PRODUCT BRIEF

OPTIMIZING ENTERPRISE AI: ACCELERATING TIME TO VALUE WITH SUPERMICRO GPU PLATFORMS WITH AMD CPUS AND GPUS

Supermicro GPU Liquid-Cooled Server: AS-4126GS-NMR-LCC

TABLE OF CONTENTS

Executive Summary 1
Enterprise AI Imperative: Time to AI Value 3
Architectural Deep Dive 4
AI Platform: Supermicro's Optimized Building Blocks 4
Accelerating Time-to-Value with Data Center Building Block Solutions (DCBBS) 5
Driving Operational Efficiency with Advanced Direct Liquid Cooling (DLC-2) 6
The Synergy Advantage 8
Economic Analysis 9
Analysis and Outlook for IT Leaders 10
Outlook 11
Appendix 11

Executive Summary

The rapid integration of Artificial Intelligence into enterprise workflows has shifted the focus of IT infrastructure procurement from acquiring powerful components to deploying complete, optimized solutions that deliver immediate value. While the market for AI accelerators has been dominated by a single supplier, a powerful alternative has emerged that warrants strategic consideration. This report provides an in-depth analysis of a total AI solution comprising AMD Instinct™ MI355X GPUs deployed on Supermicro's H14 application-optimized server platforms. This analysis concludes that this combination delivers leading performance for critical enterprise workloads and an accelerated time-to-value, directly addressing the most pressing challenges faced by enterprise IT leaders today.

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

Note: Some portions of this paper were taken from the original Signal65 paper with permission (www.signal65.com)

The key findings of this report are threefold:

1. Supermicro's Building Blocks speed server development with the latest technologies for AI enterprise deployments
2. AMD Instinct GPUs, EPYC™ CPUs, and Pollara™ NICs within Supermicro servers deliver competitive performance
3. AMD hardware, AMD's ROCm™ 7 library with AI Dev Center, enables rapid deployment of AI agents

First, the Supermicro server platform acts as a critical enabler of performance, adding value beyond standard chassis to help optimize workload performance. Supermicro's "Building Block Solutions" philosophy enables rapid integration of new technologies, ensuring the latest accelerators reach market in optimized, validated systems with minimal delay.[2] This approach accelerates deployment and simplifies the integration process, a crucial advantage in the fast-paced AI landscape.

Next, the combined AMD and Supermicro solution presents a multi-faceted and compelling economic advantage. This begins with AMD's historically competitive hardware pricing and its no-cost ROCm™ software library, compared with per-device license models from other companies. This initial cost advantage is then significantly amplified by Supermicro's platform-level efficiencies. Through its advanced Direct Liquid Cooling (DLC-2) technology, Supermicro platforms can reduce data center power consumption by up to 40% and overall TCO by up to 20% compared to traditional air-cooled deployments.[4]

Finally, in terms of time to deployment, AMD ROCm 7 delivers example microservices and blueprints that enterprises can quickly leverage for their specific applications. AMD Instinct MI355X, tested on a Supermicro AS -4126GS-NMR-LCC server, delivers leading performance across key enterprise AI benchmarks. This includes delivering up to 2x higher throughput vs. competitive options on a large model inference benchmark with Llama3.1-405B and demonstrating a 10% time-to-solution advantage in the industry-standard MLPerf LoRA fine-tuning benchmark.[1]

For IT buyers, the strategic recommendation is clear: for enterprises prioritizing deployment velocity, operational efficiency, and superior economics for mainstream AI workloads such as large-scale inferencing, fine-tuning, and agentic AI, the validated AMD-Supermicro solution represents a formidable and strategically sound alternative to the market alternatives.

Metric	AMD MI355X on Supermicro Platform Advantage
Llama2-70B Inference Throughput	Highest Total Tokens / sec. per cluster (648,248)
Platform Power Savings (DLC-2)	Up to 40% reduction vs. air cooling[4]
Data Center Footprint Reduction	Up to 60% smaller footprint with higher density[2]
Platform TCO Reduction (DLC-2)	Up to 20% lower TCO vs. air cooling[4]

Note that the performance is based on actual runs as described in Reference 1.

