SOLUTION BRIEF

SUPERMICRO AND ALGO-LOGIC OFFER AN AI-DRIVEN, HARDWARE-ACCELERATED, ULTRA-LOW-LATENCY TRADING SYSTEM

Supermicro and Algo-Logic Deliver Ultra-Low Latency execution of sophisticated trading strategies of Futures and Options. The system leverages an AI cluster, an analytics server with precise timestamping, and hardware-accelerated trade execution

Table of Contents

Background

Global commodity markets influence the prices of energy, crops, exchange rates, interest rates, and precious metals. The Chicago Mercantile Exchange (CME) is the largest platform for trading these commodities. Companies involved in these markets must constantly monitor multiple factors to determine the best prices for futures and options contracts.

Over the past 106 years, the time required to execute a trade has significantly decreased. Originally, human traders in a pit used open outcry to match orders. In the 1980s, trading shifted to electronic platforms with the rise of desktop computers. Starting in the 2000s, High-Frequency Trading (HFT) emerged, using the fastest CPUs in colocation to execute trades, reducing trade times from several seconds to less than a microsecond.

Furthermore, unexpected yet dramatic events influence global trading strategies. So-called 'Black Swan' events that once occurred only occasionally now seem to happen more often. Of the five major black swan events—the global financial crisis (2008), the flash crash (2010), the COVID-19 pandemic (2020), the Russian invasion of Ukraine (2022), and U.S. Treasury market volatility (2023-2025)—three have occurred in recent years.

Latency Drives Profit

Trading firms that respond quickly to reprice futures and options tend to profit, while those that delay may face severe financial losses. Therefore, it is vital to identify new market dynamics and trade accordingly to prevent losses and generate profits. Nowadays, trading firms can leverage AI to analyze a wide array of external factors and develop emerging trading strategies. Analytics on precisely measured market data events can also be used to test new strategies. Ultra Low Latency (ULL) trade execution platforms enable firms to capitalize on these opportunities and trade at the speeds necessary to compete with leading HFT firms.

Algo-Logic is a recognized leader offering innovative, flexible HFT solutions. Algo-Logic developed circuits that offload processing of market data, generate orders, perform Pre-Trade Risk Checks (PTRCs) that reduce losses from bad trades, and implement Tick-to-Trade (T2T) systems that generate profit by instantly sending orders to an exchange in response to market data. Algo-Logic achieves Ultra-Low Latency by implementing algorithms in logic instead of software, utilizing Gateware Defined Networking (GDN) libraries that feature pre-built and proven logic cores, which offload time-critical trading functions to hardware.

Algo-Logic provides multinational investment banks, proprietary trading firms, Independent Software Vendors (ISVs), and exchanges with hardware-accelerated trading systems. In collaboration with Supermicro, Algo-Logic provides comprehensive FPGA-accelerated solutions to deliver the benefits of low latency to trading customers worldwide.

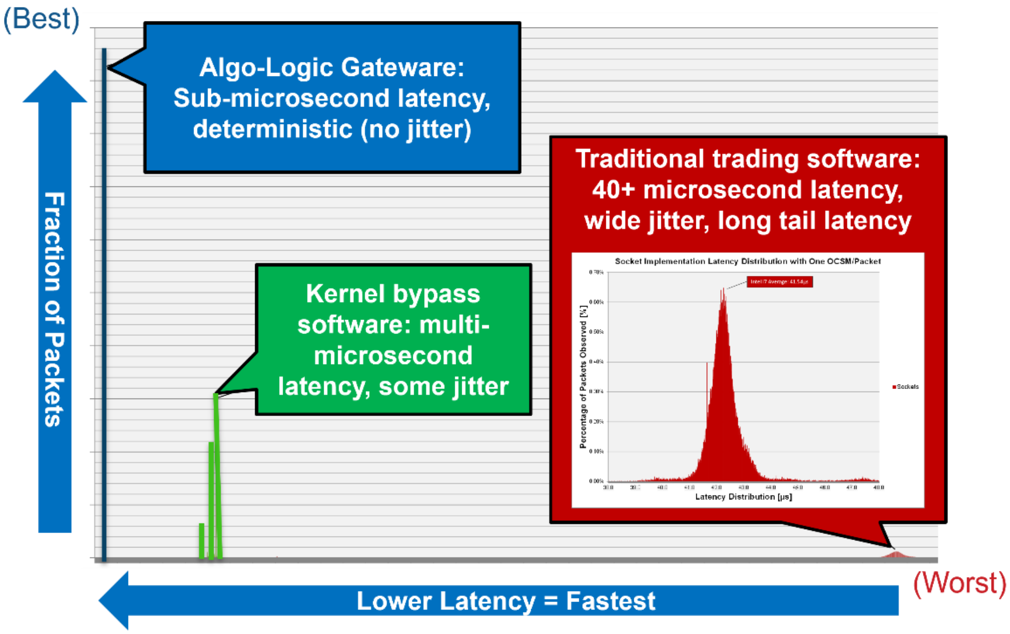


Figure 1 - Trade latency compares FPGA, Kernel Bypass, and Traditional Trading Software Comparison

Figure 1 shows the advantage of trade execution with Algo-Logic Gateware over traditional trading software implementations. Traditional software trading systems are not deterministic and exhibit a long latency tail, introducing delays of many tens of microseconds for trade execution. Even with kernel bypass software, the latency of a trade is typically several microseconds.

