VAST Data on Cisco UCS

Contents

Value statement

The VAST InsightEngine and VAST Data Platform, powered by Cisco UCS® and networking, provide a unified, scalable, and high-performance data infrastructure purpose-built for the AI era, accelerating insights and driving operational excellence.

Overview

The VAST InsightEngine and VAST Data Platform represent a revolutionary approach to managing and scaling data infrastructure for the Artificial Intelligence (AI) era. VAST AI OS is the overarching architecture including InsightEngine and serves as an AI-first operating layer that unifies storage, database, global namespace, and programmable compute/runtime to power agentic AI, including RAG (Retrieval-Augmented Generation) and event-driven pipelines. It includes recent additions such as InsightEngine for real-time RAG data preparation and AgentEngine for agent deployment and orchestration.

Underpinning VAST InsightEngine is the VAST Data Platform, the core software that provides foundational primitives: all-flash storage based on its innovative Disaggregated, Shared-Everything (DASE) architecture, a tabular database, a global namespace (DataSpace), and a serverless/event engine (DataEngine). This platform is positioned as the essential data substrate for AI factories.

The strategic partnership between Cisco® and VAST Data combines Cisco's high-performance networking and compute capabilities (such as Cisco UCS servers and Nexus® switches) with VAST Data's cutting-edge, AI-driven storage architecture. This joint solution empowers organizations to unlock the full potential of their data by providing AI infrastructure that is simple, scalable, and architected from the ground up to power deep learning and GPU-accelerated data centers and clouds. It eliminates storage tiering complexity, optimizes data-intensive workloads, and delivers enterprise-grade reliability and security, making it the new standard for enterprise AI infrastructure.

