PRODUCT BRIEF

SUPERMICRO AND AMD DELIVER SYSTEMS TO UNLOCK MILLISECOND ADVANTAGES IN QUANTITATIVE TRADING

Supermicro APU Server: AS -4145GH-TNMR

TABLE OF CONTENTS

Executive Summary

The competitive race for alpha in quantitative trading has increasingly boiled down to two core bottlenecks: latency and compute throughput. Traditional infrastructure—relying on separate CPU and GPU subsystems connected by narrow buses—is rapidly becoming an unacceptable barrier for firms seeking real-time, AI-augmented trading strategies.

This paper presents a solution blueprint built around the Supermicro AS-4145GH-TNMR server, powered by the AMD Instinct™ MI300A APU. Together they deliver a unified CPU/GPU architecture with shared high-bandwidth memory, dramatically lowering data-transfer latency and enabling new performance horizons for trading firms. Moreover, the same class of hardware has been selected for deployment in exascale-class systems, attesting to its pedigree and robustness [3][4].

By adopting this unified architecture, trading firms can achieve:

- Sub-millisecond execution latency for quantitative strategies (subject to network/exchange distance).
- Deeper, richer modeling via complex ML/AI workflows enabled by large memory bandwidth and coherent compute.
- Lower Total Cost of Ownership (TCO) and improved infrastructure efficiency—including ESG benefits from energy-optimized design [6].

Industry Challenges in Quantitative Trading

Quantitative trading today demands decision-making at microsecond levels. Algorithmic models are becoming more complex, data volumes continue to explode, and integrating machine learning and AI into trading strategies is essential. Legacy infrastructures struggle to keep pace, mainly due to a “data tax”: the latency and energy penalty incurred when data moves between discrete CPU memory and GPU memory pools.

As strategies evolve toward ML/AI, systems must provide massive memory bandwidth, high concurrency, and ultra-low latency across the entire pipeline — from data ingestion to inference to execution. Overcoming these challenges requires a unified architecture that reduces data movement and simplifies programming.

AMD Instinct MI300A: Architectural Innovation

The AMD Instinct MI300A APU integrates CPU cores and GPU compute units into a single package with shared, unified HBM3 memory and coherent interconnects, reducing or eliminating CPU↔GPU data-transfer bottlenecks that plague discrete architectures [1].

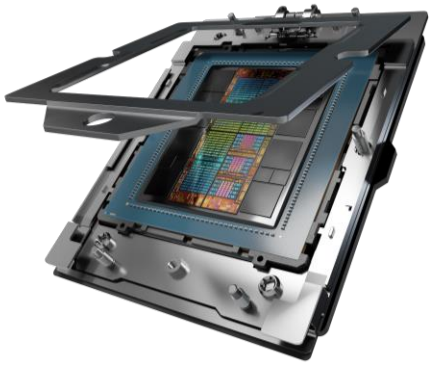


Figure 1 - AMD MI300A

Key Technical Highlights:

- 24 Zen 4 CPU cores + 228 CDNA 3 GPU compute units in one APU socket [1].
- 128 GB HBM3 unified memory per APU; shared address space (CPU+GPU) [1][4].
- Up to ~5.3 TB/s on-package peak theoretical memory bandwidth [1][4].
- 4th-Gen AMD Infinity Fabric with coherent memory [1][4].
- AMD ROCm™ 6+ with HIP and directive-based offload (OpenMP) supports the APU programming model [5].

Why this matters for quant trading:

- Eliminating host↔device copies reduces latency in inference/decision loops.
- Unified memory simplifies software pipelines so teams focus on algorithms, not data logistics.
- High bandwidth/capacity supports streaming datasets and fast model recalibration [1][5].