In contrast, FPGA trading systems offer the lowest latency, measured in just a fraction of a single microsecond — almost an order of magnitude less than even the kernel bypass. With FPGA implementations, deterministic performance enhances the trading system experience by providing low T2T latency with no tail. This incredibly fast trade execution gives trading firms a competitive edge.

Key Solution Ultra Low Latency (ULL) Trading System

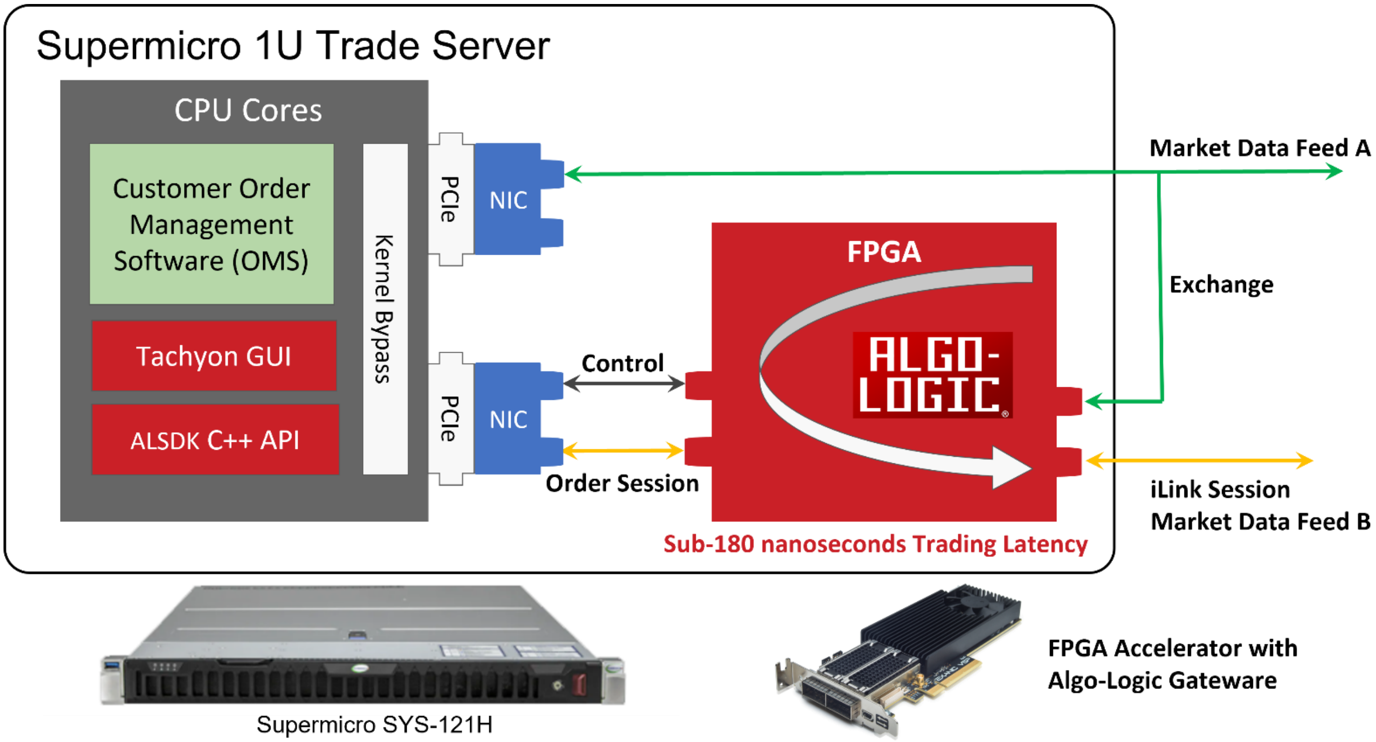


Figure 2 - Ultra Low Trade System latency workflow

Figure 2 shows that Algo-Logic has developed a trading system that utilizes an FPGA accelerator with Gateware to provide ultra-low latency trade execution. Trading firms use the Algo-Logic solution to execute their trades quickly and deterministically. Algo-Logic’s T2T system enables the implementation of trade strategies in hardware through its Software Development Kit (ALSDK) interface. This C++ library can be called from host software on a CPU core or via RESTful interfaces. ALSDK provides application programming interfaces (APIs) for pre-loading triggers and monitoring the T2T system.

Key Solution: AI-driven Hardware-Accelerated Trading System

In the past, trade strategies were painstakingly developed by quantitative analysts (quants) using cumbersome software tools. They back-tested trade strategies using historical market data by running scripts and software tools. Today, AI clusters with GPUs utilize modern analytical tools to process large volumes of market data, identifying potentially profitable trade strategies.