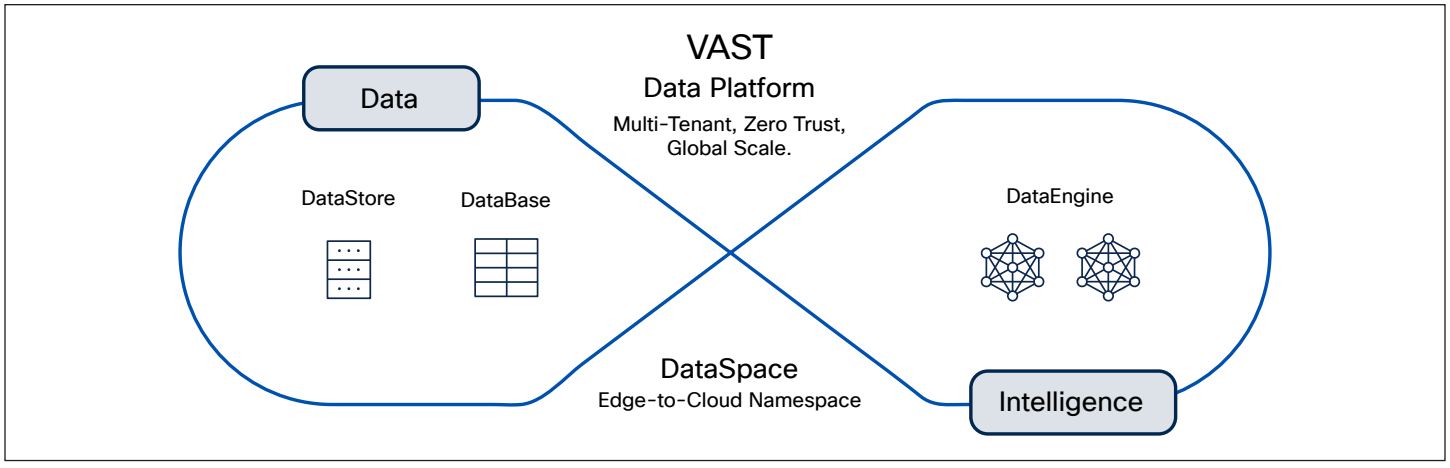

Figure 1. VAST Data Platform high-level solution

Features and benefits

Table 1. VAST InsightEngine and VAST Data Platform features and benefits

Feature	Benefit
Disaggregated, Shared-Everything (DASE) Architecture	Separates storage from compute resources for independent scaling, delivering ultra-low latency and massive bandwidth for real-time analytics and AI workloads.
Unified Global Namespace (DataSpace)	Consolidates workloads onto a single, unified platform, eliminating storage tiering complexity and simplifying management across edge, core, and cloud environments.
Multi-Protocol Data Plane (NFS, SMB, S3, NVMe/TCP)	Connects AI training/inference, MLOps, Business Intelligence (BI), and backup tools without data copying, ensuring seamless integration and data accessibility.
Flash-Optimized Erasure Coding and Similarity-Based Data Reduction	Achieves archive-like economics on all-flash storage, significantly reducing overall infrastructure costs while maintaining high performance.
VAST DataBase (Columnar, Tabular)	Provides low-latency analytics on operational data, supports feature stores, and enables vector/graph storage for real-time RAG data preparation.
VAST DataEngine (Programmable, Event-Driven Serverless)	Automates data pipelines (ingest → transform → index) and triggers jobs on data events, accelerating data processing for AI workflows.
InsightEngine (with NVIDIA NIM integration)	Keeps RAG corpora continuously current by embedding new data as it lands, supporting multimodal retrieval and accelerating AI model training.
AgentEngine	Manages and observes agent pipelines at scale, providing an application layer for AI OS that simplifies complex AI deployments.
Enterprise-Grade Reliability and Security	Offers advanced data protection features, including metadata triplication, immutable snapshots, and ransomware-resilient architecture, ensuring data integrity and rapid recovery.
Optimized for AI/ML and HPC Workloads	Provides a high-performance data pipeline that removes storage bottlenecks, ensuring GPU saturation and efficient processing of data-intensive applications.

Prominent feature

AI-First Operating Layer with Agentic AI Capabilities

VAST InsightEngine is designed from the ground up as an AI-first operating layer, unifying critical components like storage, database, and a programmable compute/runtime environment. This enables it to power agentic AI applications, including advanced RAG (Retrieval-Augmented Generation) and complex event-driven pipelines. Key innovations such as InsightEngine, which integrates with NVIDIA NIM microservices for real-time RAG data preparation, and AgentEngine, for streamlined agent deployment and orchestration, ensure that organizations can build, scale, and manage sophisticated AI workloads with unparalleled efficiency. This focus on agentic AI allows enterprises to move beyond traditional data management to truly intelligent, automated data-to-insight pipelines.

Revolutionary Disaggregated, Shared-Everything (DASE) Architecture

At the core of the VAST Data Platform is its unique Disaggregated, Shared-Everything (DASE) architecture. This innovative approach separates stateless front-end compute nodes (CNodes) from flash enclosures (DNodes), which provide persistence over NVMe over Fabrics (NVMe-oF). This disaggregation allows for independent scaling of compute and capacity resources, ensuring that any compute node can read or write any data without burdensome east-west chatter. The DASE architecture also features a global, transactional element store with Atomicity, Consistency, Isolation, Durability (ACID) semantics, eliminating per-node caches and metadata bottlenecks. Combined with flash-optimized erasure coding and Similarity-Based Data Reduction, DASE delivers breakthrough performance and efficiency, offering archive-like economics on all-flash storage.

Unified Data Platform for All Workloads

The VAST Data Platform consolidates all data types—files, objects, and tables—into a single, unified platform. Its global namespace (DataSpace) provides a consistent read/write namespace spanning edge, core, and cloud environments, complete with global read/write leases and cross-site Global Snapshot Clones for instant writable copies. This eliminates data silos and tiering complexity, allowing IT teams to manage all workloads, from traditional enterprise NAS consolidation to cutting-edge AI/ML pipelines, on one infrastructure. The platform's multi-protocol data plane (NFSv3/4.1, SMB 2.1/3.1, S3 object, Non-Volatile Memory Express [NVMe]/Transmission Control Protocol [TCP], GPUDirect/Remote Direct Memory Access [RDMA]) ensures seamless connectivity and high-throughput access for diverse applications.

