Unlock AI Innovation with a High-Performance Data Platform

Cisco UCS Infrastructure with Hammerspace Data Platform

Contents

By combining Cisco Unified Computing System™ (Cisco UCS®) infrastructure with the Hammerspace Data Platform organizations can unify distributed unstructured data, automate data movement across on-premises, hybrid, and cloud environments, and accelerate AI pipelines. The result is a validated, high-performance data platform that improves time-to-insight, increases GPU efficiency, and enables AI initiatives to scale with confidence.

Executive summary

Enterprises are racing to operationalize AI, advanced analytics, and GPU-accelerated computing across on-premises, hybrid, and multicloud environments. While compute innovation has advanced rapidly, data infrastructure remains a critical bottleneck, with valuable unstructured data fragmented across silos, locked into legacy storage systems, and difficult to access consistently at scale.

The **Hammerspace Data Platform deployed on Cisco UCS infrastructure**, delivers a unified, high-performance data foundation designed for modern enterprise AI workloads. This architecture combines Hammerspace's standards-based parallel file system architecture, global namespace, and data orchestration services with the performance, scalability, and operational consistency of Cisco UCS infrastructure.

This joint validation establishes a blueprint for high-performance data architecture, integrating Hammerspace software with Cisco UCS infrastructure to bridge the gap between metadata orchestration and high-speed data services. By utilizing Cisco® servers as high-performance metadata and data services nodes, well as ultra-low-latency Hammerspace Tier 0 clients, this architecture creates a parallel global file system with seamless multiprotocol access.

The result is a repeatable, enterprise-ready solution that eliminates data silos, maximizes GPU saturation, and accelerates the end-to-end AI pipeline—from data ingestion and preparation to model training and inference.

Audience

This paper is designed for:

- Infrastructure architects designing the next generation of data centers
- AI/ML engineering Leads frustrated by GPU underutilization
- Data-center operations managers tasked with building scalable, high-performance environments for generative AI and large-scale analytics

Benefits

Eliminate data silos and accelerate time-to-insight with a global namespace that unifies access to data across on-premises data centers, edge sites, and public clouds

Enable multi-site and hybrid-cloud agility by automating the movement of data between sites and clouds without manual data copying

Accelerate AI pipelines with high-throughput, low-latency data access using standard file-and-object protocols – no proprietary clients

Deploy enterprise AI with confidence and speed on a validated Hammerspace data platform optimized for Cisco UCS infrastructure

Enterprise AI data challenges

Enterprise AI initiatives place unprecedented demands on traditional data architectures that were not designed for such performance requirements or the need for seamless access across distributed data stores. While GPUs and accelerated compute are now widely available, organizations consistently encounter data-related obstacles that slow innovation and increased operational complexity. These problems include fragmented data silos, inconsistent performance across the AI pipeline, data performance and auditability gaps, and operational complexity at scale.

Fragmented data silos

Training data, model artifacts, checkpoints, and inference datasets are often scattered across legacy NAS systems, object stores and data lakes, multiple sites, and cloud environments. This fragmentation forces manual data copying and duplicated storage infrastructure, which adds operational complexity and costs.

Inconsistent performance across the AI pipeline

AI workloads stress storage in different ways:

- High-throughput ingest and preprocessing
- Random access during training
- Frequent checkpointing of large model states
- Low-latency access for fine-tuning, Retrieval-Augmented Generation (RAG), inference, and agentic AI workloads

Traditional enterprise storage systems are not optimized for such requirements, forcing expensive GPUs to wait for data. Such underutilization and stalled jobs add more overhead to AI projects.

Data governance and auditability gaps

As AI initiatives scale, the ability to track data lineage, access, and usage becomes a significant compliance hurdle. Traditional storage silos often lack the granular metadata required to satisfy modern regulatory requirements, leaving organizations vulnerable to data leakage or unauthorized access. Without a unified policy engine, checking to ensure that sensitive training sets are properly isolated, tracked, and audited across hybrid-cloud environments is manual and error-prone, creating significant enterprise risk.