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

Enterprise AI Imperative: Time to AI Value

While the development of massive foundational models by hyperscale companies captures headlines, the practical application of AI in most enterprises centers on leveraging these models to deliver tangible productivity gains. The primary focus for most firms is not on training from scratch but on production inferencing, including sophisticated applications such as Retrieval-Augmented Generation (RAG), the deployment of complex workflows, and the efficient fine-tuning of pretrained models with proprietary data. This shift in focus from pure research to production deployment fundamentally changes the requirements for the underlying infrastructure.

Enterprises today face a host of practical challenges that can stall AI initiatives and delay return on investment. The sheer power density of modern AI accelerators creates unprecedented thermal and power delivery challenges that legacy data center designs cannot accommodate. Integrating these complex components into a cohesive, performant, and reliable system is a non-trivial engineering task that falls outside the core competency of most enterprise IT teams. This complexity creates a significant bottleneck, where the pressure for rapid time-to-market clashes with the reality of lengthy and risky system integration cycles. A component-level procurement strategy in which GPU-accelerated servers, networking, rack infrastructure, and cooling are sourced separately exacerbates this problem, placing the immense burden of integration and validation squarely on the customer.

This operational complexity has given rise to a new paradigm for AI infrastructure[3]. This model treats AI compute not as a collection of discrete servers, but as a utility; a scalable, repeatable, and rapidly deployable resource that can be provisioned on demand. The new AI infrastructure concept prioritizes turnkey, pre-validated, rack-scale solutions that reduce the underlying complexity. This is precisely the market need that Supermicro's Data Center Building Block Solutions® (DCBBS) is designed to address.[2] By delivering fully integrated and tested clusters, DCBBS shifts the value proposition from component innovation to solution integration and deployment velocity, allowing enterprises to focus on their AI applications, not on infrastructure engineering. This evolution requires IT buyers to reframe their core evaluation criteria, Hardware Performance, Software Ecosystem, and TCO through the lens of a total solution.

- Hardware performance requires sustained, end-to-end performance of the entire system, ensuring that the accelerator is not bottlenecked by I/O, networking, or thermal throttling.
- The Software ecosystem is another critical element of the solution that provides access to hardware, enabling optimized and validated systems to rapidly come to market, ensuring that advancements can be leveraged without delay.
- Total Cost of Ownership (TCO) must be analyzed holistically, accounting not just for the initial hardware acquisition cost (CAPEX), but for the crucial operational expenses (OPEX) of power, cooling, data center space, and management overhead, all of which are platform-level considerations that can dominate the long-term cost profile of an AI deployment.

The industry's pivot towards the new model is a direct and necessary response to the physical and operational realities of deploying next-generation AI hardware. Early AI adopters, such as hyperscalers along with NeoCloud AI accelerator providers, possess specialized engineering teams capable of designing and managing bespoke, complex systems. As AI has become a mainstream enterprise imperative, many organizations lack the deep in-house resources to support it. This fact, combined with the power consumption of modern GPUs, has created a challenge that organizational IT departments and enterprise data centers were not designed to support. Solution providers such as Supermicro are filling this gap by productizing the AI data

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

center, offering integrated, validated, liquid-cooled rack-scale solutions that make large-scale AI accessible to the broader market. The new paradigm can, therefore, become an essential part of solving the AI implementation complexity.

Architectural Deep Dive

The performance of any AI solution is a function of its core computational capabilities, together with software, networking, and a platform engineered to manage these resources efficiently.

AI Engine: AMD Instinct MI355X GPU

The AMD Instinct MI355X represents a significant leap in accelerator design, pairing a next-generation compute engine with an industry-leading memory subsystem, engineered to excel at the demanding enterprise AI tasks, including model fine-tuning, RAG, and the emerging class of multi-model agentic workloads. [1]. The most significant architectural advantage of the MI355X is its massive memory subsystem. Each accelerator is equipped with 288GB of HBM3e memory, delivering a peak bandwidth of 8 TB/s. This additional memory enables enterprises to:

- Run extremely large models, such as a 405-billion-parameter model using FP4 quantization, entirely on a single GPU. This eliminates the need for tensor-parallel communications across multiple GPUs, a major source of latency and performance overhead. [1]
- Accommodate massive context windows, which are critical for the effectiveness of RAG and complex reasoning tasks. For example, a single MI355X can hold a 70-billion-parameter model's weights in FP8 along with the key/ value (KV) caches for a 128,000-token context window without spilling over to slower system RAM.[1]
- Accelerate fine-tuning by holding the full model checkpoint, along with the necessary optimizer states for techniques like Low-Rank Adaptation (LoRA), within a single GPU's memory.[1]

