# VAST Data's New AI Updates & Partnerships

STEVE MCDOWELL, CHIEF ANALYST
10/5/24

## CONTEXT

The storage world is quickly bifurcating in its approach to AI. Enterprise storage vendors sit on one side, relying on traditional all-flash storage arrays to serve data to modest-size AI infrastructure. These vendors often need third-party solutions to meet the scalability needs of big AI training clusters. On the other side sits a new breed of storage provider who focus less on storage and more on enabling an end-to-end data infrastructure.

VAST Data, Weka, and IBM, with its IBM Storage Scale, all deliver a highly scalable and performant data infrastructure designed for big data challenges like those found in AI workflows. Building on many of the ideas in HPC parallel file systems, these solutions are designed around disaggregated, software-first architectures providing a global namespace and scalable performance, operating on-prem or in the cloud.

The most forwardly aggressive of this new breed is VAST Data, which was very early in supplementing its scalable storage technology with integrated data manipulation tools. The VAST Data Platform has long had a fully integrated, full-featured database, along with AI-targeted data manipulation capabilities. Its vision is paying off, with the VAST Data Platform finding success in AI-targeted cloud services like Lambda and Coreweave.

VAST made a wide-ranging set of announcements that extend those features to directly address the needs of enterprise AI. Beyond the new features, VAST also highlighted new strategic relationships and a new user community that help bring VAST technology into the enterprise.

## NVIDIA GPU & NIM INTEGRATION

One of VAST's cornerstone announcements is a deep integration between its platform and NVIDIA's enterprise AI tools. VAST now allows NVIDIA GPUs to act as a NIM controller node (NIM is part of NVIDIA's AI Enterprise software suite)

in VAST's all-flash storage system. This enables GPUs to process data directly within the data platform, eliminating the need for data migration to external servers. Allowing the GPU to access stored data directly enhances real-time AI inference, providing a significant performance boost.

In addition to the GPU integration, VAST also embedded NVIDIA's NIM microservices software directly into its platform. NIM simplifies the deployment of AI models by providing pre-configured, containerized solutions optimized for Gen AI and large language models.

This integration allows enterprises to rapidly deploy custom and pre-trained AI models across various environments, including cloud, on-prem data centers, and edge computing environments.

## KEY ASPECTS OF NVIDIA NIM INTEGRATION

1. **NVIDIA GPUs as Controller Nodes**:
    - NVIDIA GPUs are integrated directly into the VAST Data Platform, allowing them to act as controller nodes within VAST's all-flash storage system.
    - This enables the GPUs to process data directly on the VAST platform without data migration to external GPU servers, reducing latency and improving overall efficiency.
    - The ability to perform in-situ data processing means AI models can access and analyze data in real-time, which is crucial for AI inference tasks.
2. **NIM Microservices for AI Model Deployment**:
    - VAST Data has embedded NVIDIA NIM microservices into its platform, allowing for the streamlined deployment of AI models.
    - These microservices provide containerized, pre-configured AI solutions optimized for both Generative AI and large language models.
    - The integration simplifies the deployment of custom and pre-trained AI models across different environments, such as on-premises data centers, cloud, or edge computing environments.
3. **Real-Time AI Inference**:

- By allowing NVIDIA GPUs to process data directly within the VAST Data Platform, AI inference can be conducted in real-time.
- This reduces the time it takes to generate insights from AI models, particularly for data-heavy tasks like model training, inference, and real-time data analysis.

4. **Optimized Data Workflow**:
   - Integrating NVIDIA NIM microservices allows the platform to handle the entire AI data pipeline, from data ingestion to model deployment, without needing separate infrastructure for vector databases or other AI-specific tools.
   - This reduces the complexities and overhead of traditional AI workflows, where data often needs to be copied across multiple environments and tools.

