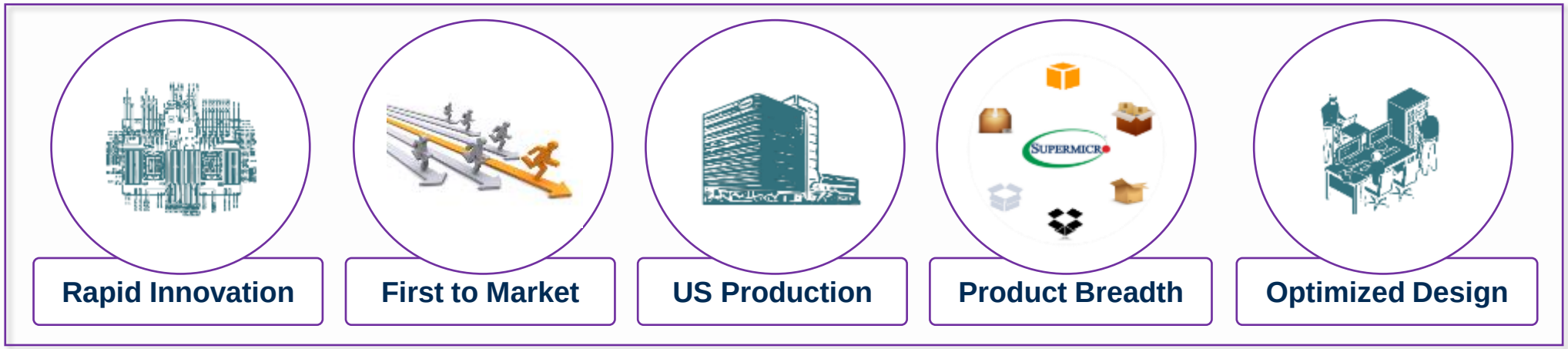
# Accelerate Everything

## NVIDIA-GPU Platform Leadership

# Supermicro Differentiators




| Competency | Supermicro | Competition |
|---|---|---|
| **Internal Hardware Design** | YES | LIMITED |
| **Broad and Optimized Portfolio** | YES | LIMITED |
| **First to Market / TTM** | YES | NO |
| **H/W Technology Leadership** | YES | LIMITED |
| **Total Solutions Hardware + Software** | YES | YES |
| **US Design and Manufacturing (Proximity to Key Technology Partners)** | YES | NO |
| **Flexible Engagement Models** | YES | LIMITED |
| **Deep Partner Relationships** | YES | YES |
| **Global Operation and Service** | YES | YES |

# The Industry's Broadest Portfolio of Servers

**Universal GPU**
Multi-Architecture Flexibility with Future-Proof Open-Standards

**PCIe GPU**
High Performance and Flexibility for AI, and 3D Simulation

**SuperBlade®**
Highest Density Multi-Node Architecture for HPC Applications

**Hyper**
Best-in-class Performance and Flexibility Rackmount Server

**BigTwin®**
Industry-leading Multi-node Architecture

**GrandTwin®**
Multi-Node Architecture Optimized for Single-Processor Performance

**FatTwin®**
Multi-node 4U Advanced Twin Architecture with 8 or 4 Nodes

**CloudDC**
All-in-one Rackmount Platform for Cloud Data Centers

**WIO**
Industry's Widest Variety of I/O Optimized Servers

**Petascale All-Flash**
High Performance, Low Latency with EDSFF E3.S and E1.S

**Enterprise Storage**
Cost Effective, High Capacity for Large-Scale Storage

**Multi-Processor**
Highest Performance and Flexibility for Enterprise

**Mainstream**
Cost Effective Systems for Everyday Applications

**Hyper-E**
Best-in-class Performance and Flexibility for Edge Data Centers

**SuperEdge**
High-Density Computing and Flexibility at the Intelligent Edge

**IoT/5G**
Compact Form Factors for 5G and Edge computing

**SuperWorkstation**
Data Center Power in Portable Form Factors

**MGX**
Modular Building Block Platform with wide CPU and GPU support

**Family Overview**

# Supermicro GPU Systems

## HGX Systems


CPU: INTEL AMD

GPU: H100 SXM

- Maximum acceleration with interconnected 8-GPU and 4-GPU per system for large AI models with pool of High-bandwidth GPU memory.
- Proven system architecture. Validated in real-world AI deployments.