To further boost performance, especially for inference, the MI355X features matrix cores that provide hardware acceleration for new quantized, lower precision data formats. While FP16 and FP8 remain the workhorses for many AI applications on the AMD Instinct platform, the latest MI355x accelerator pushes the boundaries of efficiency with the introduction of hardware support for FP4 and the new FP6 formats. This expanded support provides developers with greater flexibility to optimize their models for performance, memory usage, and power consumption.

The use of new, quantized data types (including FP8, FP4 and FP6) dramatically reduces the memory footprint and computational requirements of models. The AMD AI Tensor Engine seamlessly leverages this hardware capability for ROCm (AITER), an inference library that automatically optimizes operations for these formats, increasing processing speed and efficiency without requiring manual intervention from developers.[1]

AI Platform: Supermicro's Optimized Building Blocks

The benchmark results detailed later in this report were achieved using the AMD MI355X on a Supermicro AS -4126GS-NMR-LCC server. This platform is a direct result of Supermicro's core design philosophy: "Server Building Block Solutions®." This is a modular approach that leverages a vast portfolio of flexible, reusable components—chassis, power supplies, motherboards, and I/O modules—to build systems precisely optimized for a specific application.[2]

This philosophy provides several distinct advantages for enterprise AI deployments:

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

- Accelerated Time-to-Market: The modular design allows Supermicro to rapidly integrate and validate the latest technologies, such as new GPUs or processors, far more quickly than vendors with more monolithic design cycles. This gives customers faster access to performance-leading hardware.[2]

- Workload Optimization and Customization: Customers are not forced into a one-size-fits-all solution. The Building Block approach allows fine-tuning of system resources—balancing CPU cores, memory capacity, storage type, and network I/O — to perfectly match the demands of a given AI workload. This prevents costly overprovisioning and maximizes efficiency.[2]

- A Broad and Diverse Portfolio: Supermicro leverages this approach to offer one of the industry's widest selections of server form factors, from dense multi-node systems to large GPU servers. This ensures that an optimized platform exists for virtually any AI deployment scenario, from the edge to the data center core. [3]

The strategic importance of this modular approach cannot be overstated in a market characterized by rapid innovation cycles and persistent supply chain volatility. Custom designs can be fragile, and component issues can impact production.

Supermicro's Building Block architecture reduces the risk of this process by maintaining an inventory of standardized, interchangeable subsystems. When new key technologies like the MI355X become available, the engineering effort is focused on designing and validating their integration into an array of existing, proven chassis and power/cooling envelopes. This dramatically shortens the design-to-delivery cycle, providing a crucial competitive advantage. The advantage is not just in building high-quality servers, but in designing, building, and delivering them to customers faster than the competition.
Supermicro AS -4126-NMR-LCC Specifications:

- Dual AMD EPYC™ 9575F Processors (64 cores/socket, Base Clock 3.3 GHz, Boost All Core 4.5 GHz)
- 3.0 TB Total Memory (24x 128 GB, DDR5-6400 MT/s)
- 8x MI355X AMD Instinct™ GPUs

Accelerating Time-to-Value with Data Center Building Block Solutions (DCBBS)

Getting access to this performance leadership quickly and reliably is a major challenge for enterprises. This is where the Supermicro platform provides a decisive advantage. The "Building Block" philosophy is extended from the server level to the entire data center through the Data Center Building Block Solutions (DCBBS) initiative. DCBBS provides pre-validated, plug-and-play solutions at the rack and cluster levels, designed to serve as the organizational backbone of a modern AI-designed data center.

