VAST Data AI Reference Architecture

Version 1.0

Table of Contents

The Data Platform for the AI Era **3**

Accelerating Modern AI Pipelines **3**

AI Pipelines Accelerated 5

VAST Data Platform Overview **7**

Disaggregated, Shared-Everything Architecture (DASE) 8

VAST DataStore 8

VAST DataBase 9

VAST DataSpace 9

VAST DataEngine 10

Zero Trust Architecture 10

VAST Data Platform Deployment Options **11**

Workload Optimized 11

NVIDIA BlueField Optimized 12

Hyperscale Optimized 13

Cloud Deployments 13

Partnering to accelerate insights **14**

The Data Platform for the AI Era

As artificial intelligence (AI) becomes operationalized across every organization the limitations of the classic HPC infrastructure —complex implementation, frequent outages, and security gaps become serious impediments to adoption. AI is becoming an enterprise application and must be simple enough to be supported by IT generalists, and reliable enough for non-stop operations while supporting a growing number of applications without sacrificing performance or scale.

VAST Data has transformed the data infrastructure landscape for the AI era. Recognizing the need for organizations to prepare all their data for AI, VAST developed a unique architecture optimized for GPU-accelerated computing. By harnessing high-performance flash storage and high-speed, low-latency networks, VAST combines the speed and scalability of HPC storage systems with the reliability of enterprise solutions with efficiencies such as data reduction that deliver archive economics. In just a few years VAST has become the fastest-growing infrastructure company in history and is the foundation for AI and Deep learning at organizations from leading research institutions to GPU cloud providers.

This white paper delves into how organizations can streamline and optimize their AI pipelines by consolidating all stages of data processing and analysis onto the VAST Data Platform. From ingestion and storage to analysis and inference, learn how integrating with VAST can transform your data operations, enhance your AI capabilities, and drive unprecedented insights.

Accelerating Modern AI Pipelines

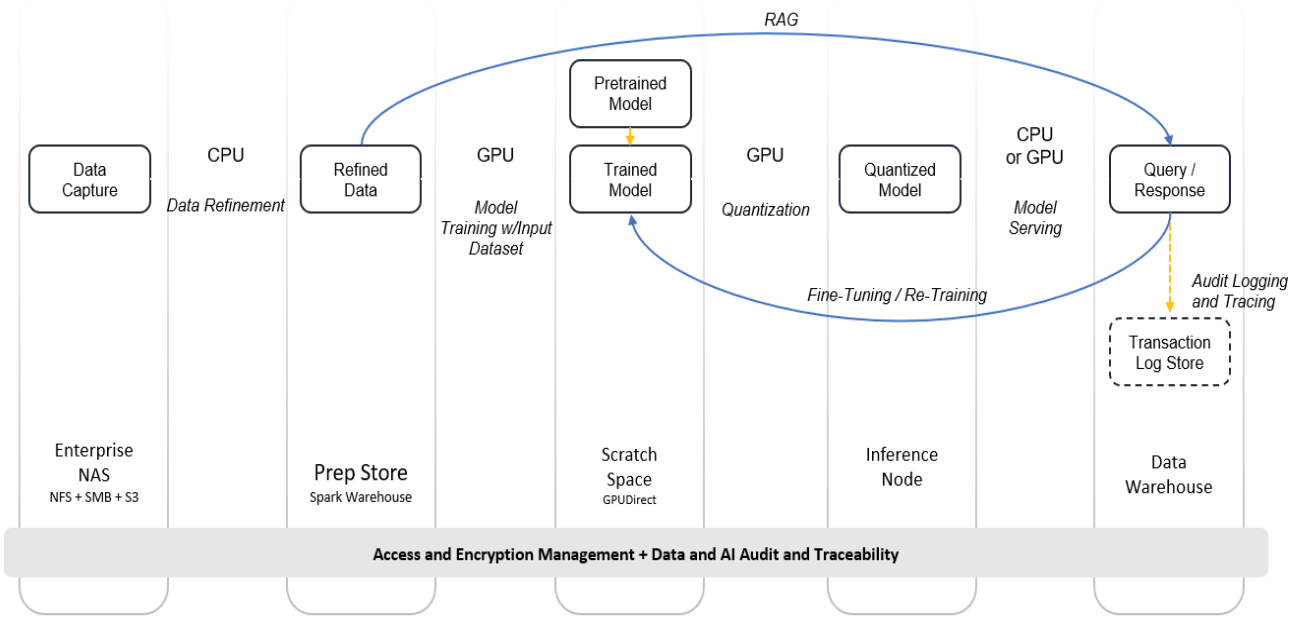
Classic HPC and AI infrastructure were designed to support single workloads, dedicating 100% of resources to a given job until its completion before providing scheduled access to the next job. Most of the attention was given to the most computationally demanding component of AI-model training. AI centers would dedicate customized data solutions for each phase requiring complex data copying and management. The new reality is that AI infrastructure must support many simultaneous workloads, ensuring proper resource allocation and secure isolation. High-performance data access is table stakes for AI infrastructure and the streamlining of AI pipelines is becoming the key factor for reducing time to insight.