## PCIe Systems


CPU: INTEL AMD

GPU: PCIe

- Versatile acceleration platforms with up to 10 PCIe GPUs for AI inference, fine-tuning, Graphical AI applications.
- H100 PCIe for maximum AI performance, L40S for cost effective AI + graphics workloads.

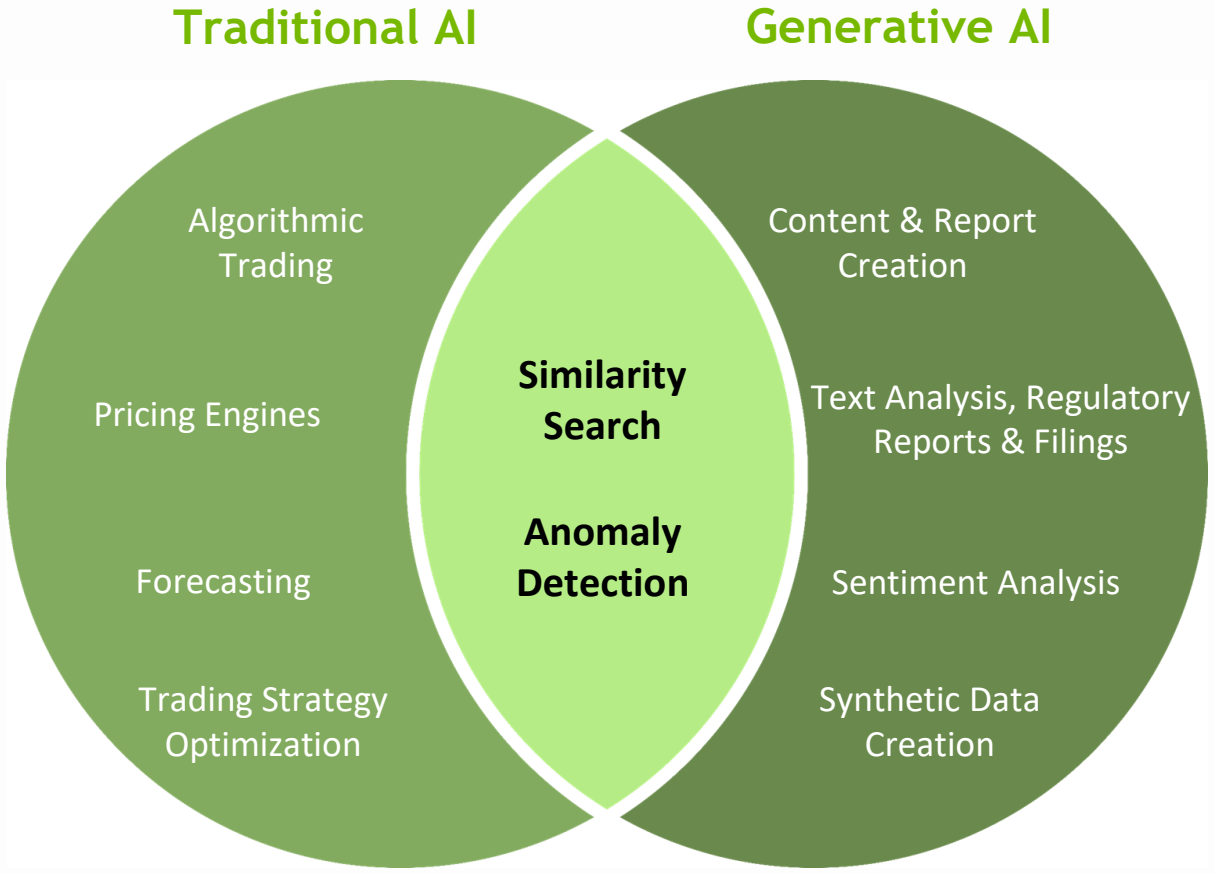