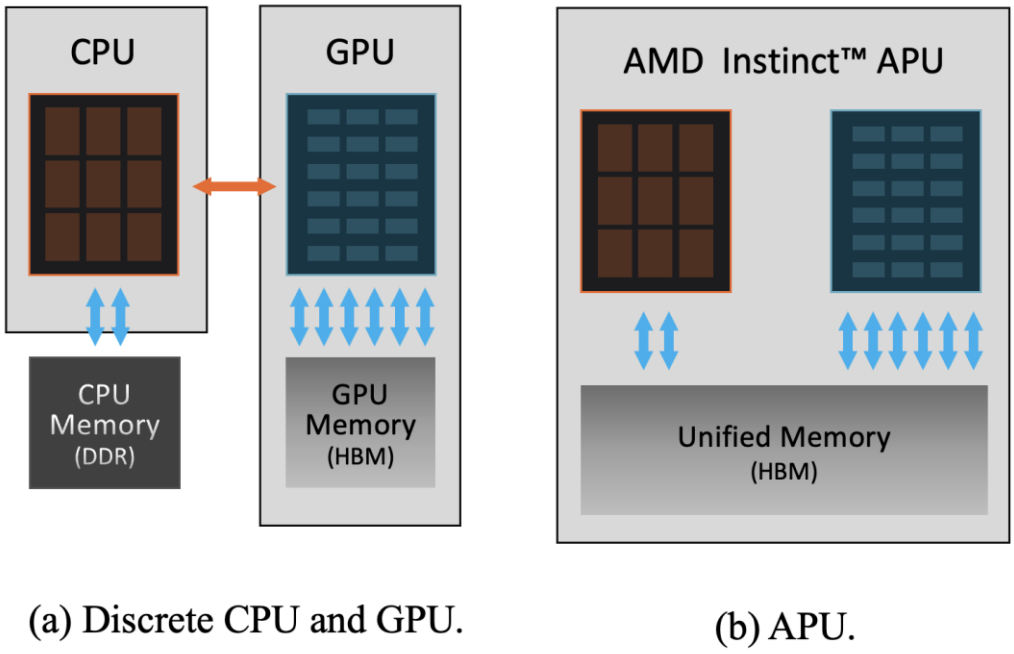


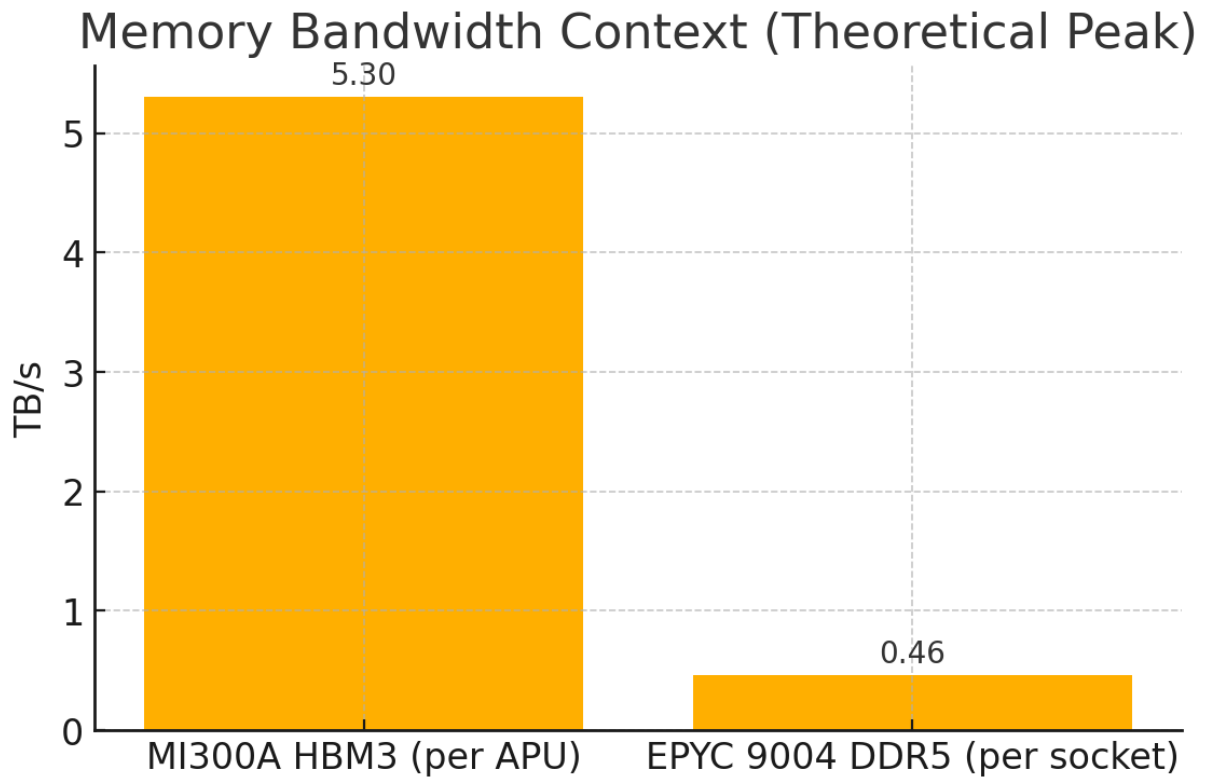
Figure 2 - Discrete CPU+GPU vs Unified MI300A APU data paths (conceptual).

