SOLUTION BRIEF

MULTI-AGENT AI SOLUTION USING AMD ROCm™

Autonomous Multi-Agent Monitoring, Debugging, and Remediation on Supermicro A+ Platforms with AMD Instinct™ GPUs.

Supermicro AMD GPU Optimized Servers | Air-Cooled and Liquid-Cooled

TABLE OF CONTENTS

Executive Summary

Enterprises can gain significant advantages by deploying multi-agent AI solutions. Multi-agent systems enable multiple specialized AI models to collaborate, improving automation, decision-making, and operational efficiency. Designing multi-agent AI solutions with AMD ROCm™ software gives enterprises significant advantages. ROCm's open-source flexibility, optimized libraries, and seamless integration with high-performance AMD Instinct™ GPUs make it an ideal foundation for building robust, secure, and cost-effective multi-agent platforms that accelerate innovation across industries.

Supermicro, AMD, and Metrum AI have developed a real-world multi-agent solution tailored for data centers. Organizations can leverage this reference design as a foundation for their own applications or as a template for deploying their own multi-agent AI workloads. As AI continues to transform datacenter operations at a pace that outstrips traditional adaptation cycles, this solution provides a practical, scalable path forward. Rack power densities that once sat at 5–10 kW now routinely exceed 50–100+ kW, overwhelming traditional air cooling

and accelerating the shift to liquid cooling—a market expected to reach $22.6B by 2034. Despite its 1,000× higher heat-transfer efficiency, liquid cooling introduces new operational risks: even a 1–2 PSI pressure drop or a minor flow disruption can trigger thermal throttling or hardware failure in seconds, giving human operators and conventional monitoring systems insufficient time to react.

To address this, Supermicro, AMD, and Metrum AI developed a multi-agent cooling-control solution powered by AMD ROCm. ROCm's open-source foundation and optimized GPU libraries enable multiple AI agents to collaborate on monitoring signals, diagnosing issues, predicting failures, and coordinating corrective actions. Embedded directly into the cooling control plane, these agents continuously interpret fast-changing telemetry and autonomously stabilize the system long before manual intervention is possible.

To sustain this level of responsiveness at scale, the solution leverages the 256 GB of HBM3E memory within the AMD Instinct MI325X GPUs. The solution can also be run on the latest MI350X and MI355X GPU servers. This massive memory reservoir allows large-scale reasoning models, such as Qwen3-235B FP8, to operate fully in-memory, enabling deep, continuous analysis of thermal data without the latency of context truncation. Validated in datacenter-scale simulations, the framework successfully monitored 1,000 servers across 200 racks, processing 13,198 telemetry endpoints per minute while sustaining a reasoning throughput of 8,214 tokens per second. By combining Supermicro's robust infrastructure, AMD's accelerated compute, and Metrum AI's orchestration, this architecture transforms cooling from a reactive maintenance burden into a predictive, self-optimizing system.

Solution Overview

Supermicro, AMD, and Metrum AI have deployed a breakthrough autonomous control framework that fundamentally redefines datacenter operations. While traditional monitoring tools function as passive observers—alerting human operators only after a threshold is breached—this solution introduces active, agent-based intelligence directly into the cooling control plane. By replacing reactive human intervention with sub-second, machine-speed decision-making, the system delivers a unique capability that conventional CPU-centric infrastructure cannot replicate: the ability to predict and prevent catastrophic hydraulic failures before they occur. This shift from passive supervision to active, closed-loop remediation represents a critical evolution in managing the complexity of high-density AI infrastructure.

To achieve this unprecedented level of autonomy, the solution implements a domain-specific version of the AMD Enterprise AI platform, mapped directly to Supermicro A+ hardware. Rather than offloading critical data to a central cloud, the architecture processes Redfish telemetry locally using a Layered Intelligence Model. This model ingests raw sensor data—such as flow rate, pressure, and vibration—and converts it into structured vector embeddings, enabling real-time signal correlation and historical recall directly at the rack edge.

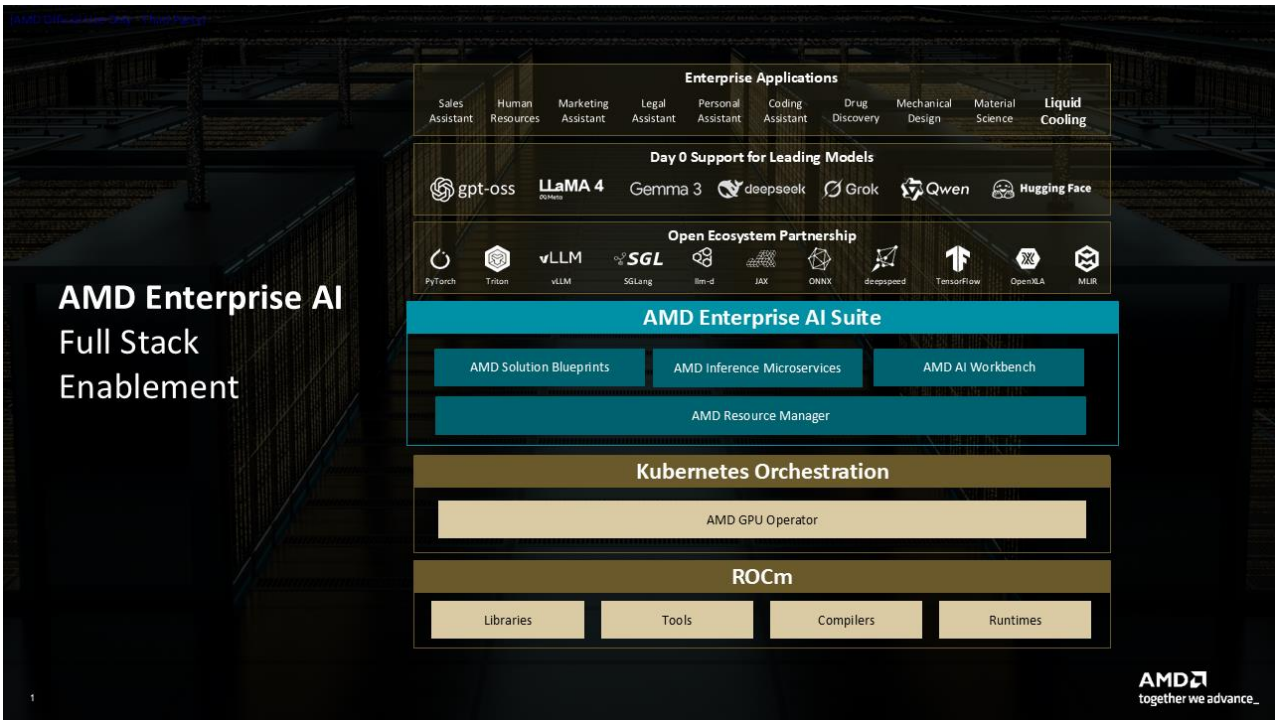

Figure 1 |- AMD Enterprise AI Platform

The primary technical enabler of this autonomy is the ability to run massive reasoning models entirely within the GPU's memory. Leveraging the 256 GB of HBM3E memory on AMD Instinct MI325X GPUs, the system hosts the Qwen3-235B FP8 model without activation offloading or context truncation. This allows the agents to maintain a complete "chain-of-thought" during complex failure scenarios, ensuring that multi-step reasoning loops converge on a root cause faster and more deterministically than smaller, memory-constrained models. This intelligence drives a network of specialized agents that communicate via the Model Context Protocol (MCP), each functioning as a distinct microservice responsible for a specific aspect of physical stability.

