

Cisco CNode-X with VAST AI OS



GPU Accelerated AI Services Where Data Lives

Cisco CNode-X with VAST AI OS extends Cisco Secure AI Factory with NVIDIA with a unified, GPU-accelerated platform for Cisco AI PODs. Built to run AI services where data lives, CNode-X accelerates the full path from data ingest to inference as a single, integrated system. It accelerates vector search and retrieval, real-time SQL analytics, scalable AI pipelines, and inference for long-context, multimodal, and multi-agent workloads. It also accelerates the complete RAG workflow by running embedding models, LLMs, vector databases, orchestration services, and serverless functions directly on the platform.

By bringing GPU acceleration into the VAST AI OS, CNode-X removes bottlenecks across ingestion, transformation, indexing, retrieval, analytics, and inference while simplifying the AI data stack. For Cisco AI PODs, that means lower latency, less data movement, faster retrieval, and higher GPU efficiency. The result is a data-centric architecture that keeps pipelines fed, keeps enterprise data continuously ready for use, and turns infrastructure into an engine for production AI. Instead of forcing customers to stitch together disconnected layers of storage, databases, and compute, CNode-X unifies data services, event-driven processing, and AI runtime services on one platform.

That lets organizations move faster, operate more simply, and scale continuous RAG, agentic AI, and multimodal inference with confidence. It gives Cisco customers a simpler path to production by collapsing complexity, accelerating outcomes, and turning data into an active advantage for AI. And it does it with the radical simplicity modern AI demands now.

Turn enterprise data into faster, simpler, scalable AI outcomes in production now.



Accelerated AI Infrastructure

Power AI pipelines with GPU-accelerated compute for demanding enterprise workloads



Built for AI Performance

Speed retrieval, analytics, and inference with data-adjacent accelerated compute.



Cisco-Validated AI Operations

Simplify deployment and scale with Cisco design, management, and support.

Type	
Processor	Dual AMD EPYC® 5th Gen processors (up to 192-cores total)
Memory	Up to 4TB DDR5 DIMM @5200MT/s
Data Connectivity	1x NVIDIA® ConnectX®-7 or BlueField®-3 DPU (200G or 400G)
Mgmt Connectivity	1x OCP 3.0 NIC slot and 1x 1GbE DC-SMC (BMC)
Boot Drive	2x 240GB or 960GB M.2 NVMe SSD
Management	Cisco Intersight via Integrated Management Controller (IMC)
VAST Certified Model	Cisco UCS C845A 4RU GPU accelerated server

Acceleration	Quantity
Model	NVIDIA® RTX PRO™ 6000 GPU Server Edition 8
Chip	5th Gen Tensor, 2nd Gen Transformer Engine, FP4
Memory	96GB GDDR7, 1.6 TB/s BW, 128MB L2 Cache