Operational complexity at scale

As environments grow:

- Storage solutions requiring proprietary client software increase friction and operational headaches.
- Manual processes to feed siloed data into compute clusters add risk and increase operational complexity.
- Bridging workloads across multi-site and hybrid-cloud using manual or scripted processes is even more error-prone, with added compliance and governance risk.

IT teams are left managing storage instead of enabling innovation.

Cisco and Hammerspace architecture

Cisco and Hammerspace address these challenges with a modern **software-defined data architecture** that separates data intelligence from the underlying storage hardware.

By combining **Cisco UCS infrastructure** with the **Hammerspace Data Platform**, organizations can build a flexible data platform for enterprise AI and accelerated workloads.

As part of this architecture depicted below, Hammerspace provides standards-based, parallel data access using industry-standard file-and-object protocols (Network File System [NFS], Server Message Block [SMB], and S3), with integrated data management capabilities that automate data protection, tiering, and data movement across sites and storage systems. Cisco provides a highly scalable, high-performance infrastructure foundation for modern data-intensive workloads, with unified management, integrated networking, and exceptional efficiency that simplifies operations and accelerates deployment.

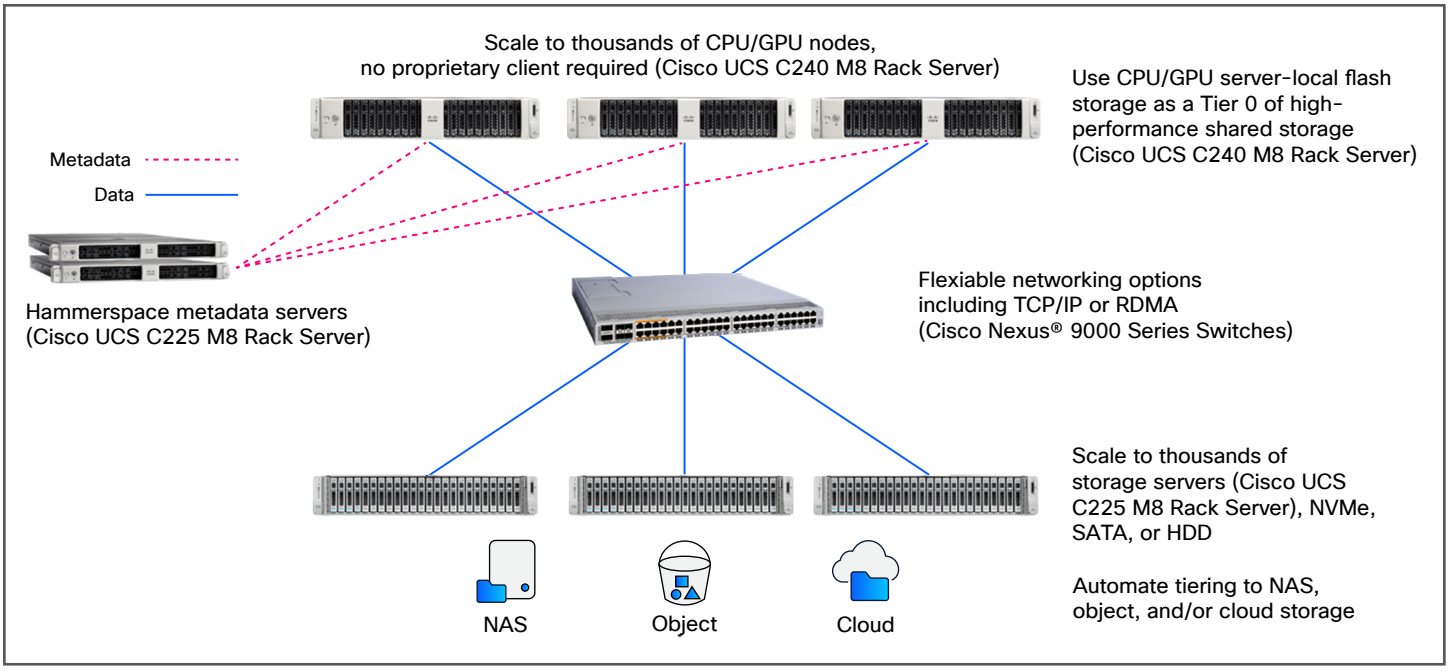

Figure 1. Cisco UCS C-Series rack servers with Hammerspace Data Platform solution architecture

The architecture includes several key layers that work together to deliver a high-performance data platform.

The storage engine: Hammerspace

Hammerspace is a software-defined data orchestration and storage solution that provides unified file access through a high-performance parallel global file system that can span different storage types from any vendor, as well as across geographic locations, public and private clouds, and cloud regions.

