SOLUTION BRIEF

SUPERMICRO, AMD, AND MYRTLE.AI CREATE AN OPTIMIZED SOLUTION TO BREAK THE MICROSECOND BARRIER IN FINANCIAL AI LATENCIES

Eliminating the Inference Gap in ML for Trading Decisions

Supermicro AMD Server - AS -2015CSTNR

TABLE OF CONTENTS

Executive Summary

In capital markets, the search for alpha has entered a new era. Modern AI models can now generate market signals beyond the reach of any traditional approach. However, using AI successfully in finance faces a familiar barrier: latency. For trading firms, even the strongest signal has little value if it arrives outside the microsecond execution window. Until now, firms have had to make a trade-off: use simpler models to preserve speed or accept slower execution to use more sophisticated intelligence. By leveraging AMD Versal™ Adaptive SoCs and Silicom's server adapter technology, this solution has set new world records in the STAC-ML™ Markets (Inference) benchmarks, providing the deterministic, ultra-low-latency performance required for modern electronic trading.

This solution brief outlines the performance breakthroughs and business value of the myrtle.ai VOLLO™ acceleration stack running on a Supermicro AS -2015CS-TNR server, as showcased in the recently released STAC ML benchmark report.

Industry silicon leaders drive CPU and GPU innovations that have made monumental strides in advanced AI inference, effectively raising the "performance floor" for general-purpose financial workloads. However, at the extreme edge of low-latency trading, the "jitter" inherent in batching-dependent workloads remains a critical obstacle. General-purpose accelerators do not have latency-optimized architectures and often struggle with "tail latency"—unpredictable performance spikes during periods of high market volatility that can lead to slippage and missed fills.

Understanding the STAC-ML™ Benchmark Results

Understanding the STAC-ML™ Benchmark Results: For the first time, a system has broken the 2-microsecond barrier for the 99th percentile LSTM inference. The Securities Technology Analysis Center (STAC®) provides objective, audited benchmarks for the financial industry. The STAC-ML Markets (Inference) benchmark suite (SUT ID: MRTL260323) specifically tests the ability of a system to perform Long Short-Term Memory (LSTM) inference on time-series market data. It provides an "apples-to-apples" comparison of latency, throughput, and efficiency.

- Three distinct LSTM models (A, B, and C) represent various levels of complexity.
- Measures "tick-to-model" latency—the time from receiving data to producing a signal—at varying throughputs.
- The audited system achieved the lowest 99th percentile (99p) latency ever reported for all models, proving its capability in "extreme" market conditions.

Why STAC-ML™ Benchmark is Relevant – The "Inference Gap"

As trading firms move from simple linear models to deep learning, they often face a "latency tax." Traditional compute (CPUs) or general-purpose accelerators (GPUs) often introduce jitter and higher latencies due to batching requirements. This solution is relevant because it eliminates the "Inference Gap." Trading firms can now deploy the predictive power of LSTMs and other, more complex models at the speed of simpler models, enabling a smarter, more responsive trading strategy.

The STAC-ML performance translates directly to real-world financial workloads:

- Algorithmic Price Prediction: Analyzing the limit order book (LOB) to predict short-term price movements.
- Liquidity Analysis: Identifying "hidden" liquidity and optimizing order routing.
- Real-time Fraud & Compliance: Checking trades against compliance models in-line, without delaying the execution path.
- Market Making: Adjusting quotes instantly based on shifting correlations across different asset classes.

Real-World Financial Workloads and Shift to AI-Driven Trading

The financial services industry is undergoing a paradigm shift from rule-based systems to machine learning-driven strategies. This benchmark is relevant because it addresses the "Inference Gap", the delay between receiving a market tick and generating a trade signal.

Benefits of using FPGAs for Modern Quants

- Parallelism: FPGAs process data streams in a truly parallel fashion, which is ideal for ML inference.
- Direct Wire-to-Inference: The Silicom Artena card (featuring the AMD Versal™ VP1802 FPGA) allows for data to move from the network interface to the ML model with minimal CPU intervention.
- Future-Proofing: As STAC-ML evolves to include larger and more diverse models, the programmable nature of the myrtle.ai VOLLO stack ensures that firms can update their strategies without replacing their hardware.