Platform support

Table 2. Platform support

Product Family	Platforms Supported
Cisco UCS Servers	Cisco UCS C225M8 AMD Servers (EBox)
Cisco Networking	Cisco Nexus 9300 Series Switches (e.g., N9K-C9332D-GX2B, N9K-C9364D-GX2A)
Management	Cisco Intersight® SaaS
Networking Connectivity	Mellanox ConnectX-7 (CX7), NVIDIA BlueField-3 DPU

Licensing

VAST Data software is licensed per usable capacity and CPU cores, and all licenses are subscription-based with terms ranging from 12 to 60 months. There is no perpetual license offering. VAST Data software licenses are portable, allowing customers to transfer existing licenses to new equipment upon retirement or upgrade, ensuring continued use without additional licensing costs. Pricing is monthly but billed upfront for the full term. The same term must be selected across all related SKUs within a given quote or configuration to ensure contract alignment.

Table 3. VAST Data Software Licensing

License Type	Description	Application Details
VAST-SW-100TB	VAST Data software subscription, quoted per 100TB. Includes DataStore, DataBase, DataEngine, and DataSpace	Select based on total usable storage per node (e.g., 1x VAST-SW-100TB per 100TB usable capacity).
VAST-SW-1CPU-CORE	VAST CPU Core License, quoted per core beyond included entitlement	First 208 cores per 1PB of licensed storage are included. Additional cores are chargeable. Use VAST Sizing Tool to determine excess core licensing.
VAST-SW-1CPU-ACC	VAST CPU Core Accelerated License	Refer to Ordering guide for details.
VAST-PS-COPILOT	VAST CoPilot service for proactive monitoring, upgrades, and system health management	Charged per cluster (not per node). Only one CoPilot service is required per deployment, regardless of node count.
VAST-PS	VAST Installation Services for VAST software	Apply 1x VAST-PS per cluster for up to 12 EBox nodes. For clusters larger than 12 nodes, add more quantities accordingly.

Product sustainability

Information about Cisco's Environmental, Social and Governance (ESG) initiatives and performance is provided in Cisco's CSR and sustainability reporting.

Table 4. Cisco Environmental Sustainability Information

Sustainability Topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries and packaging	WEEE Compliance
	Information on product takeback and reuse program	Cisco Takeback and Reuse Program
	Sustainability Inquiries	Contact: csr_inquiries@cisco.com
Power	Typical Power	650-800W/Max Power 850-1000W (per UCS C225M8 EBox)
Material	Product packaging weight and materials	Contact: environment@cisco.com

Product specifications

Table 5. Cisco UCS C225M8 EBox Configuration

Component	Specification
Form Factor	1U Single Node Rack Server
CPU	AMD Epyc 9454P (48-core, 2.75GHz)
RAM	384GB (12x 32GB) DDR5-4800
Storage (Drive Layout Options)	2x 960 GB SCMs, 8x 15TB (122TB usable per node)
2x 1.9TB SCMs, 8x 30TB (240TB usable per node)*	
2x 1.9TB SCMs, 8x 60TB*	
Connectivity	2x CX-7
Boot	2x 1TB on-board M.2 boot SSDs
Management Network	2x 10 GbE Mgmt
PSU	Typical Power 650-800W/Max Power 850-1000W

*Please check with the Cisco Compute BU for the latest drive capacity qualifications.

Table 6. Cisco EBox Performance (per EBox)

Metric	Value
Read throughput	21GB/s
Max Write throughput*	10GB/s
Sustained Write throughput	2.6GB/s

*Maximum write performance is delivered automatically and can absorb write bursts of up to 30 minutes.