One key aspect of the Algo-Logic and Supermicro complete solution is continuously comparing the actual financial performance of the system to the market’s expected performance derived from the predictive model. Although back-testing an AI-generated strategy can predict potential financial gains (alpha), real profits are only realized when live trades are executed in the market. An analytics server enables the ongoing refinement of a strategy as trades are executed in real-time. Through this process, firms can create new trading strategies, run back-testing scenarios, and verify potential enhancements in financial performance. With Supermicro and Algo-Logic, firms have the ability to update trading strategies throughout the day, rather than only on a daily or weekly basis.

The key to the success of any trading strategy is the ability to act on market data events (triggers) by sending orders to the exchange with Ultra-Low Latency. The Algo-Logic Tick-to-Trade system, with deep sub-microsecond latency, provides the speed advantage needed for small firms to compete effectively with top HFT firms. Algo-Logic supports multiple advanced triggers, enabling trading firms to implement increasingly complex strategies while maintaining very low Tick-to-Trade latencies.

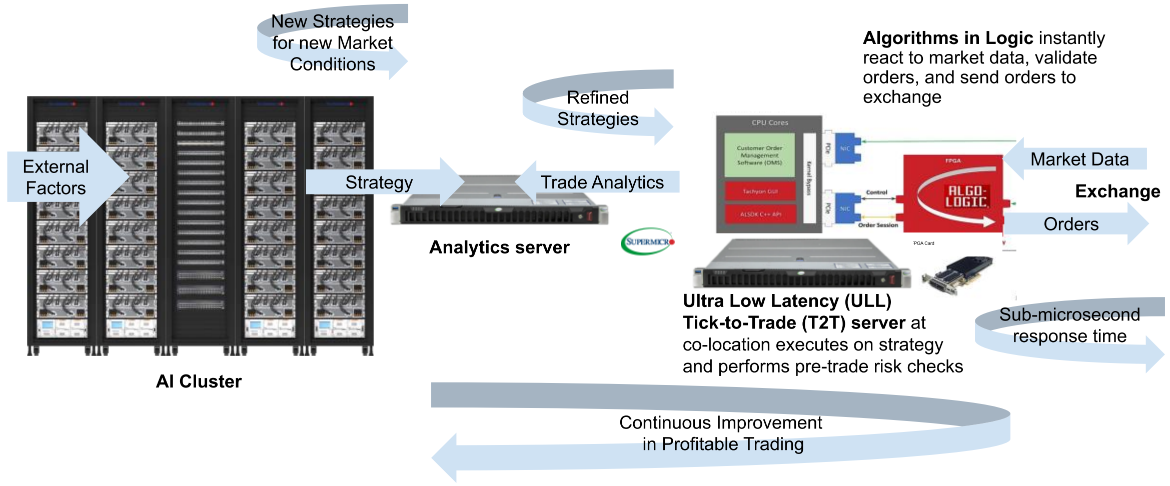

New Strategies for new Market Conditions
External Factors
Strategy
AI Cluster
Refined Strategies
Trade Analytics
Analytics server
SUPERMICRO
Algorithms in Logic instantly react to market data, validate orders, and send orders to exchange
CPU Cores
Customer Order Management Software (OMS)
Tachyon GUI
ALSDK C++ API
Kernel Bypass
PCIe
NIC
Control
Order Session
FPGA
ALGO-LOGIC
FPGA Card
Market Data
Exchange
Orders
Ultra Low Latency (ULL) Tick-to-Trade (T2T) server at co-location executes on strategy and performs pre-trade risk checks
Sub-microsecond response time
Continuous Improvement in Profitable Trading