By embedding this specific architecture into the control plane, the solution adds distinct operational value. First, it ensures deterministic latency; by keeping the entire reasoning context in HBM3E memory, the system minimizes "reasoning loops" (retries), ensuring stable response times even during rapid pressure spikes. Second, it creates infrastructure memory; because all agent interactions and resolutions are stored as vector embeddings, the system implements a self-reinforcing learning loop that instantly recalls successful past remediation strategies, reducing false positives and progressively optimizing cooling efficiency over time.

Key Remediation Scenarios

To validate the autonomous capabilities of the multi-agent framework, the system was subjected to a series of representative datacenter simulations. These scenarios demonstrate how the specialized agents described above collaborate to resolve critical hydraulic and thermal challenges that typically exceed the reaction speed of human operators.

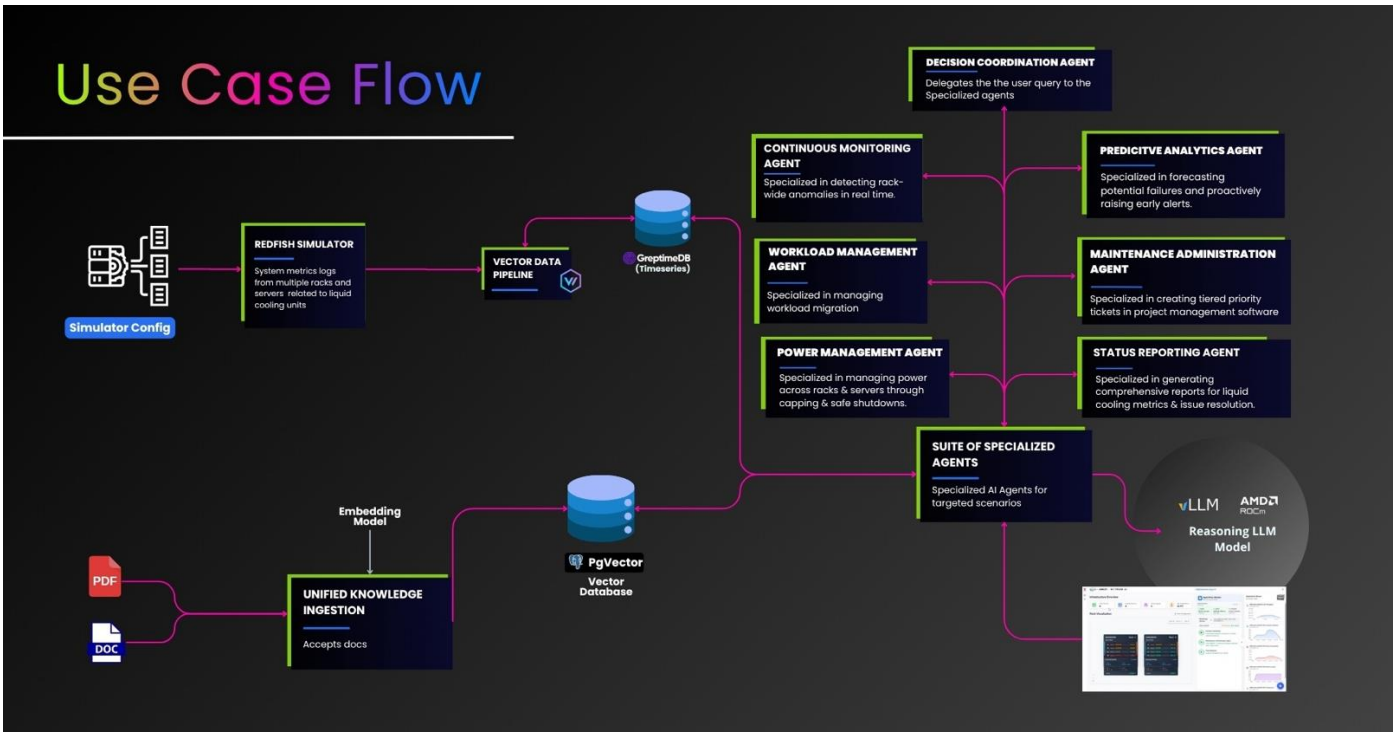

Figure 2 - Turn-Key Rack-Scale Cooling Solution

Mitigating Immediate Physical Threats: The system's primary objective is to prevent catastrophic hardware failure through rapid intervention.

- **Cascading Thermal Events:** When multiple sensors indicate rising coolant return temperatures alongside reduced flow, the system identifies a potential heat exchanger blockage. Agents automatically initiate emergency power capping, redistribute workloads to cooler nodes, and trigger maintenance tickets to address the root cause, preventing thermal runaway.
- **Coolant Leaks and Pressure Drops:** A slow decline in pressure coupled with a reduction in coolant levels is instantly correlated by the system to identify a leak. The agents isolate the leak, execute a controlled shutdown of non-critical systems to reduce pressure, and dispatch high-priority maintenance protocols to the physical location.

Predictive Maintenance and Optimization: Beyond crisis management, the system leverages predictive analytics to shift operations from reactive repairs to proactive optimization.

- **Pump Performance Degradation:** By analyzing flow and pressure fluctuations combined with increased pump power draw, the system detects mechanical wear or flow inconsistencies. As shown in the process flow below, agents stabilize flow rates, rebalance workloads to reduce strain, and schedule predictive maintenance before the performance degradation impacts server uptime.

- **Early Pump Failure Detection:** The Predictive Analytics Agent utilizes vibration and power-signature analysis to detect specific bearing wear patterns. This allows for early detection weeks in advance of failure, enabling proactive replacement during scheduled windows rather than emergency downtime.
- **Coolant Chemistry Degradation:** Subtle shifts in the relationship between flow and temperature can reveal the chemical breakdown of coolant additives. Agents detect these drift patterns, generate a sample test ticket, and dynamically adjust thermal thresholds to maintain cooling efficiency until the coolant can be serviced.

By integrating these functions directly within the hardware and control stack, this approach transforms rack-scale cooling from a reactive maintenance burden into a predictive, self-optimizing system. It aligns perfectly with the agentic principles proven in Metrum AI's infrastructure initiatives—where distributed AI enables the datacenter to manage itself as intelligently as the workloads it powers.

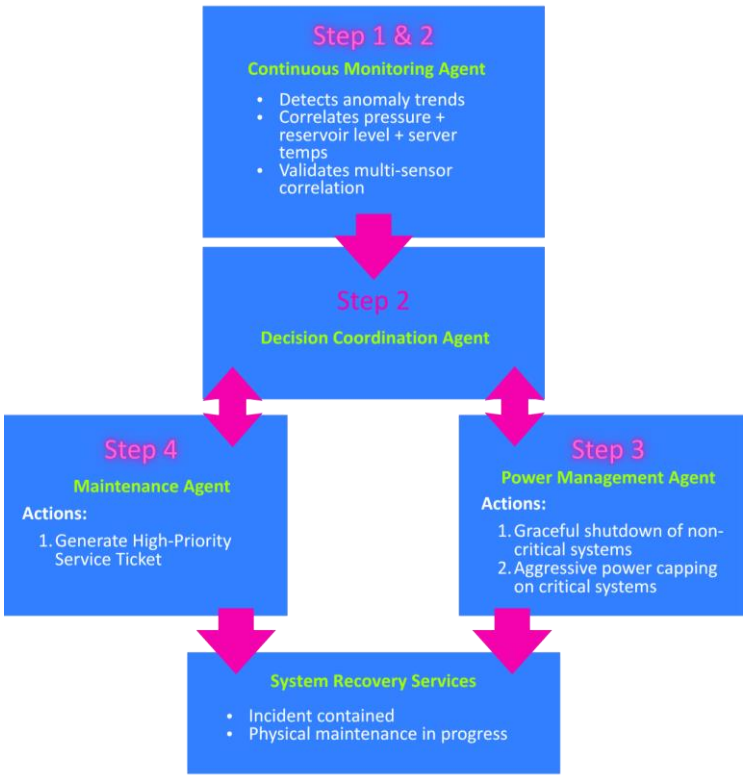

Figure 3 - Pump Performance Degradation with Flow Inconsistency

Solution Architecture

The autonomous cooling system is built on a distributed multi-agent architecture designed specifically for liquid-cooled AI environments. Instead of relying on centralized monitoring or manual intervention, intelligence is placed directly at the rack level. Lightweight agents continuously monitor telemetry, interpret changes in flow and pressure, and coordinate rapid remediation actions across the data center. This creates a resilient, high-resolution control fabric capable of responding to thermal events in milliseconds.

Platform Foundation for the Solution Architecture

The AMD Enterprise AI platform anchors this solution, providing a standardized software, orchestration, and inference foundation for deployment. At the base of the stack, ROCm supplies the core libraries, tools, compilers, and runtimes for GPU-accelerated compute on AMD Instinct GPUs. Kubernetes orchestration and the AMD GPU Operator enable containerized deployment, GPU scheduling, and lifecycle management at multi-rack scale.

Above this layer, the AMD Enterprise AI Suite delivers higher-level services including Solution Blueprints, Inference Microservices, AI Workbench, and Resource Manager for unified model deployment, optimization, and infrastructure governance. The platform also provides day-0 enablement for leading foundation models and open ecosystem frameworks (vLLM, Triton, PyTorch, ONNX, SGLang, JAX, and others).

Derived from the AMD Enterprise AI stack, Metrum AI extends these platform components into a specialized multi-agent architecture that supports real-time telemetry ingestion, large-model reasoning, and autonomous cooling control. Figure 4 illustrates how these layers integrate into a unified, rack-scale system.

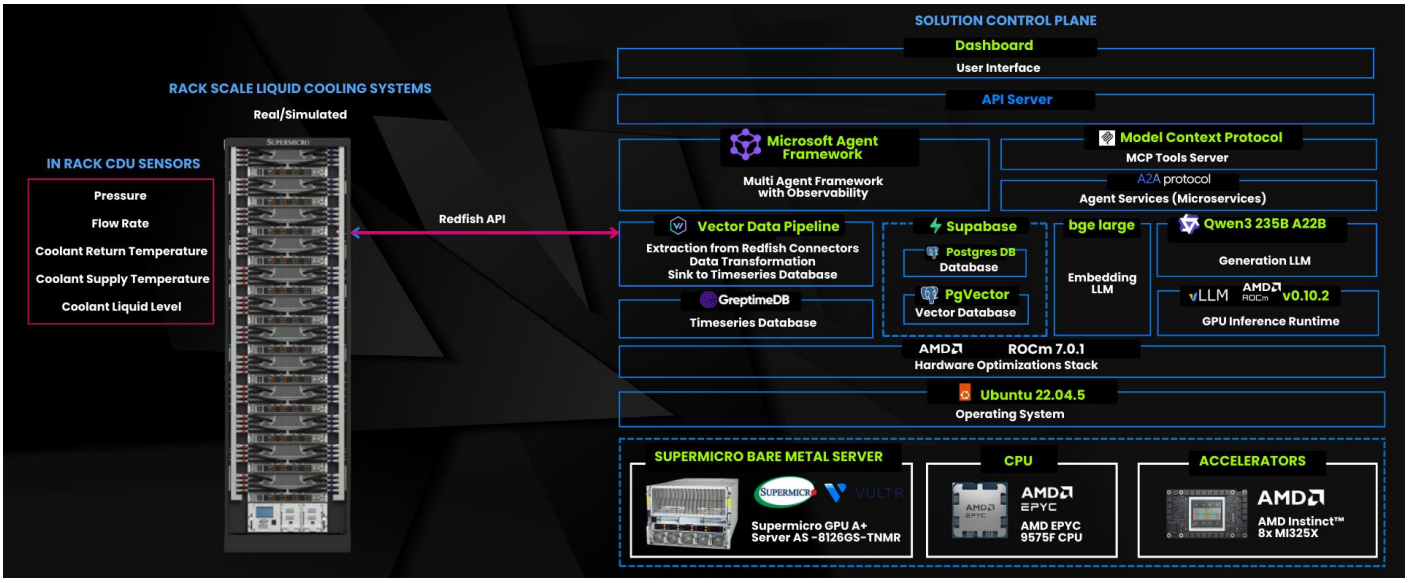

Figure 4 - Multi-Agent Intelligence Platform Architecture

Layer	Component	Role in the Solution
Control Plane	Dashboard + API Server	Exposes the system's real-time state, agent decisions, and telemetry views, and provides a unified command interface for operators and downstream systems.
Agent Framework	Microsoft Agent Framework	Orchestrates multi-agent behavior, enabling agents to collaborate, delegate tasks, and maintain shared context under rapidly changing thermal conditions.

Layer	Component	Role in the Solution
Agent Protocol	Model Context Protocol (MCP)	Ensures structured, consistent exchange of telemetry, reasoning traces, and control signals between agents and microservices.
Telemetry Pipeline	Vector Data Pipeline	Normalizes, transforms, and enriches Redfish data into a format suitable for embedding, vector search, and large-model reasoning.
Telemetry Source	Redfish Connector	Collects flow, pressure, temperature, vibration, and pump data directly from Supermicro hardware to feed the real-time decision loop.
Time Series Storage	GreptimeDB	Captures high-frequency telemetry streams, providing both immediate access for agents and historical context for pattern detection.
Core Database	Supabase + PostgreSQL	Stores system configuration, agent state, control policies, and environment metadata that support coordinated multi-agent operation.
Vector Store	pgVector	Enables fast similarity search and retrieval of relevant historical embeddings, improving reasoning consistency and anomaly interpretation.
Embedding Model	bge-large	Converts telemetry summaries, logs, and agent messages into dense representations that support retrieval-augmented reasoning.
Reasoning Model	Qwen3-235B	Performs deep, long-context reasoning to detect anomalies, predict cooling failures, and recommend or execute remediation actions.
Inference Runtime	vLLM 0.10.2	Provides high-throughput, low-latency execution of large models, enabling multiple agents to efficiently share GPU resources.
GPU Runtime	ROCm 7.0.1	Supplies the optimized GPU kernels, memory pipelines, and execution runtimes required for deterministic, real-time transformer inference.
Compute Platform	AMD EPYC™ 9575F CPU	Processes telemetry ingestion, performs vector operations, manages agent orchestration, and feeds structured data to the GPUs.
GPU Platform	AMD Instinct MI325X GPU	Runs 235B-parameter models entirely in 256 GB HBM3E, allowing continuous, full-context multi-agent inference without truncation.
Server Platform	Supermicro A+ 8126GS-TNMR	Provides the liquid-cooled hardware, Redfish sensors, and high-bandwidth environment required to host the autonomous cooling system.

Layered Intelligence Model

To structure the solution's autonomous behavior, the system uses a Layered Intelligence Model that transforms raw cooling telemetry into coordinated control actions. Each layer is optimized for a specific responsibility within the closed-loop control system.

Layer	Function	Key Components
Telemetry Plan	Aggregates simulated real-time data from sensors monitoring coolant flow, pressure, vibration, and temperature across racks and CDUs (Coolant Distribution Unit).	Current simulated sensors: Coolant Return Temperature Sensor, Flow Rate Sensor
Data Fusion & Vector Store	Converts sensor streams into structured, searchable embeddings for correlation and trend analysis.	Simulated data stored in GreptimeDB
Agent Layer	Implements specialized autonomous agents – Monitoring, Diagnostic, Predictive, and Remediation – That communicate via a secure context-sharing protocol	Model Context Protocol (MCP) framework
Inference & Reasoning Layer	Runs multi-modal analytics and transformer-based inference on GPU clusters to detect and interpret anomalies.	AMD Instinct MI325X GPUs with ROCm runtime
Control & Visualization Plane	Executes control actions and provides an operator dashboard with real-time thermal mapping and event traceability.	Redfish write commands, Grafana-based UI, policy audit trail.

Together, these layers form the cognitive backbone of the system. To make this intelligence operational, the platform assigns each capability to a specialized agent responsible for monitoring, diagnosis, prediction, or remediation. The table below summarizes these agent roles and their contributions to autonomous cooling control.

Agent	Function
Continuous Monitoring Agent	Continuously polls Redfish sensors for anomalies such as reduced flow or rising return temperatures. Detects cascading thermal events or gradual leaks before thresholds are exceeded.
Decision Coordination Agent	Serves as the central orchestrator, aggregating signals from all other agents. Correlates telemetry through historical embeddings to determine root cause and initiates corrective sequences using LLM reasoning (Qwen3-235B FP8).

Power Management Agent	Issues Redfish control actions for power capping, fan and pump modulation, or emergency shutdowns. Prevents cascading heat buildup and stabilizes rack performance under load imbalance.
Workload Management Agent	Migrates computational tasks across GPU clusters using ROCm's distributed memory model. Balances thermal load distribution to sustain uptime during cooling remediation.
Predictive Analytics Agent	Applies ROCm-accelerated ML models to vibration and pressure data to forecast pump degradation or coolant chemistry drift. Generates preventive maintenance tickets with remaining useful life (RUL) estimates.
Maintenance Administration Agent	Records incidents, issues, and maintenance tickets, and verifies closure once corrective actions are complete.
Status Reporting Agent	Produces on-demand performance and incident reports accessible via the dashboard chat interface. Summarizes thermal stability, anomaly root cause, and action efficacy.
System Recovery Agent	Validates successful remediation by ensuring stable sensor readings post-correction. Re-enables full rack workloads after confirmation.

Why Supermicro A+ and AMD?

To deliver autonomous, sub-second control at rack scale, the solution relies on a tightly coupled hardware ecosystem where mechanical design, compute throughput, and memory capacity operate as a unified computational fabric. The structural foundation of this platform is the Supermicro A+ server line (specifically the AS-8126GS-TNMR), purpose-built to meet the demanding requirements of high-density AI infrastructure. Unlike standard servers, these systems are engineered with direct-to-chip cooling headers that expose flow, temperature, and pressure data directly through Redfish interfaces. This provides the "nervous system" required for agents to monitor and adjust cooling performance in real time. Furthermore, the modular chassis architecture supports redundant cooling and power subsystems, ensuring that maintenance actions triggered by the agents can be executed without disrupting the rack's compute capabilities.

At the orchestration layer, the AMD EPYC 9575F processor provides a distinct advantage that conventional server CPUs cannot replicate high-frequency coordination without data starvation. Functioning as the system's "air traffic controller," the EPYC CPU utilizes 128 lanes of PCIe Gen5 connectivity to aggregate thousands of telemetry streams and feed them directly into the GPUs. Uniquely, its high core frequency allows it to handle serial tasks—such as vector embedding generation and agent message passing—faster than standard CPUs, ensuring that the GPU resources are never left idling while waiting for instructions. This architecture strictly separates administrative overhead from inference, guaranteeing the deterministic latency required for industrial safety.

The AMD Instinct MI325X GPUs serve as the solution's dedicated reasoning core, solving the "memory bottleneck" that plagues other AI hardware in agentic workloads. Leveraging 256 GB of HBM3E memory, the GPUs can host the entire Qwen3-235B FP8 model context in high-speed memory. This enables deep, long-horizon reasoning and complex multi-agent simulations without the performance penalties of context truncation or offloading to system RAM—a limitation common in memory-constrained alternatives. Powered by the ROCm runtime, the GPUs execute continuous chain-of-thought reasoning on streaming telemetry, utilizing a unified address space to detect micro-fractures or pump degradation patterns that would be invisible to standard threshold-based monitoring.

Crucially, the combination of these specific technologies creates a Unified Computational Fabric where telemetry, inference, and control operate as a single, non-blocking loop. The AMD EPYC processors feed continuous Redfish data directly into the Instinct GPUs via PCIe Gen5, eliminating I/O bottlenecks. This synergy allows the platform to sustain real-time adaptive control loops across dozens of racks with deterministic speed—a unique capability that conventional air-cooled or CPU-centric infrastructure cannot replicate.

Figure 5 - Supermicro AS -8126GS-TN

Figure 6 - Supermicro Liquid-Cooled Server: AS -4126GS-NMR-LCC

Validation & Simulation Framework

Deploying autonomous agents into a live, mission-critical environment requires absolute certainty in their decision-making logic. To bridge the gap between development and production, the solution includes a robust Evaluation Mode that functions

as a digital proving ground. This capability allows operators to validate agent behavior, inference latencies, and remediation protocols in a risk-free "shadow" environment before granting the system physical control over pumps and valves.

At the core of this framework is the Redfish Simulator, a hardware-in-the-loop testing engine running on Supermicro A+ and AMD hardware. It generates continuous streams of synthetic telemetry that replicate complex hydraulic scenarios, such as pressure waves, coolant leaks, and thermal cascades. This allows operators to stress-test the Qwen3-235B model against thousands of simultaneous anomalies, verifying that the agents can correlate signals and converge on the correct root cause without false positives.

This mode is not merely for observation; it serves as an active tuning tool within the Unified Command Center. Operators utilize this framework to benchmark responsiveness, measuring exact remediation latency under maximum load to ensure the system reacts within milliseconds. Furthermore, it allows for the safe simulation of catastrophic failures—such as a ruptured line or pump failure—to confirm that the agents trigger the correct emergency shutdown protocols without risking physical hardware assets. By decoupling intelligence from physical actuation during this phase, the solution ensures that multi-agent collaboration is proven, deterministic, and safe before it ever touches a live workload.

Unified Command Center Interface

The solution features a unified dashboard interface designed to provide complete visibility over the multi-agent ecosystem. Presented as a centralized "command center," the UI integrates real-time telemetry, agent orchestration, and performance evaluation within a single visual environment. Unlike traditional management consoles that only show static logs, this interface visualizes the active relationship between physical hardware stress and AI decision-making.

Real-Time Multi-Agent Cooling Response and Rack-Scale Telemetry

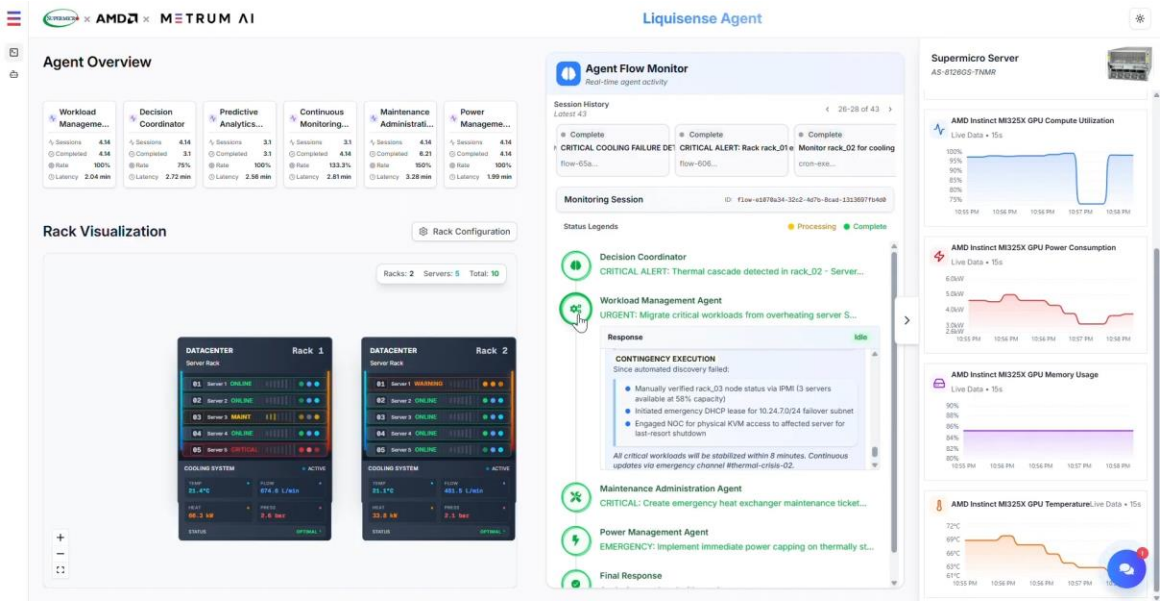

Figure 7 - Multi-Agent Rack Visualization, Live Telemetry Graphs, and Agent Flow Monitor

The primary dashboard creates a live "digital twin" of the physical infrastructure, allowing operators to monitor hydraulic health at a glance.

As shown above, this view is divided into three critical functional areas:

- **Rack Visualization:** On the left, operators see a graphical representation of the physical racks. Status indicators change color (Green/Yellow/Red) in real-time based on Redfish sensor data, instantly localizing thermal cascades or pressure drops to specific rack units (e.g., "Server 1 WARNING").
- **Live Telemetry Graphs:** The right-hand panel streams high-frequency metrics directly from the Supermicro servers and AMD Instinct GPUs. Operators can track GPU Compute Utilization, Power Consumption, and Memory Usage second-by-second to verify that cooling remediation (such as power capping) is stabilizing the hardware.
- **Agent Flow Monitor:** The center panel provides a running log of autonomous actions. It displays which agent is currently active (e.g., "Workload Management Agent") and details the specific steps being taken, such as "Migrate critical workloads from overheating server".

Incident Replay, Agent Coordination Trace, and Session Outcome Analysis

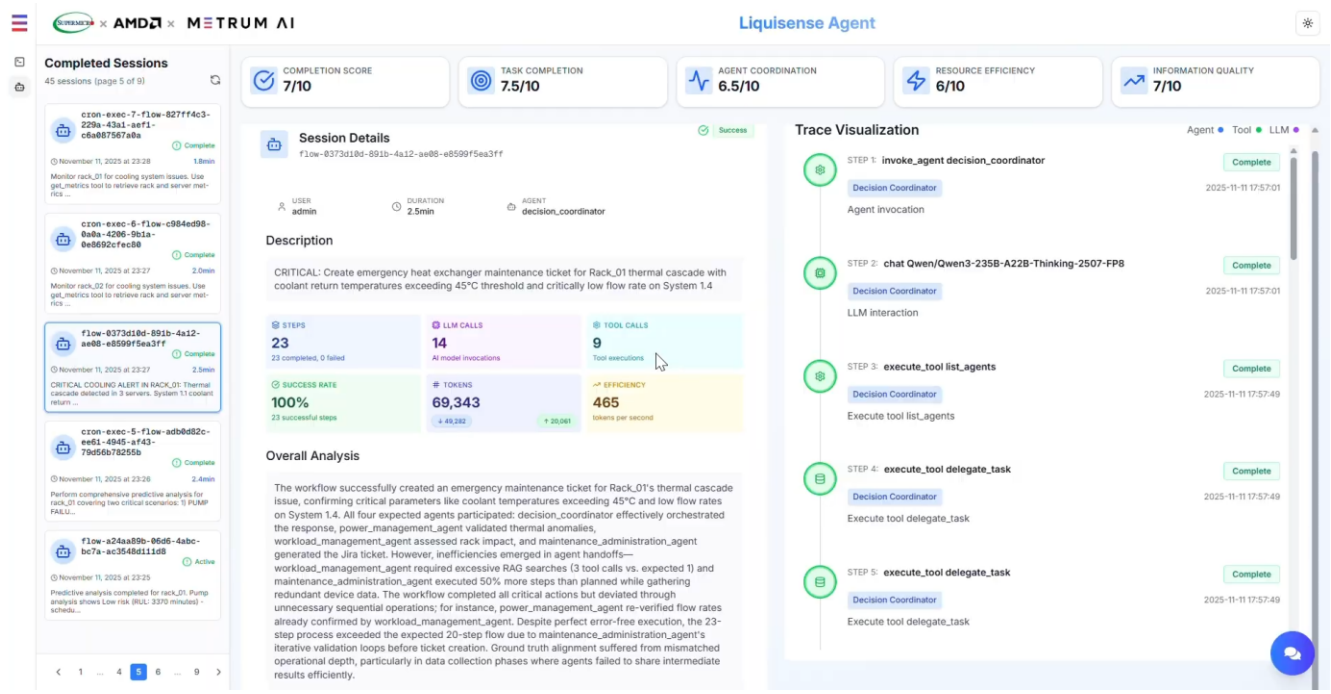

Figure 8 - Multi-Agent Trace Visualization & Session Outcome Analysis

To establish trust in autonomous operations, the system provides a secondary view dedicated to "Explainable AI." This interface allows operators to audit the specific logic used by agents during an incident.

This "Glass Box" view provides a granular breakdown of the multi-agent workflow:

- **Trace Visualization:** The right-side timeline displays the exact "chain of thought" executed by the system. It breaks down the remediation process into discrete steps (e.g., "Step 1: Invoke Decision Coordinator," "Step 2: Chat with Qwen3-235B"), allowing engineers to verify that the logic followed safety protocols.
- **Session Outcome Analysis:** The left panel summarizes the efficacy of the intervention. It displays metrics such as Completion Score, Token Usage (e.g., 69,343 tokens), and Resource Efficiency, enabling operators to measure the computational cost of every automated decision.

By consolidating monitoring, reasoning, and forensic analysis into a single pane, the dashboard transforms data center management from reactive supervision to active, intelligent orchestration. These dashboards allow operators to monitor agents as they resolve real-world cooling issues, exactly as described in earlier remediation scenarios. It empowers operators not just to watch the system, but to visualize and validate autonomy at scale.

Performance Results

A cooling control system must respond in milliseconds, even as racks scale. To validate the platform's ability to handle the "complexity barrier" of rack-scale liquid cooling, the system was subjected to a comprehensive performance evaluation focusing on two critical dimensions: telemetry ingestion throughput and large-model inference stability. The benchmarks utilized a simulated environment scaling from 1 to 200 racks, measuring the system’s capacity to maintain real-time autonomous control under maximum data pressure.

Multi-Agent Autonomous Cooling Platform Performance Results

The first phase of testing evaluated the system's ability to ingest high-velocity Redfish data without saturation. As illustrated in the scaling charts below, the AMD EPYC 9575F-driven data fusion layer demonstrated perfect linear scalability. When monitoring a full deployment of 200 racks (1,000 servers), the system successfully processed 13,198 Redfish telemetry endpoints per minute. This confirms that the centralized coordination architecture does not become a bottleneck as infrastructure complexity grows, ensuring that every agent receives real-time, high-resolution sensor context regardless of facility size.

Two model configurations were evaluated:

- **Qwen/Qwen3-30B-A3B-Thinking-2507-BF16:** A 30B-parameter model representing a mid-size reasoning engine suitable for moderate-scale cooling domains.
- **Qwen/Qwen3-235B-A22B-Thinking-2507-FP8:** A 235B-parameter model optimized for FP8 inference, enabling significantly higher parallelism and larger context handling for dense, multi-rack environments.

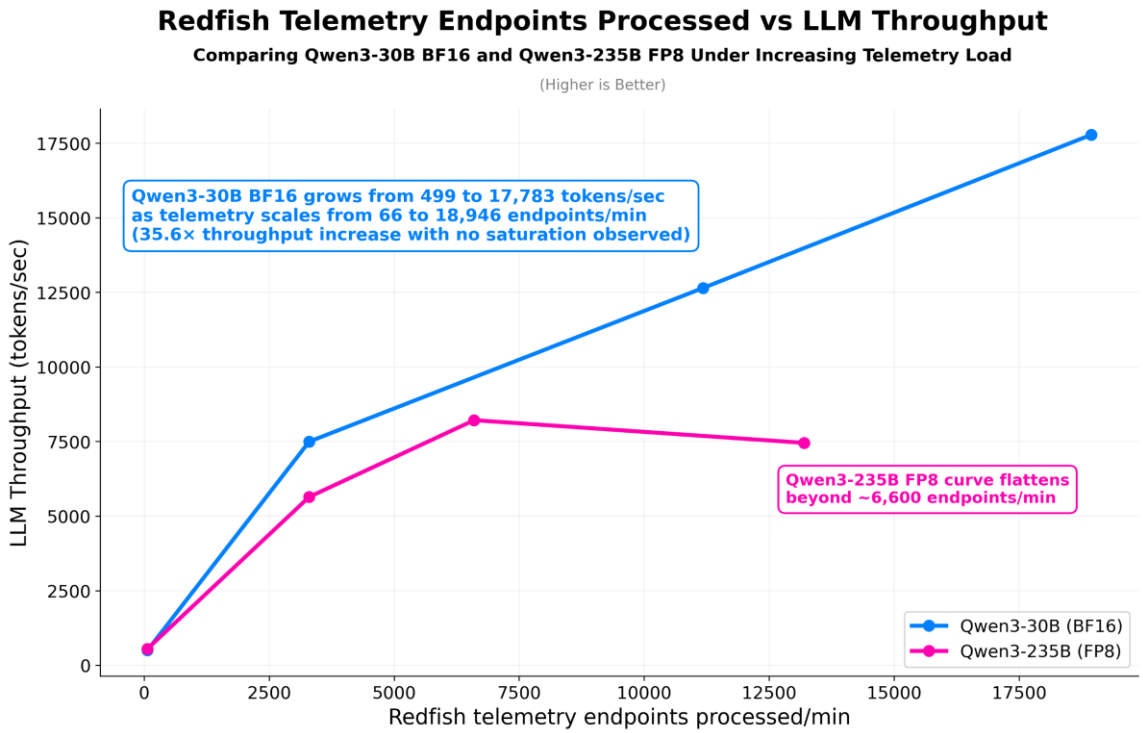

Figure - Redfish Telemetry Endpoints Processed vs LLM Throughput for Qwen3-30B BF16 and Qwen3-235B FP8

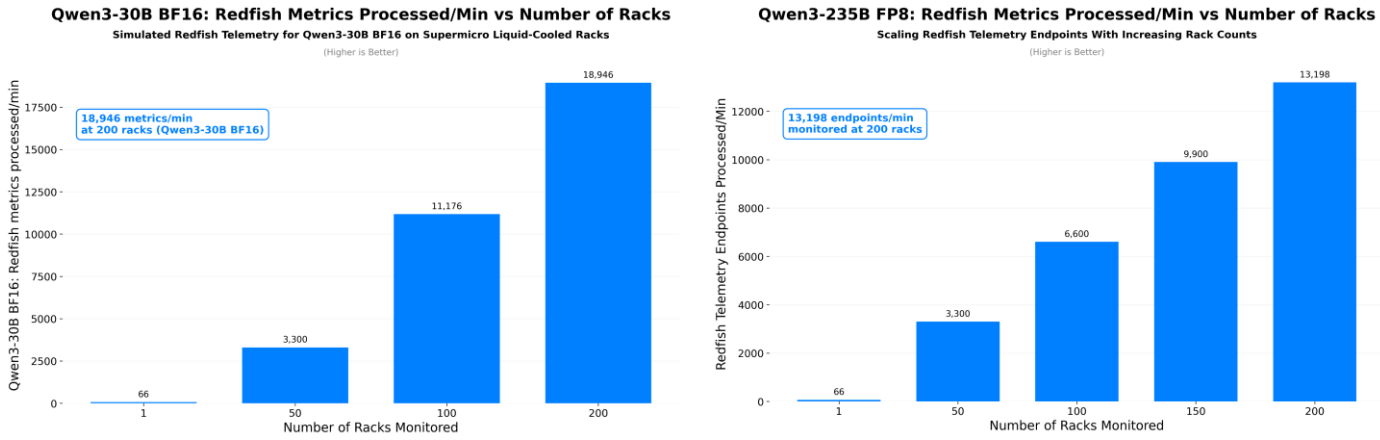

Figure 10 - Redfish Telemetry Scaling Across Rack Counts: (Left) Qwen3-30B BF16 and (Right) Qwen3-235B FP8

As telemetry load increases across larger rack counts, the system maintains stable ingestion through the AMD EPYC-driven data fusion layer, ensuring every agent receives real-time, high-resolution sensor context. With Redfish throughput scaling cleanly from single-rack to 200-rack environments, the next step is evaluating how large-scale reasoning workloads behave under the same system pressure. The following results measure LLM performance—both for Qwen3-30B BF16 and Qwen3-235B FP8—running on AMD Instinct MI325X GPUs as rack count, concurrency, and token complexity increase.

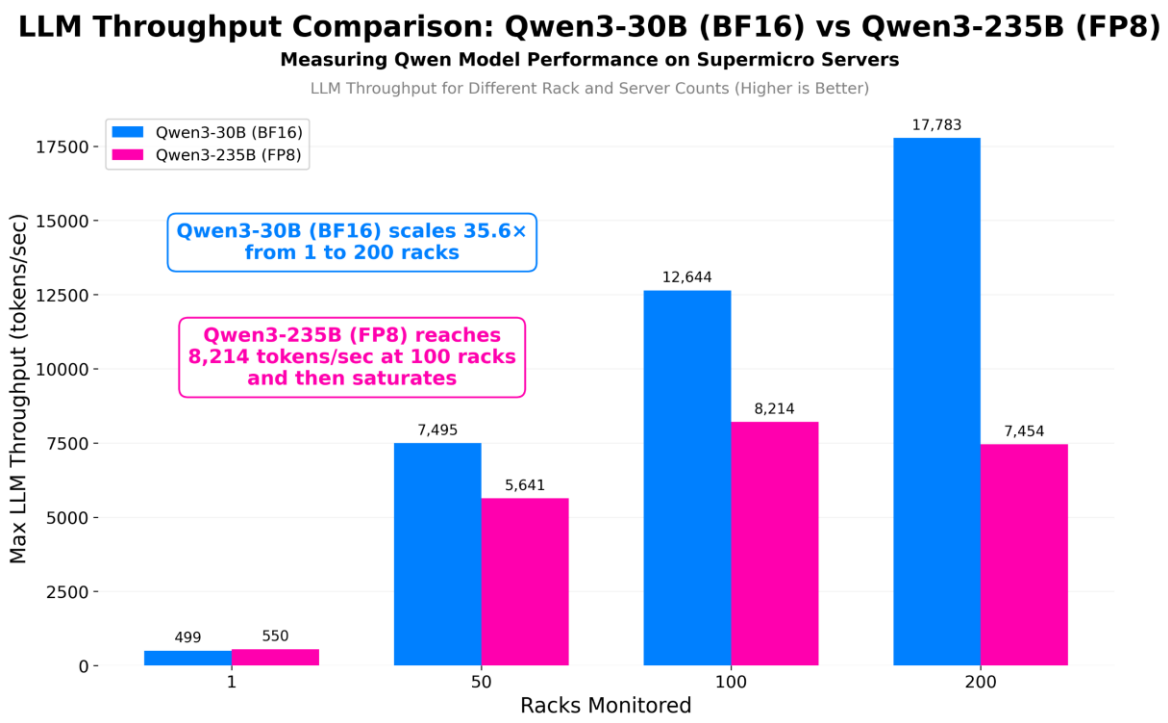

Figure 11 - Max LLM Throughput for Qwen3-30B BF16 and Qwen3-235B FP8 Across 1–200 Racks

Across all benchmarks, the integrated solution demonstrates stable, real-time, end-to-end autonomous operation under datacenter-scale load. The integrated system sustains linear Redfish telemetry ingestion up to 13,198 endpoints per minute while simultaneously maintaining 8,214 tokens/sec of multi-agent large-model reasoning across four replicas of Qwen3-235B FP8 on 8× MI325X GPUs. These results confirm that telemetry ingestion, data fusion, agent coordination, and large-scale inference operate as a unified, non-blocking control loop—validating the platform's ability to support autonomous, rack-scale liquid-cooling intelligence with deterministic latency and full-context execution.

Model Inference Throughput and Concurrency Scaling

To complement the end-to-end, solution-level performance results, additional model-level inference benchmarking was conducted to characterize raw throughput and concurrency behavior under controlled conditions. These tests isolate large-model execution on AMD Instinct MI325X GPUs using vLLM and ROCm, measuring sustained token throughput across a wide range of concurrent request levels and input/output token lengths. Both short-context (128 input / 128 output tokens) and long-context (2048 input / 2048 output tokens) configurations were evaluated for the Qwen3-30B BF16 and Qwen3-235B FP8 models to quantify scaling behavior under increasing parallel load.

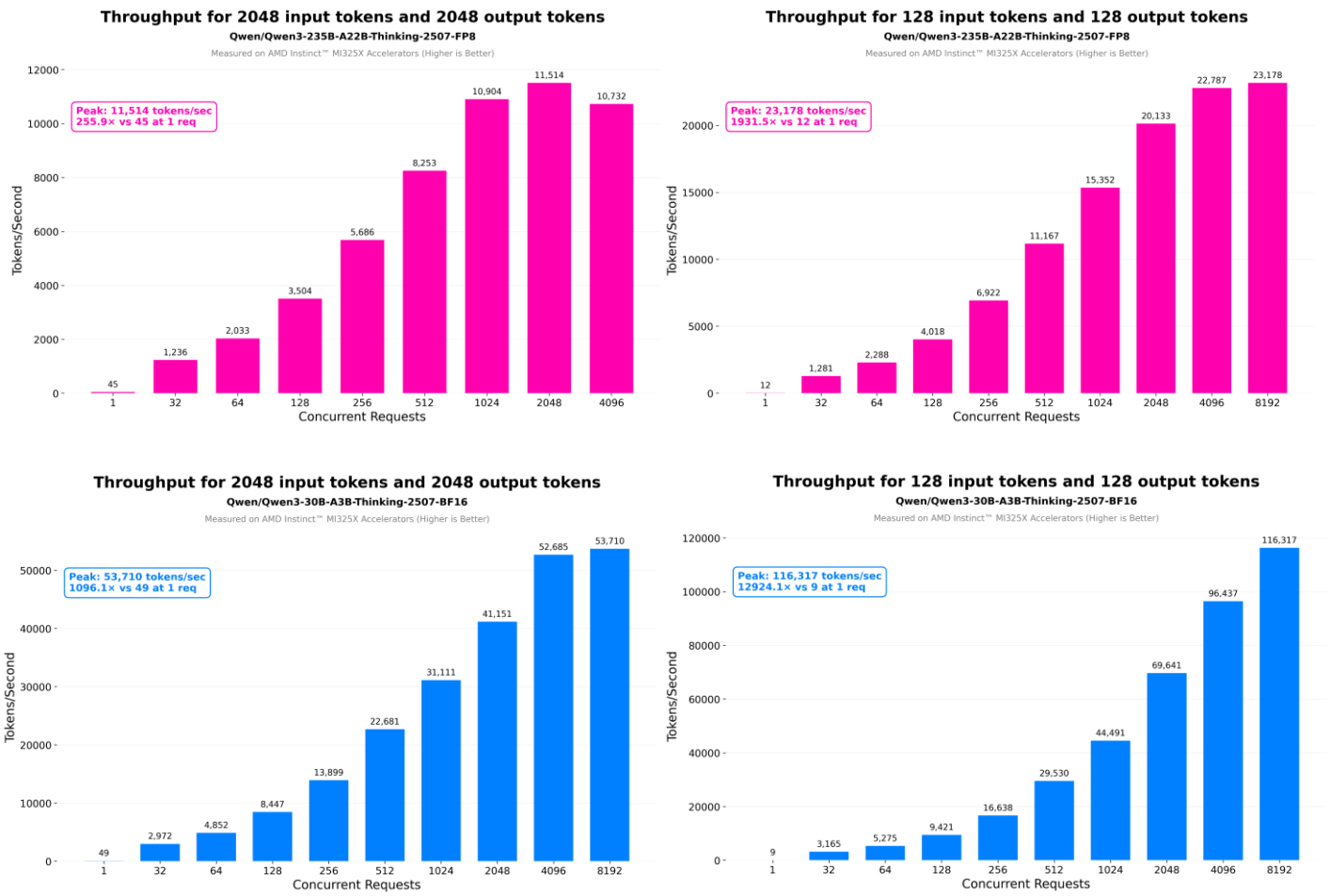

Figure 12 - Concurrency Scaling for Qwen3-235B FP8 and Qwen3-30B BF16 Across Long- and Short-Context Token Configurations

Isolated model-level benchmarking on AMD Instinct MI325X GPUs shows that both Qwen3-30B BF16 and Qwen3-235B FP8 scale predictably under increasing concurrency across short- and long-context workloads. Qwen3-30B BF16 achieves higher absolute throughput under high request parallelism, while Qwen3-235B FP8 exhibits superior stability for long-context reasoning enabled by full in-memory execution within the MI325X's 256 GB HBM3E. No throughput cliffs, context truncation, or concurrency instabilities were observed across the tested regimes, confirming that MI325X sustains both high-throughput and deep-reasoning workloads under dense multi-request conditions.

Multi-Agent Autonomous Cooling Platform Key Insights

- **Linear and Predictable Telemetry Scaling:**
  Redfish telemetry ingestion increases proportionally with rack count—from 1 to 200 racks—demonstrating stable, linear scaling across ingestion, fusion, and routing layers.
- **Validated Datacenter-Scale Throughput:**
  The system successfully processed 13,198 Redfish telemetry endpoints/min while simultaneously sustaining 8,214 tokens/sec of large-model reasoning across four replicas of Qwen3-235B FP8 on 8× MI325X GPUs.
- **High-Memory LLM Performance at Scale:**

The 256 GB HBM3E on each AMD Instinct MI325X enabled larger context windows, deeper per-token reasoning, and stable performance under concurrent multi-agent loads—without activation offloading or window truncation.

- **Superior Multi-Agent Convergence:**
  Qwen3-235B FP8 exhibited smoother convergence, fewer retries, and more consistent decision-making under telemetry pressure due to improved semantic depth and FP8 optimization on AMD MI325X GPUs.
- **Memory-Enabled Replica Density:**
  The large memory footprint allowed the platform to host four concurrent replicas of a 235B-parameter model per system, enabling multi-agent orchestration and long-horizon reasoning in dense multi-rack environments.