Customer requirements

Designed for technical and business leaders within the Financial Services Industry (FSI) who prioritize execution speed:

- Quantitative Traders: Seeking to deploy more sophisticated ML models without increasing latency budgets.
- Electronic Trading Desks: Aiming to reduce slippage and improve fill rates in competitive markets.
- Risk Managers: Requiring real-time, intra-day risk assessments that can process massive data streams without lag.
- CTOs & Infrastructure Architects: Designing high-density co-location deployments where power and space are at a premium.
- Extreme Determinism: requires consistent latency even at high-percentile tail ends (99th and 99.9th), ensuring predictable performance during "flash" market events.
- Sub-Microsecond Latency: as low as 1.5 microseconds for LSTM_A, users should be able to identify signals and react before the broader market can process the same information.

The Supermicro Solution and System Under Test (SUT)

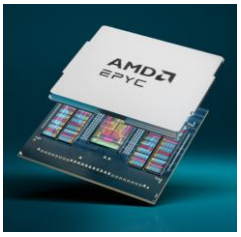
Supermicro CloudDC Server:

- 1U/2U platforms optimized for "Tick-to-Trade" performance.
- Form Factor: 2U Rackmount.
- CPU: 1x AMD EPYC 9575F (64-core).
- Memory: 12 x 16GB DDR5 DIMMs 6000MT/s (192GiB Total)
- Expansion: Dual PCIe 5.0 x16 slots utilized for the Silicom Artena accelerators.
- PCIe Gen 5 for maximum bandwidth to NICs and accelerators
- Cooling Advantage: optionally utilizing EVAC heat sink to maintain peak turbo frequencies without thermal throttling.

AMD Versal™ Premium Series Adaptive SoC

- High logic density for differentiation, adaptability, and faster time to insight
- Ultra-High Bandwidth networking with up to 5 Tb/s to handle market data surges
- Low-latency SERDES to support fast order execution
- Programmable logic supporting AI/ML for network intelligence, e.g., anomaly detection and self-provisioning
- PCIe Gen5 for lower system latency and higher throughput
- Engineered for thermal efficiency, industry-leading performance per watt

AMD EPYC™ 9005 Series Processors:



- High-frequency "F" SKUs provide raw, single-threaded clock speeds that are critical for low-latency stacks.

AMD/ Silicom PCIe accelerator card:

- FBAP4@VP18-2L0S from Silicom, containing an AMD Versal™ Premium series VP1802 Adaptive SoC

AMD Solarflare™ X4 Adapters:

- Achieves sub-microsecond latency with up to 40% lower latency than the previous generation
- Offload the network stack from the CPU to drastically reduce latency, features kernel-bypass Onload technology, maximizes CPU efficiency, reduces overhead

Software Stack – myrtle.ai VOLLO™: VOLLO is the intelligence that bridges the gap between AI researchers and FPGA hardware. Customers can use it to compile standard ML models (such as those from PyTorch or TensorFlow) for execution on highly optimized FPGA hardware without requiring FPGA programming expertise.

- No RTL Required: VOLLO allows quants to take models directly from PyTorch and deploy them to FPGAs using a standard library.
- Architectural Optimization: It uses a specialized data-flow architecture optimized for time-series inference, minimizing data movement and maximizing parallel processing.
- Flexibility: It supports multiple model instances and varied configurations, allowing a single FPGA to serve multiple trading strategies simultaneously.

Hardware Acceleration –The hardware foundation of this solution is housed on the Silicom Artena (FBAP4@VP18-2L0S) platform, which hosts the AMD Versal™ Premium Adaptable SoC.

- AMD Versal™ Premium VP1802: This Adaptable SoC combines traditional FPGA logic with AI Engines (AIE) and hardened memory controllers. It provides the massive memory bandwidth necessary for low-latency LSTMs.
- Silicom Engineering: The Artena card provides high-speed PCIe Gen5 connectivity and optimized thermal dissipation, ensuring the AMD SoC operates at peak performance within the Supermicro chassis.

The Compute Engine – AMD EPYC™ Processors: While the FPGA handles the inference, the AMD EPYC™ 9005 Series (Turin) processor acts as the high-speed orchestrator.