Figure 3: AI-driven Hardware-accelerated Trading System workflow

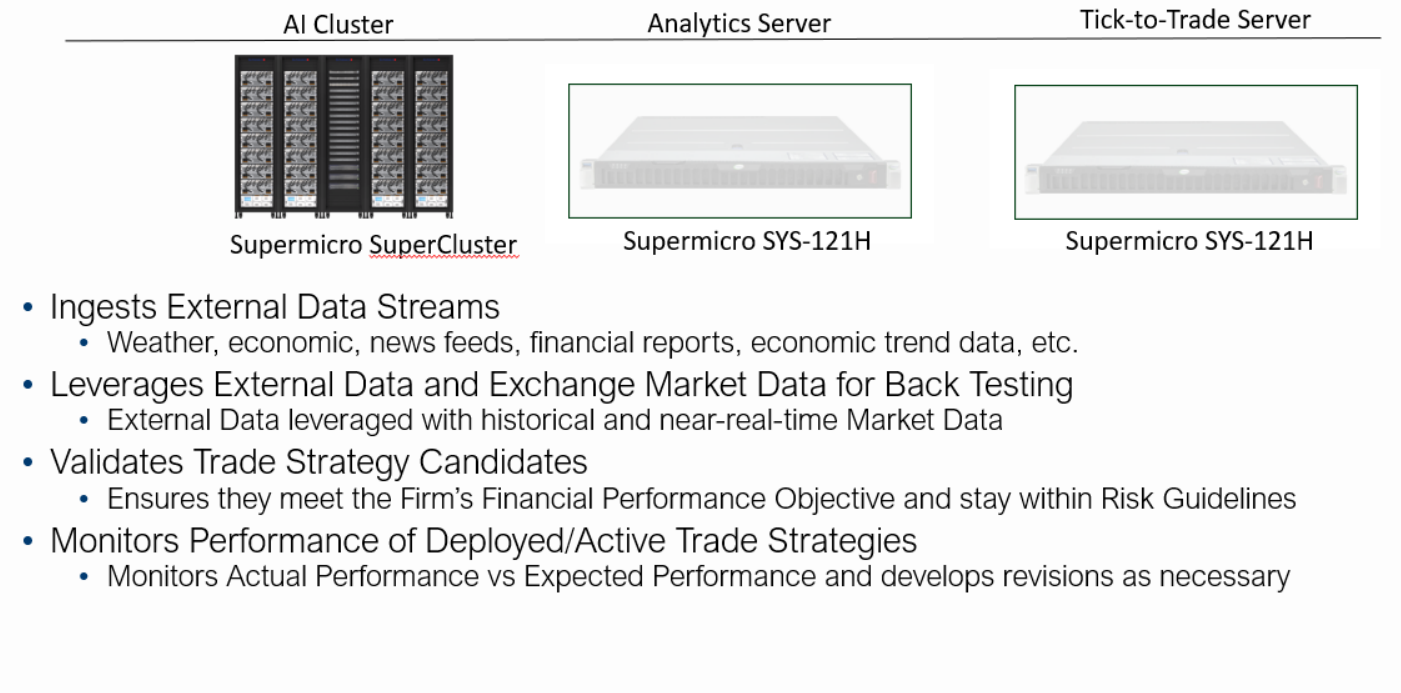
Figure 3 illustrates the workflow for the Supermicro and Algo-Logic AI-driven Hardware-accelerated Trading System.

Key Elements of an AI-driven Hardware-accelerated Trading System

There are three significant elements: the Supermicro Al Cluster, the Supermicro SYS-121H Analytic Server, which is located in the exchange colocation, and the Supermicro SYS-121H Tick-to-Trade Server, which is situated in the ultra-low latency rack within the exchange colocation.

	AI Cluster Supermicro SuperCluster	Analytics Server Supermicro SYS-121H	Tick-to-Trade Server Supermicro SYS-121H
Location	On-Prem/Cloud/Data Center	Exchange Colo	Exchange Colo (ULL Rack)
Inputs	Many External Data Feeds Historical Market Data Near-Real-Time Market Data Order Execution Analytics	Real-Time Market Data Candidate Trade Strategies	Real-Time Market Data Trade Strategy Parameters
Focus	Validate/Backtest Trade Strategies Monitor Actual vs Expected Performance Monitor Risk Levels and Performance	Monitor Actual vs Expected Performance Monitor Exchange Latency Profile Refine Trade Strategy Candidates Define Trade Strategy and Parameters	ULL Trade Strategy Execution Pre-Trade Risk Check
Hardware	GPU Cluster Data Store	GPU for AI at the Edge FPGA for Real Time Data Analytics	FPGA for Market Data Analysis, Order Execution and Pre-Trade Risk Analysis
Delivers	Kubernetes Packaged Trade Strategies	Trade Strategy and FPGA Parameters	Order Execution

AI Clusters



AI Cluster

Analytics Server

Tick-to-Trade Server

Supermicro SuperCluster

Supermicro SYS-121H

Supermicro SYS-121H

- Ingests External Data Streams
  - Weather, economic, news feeds, financial reports, economic trend data, etc.
- Leverages External Data and Exchange Market Data for Back Testing
  - External Data leveraged with historical and near-real-time Market Data
- Validates Trade Strategy Candidates
  - Ensures they meet the Firm’s Financial Performance Objective and stay within Risk Guidelines
- Monitors Performance of Deployed/Active Trade Strategies
  - Monitors Actual Performance vs Expected Performance and develops revisions as necessary

The AI Cluster ingests recent or historical market data along with data streams from external sources. These streams can include weather data, business news feeds, financial reports, and economic trends. The AI Cluster builds a predictive model that anticipates market performance in response to these external factors. The model forecasts how the strategy should perform to achieve the firm’s financial objectives while maintaining an acceptable risk profile.

Additionally, the cluster monitors the actual performance of the deployed trade strategy on the Tick-to-Trade servers to determine whether it aligns with the forecasted or anticipated performance. By monitoring the strategy, the system identifies shortfalls and evaluates potential enhancements that will improve financial performance.

The AI Cluster packages the candidate trade strategies provided to the analytics server. It is built using Rack Scale Compute, Networking, Cabling, and cooling Infrastructure from Supermicro.