This approach directly addresses the most significant deployment pain points for enterprise IT. Instead of grappling with complex network topologies, high-amperage power delivery schemes, and advanced thermal management, customers receive a turnkey solution. DCBBS can reduce the time required to move from design to a fully operational AI cluster, providing a critical time-to-market and time-to-online advantage.[3] Offerings like the 256-node defined scalable unit provide a pre-engineered, turnkey solution for large-scale deployments, enabling predictable scaling and rapid expansion.[5]

Driving Operational Efficiency with Advanced Direct Liquid Cooling (DLC-2)

The extreme power density of modern AI accelerators, with TDPs exceeding 1 kW, has made traditional air cooling an inefficient and unsustainable solution for at-scale deployments. Direct-to-chip liquid cooling is no longer a niche technology for academic

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

supercomputers; it is a mainstream necessity for enterprise AI. Supermicro's next-generation Direct Liquid Cooling solution, DLC-2, is a comprehensive architecture designed to manage these thermal loads with maximum efficiency. The system uses cold plates to directly cool all major heat-generating components—including the CPUs, GPUs, DIMM memory modules, and voltage regulators (VRMs)—and circulates coolant through in-rack Coolant Distribution Units (CDUs) and space-saving vertical Coolant Distribution Manifolds (CDMs).[7]

This advanced thermal management system delivers dramatic and quantifiable improvements in data center efficiency and TCO.

Metric	Claimed Improvement (vs. Air Cooling)	Enabling Technology/Reason
Data Center Power Savings	Up to 40% [7]	Eliminates need for energy-intensive CRAC/CRAH units.
Overall TCO Reduction	Up to 20% [4]	Compounded savings from power, space, and hardware efficiency.
Data Center Footprint Reduction	Up to 60% [2]	Higher compute density allows more servers per rack and per square foot.
Water Consumption Savings	Up to 40% [7]	Supports warm water cooling (up to 45°C inlet), eliminating chillers.
System Heat Capture	Up to 98% [7]	Comprehensive cold plate coverage on all major components.
Noise Reduction	Down to ~50dB [7]	Drastically reduced fan speeds create a "library quiet" data center.

AMD Instinct GPU Performance Overview

The generational improvements in AMD's Instinct GPU line have delivered significant performance gains in just a few years, moving from the MI300X to the MI325X and now the MI355X in less than two years. One of the leading methods for comparing AI performance is to use industry-standard benchmarks, such as MLCommons, MLPerf Data Center Training, and Inference test results. AMD has submitted multiple results for these workloads, which provide an independent way to assess performance in real-world AI use cases. In Figure 1 below, we compare the MI300X, MI325X, and MI355X on the same benchmark, the MLPerf Llama2-70B offline workload, with results shown as Tokens per second.

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

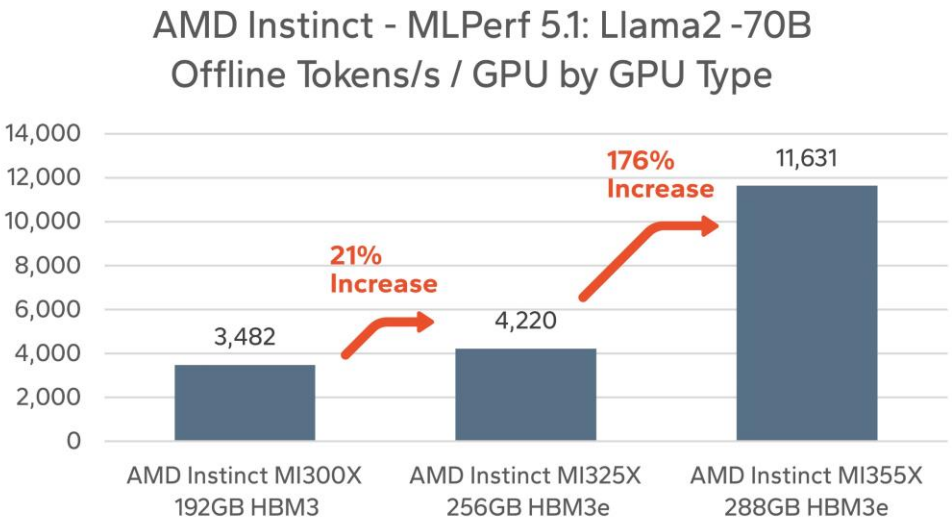

Figure 1 - MLPerf 5.1 Comparison of AMD Instinct GPUs using MLPerf v5.1 – Higher is Better

Note: The results for this and other benchmarks are provided in the Appendix.