The **Hammerspace Data Platform** is designed to solve enterprise AI data challenges by unifying data access, performance, and orchestration into a single, software-defined platform.

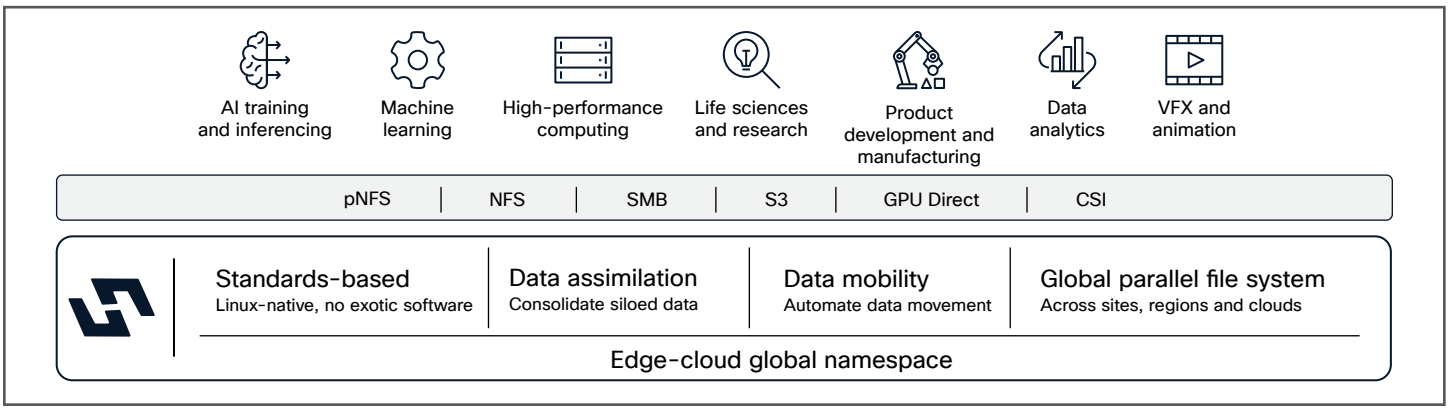

Figure 2. Logical view of the Hammerspace Data Platform

Hammerspace architecture attributes

High-throughput, low-latency performance for HPC and AI To process large unstructured data sets, maximize GPU/CPU utilization, and accelerate pipelines and workflows	**Linear scalability to thousands of nodes** To handle massive unstructured data volumes without a performance drop-off	**Standards-based connectivity – no proprietary clients** To simplify deployment and integration into existing pipelines and workflows
Separate metadata and data paths To enable metadata-driven policies and control with the speed of a direct data path between client and storage	**Open architecture – use any file or object storage** To enable the use of existing storage infrastructure plus flexibility when building new storage	**Span multiple sites and clouds with a global file system** To enable hybrid-cloud agility and multi-site collaboration at scale

Hammerspace node types

Hammerspace software is deployed with a scale-out architecture in a cluster for each site, comprising two node types that work together as a single system:

- **Anvil metadata services nodes,** which house the metadata control plane and drive the intelligence of the system.
  - No file I/O passes through the Anvil nodes
  - Anvil nodes are typically deployed as High-Availability (HA) pairs in production.
- **DSX nodes (data services nodes),** which handle all I/O operations, replication, data movement, etc., and are designed to scale out when needed to accommodate any level of performance requirements.

