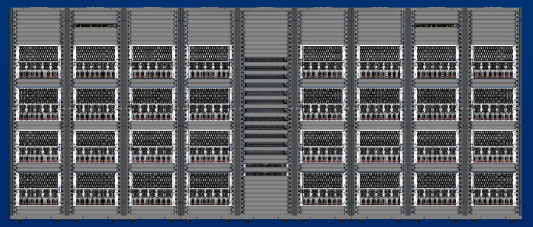


SU # 3



...

SU # 12



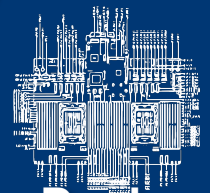
Increment scalable units to scale to ONE MEGA AI CLUSTER

Double-Density Liquid-Cooled AI SuperCluster

- 32 x 4U 8GPU systems
- 64 x NVIDIA H100/ H200 GPUs
- Identical Infiniband Compute Fabric
- 80KW per rack handled by SMC Direct to Chip LC solution
- AC vs. LC: Double density, halved Data Center space

Building Rack-Scale Solutions: Key Steps

1. Consultation


2. Design


3. Assembly


4. Testing


5. Deployment


6. Support

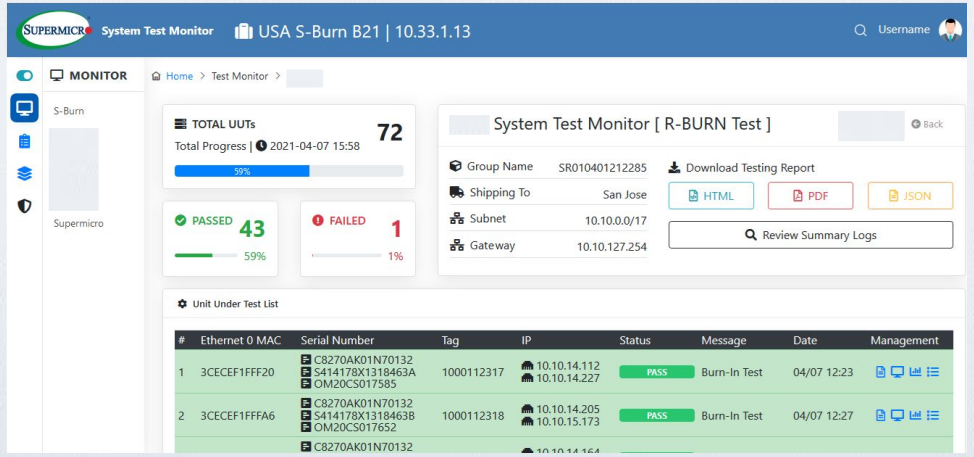

- One stop shop for complete plug & play rack solutions.
- Initial design to post-deployment
- POC (Proof of Concept) available for large scale projects
- L11/ L12 Solution Validation Lab: Complete cluster level testing to deliver PnP SuperClusters

SMC L11/ L12 Solution Validation Lab

L11: Node Level Assembly, configuration, testing

L12: L11 + Application Loading, Validating, Optimization

- ✔ L11: More than **30** tests can be performed by Rack mass production **(Full Automation)**, reports ready after test.
- ✔ L12: More than **56** different workloads/benchmarks can be performed by cluster level test.
- ✔ Flexibility to accommodate customers' specific testing requirements.



GPU Test Examples:

NVQual, cuda_mem, GPU validation, GPU linpack, HPGC, NCCL, RCCL, GEMM, Bandwidth, GPU MemoryPerf, DeviceQuery, MultiCopy, KFDTest, RVS, TransferBench, GST and gpu_burn, ResNet-50/101 and ImageNet etc.

A Quick Recap Supermicro's AI Building Blocks

- One of The Largest GPU Server Portfolio in the Industry
- Infrastructure Design Deployment And Management
- Peace of Mind with NVIDIA-Certified Modern AI Building Blocks
- A Team of Experts to Close Experience Gaps
- One Stop AI Infrastructure Place

www.supermicro.com