The AMD Instinct GPU lineup has evolved rapidly over the past few years, with each successive generation showing significant improvements on a range of workloads. In particular, the MI355X shows more than a 2X increase compared to the previous generation MI325X on the MLPerf 5.1 Llama2-70b Offline Inferencing workloads.

Scalability is also an important consideration for both fine-tuning and inferencing workloads. To show the scalability of AMD's GPUs while performing an inferencing workload, Figure 2 shows the AMD MI355X results for the MLPerf Llama2-70B as the solution scales from 1 to 4, then 8 nodes (with 8, 32, and 64 total GPUs). Notice that the scaling is nearly linear, indicating that optimized hardware and software stacks can achieve linear scaling.

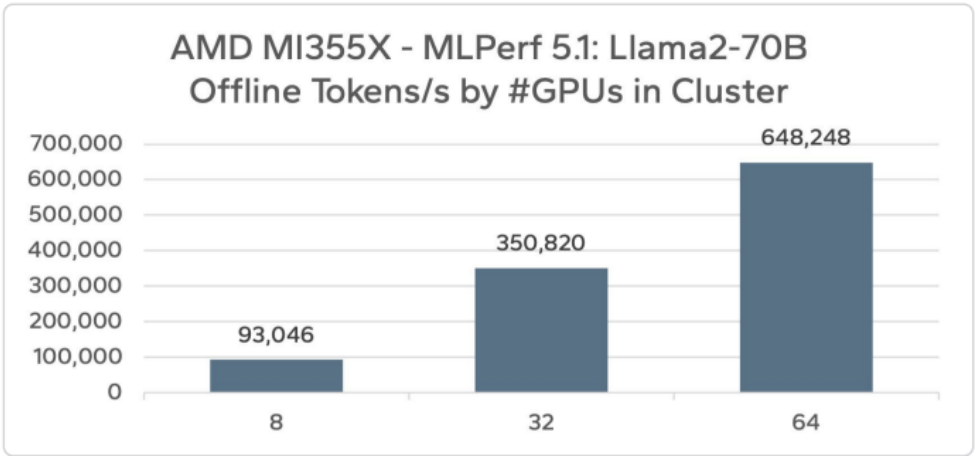

Figure 2 - MLPerf 5.1 Scaling of MI355X GPUs using MLPerf v5.1 – Higher is Better

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

Fine-Tuning, Another Key Enterprise Workload

The MI355X architectural advantages, particularly its large HBMe capacity and high memory bandwidth, become even more apparent in fine-tuning and inference workloads, where latency and throughput are critical. The following test case highlights these advantages.

Figure 3 shows a comparison of an MI300X to an MI355X for an MLPerf Llama2-70B LoRA Fine-Tuning workload. The results show that a single node of 8 MI355X GPUs can outperform a 3-node cluster of 32 – MI300X GPUs for this workload. This shows the generational improvements achieved in only two generations of AMD Instinct GPUs combined with ROCm software enhancements.

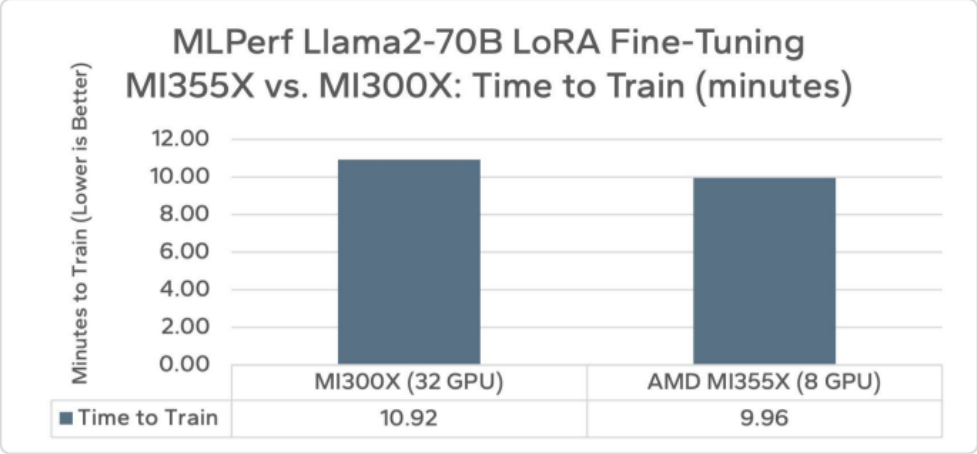

Figure 3 - MLPerf 5.1 Scaling of MI355X GPUs running MLPerf v5.1 – Lower is Better

The Synergy Advantage

The performance gains shown highlight the gains possible due to the synergy between the accelerator's architecture and a platform designed to unlock its full potential, deliver it to customers rapidly, and operate it with maximum efficiency.