Supermicro Scalable AI SuperCluster

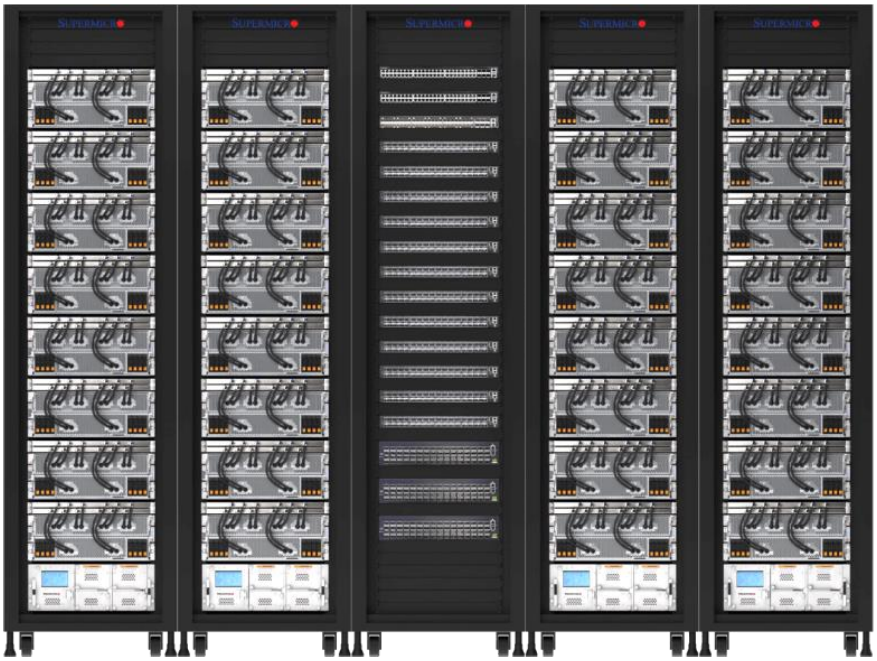


Figure 4 - Supermicro AI SuperCluster

The Supermicro SuperCluster (Figure 4), powered by the NVIDIA Blackwell GPUs, advances the latest AI platforms and introduces new breakthroughs, including evolving scaling laws and the emergence of reasoning models. It provides essential infrastructure components to scale the NVIDIA Blackwell Platform and achieve optimal AI training and inference performance. It simplifies AI infrastructure design and deployment by offering a fully validated, liquid-cooled AI cluster with plug-and-play deployment capabilities, using the Supermicro Data Center Build Block (DCBBS) solution.

The Supermicro SuperCluster offerings are available in 42U, 48U, or 52U configurations, and include upgraded cold plates and a 250kW coolant distribution unit (CDU), doubling the cooling capacity over previous generations. The SuperCluster is built using the 4U dual-processor liquid-cooled servers (SYS-422GA-NBRT-LCC – 4U/ SYS-421GE-NBRT-LCC) with NVIDIA HGX™ B200 8-GPU, dual Intel Xeon 6 CPUs, and 8 NVIDIA Blackwell B200 GPU processors.

Supermicro NVIDIA HGX platform powers some of the world's largest liquid-cooled AI data centers. The new 4U NVIDIA HGX B200 8-GPU system features new cold plates and a tubing design that further enhances efficiency and serviceability. The new system features 8 NVIDIA Blackwell GPUs, each with 180GB HBM3e memory. The GPUs are interconnected at 1.8TB/s through the latest NVIDIA NVLink™ with 1.4TB of GPU memory capacity per system. The SuperCluster creates a massive pool of GPU resources that act as one AI supercomputer, featuring 1:1 networking to the GPU with 8x 400GB/s NVIDIA ConnectX®-7 adapters or NVIDIA BlueField®-3 SuperNICs and 2 NVIDIA BlueField®-3 DPUs per system.

Please refer to the product details in Appendix A.

The SuperCluster features NVIDIA Quantum InfiniBand or NVIDIA Spectrum networking in a centralized rack. The architecture enables a non-blocking, 256-GPU scalable unit in only five racks or an extended 768-GPU scalable unit in only nine racks.