- PCIe Gen5 Leadership: The EPYC processor provides the industry-leading number of PCIe Gen5 lanes, ensuring zero bottlenecks between the network, the CPU, and the Silicom FPGA card.

- Memory Throughput: With 12-channel DDR5 memory, the processor ensures that the "pre-processing" and "post-processing" of market data keep pace with the FPGA's sub-microsecond speeds.

The audited server is a Supermicro AS -2015CS-TNR, a 2U, single-processor H13 CloudDC system designed for maximum flexibility. This server is optimized for high-performance edge and data center deployments. Its thermal design allows the Silicom FPGA cards to run high-frequency bitstreams without throttling, which is critical for maintaining the deterministic latency highlighted in the STAC-ML report.

Record-Breaking Latency

This system achieved the lowest 99th-percentile (99p) latency ever reported across all three benchmark models (LSTM_A, LSTM_B, and LSTM_C). Sub-Microsecond Frontiers: New World Records – 99th Percentile Latency.

- LSTM_A: First system to break the 2 µs barrier
- LSTM_B: First system to break the 3 µs barrier
- LSTM_C: First system to break the 8 µs barrier

Deterministic Performance - Unlike general-purpose compute, this FPGA-based solution maintains consistent, low-jitter performance even as the number of model instances increases.

Transition - The transition from rule-based systems to ML-driven strategies is the current battleground of algorithmic trading. While the market continues to leverage AI Solutions with advanced CPU and CPU designs, it has also been realized that specialized solutions are best suited to address tail latencies caused by inherent workload jitter.

The integration of myrtle.ai's VOLLO with Supermicro Servers, using the AMD Versal™ Premium Adaptable SoC on Silicom's Artena accelerator card, breaks the microsecond barrier. The Supermicro + myrtle.ai solution delivers three decisive advantages:

- Competitive Execution Edge: 99th-percentile latencies as low as 2 microseconds enable firms to process signals through complex machine learning models and execute trades faster than competitors relying on CPU or GPU inference—reducing slippage and capturing more alpha in volatile markets.
- Increased Model Sophistication: Traders can now deploy accurate LSTM, SSM, CNN, MLP, etc. models at speeds previously reserved for simple linear regressions—no more forced compromises.
- Operational Efficiency: Record throughput in a compact 2U footprint reduces rack space and power consumption in expensive co-location facilities.

For More Information

- Supermicro Financial Services: https://www.supermicro.com/en/solutions/ai/finance
- AMD EPYC Processors: https://www.amd.com/en/products/processors/server/epyc.html
- myrtle.ai: https://myrtle.ai/products/vollo-for-capital-markets/

Appendix

Reference: SUT ID: MRTL260323 STAC-ML™ Markets (Inference) Benchmarks Tacana Suite

While this solution uses FPGAs, Supermicro also delivers the industry's top AI Inference performance benchmarks for both GPU- and FPGA-based solutions.

	GPU Accelerated	FPGA Accelerated
Use Case	Mid-frequency, Alpha Gen, Sentiment, Risk.	HFT/MFT, Tick-to-Trade, Order Execution.
Winning Metric	Throughput & Versatility	Deterministic Ultra-Low Latency
Model Complexity	Large models LLMs, Deep LSTMs, Transformers).	Smaller, highly optimized models (e.g., SSMs, RNNs, LSTMs, MLPs, CNNs).
Development Dependencies	Python, PyTorch, CUDA	VOLLO SDK compiles models from PyTorch and ONNX
Jitter/Variance	Minimal, software-dependent.	Virtually zero (Hardware-level determinism)
Deterministic Model Performance Characteristics	High Throughput, Low Latency Inference	Ultra Low Latency Inference

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

For more than 50 years AMD has driven innovation in high-performance computing, graphics and visualization technologies. Billions of people, leading Fortune 500 businesses and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible.

Learn more at www.amd.com

MYRTLE.AI

Myrtle.ai is an AI/ML software company that delivers world-class inference accelerators on FPGA-based platforms from all the leading FPGA suppliers. With broad neural network expertise, myrtle.ai has delivered accelerators for applications including fintech, wireless telecoms, LLMs, speech processing, and recommendation. Visit myrtle.ai