The storage foundation: Cisco UCS C225 M8 Rack Server

For validation, both Anvil and DSX nodes were deployed on Cisco UCS C225 M8 Rack Servers. However, as the Hammerspace platform is software-defined and hardware-agnostic, these services can run on other suitable servers within the Cisco UCS portfolio based on workload and performance requirements.

The **Cisco UCS C225 M8 Rack Server** is a high-density, single-socket rack server that delivers industry-leading performance and efficiency for your AI workloads.

- **Unprecedented processing power:** By utilizing 5th Gen AMD EPYC processors with up to 192 cores, these servers help to ensure that your storage software never starves for CPU cycles.
- **Next-Gen I/O and memory:** With the introduction of PCIe Gen 5 for high-speed I/O and a DDR5 memory bus, data moves between the CPU and the network without internal contention.
- **Storage density:** Each storage node contains high-performance drives, providing a massive, low-latency pool of ~840 TB raw capacity in a single global namespace.
- **High-speed connectivity:** Each node in your storage cluster features high-speed network adapters that provide the ultra-low latency and extreme throughput needed for AI model training and inference.

The AI compute: Cisco UCS GPU-accelerated infrastructure

Enterprise AI workloads require massive compute capability combined with extremely high-throughput data access. Cisco UCS provides a broad portfolio of GPU-accelerated systems designed to support the full AI lifecycle.

Cisco UCS GPU platforms support a wide range of configurations, from PCIe-based GPU servers for scalable AI workloads to high-density systems based on NVIDIA HGX, NVIDIA MGX, and AMD OAM accelerator architectures.

When combined with Hammerspace's Tier 0 architecture, Hammerspace can utilize server-local NVMe on the Cisco UCS GPU nodes as a "Tier 0" storage layer. This places the most frequently accessed data ("hot data") as close to the GPU as possible, effectively eliminating network hops for active training sets.

The unified management: Cisco Intersight

Managing a high-performance AI cluster should not require a dozen different tools. Cisco Intersight® is a lifecycle management platform that unifies your experience across the entire Cisco Unified Computing System.

- **One consolidated dashboard:** Whether your infrastructure resides in the enterprise data center, at the edge, or in a remote site, Intersight gives you a single view of your real-time status and interdependencies.
- **Intersight Standalone Mode (ISM):** Your Cisco UCS servers are managed in standalone mode, providing centralized, cloud-powered management and strict policy enforcement.
- **Automation at scale:** You can manage your entire system as a single logical entity through an intuitive GUI, or you can automate complex deployments and configurations using a robust API. This allows you to deploy a secure, multitenant AI data cloud in minutes rather than weeks.