Supermicro AS -4145GH-TNMR: System Overview

The Supermicro AS-4145GH-TNMR is engineered for dense, high-performance MI300A deployments [2]. System highlights include:

- 4× AMD Instinct MI300A APUs per 4U chassis; 512 GB unified HBM3 (128 GB/APU) [2].
- PCIe Gen5 expansion and AIOM/OCP NIC 3.0 for high-speed networking [2].
- Up to 24× 2.5" SAS3/SATA, 8–16× NVMe bays (config-dependent); dual M.2 NVMe [2].
- 10 heavy-duty fans; 4× 2700W redundant Titanium-level PSUs [2].
- ENERGY STAR certified configuration available [6].

Implication: exceptional compute density and I/O enable real-time market-data ingestion, rapid feature generation, and co-located inference. Unified memory reduces software complexity for quant teams while maintaining peak throughput [1][2][5].



Sources: MI300A data sheet (5.3 TB/s) [1]; EPYC 9004 DDR5-4800 ~460.8 GB/s ≈ 0.4608 TB/s [12].

Figure 3 - Memory bandwidth context (MI300A HBM3 vs EPYC 9004 DDR5 per socket) — theoretical peaks for architectural context [1][12].

Greenfield Quant Solution Design

Reference architecture for Tier-1 FSI firms combining supercomputer-grade hardware with low-latency market connectivity.

- Compute: 8-node cluster; each node 4× MI300A APUs → 32 APUs total; ~4 TB aggregated unified HBM3 [1][2][4].
- Software: ROCm 6.x; HIP + OpenMP offload; PyTorch & TensorFlow for research-to-production pipelines [5].
- Storage: NVMe-based primary tier for hot tick data; scalable object store for history & features.
- Networking: AIOM/OCP NIC 3.0; co-location or ultra-proximate interconnect to exchanges.

- Execution Path: streaming ingestion → feature calc → model inference → order decision — all in unified memory to avoid copies [1][5].
- Observability: watts-per-trade, trades-per-second, cost-per-latency-bucket KPIs for TCO and ESG tracking [6].
- Resilience & Security: redundant PSUs/fans; firmware hardening; out-of-band management; role-based access control.

Competitive Advantage for FSI Providers

Latency as a strategic differentiator:

- Unified memory removes host↔device copies in critical loops; supports sub-ms inference (network permitting) [1][5].
- Larger on-package bandwidth/capacity enables more complex features and real-time models [1][4].
- Independent finance benchmarks validate the class of GPU acceleration for trading workloads (examples): STAC-ML Markets audited inference records on Supermicro GH200; STAC-A2 audits show major Monte Carlo speedups across vendors [7][8][9].

Infrastructure-level advantage:

- High density + ENERGY STAR configuration + simplified code paths → lower TCO and faster time-to-market [2][5][6].
- ESG alignment via energy-efficient design and certified configurations [6].

Conclusion

The era of discrete CPU and GPU systems—plagued by latency penalties and software complexity—is ending. The Supermicro AS-4145GH-TNMR, powered by AMD Instinct MI300A, represents a transformative shift for latency-sensitive, AI-driven trading. For firms aiming to stay ahead, the question is no longer "if" to adopt unified compute, but how quickly to deploy it. This architecture provides the toolset to build a future-ready quantitative platform, secure a lasting strategic advantage, and lead the market.