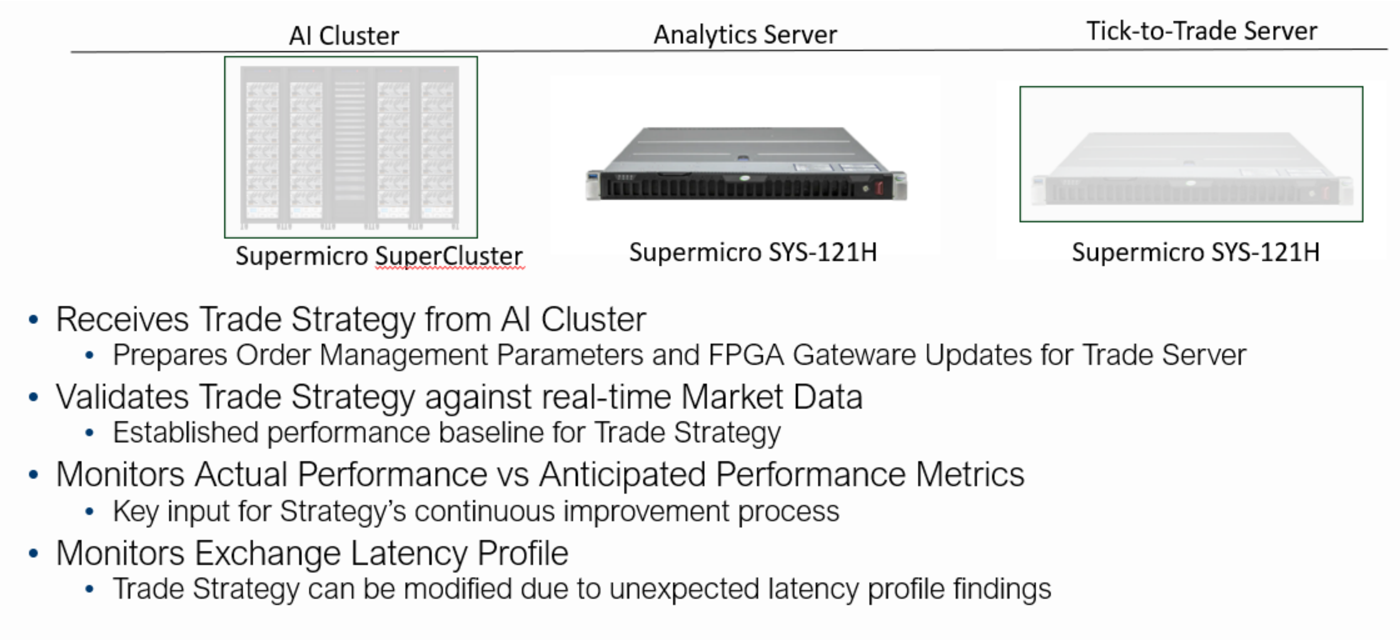
Management Software:

Supermicro’s SuperCloud Composer provides management tools to monitor and optimize air- or liquid-cooled infrastructure, overseeing all data center racks—including compute, storage, and networking—in a single unified dashboard. SuperCluster natively supports NVIDIA AI Enterprise software, accelerating the transition to production AI. NVIDIA NIM™ microservices enable organizations to easily access and deploy the latest AI models and agents fully optimized for the new NVIDIA Blackwell Platforms.

Services:

Supermicro’s on-site rack deployment helps enterprises build a data center from the ground up, including the planning, designing, powering up, validating, testing, installing, and configuring racks, servers, switches, and other networking equipment to meet the organization’s specific needs.

Supermicro SYS-121H Analytics Server



AI Cluster | Analytics Server | Tick-to-Trade Server

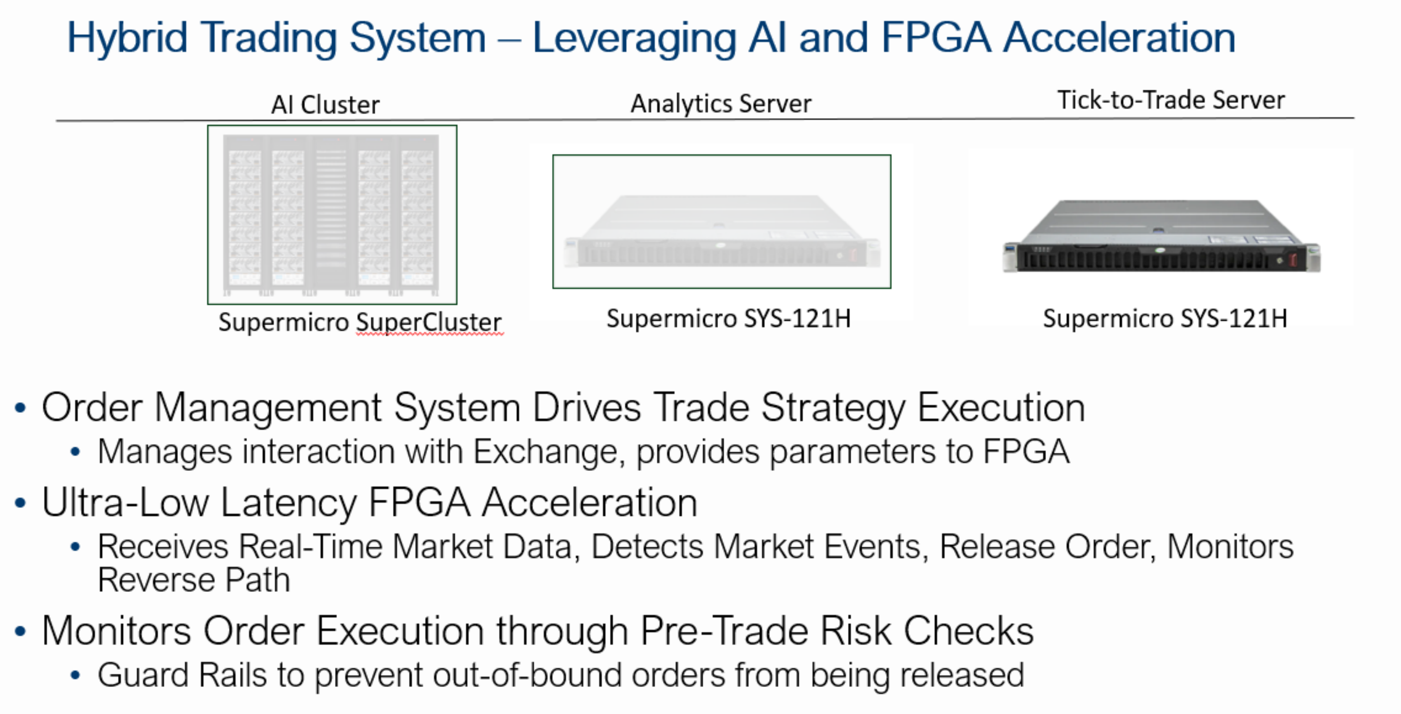
Supermicro SuperCluster | Supermicro SYS-121H | Supermicro SYS-121H