## MGX Systems


CPU: ARM INTEL

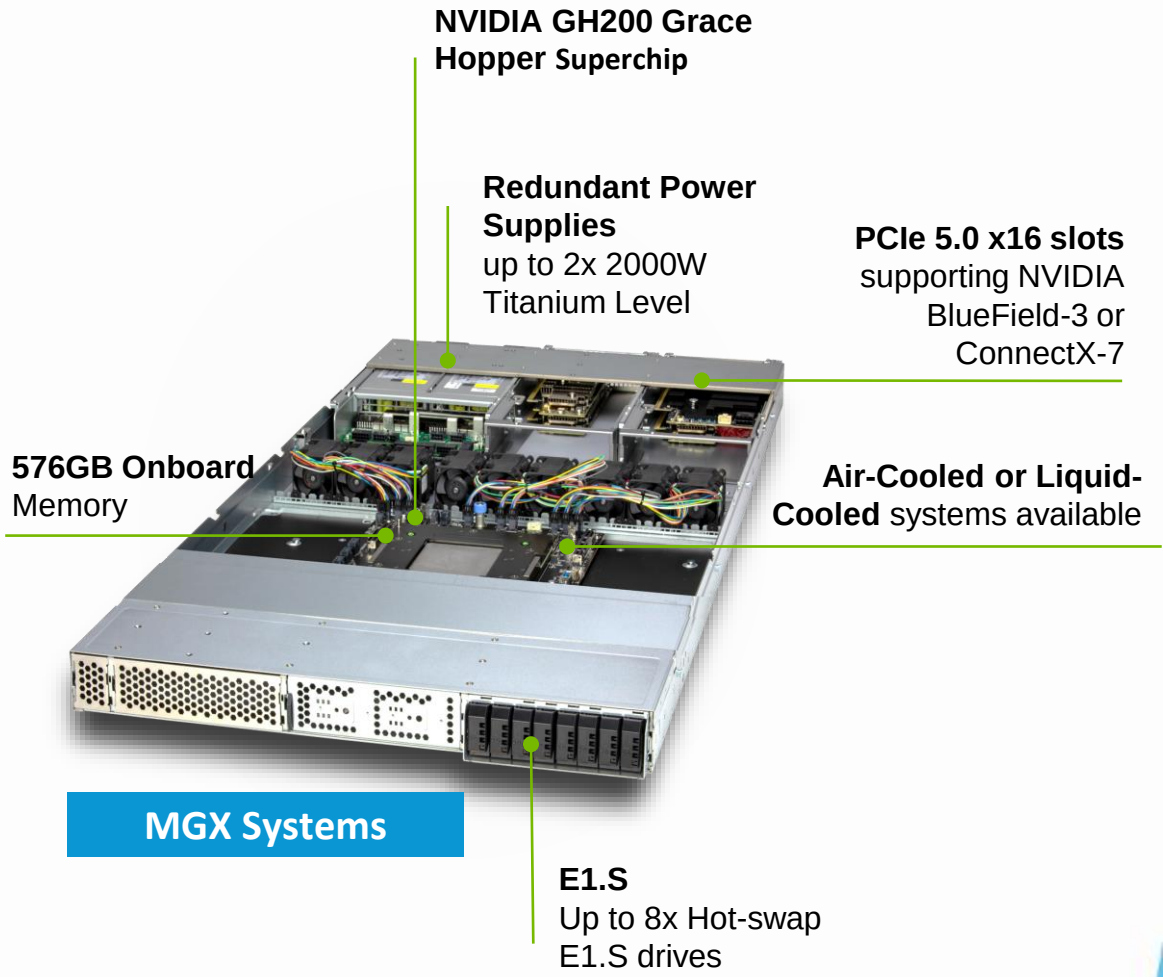
GPU: GH200 PCIe

- Grace Hopper Superchip systems with large shared memory for large AI model or high-volume inference.
- Modular PCIe GPU platform supporting ARM and x86 CPUs.
- Compact form factors for high computing density and scalability.