Glossary

- APU (Accelerated Processing Unit): A chip that integrates CPU and GPU cores into a single package.
- HBM3 (High-Bandwidth Memory Gen3): A high-speed, low-latency memory technology closely coupled with compute units.
- FSI (Financial Services & Insurance): Industry segment focusing on financial products, trading, and risk management.
- ROCm (Radeon Open Compute): AMD's open software platform for heterogeneous CPU/GPU computing.
- AIOM (Advanced I/O Module): A flexible, high-speed networking expansion interface (e.g., OCP NIC 3.0).

For More Information

Supermicro AMD APU https://www.supermicro.com/en/products/system/gpu/4u/as%20-4145gh-tnmr

References:

[1] AMD Instinct MI300A APU Data Sheet — https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300a-data-sheet.pdf
[2] Supermicro AS-4145GH-TNMR Datasheet — https://www.supermicro.com/en/products/system/datasheet/as-4145gh-tnmr
[3] LLNL: El Capitan verified as world's fastest (MI300A APUs) — https://www.llnl.gov/article/52061/lawrence-livermore-national-laboratorys-el-capitan-verified-worlds-fastest-supercomputer

[4] LLNL HPC: Using El Capitan Systems — Hardware Overview — https://hpc.llnl.gov/documentation/user-guides/using-el-capitan-systems/hardware-overview
[5] ROCm Blog: MI300A — Exploring the APU Advantage — https://rocm.blogs.amd.com/software-tools-optimization/mi300a-programming/README.html
[6] ENERGY STAR: Supermicro AS-4145GH-TNMR Listing — https://www.energystar.gov/productfinder/product/certified-enterprise-servers/details/4005728
[7] STAC-ML Markets: New record with Supermicro GH200 submission — https://stacresearch.com/news/smc250910/
[8] STAC Report: Intel GPUs under STAC-A2 (derivatives risk) — https://docs.stacresearch.com/news/INTC230927
[9] STAC News: Oracle Cloud + NVIDIA GPUs under STAC-A2 — https://docs.stacresearch.com/news/NVDA231026
[10] Budish, Cramton & Shim (2015), QJE: The HFT Arms Race — https://academic.oup.com/qje/article/130/4/1547/1916146
[11] Bartlett & McCrary (2019): SIP latency microseconds (working paper) — https://www.law.berkeley.edu/wp-content/uploads/2019/10/bartlett_mccrary_latency2017.pdf
[12] AMD EPYC 9004 bandwidth (Supermicro) — https://www.supermicro.com/en/support/resources/cpu-amd-epyc-9005-9004-7003

Appendix

A. System Configuration Example:

A representative deployment: 8-node cluster of AS-4145GH-TNMR servers, each housing 4 MI300A APUs → 32 APUs total. With 128 GB HBM3 per APU, the cluster provides ~4 TB of aggregated unified HBM3. Pair with ~1 PB NVMe storage for high-throughput, low-latency data pipelines.

B. Performance Benchmarks (Illustrative)

- Sub-millisecond latency for trade execution workflows is achievable when leveraging the unified memory architecture (actual results depend on network, exchange distance, and algorithm complexity).
- Model training/iteration cycles can be up to ~3× faster than legacy discrete CPU+GPU systems, due to unified memory and fewer data-movement bottlenecks [1][5].
- Power consumption may be ~40% lower on a per-trade basis, reflecting higher compute efficiency and reduced overhead (validate in your environment).

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, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

AMD

AMD is the high performance and adaptive computing leader, powering the products and services that help solve the world's most important challenges. Our technologies advance the future of the data center, embedded, gaming and PC markets.

Founded in 1969 as a Silicon Valley start-up, the AMD journey began with dozens of employees who were passionate about creating leading-edge semiconductor products. AMD has grown into a global company setting the standard for modern computing, with many important industry firsts and major technological achievements along the way. Visit www.amd.com