## BENEFITS OF THE INTEGRATION

- **Reduced Latency**: Eliminating data movement between storage and compute layers significantly reduces latency, allowing for faster AI inference and training cycles.
- **Seamless AI Model Deployment**: With NIM microservices embedded, enterprises can quickly deploy AI models without complex configuration, speeding up AI development and scaling.
- **Scalability**: NVIDIA GPU integration provides scalable AI processing, allowing the platform to handle large volumes of data and support real-time AI workflows across enterprise environments.

# NEW VAST INSIGHTENGINE

While its tighter integration with NVIDIA NIM enables faster AI workflows, VAST's new InsightEngine brings efficiency to AI data handling. The VAST InsightEngine leverages the power of NVIDIA GPUs and NIM to provide real-time AI data processing.

The core new features of the VAST InsightEngine include:

### 1. NVIDIA GPU Integration as Controller Nodes

- **Direct Data Processing on Storage**: InsightEngine allows NVIDIA GPUs to function as controller nodes, enabling the GPUs to process data stored within the VAST storage array directly. This eliminates the need to move data to external GPU servers for processing, reduces latency, and improves efficiency.
- **Real-Time AI Inference**: The NVIDIA integration accelerates real-time AI inference by enabling AI models to be run on data immediately as it is ingested, providing faster insights and quicker decision-making.

**2. NIM Microservices Integration**

- **AI Model Deployment**: NVIDIA's NIM microservices are now natively embedded within the VAST InsightEngine, allowing for the streamlined deployment of AI models, including both custom and pre-trained models, across multiple infrastructures such as data centers, clouds, and workstations.
- **Containerized AI Models**: NIM provides AI models as optimized containers, simplifying and speeding up AI deployment for enterprises.

**3. Real-Time Vector and Graph Embeddings**

- **Semantic Understanding of Incoming Data**: InsightEngine uses NVIDIA-powered models to generate vector and graph embeddings in real-time as data is ingested. This enables the immediate transformation of raw data into a format ready for advanced AI retrieval and inference.
- **Enhanced Retrieval-Augmented Generation (RAG)**: These vector and graph embeddings are crucial for RAG, where AI systems use proprietary data to inform LLM queries.

**4. Integrated Data Engine for Triggering AI Processes**

- **Automatic AI Process Triggering**: The VAST DataEngine can trigger AI processes (like vector embedding) when new data is written to the system. This automation ensures that data is always AI-ready in real-time, without manual intervention or delays.
- **Instant Availability for Retrieval**: As soon as data is ingested, InsightEngine prepares it for immediate AI retrieval and inference operations, enabling continuous, real-time analytics.

**5. Massive Scale and Storage for AI Workloads**

- **Exabyte-Level Data Storage**: VAST InsightEngine supports the storage of exabytes of both structured and unstructured data, accommodating large-scale enterprise datasets.
- **Trillions of Embeddings**: The system can store and manage trillions of vector and graph embeddings, supporting massive knowledge graphs and vector spaces for AI processing.
- **Real-Time Similarity Search**: InsightEngine enables real-time similarity searches across these large-scale vector spaces and knowledge graphs, facilitating advanced AI analysis and decision-making.

**6. Unified AI and Data Infrastructure**

- **No Need for Separate Data Lakes**: VAST's architecture eliminates the need for external data lakes or third-party SaaS platforms, integrating AI-ready data processing within the same storage system.
- **Secure Data Handling**: The InsightEngine ensures data consistency and security by syncing file system or object storage updates with the vector database and indices, providing robust data access management across multi-tenant environments.

**7. AI-Enabled Decision-Making Across Data Types**

- **Supports Structured and Unstructured Data**: InsightEngine processes both structured (like databases) and unstructured (like files and streaming data) data in real time, making it a versatile solution for various AI workloads.
- **Enterprise-Ready AI Infrastructure**: The platform is designed to be a universal AI infrastructure, offering a unified view of all enterprise data for advanced AI-enabled decision-making.