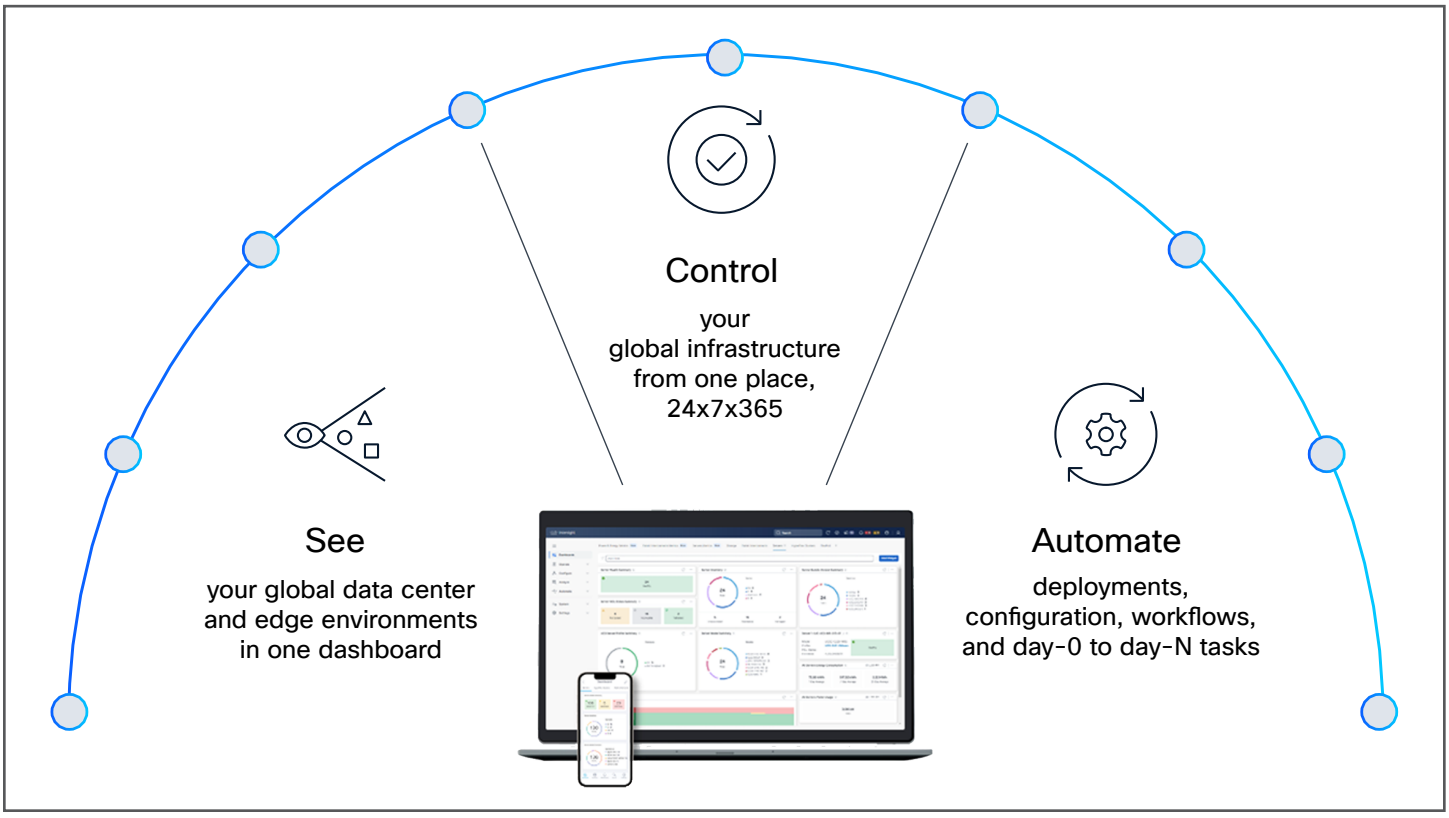

Figure 3. Cisco Intersight. IT operations – simplified

The networking fabric: Cisco Nexus 9000 Series Switches

High-performance AI workloads require a network fabric capable of delivering high throughput and low latency. Cisco Nexus 9000 Series switches provide the high-bandwidth Ethernet infrastructure needed to connect compute clusters, storage systems, and AI data pipelines.

The latest Nexus platforms support Ethernet speeds up to 800 GbE, enabling organizations to build scalable fabrics capable of supporting large GPU clusters and data-intensive workloads. These switches are designed to deliver high performance, energy efficiency, and operational scalability for modern AI and machine learning infrastructure.