VAST Data with multiprotocol access to all data and support for both structured and unstructured data eliminates the need to make additional copies of data or manage data over silos. This simplicity is a key benefit for organizations like cloud service providers and enterprise customers. With traditional AI infrastructure, each step of the pipeline requires a unique storage system, fine tuned for a specific step. Slow, low-cost archive-tier storage for raw data, while lower capacity, high performance flash is reserved only for the most performance-intensive processes. VAST now makes it possible to use a single tier of affordable high performance flash for the entire pipeline. This radical new approach not only provides peak performance for every step but eliminates the complex and time-consuming copying of data from tier to tier. By storing all data on high-performance flash storage, VAST Data's architecture enables AI workflows to be significantly accelerated, freeing up valuable time for model training and discovery rather than cumbersome data management tasks.

Data security concerns are now top of mind for AI pipelines. Nefarious actors may seek to sabotage data pipelines with data poisoning or exploit these massive data sets to violate the privacy of individuals. VAST is built as a zero trust data platform with support Role-Based Access Control (RBAC), Attribute-Based Access Control (ABAC), end-to-end encryption, and robust auditing. The consolidation of disparate systems needed to support AI in the past to a single platform reduces the attack surface and eliminates the complexity of managing permissions across multiple storage clusters.

Data Requirements for AI Pipelines End-to-End

VAST supports every stage of the AI pipeline



AI Pipelines Accelerated

Streamline AI pipelines by consolidating training data, feature stores, model artifacts and inference to a single platform.

Data Capture

The first step involves gathering raw data from various sources, such as databases, sensors, online transactions, social media platforms, imagery, and more. The ability to ingest data from file, object, or streaming sources provides the flexibility to gather data from any desired source.

The sheer quantity of data at this step requires a data platform capable of massive scale as well as efficiency to make retaining data affordable.

Data Refinement

Raw data often contains errors, inconsistencies, or missing values. This step involves cleaning the data by removing or correcting anomalies and preprocessing it (e.g., normalization, feature extraction) to transform the data into a format suitable for analysis. This step is typically performed by copying data to the local storage of high-performance servers. VAST's high-performance file and object performance allow these servers to read and process data from shared storage, eliminating the need for cumbersome bulk copy and data management. Furthermore, integrating NVIDIA RAPIDS with the VAST Data Platform provides transparent GPU acceleration for data analytics workloads using Apache Spark. Organizations can execute their end-to-end data pipelines entirely on GPUs while getting further performance benefits from VAST's query filtration capabilities.

Model Training

In this phase, machine learning algorithms use the refined data to learn and identify patterns. This process involves selecting an algorithm, feeding it the training data, and iteratively adjusting the model to minimize errors. Training is compute intensive and requires the highest levels of performance. In legacy data pipelines, refined data would be copied to "scratch space" on finely tuned parallel file storage systems. VAST supports this high level of performance for all data without the need to copy or take an outage to tune the system. Data scientists can experiment and iterate across massive data sets to optimize model performance rather than waiting for data to be copied.