- Receives Trade Strategy from AI Cluster
  - Prepares Order Management Parameters and FPGA Gateware Updates for Trade Server
- Validates Trade Strategy against real-time Market Data
  - Established performance baseline for Trade Strategy
- Monitors Actual Performance vs Anticipated Performance Metrics
  - Key input for Strategy’s continuous improvement process
- Monitors Exchange Latency Profile
  - Trade Strategy can be modified due to unexpected latency profile findings

The Analytics Server is located in the exchange, where rack space and power are limited. The analytics service can be set up in a standalone chassis or combined within the same chassis as the Tick-to-Trade server. The analytics server processes real-time market data, refines the trade strategy, and loads triggers onto the Tick-to-Trade server with the correct parameters. Additionally, it monitors the exchange gateway latency profiles to determine if routing the order to a faster gateway offers an advantage or if trading should be avoided due to delayed (stale') market data. Finally, the analytics server captures packets with precise timestamps to track the actual performance of the trade strategy compared to its expected performance.

Please refer to product details in Appendix B.

Supermicro SYS-121H Tick-to-Trade Server



Hybrid Trading System – Leveraging AI and FPGA Acceleration

AI Cluster — Supermicro SuperCluster

Analytics Server — Supermicro SYS-121H

Tick-to-Trade Server — Supermicro SYS-121H

- Order Management System Drives Trade Strategy Execution
  - Manages interaction with Exchange, provides parameters to FPGA
- Ultra-Low Latency FPGA Acceleration
  - Receives Real-Time Market Data, Detects Market Events, Release Order, Monitors Reverse Path
- Monitors Order Execution through Pre-Trade Risk Checks
  - Guard Rails to prevent out-of-bound orders from being released

The Tick-to-Trade server houses the Order Management System (OMS) and Tick-to-Trade FPGA card, which is the heart of the Ultra-Low Latency trade execution server. The fast algorithms for trade are implemented in logic on an FPGA card using Algo-Logic's GDN libraries to provide deep sub-microsecond latency orders in response to market data with virtually zero jitter.

Refer to the product details in Appendix C.

Algo-Logic's default system provides six standard Tick-to-Trade Triggers that release orders when specific market data events occur. The Algo-Logic system integrates existing Order Management System (OMS) software through C++ Application Programming Interfaces for dedicated Tick-to-Trade algorithms or via RESTful interfaces for trading.

Benefits of AI-Driven, Hardware-Accelerated, Ultra-Low-Latency Trading System

The Tick-to-Trade "Race to Zero" continues, and reducing the latency by nanoseconds has become increasingly more challenging. Algo-Logic continues to focus on this metric, and the company is working closely with its customers to develop new, more sophisticated triggers that allow orders to be fully executed in less than a microsecond, far surpassing the many microseconds required for trading by software. This speed advantage yields greater profits for the trading firms.

Competitive Differentiation

The system's competitive advantage is speed. Not only is the Ultra Low Latency (ULL) and fast Order Execution enabled by FPGA, but also the AI cluster and analytics server provide the ability to define, validate, and back-test new trade strategies and quickly move them into production. Rapidly responding to fast-changing markets is the key to winning.

Revenue Opportunity

Existing customers using Supermicro servers with the Algo-Logic T2T trading solution report that they can quickly cancel orders (Cancel on Behalf), which gives them the confidence to take a more aggressive approach and capture more profits. This strong defensive capability to minimize losses from getting “picked off”, combined with the fast order execution and sophisticated triggers for offensive trading strategies, is a ‘one-two’ punch that allows firms to minimize losses in turbulent times and capture profit as soon as new opportunities emerge.

Conclusion

By combining Algo-Logic’s Ultra Low Latency trading logic with Supermicro’s market-leading server technology, high-performance, pre-configured trading systems can be rapidly deployed. Trading firms can leverage AI to develop trading strategies that ultimately run in logic. This innovative system not only delivers Ultra-Low Latency Tick-to-Trade capabilities but also leverages AI capabilities to develop, back-test, and continuously refine trading strategies.