Table 7. Typical Data Reduction Rates (DRR)

Workload	DRR Rate
AI	2:01
Media	1.5:1
HPC	2:01
General file and object	3:01

Table 8. Cisco UCS Ebox Raw Capacity Scaling (Usable TB)

Number of EBoxes	Cisco UCS 122 (8x 15.3TB)	Cisco UCS 244 (8x 30TB)	Cisco UCS 430 (8x 61.4TB)
8x EBoxes (min)*	638 TB	1276 TB	2233 TB
9x EBoxes	754 TB	1507 TB	2638 TB
10x Eboxes	868 TB	1737 TB	3040 TB
11x Eboxes	982 TB	1,965 TB	3439 TB
12x EBoxes	1,096 TB	2,192 TB	3837 TB
13x EBoxes	1,209 TB	2,419 TB	4234 TB
14x EBoxes	1,322 TB	2,645 TB	4630 TB
15x EBoxes	1,435 TB	2,871 TB	5025 TB
Each additional EBox adds at least**	113 TB	225 TB	393 TB
25x EBoxes	2,558 TB	5,117 TB	8956 TB
50x EBoxes	5,355 TB	10,710 TB	18743 TB
100x EBoxes	10,868 TB	21,736 TB	38039 TB

* When you select fewer than 12 EBoxes for a cluster, this reduces the maximum write performance because the system sets the write buffer stripe width at initial installation, and it cannot be changed later.

- Design with 8–10 EBoxes is resulting in approximately 15% less write bandwidth.
- Design with 11 EBoxes is resulting in approximately 6% less write bandwidth.
- 12 EBoxes or more deliver full write bandwidth.

For optimal performance, start with 12 or more EBoxes. If you're unsure how this impacts your cluster design, please reach out to the Cisco or your VAST sales contacts for guidance.

**Capacity scaling is slightly better than linear.

System requirements

Table 9. System requirements

Feature	Description
Minimum Cluster Size	8 Cisco UCS C225M8 EBoxes (12 EBoxes recommended for base cluster size)
Networking	Minimum 2 Nexus Switches (9332D-GX2B or 9364D-GX2A as leaf switches)
Management Network	1G/10G management switches (e.g., 93108TC-FX3)
Cabling	Copper transceivers (QDD-2Q200-CU3M) for lengths up to 3m. Fiber cables (SFP-200G-SR4 or QDD-400G-SR8 with breakout) for longer lengths
Software Version	VAST Data 5.3 FW (generally available NOW)

Ordering information

The VAST InsightEngine and VAST Data Platform, including Cisco UCS hardware and VAST Data software, can be ordered via Cisco Commerce Workspace (CCW).

Ordering VAST Data Software Only (SolutionsPlus):

To order VAST Data software only, search for VAST-DATA-SPLUS in CCW. This ATO (Approved To Order) contains the VAST Data software licenses (VAST-SW-100TB, VAST-SW-1CPU-CORE, VAST-SW-1CPU-ACC) and services (VAST-PS-COPILOT, VAST-PS). Ensure to select the appropriate quantities and a consistent subscription term (12-60 months).

Ordering Cisco UCS Hardware and VAST Software (MLB):

For a single integrated bundle of Cisco UCS infrastructure and VAST Data software, use the Multi-Line Bill of Materials (MLB): VAST-DATA-MLB.