Unlocking Performance with Superior GPU Memory Architecture

The performance advantages of the AMD MI355X are a direct result of deliberate architectural decisions centered on memory capacity and bandwidth. This hardware superiority translates into tangible benefits for key enterprise use cases:

- Faster Fine-Tuning: The 288 GB of HBM3e is the key to the MI355X's leadership in LoRA fine-tuning. It allows the full model weights, activations, and the memory-intensive optimizer states to coexist within a single GPU's memory, eliminating bottlenecks and allowing enterprises to iterate on custom models more rapidly. [1]

- Denser Long-Context Chat: For applications like RAG and advanced summarization, user density per GPU is a critical economic metric. A single MI355X can maintain the full KV cache for two simultaneous 128,000-token chat sessions with

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

a 70-billion-parameter model. This effectively doubles the number of users that can be served per GPU compared to accelerators with less memory, which would be forced to resort to performance-sapping memory paging techniques.[1]

- Accelerated RAG Pipelines: In RAG workflows, the speed of embedding and indexing new data is crucial. The MI355X's larger HBM allows an embedding instance to pin a larger portion of the vector index directly in high-speed GPU memory alongside the encoder model. This avoids streaming data over the much slower PCIe bus, resulting in a measured 2.1x speedup in index-build time compared to GPUs with smaller memory capacities. [1]

Future-Proofing Investments with a Green, Modular Approach

The operational efficiencies delivered by Direct Liquid Cooling-2 (DLC-2) from Supermicro align directly with the growing corporate imperative for sustainable or "Green Computing". By dramatically lowering a data center's Power Usage Effectiveness (PUE), the solution reduces not only operational costs but also the environmental impact associated with power generation.

This commitment to sustainability is further reinforced by Supermicro's disaggregated server design. This modular architecture fundamentally reduces electronic waste (e-waste). Instead of a "rip-and-replace" model where entire servers are discarded every few years, customers can independently upgrade key subsystems. For example, the compute module (containing the CPU, GPU, and memory) can be refreshed to take advantage of new technology, while the existing chassis, networking, power supplies, and liquid cooling infrastructure are retained. This approach can reduce the capital expenditure for a refresh cycle by 45% to 65% and minimize operational costs.

The combination of DCBBS and DLC-2 represents a fundamental de-risking of AI infrastructure investment for enterprises. The traditional procurement model forced customers to undertake a high-risk, complex science project, fraught with uncertainty: questions such as whether power requirements were sufficient, whether cooling was adequate, and whether IT resources were available to integrate and install equipment.

Supermicro's total solution approach effectively absorbs this risk. It transforms the procurement process from buying components with uncertain outcomes to consuming a finished product, a pre-validated, pre-cabled, liquid-cooled cluster with a guaranteed performance envelope and predictable operational characteristics. This "productization" of the AI data center is a massive value proposition that lowers the barrier to entry for deploying large-scale AI, making it accessible to a much broader range of enterprises.

Economic Analysis

A simplistic comparison of GPU list prices is a dangerously incomplete method for evaluating AI infrastructure. An actual Total Cost of Ownership (TCO) analysis must encompass the entire lifecycle cost of the solution, from initial acquisition and software licensing to multi-year operational expenses of the data center and future technology refreshes. When viewed through this holistic lens, the combined AMD-Supermicro solution offers a powerful, compounding economic advantage. Due to wide variations in AI hardware pricing and acquisition costs impacted by many factors, our analysis does not include Acquisition costs, which are typically associated with capital acquisition expenses (CAPEX).

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

Operational Costs

The initial CAPEX and software savings are then amplified by profound operational efficiencies delivered by the Supermicro platform.