Appendices

Appendix A

AI Server specifications: SYS-422GA-NBRT-LCC – 4U/ SYS-421GE-NBRT-LCC , Liquid-Cooled System with NVIDIA HGX™ B200 8-GPU

Fully integrated liquid-cooled AI Cluster 8 Node Rack, Scalable to 32 or 96 Node Scalable Unit	
Compute	SYS-422GA-NBRT-LCC – 4U/ SYS-421GE-NBRT-LCC Liquid-Cooled System with NVIDIA HGX B200 8-GPU
CPU	Dual Intel® Xeon® 6900 series processors with P-cores (SYS-422GA-NBRT-LCC) Dual 5th/4th Gen Intel® Xeon® Scalable processors (SYS-421GE-NBRT-LCC)
Memory	24 DIMMs, up to DDR5-6400 (SYS-422GA-NBRT-LCC) 32 DIMMs, up to DDR5-5600 (SYS-421GE-NBRT-LCC)
GPU	NVIDIA HGX B200 8-GPU (180GB HBM3e per GPU) 1.8TB/s NVIDIA NVLink™bandwidth with NVSwitch
Networking	8 single-port NVIDIA ConnectX®-7 NICs or NVIDIA BlueField®-3 SuperNICs, up to 400Gbps 2 dual-port NVIDIA BlueField®-3 DPUs
Storage	8 front hot-swap 2.5" NVMe drive bays 2 M.2 NVMe slots
Power Supply	4x 6600W Redundant Titanium Level power supplies
Networking	
Compute and Storage Networking	NVIDIA Quantum-2 400G InfiniBand switches or NVIDIA Spectrum-4 400GbE Ethernet switches
In-band Management Switch	NVIDIA Spectrum SN4600 100GbE Ethernet
Out-of-band Management Switch	48-port 1Gbps Ethernet ToR management switch
Liquid Cooling	Supermicro 250kW capacity Cooling Distribution Unit (CDU) with redundant PSU and dual hot-swap pumps Vertical Cooling Distribution Manifold (CDM)
Rack	48U 750mmx1200mm
PDU	208V 60A 3Ph
Management Software	Supercloud Composer
Note : Supermicro can customize an AI cluster for Liquid Cooled or Air Cooled environments, and offer an option for other industry-standard CPUs, GPUs, Networking, and Racks	

Appendix B

Analytics Server specifications: Supermicro SYS-121H

Analytics Server	
Server	Hyper SuperServer SYS-121H-TNR
CPU	5th Gen Intel® Xeon® / 4th Gen Intel® Xeon® Scalable processors
Memory	32 DIMMs, up to DDR5-5600
FPGA	Cisco Nexus X40
GPU	Up to 1 double-width or 3 single-width GPUs
Networking	NVIDIA ConnectX-6 Dx EN 100GbE Adapter Card PCIe 2-port, 200Gb/s InfiniBand ; Ethernet Adapter Card, QSFP56
Storage	8 front hot-swap 2.5" NVMe*/SAS*/SATA drive bays
Power Supply	2x 1200W Redundant Titanium Level (96%) Hot-plug power supplies
Management	SuperCloud Composer® Supermicro Server Manager (SSM) Supermicro Update Manager (SUM) SuperServer Automation Assistant (SAA)
Note: Supermicro can customize the Server based on application requirements and offer options for other industry-standard CPUs, GPUs, Networking, and Racks	

Appendix C:

Tick-to-Trade Server specifications: Supermicro SYS - 121H

Tick to Trade Server	
Server	Hyper SuperServer SYS-121H-TNR
CPU	5th Gen Intel® Xeon® / 4th Gen Intel® Xeon® Scalable processors
Memory	32 DIMMs, up to DDR5-5600
FPGA	Cisco Nexus V5P with Ultrascale+ VU5P with Algo-Logic T2T Gateware
Networking	Intel 100G network adapter for capture
Storage	8 front hot-swap 2.5" NVMe*/SAS*/SATA drive bays
Power Supply	2x 1200W Redundant Titanium Level (96%) Hot-plug power supplies
Management	SuperCloud Composer® Supermicro Server Manager (SSM) Supermicro Update Manager (SUM) SuperServer Automation Assistant (SAA)
Note: Supermicro can customize the Server based on application requirements and offer options for other industry-standard CPUs, GPUs, Networking, and Racks	

Learn More

For More Information:

www.supermicro.com/financial_services

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.

Learn more at www.supermicro.com

ALGO-LOGIC

Algo-Logic Systems is a recognized leader delivering innovative, flexible solutions, from Pre-Trade Risk Checks to fully automated Tick-to-Trade systems delivering ultra-low latency solutions implemented in Field Programmable Array (FPGA) logic. Algo-Logic's rapid time-to-market is due to our Gateware Defined Networking (GDN) libraries featuring pre-built and proven IP cores. We deliver value to our customers by providing Ultra-Low Latency solutions enabled by FPGA Acceleration for High Frequency Trading and Real-Time Data management.

Learn more at www.algo-logic.com