InsightEngine makes it easier for companies to scale AI initiatives while maintaining real-time performance and data security by removing bottlenecks in data processing and model deployment. Its real-time processing capabilities, seamless NVIDIA integration, and ability to handle large-scale AI workloads make it a compelling solution.

## STRATEGIC PARTNERSHIPS WITH EQUINIX & CISCO

In a move that will help expand further into the enterprise market, VAST Data also announced new strategic partnerships with Cisco and Equinix. These

partnerships align with VAST's goal of streamlining AI adoption by integrating its data platform with industry-leading infrastructure and services.

## CISCO PARTNERSHIP

VAST partnered with Cisco to bundle its software with Cisco's UCS servers. This move serves several purposes:

- **Integrated Hardware and Software Solution**: By offering VAST's software with Cisco UCS servers, customers can deploy a turnkey solution that combines VAST's all-flash storage and AI-optimized software stack with Cisco's computing infrastructure.
- **AI Workload Optimization**: Cisco UCS servers are optimized for data-heavy workloads, which makes them an ideal match for VAST's AI data processing capabilities.
- **Simplified Deployment**: Bundling VAST's software with UCS servers simplifies customers' procurement and deployment process, as they can acquire a ready-made infrastructure stack from two trusted vendors.

## EQUINIX PARTNERSHIP

VAST has also partnered with Equinix, leveraging the company's IBX co-location sites to expand its market presence and accessibility:

- **Global Infrastructure Access**: By deploying VAST's platform in Equinix IBX co-location sites, VAST's solutions become more accessible to enterprises worldwide. Equinix's global network of data centers allows customers to deploy VAST's AI infrastructure closer to their data, reducing latency and improving performance.
- **Hybrid Cloud and Co-Location**: Equinix's infrastructure is widely used for hybrid cloud and edge computing environments. By partnering with Equinix, VAST can target enterprises looking for a seamless blend of on-premises and cloud-based AI deployments.
- **Cross-Connect with Ecosystem Partners**: Equinix's co-location sites are hubs for many technology partners, including cloud providers, enterprises, and network service providers. VAST's presence in these locations enables easy cross-connection with other AI technologies, enhancing the overall ecosystem and enabling smoother integrations with other platforms and services.

## ANALYSIS

VAST Data's integrated data storage and management approach optimizes every stage of the AI pipeline, from data ingestion and preprocessing to model training, evaluation, and deployment. Its most recent announcement also includes RAG and related enterprise AI workflows, ensuring faster and more reliable AI development cycles.

The capabilities of the VAST Data Platform are significantly differentiated from competing solutions. Organizations using the integrated NVIDIA software services and VAST data manipulation capabilities will experience new, likely unparalleled, levels of efficiency. It's a compelling story, one that recently led storage industry stalwart NetApp to announce an architecture built on many of these same principles (though without a corresponding product announcement). VAST should see this as a strong validation of its approach.

The challenge for VAST, and others following the same path, is that enterprise IT has historically treated storage and applications as different domains. Adopting the VAST Data Platform means accepting a level of lock-in and dependency that doesn't exist in the more traditional legacy storage world.

For many customers, the gains in performance and efficiency will be worth it. Others will want to manage these workflows separately. It's reminiscent of a similar challenge that occurred with hyperconverged infrastructure a decade ago.

The Cisco and Equinix partnerships open new channels for VAST to sell its AI-focused storage and data infrastructure solutions, giving it access to larger enterprise customers and simplifying adoption. These partnerships enable VAST to offer more flexible deployment options through integrated hardware solutions (with Cisco UCS) or global data center infrastructure (with Equinix).

There's little question that VAST Data is well-positioned as a leader in AI infrastructure, offering enterprises the tools they need to efficiently and effectively scale AI workloads. This is the right time for the newly announced capabilities and relationships. VAST Data finds itself delivering enterprise AI features when enterprises are starting to build AI infrastructure. It's a powerful position.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.