- Power and Cooling Savings: As detailed previously, the Supermicro DLC-2 solution can reduce overall data center power costs by up to 40%. For a multi-megawatt AI cluster, this translates into millions of dollars in annual electricity savings, directly impacting OPEX.
- Density and Footprint Savings: The ability to achieve up to 60% higher compute density reduces the physical footprint of the deployment.2 For enterprises utilizing colocation facilities, this translates directly to lower monthly recurring costs for rack space. For those building new data centers, it means substantially lower capital costs for construction, as a smaller facility is required to house the same amount of compute power.
- Reduced Technology Refresh Costs: The disaggregated server architecture provides a long-term TCO advantage. By allowing independent upgrades of compute modules, the CAPEX required for a technology refresh can be reduced by 45% to 65% compared to a traditional model that requires replacing the entire server. Over a typical 3 to 5-year asset lifecycle, these savings become highly significant

Market Validation

This powerful economic value proposition is not merely theoretical; the market is actively validating it at an unprecedented scale. Supermicro has experienced extraordinary revenue growth, with a significant increase from $14.9 billion in revenue in FY2024 to $22.0 billion in FY25. This growth is evidence that enterprise and hyperscale customers are embracing the company's strategy of delivering optimized, first-to-market solutions with a superior TCO. The market is rewarding the ability to deliver not just components, but complete, efficient, and rapidly deployable AI solutions.

The combined economic model of the AMD-Supermicro solution creates a compounding TCO advantage. The savings are not merely additive; they are multiplicative. The process starts with a lower CAPEX baseline for the hardware. It then eliminates a significant annual OPEX line item in software licensing. Next, it drastically reduces the largest remaining operational cost, power and cooling, by up to 40%. It further reduces costs by shrinking the required physical space. Finally, it lowers future CAPEX for technology refreshes by up to 65%. Each stage of savings reduces the base cost used to calculate the next stage of operations, creating a powerful compounding effect throughout the equipment's lifecycle. This results in a TCO that is substantially lower than a component-level analysis would ever suggest, creating an economic moat that is difficult for a premium-priced, monolithic competitor to overcome in the mainstream enterprise market.

Analysis and Outlook for IT Leaders

Findings

- The AMD Instinct MI355X, with its industry-leading 288 GB of HBM3e memory, demonstrates clear and often decisive performance leadership in memory-intensive enterprise AI workloads like large model inference and fine-tuning.
- The Supermicro platform is a critical enabler of this performance. Its Building Block philosophy accelerates time-to-market, while its Data Center Building Block Solutions and advanced DLC-2 liquid cooling technology provide unmatched operational efficiency, density, and deployment velocity.

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

- The combined solution offers a multi-layered, compounding TCO advantage derived from competitive hardware pricing, zero-cost software licensing, dramatic reductions in power and cooling costs, and lower long-term refresh expenses.

Strategic Recommendations for IT Leaders

1. Evaluate Workloads, Not Just Specifications: IT leaders should move beyond comparing theoretical peak specifications on datasheets and instead benchmark solutions against their own specific production workloads. If the primary use cases involve inferencing with large models, leveraging long context windows for RAG, or frequent fine-tuning, the architectural memory advantage of the MI355X is highly likely to provide a decisive performance and TCO edge.
2. Adopt a "Total Solution" Procurement Model: The complexity and risk associated with deploying AI at scale demand a shift away from component-level RFPs. Leaders should prioritize integrated, rack-scale solutions from vendors who can deliver validated, turnkey clusters, accelerate time-to-value, and provide the predictability required for accurate financial planning.
3. Plan for Liquid Cooling as the Default: For any new, at-scale AI deployment, direct-to-chip liquid cooling should be the default planning assumption. The economic, density, and sustainability benefits are too significant to ignore. Engaging with vendors who have proven, data-center-scale liquid cooling solutions is no longer optional; it is a strategic necessity for building efficient and future-proof AI infrastructure.

Outlook

The industry trends are clear: AI models will continue to grow, context windows will expand, and multi-model agentic pipelines will become more common. Each of these trends increases the pressure on the memory subsystem, suggesting that the MI355X's architectural advantage is poised to become even more critical over time. The powerful partnership between AMD's performance-leading accelerator technology and Supermicro's expertise in direct liquid-cooling, scalable system designs, and rapid deployment is now a leading provider of IT systems, offering credible, high-performance, and economically efficient alternatives that give enterprises a powerful new option for building their AI future.

Appendix

The following hardware configurations are the same for each test scenario, except as otherwise stated.