# Combining Traditional HPC and Generative AI in Capital Markets




**Competitive Advantage Use Cases**



# Grace-Hopper Superchip

## One Stop Platform for HPC + AI Accelerating Capital Markets

### AI (Neural nets) – LLM/GenAI

- AI Unstructured Data using NLP with LLMs, Other Systematic Trading Algos
- Framework – PyTorch/TensorFlow, NVIDIA NIM e.g. RAG based retrieval, NVIDIA AI Enterprise

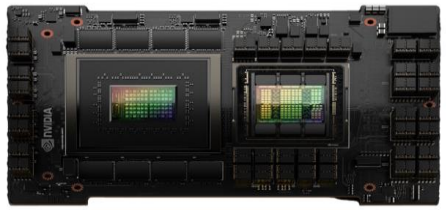
### Quant Finance / HPC

- Pricing, Risk (MC Sim, Margin, FRTB, CVA, SIMM, XVA) & Back testing
- Framework – CUDA C/C++, Parallel Algorithms C++, NVIDIA Accelerated Python – RAPIDS, Open ACC

### Data Processing – ETL/ML

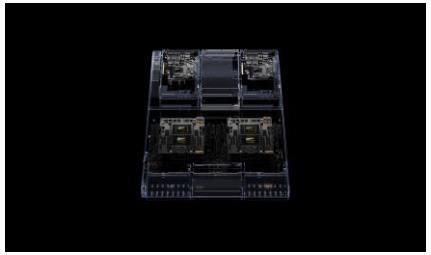
- Feature Engineering, Data Prep, & Data Science (e.g., XGBOOST)
- Framework – NVIDIA Accelerated Python on Steroids - RAPIDS, Spark on GPU

NVIDIA Grace Hopper GH100 CPU + GPU for LLMs and inference



7X Faster data transfers and queries than PCIe Gen 5

NVIDIA Grace Hopper GH200 CPU+ GPU for LLMs and inference



**3.5x** more memory and **3x** more bandwidth than the current generation

# TCO, ROI and Productivity in Customer cases

Resources and NVIDIA GTC Customer stories

## Capital One Example

**Before GPUs:**
26-Hour
Run Time

**100x**
Training Time
Improvement

RAPIDS

**98%**
Decrease in
total costs

dask

**After GPUs:**
15-Minute
Run Time

## Generic Savings Calculator Per Year

| | | | |
|---|---|---|---|
| No. of Data Science models | 50 | 400 | 5000 |
| No. of AI models (e.g., LLMs) | 5 | 40 | 500 |
| **Estimated Cost Savings** | $0.5 M | $4-5 M | $65M |

## Reduce TCO by 71%

- Reduce No of Servers (e.g., CVA Credit valuation) — 100 to 4 servers and 7X speeds up on Grace Hopper
- Baseline cost — Before 100 Servers @15,000 = 1,500,000
- After - 4 Servers @ 443,929
- Reduced TCO saving 71% (1,056,071) for one use case
- Avoid costly Maintenance, Reduced Admin server costs

## Increase ROI - Do More with Performance/Power

- Do Generative AI + Data Science + Quant Finance
- Multiple accelerated workloads > 20x the Speeds for Quant Finance and Data Science with lower power
- **Leveraging Spark 3 and NVIDIA's GPUs to Reduce Cloud Cost by up to 70% for Big Data Pipelines - Paypal**

## Accelerated Software Productivity

**Choice - Language, Framework**

- NVIDIA HPC SDK, C/C++
- CUDA, ISO C++ parallelism,
- RAPIDS in Python, NVIDIA SPARK on GPU, RAPIDS / DASK with single line code changes
- NVIDIA AI Enterprise (DL Model and Framework) and NVIDIA NIM inferencing

Other Resources
STAC A2 blog, STAC A3 blog
Book limit order prediction blog
JPMC Risk Calculations

Partners
**Example ISVs – Murex, KX, Kore.ai**
Algo Trading, CFO GPT,
KYC/AML/Fraud, NVIDIA AI

NVIDIA GTC Sessions:
- Capital One
- Cohen & Steers
- Wells Fargo
- Citibank NN For Exotic
- CBOE
- Bank of America

NOT STAC BENCHMARKS

# Liquid-Cooled Rack: Sample Configurations

Supermicro NVIDIA
MGX Systems with GH200

Supermicro NVIDIA
HGX 4U 8-GPU Systems

Supermicro NVIDIA
HGX 8U 8-GPU Systems

Which NVIDIA liquid cooled rack level solutions work best for your needs?

**DISCLAIMER**

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION