PRODUCT BRIEF

Supermicro's High-Density FlexTwin™ for Agentic AI Workloads

High-density design excels at agentic AI workloads

OCTOBER 2026

EIA FlexTwin and ORv3 FlexTwin

Table of Contents

Executive Summary 2
Highlights 2
Target Customer and Use Cases 3
Key Features 3
The Agentic AI Infrastructure Shift 4
Solution Overview / Differentiation 5
Large-Scale Agentic AI Environments 6
Conclusion 7
For More Information 7

EXECUTIVE SUMMARY

Global data centers are at a critical inflection point as cloud service providers, enterprises, and emerging Agentic AI deployments demand more general-purpose computing within constrained facility power and cooling envelopes. Unlike single-pass inference, Agentic AI operates as a multi-step system that repeatedly plans, retrieves data, invokes tools, executes code, verifies results, and applies governance. These CPU-intensive control and execution stages increase demand for dense, memory-rich, energy-efficient compute alongside accelerators.

The Supermicro FlexTwin™ multi-node architecture, paired with Intel® Xeon® 6 processors, is designed to alleviate these infrastructure bottlenecks. The Supermicro FlexTwin design delivers high compute density, memory bandwidth, and advanced liquid cooling in two rack standards: the EIA 19-inch (2U 4-node) for established data-center footprints, and the Open Compute Project ORv3, 21-inch (1OU 2-node) for denser, open-rack deployments. To address these demands, data center operators can leverage Intel® Xeon® 6900 Series P-core configurations for latency-sensitive control paths and performance-critical workflows, or deploy Intel Xeon 6+ E-core configurations, boasting up to 288 cores per socket and 576 cores per node, to maximize concurrent agent runtimes, tool-execution workers, and RAG pipelines per rack.

Built for maximum efficiency, FlexTwin scales to 46,080 cores in a single 44OU ORv3 rack or 55,296 cores in a 48U EIA rack, supported by integrated Direct Liquid Cooling (DLC) that captures up to 95% of heat from CPUs, memory, VRMs, and I/O, along with optional in-rack/in-row CDUs and ORv3 busbar power shelves. By maintaining tested CPU-to-GPU ratios (1:1 to 1.4:1) and providing front-accessible, hot-pluggable serviceability, FlexTwin ensures that context preparation, routing, retrieval, and governance keep pace with high-throughput accelerator clusters, empowering operators to scale always-on agent fleets, multi-tenant enterprise copilots, and high-density cloud-native workloads while delivering peak performance and a lower total cost of ownership (TCO) within strict facility limits.

EIA FlexTwin and ORv3 FlexTwin

HIGHLIGHTS

Choice of facility: One FlexTwin family, two rack standards. Deploy the same liquid-cooled multi-node architecture in EIA 19-inch environments (2 U 4-node) or OCP ORv3 21-inch environments (1 OU 2-node) without changing the operational model.

Match the CPU architecture to the task: Intel Xeon 6900 Series P-core configurations target performance-sensitive HPC and AI workloads, while Intel Xeon 6+ E-core configurations target high agent density, cloud-native services, and per-watt throughput.

Maximize cores for agents: Fit more independent dual-socket nodes and E-cores per rack while sustaining performance under strict power limits with Intel Xeon 6+ (up to 288 cores per socket, 576 cores per node).

Improve rack-scale efficiency: Integrated with direct liquid cooling, optional in-rack or in-row CDUs, and (on ORv3) power busbars with unified power shelves reduce distribution losses and cooling overhead.

Future-ready operations: Front-accessible, hot-pluggable nodes, storage, and I/O simplify cold-aisle service on both EIA and ORv3 deployments.

TARGET CUSTOMER AND USE CASES

The Supermicro FlexTwin system with Intel Xeon 6+ processors is designed for operators who need maximum concurrent CPU workload per rack: CSPs, enterprise AI platform teams, financial services, public sector, and research organizations running Agentic AI control planes alongside GPU or accelerator clusters.

Complementary P-core configurations (Intel Xeon 6900 Series)

Where latency-sensitive control flow or single-thread performance matters, the Supermicro FlexTwin supports Intel Xeon 6900 Series processors with P-cores. These SKUs remain available in the same multi-node, liquid-cooled family for mixed agent and traditional high-performance workloads.

Targeted P-core use cases include:

- Enterprise copilots and workflow agents
- RAG, knowledge agents, vector search, and document processing
- Code-generation agents that compile, test, and validate software
- Financial-services agents for regulatory review, fraud investigation, risk analysis, and controlled automation
- AI platform services for routing, memory, state, guardrails, and telemetry

Always-on agent and throughput-per-watt workloads (E-core optimized)

The Intel Xeon E-core-optimized FlexTwin configuration is ideal when maximum cores per rack are required and when better performance per dollar per watt is needed under tight facility power limits, helping reduce total cost of ownership.

Targeted E-core use cases include:

Efficiency / Density-sensitive Agentic AI workloads

- High-concurrency multi-agent runtimes. Large numbers of agents performing orchestration, state handling, retries, and lightweight tool tasks in parallel.
- Scale-out tool-execution workers. API calls, workflow steps, isolated containers, and short-lived sandbox jobs distributed across many cores.
- RAG ingestion and document-processing tiers. Parallel parsing, chunking, metadata extraction, indexing preparation, and data movement.
- Always-on enterprise agent services. Multi-tenant copilots, approval workflows, service agents, and background automation requiring throughput per watt.
- Observability, policy, and governance services. Telemetry collection, audit processing, identity checks, policy enforcement, and guardrail pipelines.
- Cloud-native Agentic AI platform services. Containerized orchestration, memory/state services, queues, gateways, and microservices that benefit from core density.

KEY FEATURES

Each FlexTwin node is an independent dual-socket server. The FlexTwin system has liquid-cooled cold plates for CPU only, with 61% heat capture, and advanced cold-plate configurations deliver heat-capture rates up to 98% by removing heat from the CPU, memory, VRM, and I/O.

- Dense form factors. EIA: 2U chassis with 4 dual-socket nodes for standard 19-inch racks. ORv3: 1 OU chassis with 2 dual-socket nodes for 21-inch Open Rack v3.
- Processor choice. Intel Xeon 6+ with up to 288 E-cores per socket and 576 cores per dual-socket node. Optional Intel Xeon 6900 Series with P-cores for latency-sensitive paths.
- Memory. Up to 24 DIMM slots per node; large DDR5 capacity for context assembly, caches, and multi-tenant agent state.
- Liquid cooling. DLC systems with in-rack and in-row CDU options; ORv3 designs add integrated manifolds and busbar power.
- Storage. Front hot-swap E1.S NVMe bays and internal M.2 options per node for local scratch, images, and agent artifacts.
- I/O. AIOM / OCP 3.0 networking plus available PCIe expansion on EIA 2U 4-node systems for fabric adapters and accelerators used by tool workers.
- Rack standards. EIA 19-inch enterprise racks and OCP ORv3 21-inch racks with integrated power shelves and blind-mate power and cooling.

THE AGENTIC AI INFRASTRUCTURE SHIFT

Traditional inference sends a request to a model and returns an answer. Agentic AI repeatedly plans, acts, observes, evaluates, and retries until a goal is completed. Each iteration can invoke databases, vector stores, APIs, code sandboxes, files, enterprise applications, and additional models.

Agentic Stage	Primary Infrastructure Work	FlexTwin Value
Plan and route	Orchestration, state, scheduling, retries, model and tool selection	High CPU concurrency and memory bandwidth
Retrieve and prepare	RAG, vector search, parsing, indexing, databases, and data movement	Large memory capacity and high-speed I/O
Act and execute	APIs, workflow engines, code execution, containers, and sandboxes	Dense cores for concurrent agents and workers
Evaluate and govern	Guardrails, verification, policy, identity, telemetry, and audit	Common x86 security and manageability

Representative Workflow Proof Points

- In a reported financial-services anomaly-detection workflow, CPU-side work consumed most pipeline time; web-search enrichment took about three times as long as LLM inference.
- In a 2,140-problem code-generation test, GPU code generation completed in 62.8 seconds while CPU sandbox execution consumed 64.1 seconds, making CPU execution the limiting factor in that configuration.
- The design implication is to optimize the full pipeline and provision enough CPU capacity for pre-work, tool-work, and post-inference work.

Deployment Guidance

Start with the workflow, then size the rack. Agentic AI infrastructure should be evaluated as a coordinated system rather than as isolated CPU and GPU components.

Step	Guidance
1. Characterize the agent loop	Map planning, model calls, retrieval, tool execution, document or code processing, verification, and governance.
2. Measure CPU-side demand	Capture agent density, p95 latency, wall-clock time, throughput, memory use, I/O, and accelerator idle time.
3. Select the node profile	Use P-core systems for performance-sensitive or compute-heavy work and E-core systems for concurrent scale-out services.
4. Balance CPU and GPU	Use 1:1 to 1.4:1 CPU-to-GPU as tested current guidance, not as a universal rule. Validate with the actual model, framework, and tool chain.
5. Scale repeatable blocks	Add FlexTwin nodes or racks as demand for orchestration, retrieval, execution, and governance grows.

Infrastructure Metrics for Agentic AI

- Task success rate: confirm the workflow completes correctly before optimizing speed.
- Agent density: identify how many concurrent agents each CPU resource can sustain before saturation.
- p95 latency and wall-clock time: measure both user experience and end-to-end completion time.
- Throughput and cost per completed task: evaluate fleet capacity and economic efficiency.

CPU-to-GPU Sizing Guidance

Intel testing cited in the companion white paper recommends a current CPU-to-GPU ratio of 1:1 to 1.4:1 for the tested Agentic AI and code-generation configurations, equivalent to approximately 86 to 120 CPU cores per GPU. The required ratio varies with model size, token throughput, hardware generation, workflow design, and CPU-side work. Profile the actual deployment before final sizing.

SOLUTION OVERVIEW / DIFFERENTIATION

FlexTwin is a flexible, liquid-cooled multi-node design delivered either as individual EIA systems or as a fully integrated ORv3 rack-scale solution.

Illustrative rack-scale density with Intel Xeon 6+ (up to 288 cores/socket)

Configuration	Systems/rack	Compute nodes	Processor cores*
44OU OCP ORv3 — 1 OU 2-node FlexTwin	Up to 40	Up to 80	Up to 46,080
48U EIA — 2U 4-node FlexTwin	Up to 24	Up to 96	Up to 55,296

**Theoretical maximum using dual-socket nodes at 288 cores per socket. Actual populated density depends on CDU placement, power budget, and facility constraints.*

The 21-inch ORv3 rack infrastructure integrates power shelves with blind-mate power and cooling, eliminating PSU/PDU complexity at the system level while enabling higher power density and more compute per rack. This reduces the number of power components, lowers cost, and can improve MTBF compared with conventional 19-inch power distribution. ORv3 racks include integrated power busbars and DLC manifolds that support both in-rack and in-row CDUs. Power is delivered through the rack so each 1OU server does not require its own power supply unit, freeing space for compute.

EIA 2U 4-node FlexTwin systems use shared, redundant Titanium-level power supplies in a standard 19-inch chassis. They fit existing enterprise and colocation halls that have not adopted Open Rack, while still providing front-serviceable nodes and the same DLC approach. Operators can standardize on FlexTwin architecture across mixed campuses: ORv3 where new high-density halls are built, EIA where 19-inch rows already exist.