Test Procedures and Metric Calculations:

- Test procedure is the same as described in the “MI355X: MLPerf Training Llama-2-70B Lora 8-GPU Training Score”: https://signal65.com/wp-content/uploads/2025/06/Signal65-Insights_AMD-Instinct-MI355X-Examining-Next-Generation-Enterprise-AI-Performance.pdf
- Test procedure for 4 node MI300X submission by MangoBoost: https://www.mangoboost.io/resources/blog/mangoboost-sets-a-new-standard-for-multi-nodes-llama2-70b-lora-on-amd-mi300x-gpu

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

MLPerf Results

					Llama2-70B 99.9	
					Offline	Offline/GPU
ID	Submitter	Nodes	Accelerator	Total Accel.	Tokens/s	Token/s
5.1-0001	AMD	1	AMD Instinct MI300X 192GB HBM3	8	27,803.90	3,475.49
5.1-0066	MangoBoost	1	AMD Instinct MI300X 192GB HBM3	8	27,854.40	3,481.80
5.1-0090	Supermicro_ MangoBoost	2	AMD Instinct MI325X 256GB HBM3e	16	65,320.10	4,082.51
5.1-0091	Supermicro_ MangoBoost	3	AMD Instinct MI325X 256GB HBM3E (x16), AMD Instinct MI300X 192GB HBM3 (x8)	24	92,158.20	3,839.93
5.1-0095	Vultr	1	AMD Instinct MI325X 256GB HBM3e	8	33,762.50	4,220.31
5.1-0099	AMD	1	AMD Instinct MI355X 288GB HBM3e	8	93,045.80	11,630.73
5.1-0103	AMD_ MangoBoost	4	AMD Instinct MI355X 288GB HBM3e	32	350,820.00	10,963.13
5.1-0105	AMD_ MangoBoost	8	AMD Instinct MI355X 288GB HBM3e	64	648,248.00	10,128.88

References

1. Signal65-Insights_AMD-Instinct-MI355X-Examining-Next-Generation-Enterprise-AI-Performance.docx https://signal65.com/wp-content/uploads/2025/06/Signal65-Insights_AMD-Instinct-MI355X-Examining-Next-Generation-Enterprise-AI-Performance.pdf

2. Data Center Building Block Solutions® (DCBBS) | Supermicro, accessed September 15, 2025 https://www.supermicro.com/en/solutions.dcbbs

3. Supermicro Begins Volume Shipments of NVIDIA Blackwell Ultra Systems and Rack Plug-and-Play Data CenterScale Solutions, accessed September 15, 2025 https://www.supermicro.com/en/pressreleases/supermicro-beginsvolume-shipments-nvidia-blackwell-ultra-systems-and-rack-plug-and

4. Supermicro's DLC-2, the Next ... - Super Micro Computer, Inc., accessed September 15, 2025, https://ir.supermicro.com/news/news-details/2025/Supermicros-DLC-2-the-Next-Generation-Direct-Liquid-Cooling-Solutions-Aims-toReduce-Data-Center-Power-Water-Noise-and-Space-Saving-on-Electricity-Cost-by-up-to-40-and-Lowering-TCOby-up-to-20/default.aspx

© 2025 Copyright Super Micro Computer, Inc. All rights reserved

5. Supermicro Unveils DCBBS and DLC-2 to Power Liquid-Cooled AI Data Centers, accessed September 15, 2025, https://www.storagereview.com/news/supermicro-unveils-dcbbs-and-dlc-2-to-power-liquid-cooled-ai-data-centers

6. Supermicro Begins Volume Shipments of NVIDIA Blackwell Ultra Systems and Rack Plug-and-Play Data Center Scale Solutions, accessed September 15, 2025, https://ir.supermicro.com/news/news-details/2025/SupermicroBegins-Volume-Shipments-of-NVIDIA-Blackwell-Ultra-Systems-and-Rack-Plug-and-Play-Data-Center-ScaleSolutions/default.aspx

7. Supermicro DLC-2 architecture reduces data center power, space, water, and costs, accessed September 15, 2025, https://www.supermicro.com/white_paper/White_Paper_Supermicro_DLC-2.pdf

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

© 2025 Copyright Super Micro Computer, Inc. All rights reserved