Architectural synergy: the technical intersection of Cisco UCS and Hammerspace

The integration of the Hammerspace Data Platform with the Cisco Unified Computing System (Cisco UCS) represents more than a simple hardware-software validation; it is a symbiotic architectural fit designed to solve the "data gravity" problem in high-performance computing and AI/ML pipelines. By marrying Hammerspace's metadata-driven orchestration with Cisco's fabric-centric infrastructure, organizations can decouple data from underlying storage silos, creating a high-performance, globally accessible namespace.

1. Unified control and data-plane orchestration

The validation utilized Cisco UCS C-Series rack servers to host both the Hammerspace metadata service (Anvil) and the data routing (DSX) nodes.

- **Metadata efficiency:** Hammerspace offloads the metadata path from the data path. By running the Anvil metadata service on Cisco UCS nodes equipped with high-frequency processors and low-latency NVMe, the system can handle billions of files with sub-millisecond response times.
- **Cisco Intersight integration:** The Hammerspace stack is complemented with Cisco Intersight. This allows for "cloud-like" operations in on-premises data centers, enabling automated deployment, proactive monitoring, and lifecycle management of the storage cluster through a single API-driven dashboard.

2. High-velocity data pipelines through Cisco VICs and Cisco Nexus

A critical component of this technical solution is the Cisco Virtual Interface Card (VIC). Hammerspace thrives on high-bandwidth, low-latency networking to maintain its parallel file system performance, and the VIC allows deployment of a high-performance low-latency fabric from Intersight.

- **Linear scalability:** Testing proved that as Cisco UCS nodes are added to the Hammerspace cluster, performance scales linearly. This is facilitated by the Cisco Nexus 9000 Series Switches, which provide the high-radix, non-blocking fabric necessary to support "line-rate" data ingestion.
- **Parallel NFS (pNFS) optimization:** Hammerspace utilizes pNFS to allow clients to communicate directly with storage nodes. Cisco's Nexus fabric helps to ensure that these multi-path data streams are load-balanced and optimized, preventing the network bottlenecks typical of legacy NAS architectures.
- **Fabric observability with Cisco Nexus Dashboard:** By integrating the fabric with the Cisco Nexus Dashboard, IT teams gain comprehensive, single-pane-of-glass visibility and automation across the entire network. This simplifies the management of the high-speed data paths required for Hammerspace, ensuring that performance metrics and network health are continuously optimized to meet the rigorous demands of AI training and inference workloads.

3. Tier 0 performance: harnessing local NVMe and GPUs

The validation specifically targeted the most demanding AI/ML workloads by utilizing Cisco UCS GPU-ready servers as high-performance clients.

- **Hammerspace Tier 0 architecture:** To satisfy the extreme IOPS requirements of GPU training, Hammerspace can utilize server-local NVMe on Cisco UCS nodes as a Tier 0 storage layer. This places the most frequently accessed data ("hot data") as close to the GPU as possible, effectively eliminating network hops for active training sets.
- **Automated data placement:** Hammerspace's objective-based policies automatically move data from massive secondary storage (S3 or legacy NAS) onto the Cisco local NVMe tier when a job starts and migrates it back when the job completes. This ensures that expensive GPU resources are never "starved" for data.

Conclusion: Enabling scalable, data-centric AI infrastructure

As enterprises scale their AI initiatives, the limitations of traditional data architectures become increasingly evident. Siloed storage, inconsistent performance, and operational complexity continue to hinder the ability to fully utilize modern GPU-accelerated environments.

By combining the strengths of **Cisco UCS infrastructure** with the **Hammerspace Data Platform**, organizations can adopt a more flexible, software-defined approach to data management. This architecture brings data closer to compute, streamlines access through a global namespace, and automates data placement across environments – helping eliminate bottlenecks that slow AI workflows.

The result is a high-performance, scalable data platform that supports the full AI lifecycle while improving infrastructure efficiency and simplifying operations. With enterprise-ready design, this solution provides a strong foundation for organizations looking to operationalize AI at scale and accelerate time-to-value from their data.

Learn more

- Cisco UCS C-Series Rack Servers
- Hammerspace

C11-6039901-00 04/26