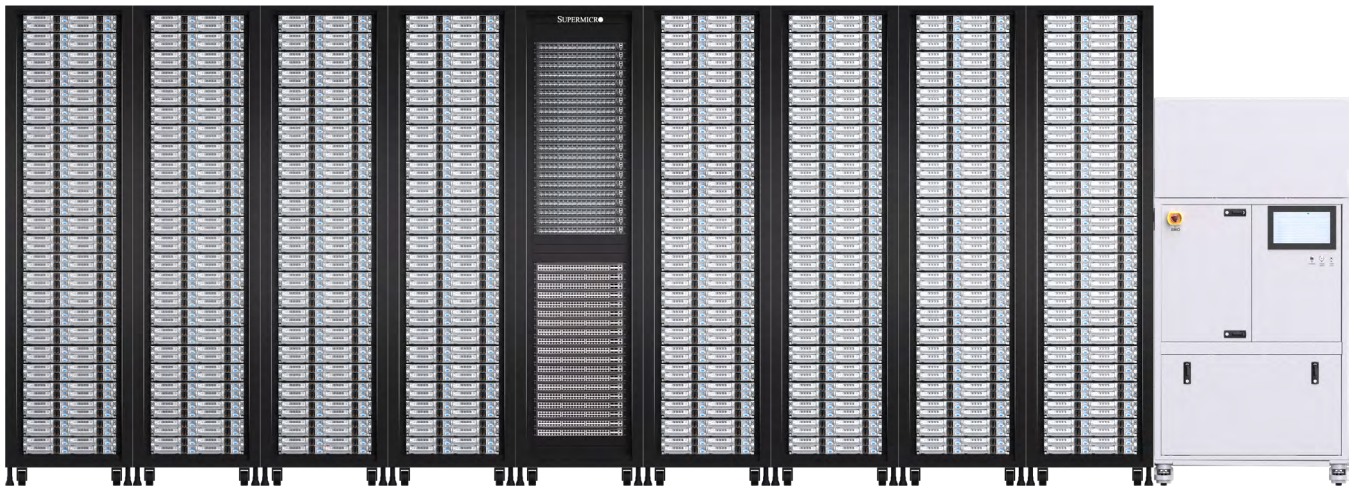
FlexTwin

EIA FlexTwin Rack vs ORv3 FlexTwin Rack Picture]

LARGE-SCALE AGENTIC AI ENVIRONMENTS

Many teams and tenants share agentic platforms. Individual agent graphs may use tens to hundreds of cores for orchestration, retrieval, and tool sandboxes; a single FlexTwin rack can present tens of thousands of cores. Clusters of FlexTwin racks provide the CPU plane that keeps accelerators fed: context preparation, routing, verification, and policy no longer stall GPU-bound inference.

Using an in-row CDU can increase the number of FlexTwin systems per square foot of floor space. Combined with high core counts in the Intel® Xeon® 6+ processors, this density is intended for always-on agent fleets, multi-tenant copilots, and scale-out document and tool pipelines rather than only traditional tightly coupled HPC jobs.



EIA FlexTwin Rack vs ORv3 FlexTwin Rack

CONCLUSION

Modern infrastructure teams must deliver performance, efficiency, density, and serviceability together. Agentic AI adds a further requirement: sufficient CPU capacity for the work surrounding model inference. Context preparation, orchestration, retrieval, tool execution, sandboxing, policy, security, and evaluation create large and variable CPU workloads. Under-provisioning these stages delays the full workflow and leaves accelerators waiting.

Supermicro FlexTwin provides a high-density, liquid-cooled Intel Xeon foundation for this balance of compute. With Intel Xeon 6 E-core SKUs for throughput per watt, optional P-core SKUs for latency-sensitive control, EIA and ORv3 rack choices, large memory, high-speed I/O, and front-access serviceability, FlexTwin enables customers to scale Agentic AI—and complementary high-performance workloads—within demanding facility constraints.

FOR MORE INFORMATION

Supermicro FlexTwin: https://www.supermicro.com/en/products/twin/dual-processor

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

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continually work to advance the design and manufacturing of semiconductors, helping to address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge, and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com. Visit: www.intel.com

©Super Micro Computer, Inc. Specifications subject to change without notice. All other brands and names are property of their respective owners. All logos, brand names, campaign statements and product images contained herein are copyrighted and may not be reprinted and/or reproduced, in whole or in part, without express written permission by Supermicro Corporate Marketing.

Printed in USA ♻ Please Recycle FlexTwin-Solution-Brief-2026-10-R03 **OCTOBER 2026**