Model Inference Throughput and Concurrency Scaling Key Insights

- **Throughput vs. Reasoning Tradeoff:**
  Qwen3-30B BF16 consistently achieves higher peak tokens/sec under high concurrency, making it well-suited for throughput-optimized inference workloads, while Qwen3-235B FP8 prioritizes deeper reasoning stability under dense parallel execution.
- **Long-Context Stability at Scale:**
  Qwen3-235B FP8 maintains stable throughput across long-context (2048 input / 2048 output tokens) workloads without activation offloading or context truncation, enabled by full in-memory execution within the MI325X's 256 GB HBM3E.
- **Predictable Concurrency Scaling:**
  Both Qwen3-30B BF16 and Qwen3-235B FP8 scale smoothly as concurrent request levels increase, with no observable throughput cliffs or inference instability across short- and long-context regimes.
- **FP8 Efficiency for Large Models:**
  FP8 execution on AMD Instinct MI325X enables large-parameter models to sustain high utilization under concurrent load while preserving semantic depth and long-horizon reasoning continuity.
- **Concurrency Without Memory Saturation:**
  Even under thousands of concurrent requests, neither model exhibits memory-induced degradation, confirming that MI325X provides sufficient bandwidth and capacity for dense, multi-request inference pipelines.

These performance results demonstrate how the Supermicro A+ platform, powered by AMD EPYC processors for telemetry fusion and AMD Instinct MI325X GPUs for large-model inference, maintains consistent responsiveness even under extreme telemetry pressure. The system scaled to ingest up to 13,198 Redfish telemetry endpoints per minute while simultaneously sustaining 8,214 tokens/sec across four replicas of the Qwen3-235B FP8 model—validating the platform's ability to support autonomous, rack-scale liquid-cooling intelligence that must operate in real time, with both ingestion throughput and deep multi-agent reasoning.

Benefits

The shift to liquid cooling introduces a fundamental operational gap: hydraulic dynamics move faster than human operators can react. By bridging this gap with an autonomous, multi-agent control plane, the collaboration between Supermicro, AMD, and Metrum AI delivers a solution that transforms cooling from a source of risk into a competitive advantage.

The combination of Supermicro A+ infrastructure and the AMD Enterprise AI platform delivers four high-impact benefits that directly address the challenges of modern high-density datacenters:

- **From Reactive to Predictive Reliability:** Traditional monitoring reacts only *after* a failure occurs. In contrast, this solution utilizes AMD Instinct GPUs to analyze telemetry streams in real-time, detecting micro-fractures, pump degradation patterns, and coolant chemistry drift weeks before they trigger a failure. This shifts the operational model from reactive crisis management to proactive intervention, preventing catastrophic leaks and ensuring maximum uptime for critical AI workloads.
- **Precision Energy Efficiency:** Static cooling policies waste energy by over-cooling idle racks. By dynamically adjusting pump speeds and fan curves based on real-time CPU/GPU telemetry, the multi-agent platform ensures that energy is expended only where thermally necessary. AMD's high-performance-per-watt architecture further amplifies this efficiency, minimizing the power overhead of the control plane itself to improve overall Power Usage Effectiveness (PUE).
- **Deterministic Scalability:** A major challenge in liquid-cooled clusters is maintaining control visibility as rack counts grow. The modular design of Supermicro A+ servers, combined with ROCm's distributed-memory model, ensures that inference performance remains consistent at all scales. As validated in the performance benchmarks, the system maintains sub-second reasoning latency whether monitoring a single rack or a 200-rack cluster, providing a future-proof foundation for expansion.
- **Operational Continuity:** Hardware maintenance typically requires costly downtime. The solution addresses this by combining predictive intelligence with physical modularity. Agents identify exactly which component is degrading, and Supermicro's hot-swappable architecture (fans, PSUs, pump modules) allows operators to replace that specific component while the system remains online. This extends the infrastructure lifecycle and reduces the frequency of full-system refresh cycles.
- **High-Bandwidth Operational Intelligence:** The volume of data generated by liquid-cooling sensors often overwhelms standard control loops. AMD EPYC processors provide the massive I/O bandwidth needed to ingest these uninterrupted streams, while AMD Instinct GPUs rapidly process the multivariate data. This creates a "self-optimizing fabric" in which the cooling subsystem continuously learns from its environment, turning raw sensor noise into actionable operational intelligence.

Conclusion

The rapid transition to high-density liquid cooling presents a fundamental paradox: while essential for modern AI workloads, the hydraulic dynamics of these systems move faster than human operators or traditional tools can manage. The collaboration between Supermicro, AMD, and Metrum AI resolves this "complexity barrier" by transforming cooling from a passive maintenance burden into an active, intelligent control layer.

By integrating Supermicro's purpose-built A+ infrastructure with AMD's computational power, the solution delivers sub-second decision-making to ensure safety at scale. Powered by AMD EPYC™ processors for high-throughput data fusion and AMD Instinct MI325X GPUs for deep reasoning, the architecture has been proven to ingest 13,198 telemetry endpoints per minute and sustain 8,214 tokens/sec of multi-agent reasoning. Crucially, this capability is unlocked by the MI325X's 256 GB of HBM3E memory, which allows agents to maintain the full historical context needed to predict failures weeks in advance—a feat impossible on memory-constrained hardware.

The result is a data center that is adaptive, resilient, and self-optimizing. It supports the massive compute intensity of next-generation AI workloads while proactively managing the thermal and physical risks that come with them. This solution demonstrates that the ecosystem is now fully prepared to support autonomous platforms. With ROCm providing an open, high-performance foundation and Supermicro delivering liquid-cooled density, organizations can now deploy the intelligent infrastructure necessary to support the AI era.

We invite partners, developers, and innovators to bring advanced agentic workloads—monitoring, remediation, optimization, and complete closed-loop control—to life on Supermicro and ROCm. The infrastructure is ready. The models are ready.

For More Information

Email a_plus_server_taskforce@supermicro.com

Disclaimer — Benchmarks were conducted by Metrum AI on hardware and software using the configurations, datasets, and model servers described. Performance may vary based on model selection and version, framework and library revisions, drivers and firmware, precision settings, data preprocessing, and application complexity. Because these components evolve over time, observed performance may change as software and firmware mature. Results are provided for informational purposes only and do not constitute a guarantee of future performance. Users should conduct their own validation testing on production-representative workloads and configurations.

References

[1] Precedence Research, "Data Center Liquid Cooling Market Size 2025 to 2034," 2024. [Online]. Available: https://www.precedenceresearch.com/data-center-liquid-cooling-market. Accessed: Dec. 2025.

[2] International Data Corporation (IDC), "Data Center Liquid Cooling Market," industry analysis, summarized by Virtue Market Research. [Online]. Available: https://virtuemarketresearch.com/report/data-center-liquid-cooling-market. Accessed: Dec. 2025.

[3] Boyd Corporation, "Energy Consumption in Data Centers: Air versus Liquid Cooling," Boyd Corp. Blog, 2023. [Online]. Available: https://www.boydcorp.com/blog/energy-consumption-in-data-centers-air-versus-liquid-cooling.html. Accessed: Dec. 2025.

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

METRUM AI

Metrum AI delivers end-to-end artificial intelligence solutions designed to accelerate industry transformation. The product suite includes Metrum Insights, which benchmarks AI model and agent performance and accuracy; Metrum Infrastructure Agents, providing root cause analysis and automated remediation for datacenter infrastructure; and Metrum AI Agents Factory, a platform for building purpose-built AI agents tailored to specific industry needs. Metrum AI solutions drive impact across IT, telco, finance, insurance, retail, manufacturing, healthcare, and more.

Learn more at: www.metrum.ai

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