1. In CCW, search for and add VAST-DATA-MLB to your configuration.
2. Click “Select Options” to customize the MLB.
3. Select Cisco UCS C-Series hardware (e.g., UCSC-C225M8N-EBOX for new clusters, UCSC-C225M8N-EXP for expansion/RMA).
4. Configure PCIe NIC (Mellanox ConnectX-7 or NVIDIA BlueField-3 DPU) and country-specific selections for each EBox.
5. Add corresponding VAST Data software Product IDs (PIDs) (VAST-SW-100TB, VAST-SW-1CPU-CORE, VAST-PS-COPILOT, VAST-PS) and configure quantities and subscription terms.
6. Add networking components: Select Nexus Switches (e.g., N9K-C9332D-GX2B, N9K-C9364D-GX2A) and necessary transceivers/cables. Ensure to include SKU: DCN-AI for AI intent.
7. Add Cisco Intersight for cloud-based management: Select DC-MGT-IS-SAAS-AD and specify quantity (typically 1 per UCS server).

Table 10. Ordering information

Cisco PID	PID Description
VAST-SW-100TB	VAST Data Subscription (quoted per 100TB), 12–60 months term. Includes DataStore, DataBase, DataEngine, DataSpace
VAST-SW-1CPU-CORE	VAST Data CPU CORE (quoted per core), 12–60 months term
VAST-SW-1CPU-ACC	VAST Data CPU CORE Accelerated (quoted per core), 12–60 months term
VAST-PS-COPILOT	VAST Data CoPilot for system monitoring, upgrades, expansions, and overall system health management
VAST-PS	VAST Data Installation Services for VAST software (per cluster up to 12 EBoxes)

Table 11. Core Components of Cisco + VAST Data Solution Offering

MLB	Category Name	ATO	PID Description
VAST-DATA-MLB	VAST Data Software and Cisco Hardware MLB		
	CISCO UCS C225 M8	UCSC-C225M8N-EBOX	UCS C225 M8 1U Rack Server for VAST with 15.3 TB drives
		UCSC-C225M8N-EXP	UCSC-C225-M8N-1P VAST expansion/RMA Server, 15.3 TB drives
	VAST DATA SOFTWARE	VAST-DATA-SPLUS	VAST Data Solution Plus Offer
	CISCO INTERSIGHT	DC-MGT-SAAS	Cisco Intersight SaaS
		CNDL-DESELECT-D	Conditional Deselect
	NEXUS SWITCHES	N9K-C9332D-GX2B	Nexus 9300 Series, 32p 400G Switch
		N9K-C9364D-GX2A	Nexus 9300 Series, 64p 400G Switch

Deal Registration Process: All opportunities must be registered via VAST Data's partner portal (partners.vastdata.com). Use your cisco.com email address and associate yourself with "Cisco Systems Inc." VAST registered deals are valid for 180 days from approval if submitted through Cisco and can be extended upon demonstration of active engagement. VAST will review and respond to submissions within two business days.

Warranty information

Cisco hardware components included in the VAST Data solution carry standard Cisco hardware warranties. For VAST Data software, customers with a current (non-expired) VAST Data license subscription will have access to the latest software updates and patches. Specific warranty details for VAST Data software should be confirmed with VAST Data or Cisco sales representatives.

Cisco and partner services

Cisco and its certified partners offer comprehensive services to support the deployment, management, and optimization of the VAST InsightEngine and VAST Data Platform. Cisco Solution Support is available for VAST Data software, with Cisco Technical Assistance Center (TAC) providing Level 1 support and engaging VAST Data engineering for software-related issues. Customers also have the option to contact VAST Data directly for support. Services include planning and design, implementation, and ongoing optimization to ensure peak performance and accelerate business innovation. Renewals of VAST software originally purchased through Cisco's Global Price List (GPL) shall be transacted through Cisco's GPL.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. Learn more.

Learn more

Unlock the Full Potential of Your Data for the AI Era.

Are you ready to transform your data center into an AI-ready powerhouse? The Cisco + VAST Data solution provides the simplicity, scalability, and performance required to accelerate your AI, machine learning, and data-intensive workloads. Eliminate complexity, reduce costs, and gain actionable insights faster than ever before. For more information, please contact your Cisco sales representative or visit: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/ai-pods.html.

C78-5415801-00 09/25