Model Quantization

Model quantization reduces the precision of a model's parameters (e.g., from 32-bit floating-point to 8-bit integers) to decrease its size and speed up inference, with minimal impact on accuracy. VAST Data's high-performance and scalable storage architecture is optimal for quantization tasks because it provides the speed and bandwidth to handle the large datasets and intensive I/O operations, essential for the iterative nature of model optimization and quantization processes.

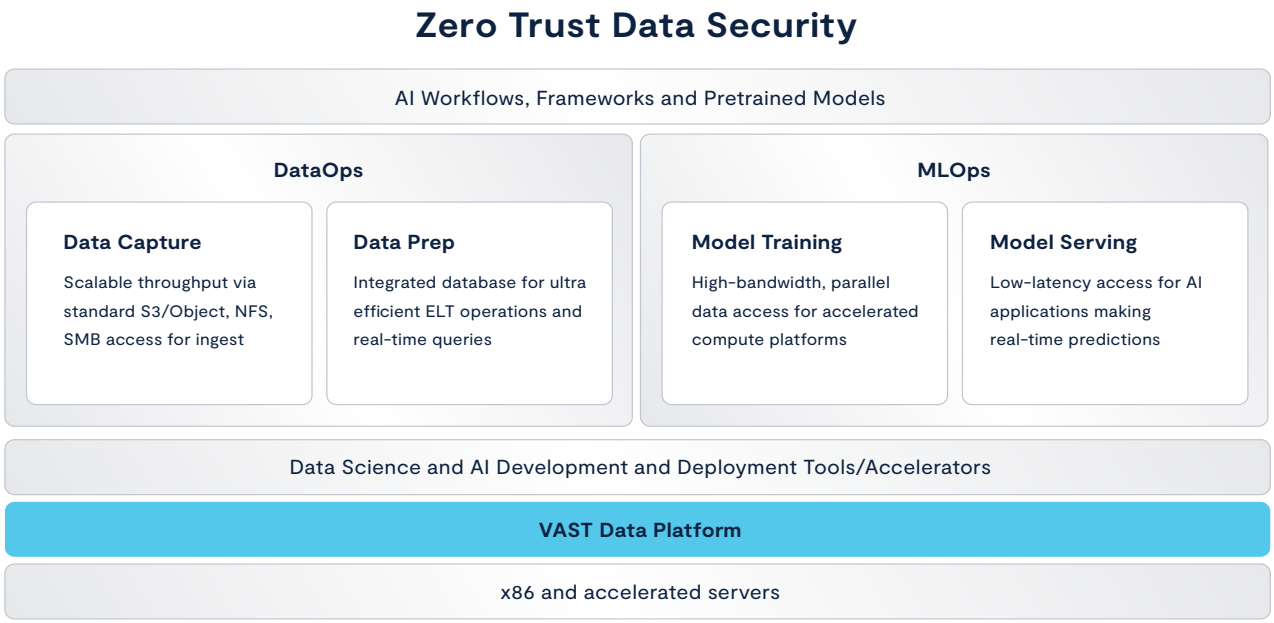
Model Serving

Model serving or inference is the phase where a trained AI model is used to make predictions on new data or answer the queries of LLM user prompts. It's critical for AI applications in real-world scenarios, requiring fast, efficient processing to deliver timely results. Customers can leverage NVIDIA Triton™ Inference Server, available with NVIDIA AI Enterprise, to standardize AI model deployment and optimize performance for many query types, including real time, batched, ensembles and audio/video streaming.

Furthermore, the unified VAST Data Platform simplifies the implementation of Retrieval-Augmented Generation (RAG) techniques. RAG enhances large language models by providing them access to a broader range of data (stored on VAST), which increases the accuracy and relevance of the model's responses. This approach improves the model's output without the need for retraining, making it more versatile and useful across diverse contexts, particularly for applications requiring up-to-date or domain-specific information. By enabling easy access to a comprehensive data set, VAST empowers language models to leverage RAG effectively, delivering more accurate and contextually relevant responses.

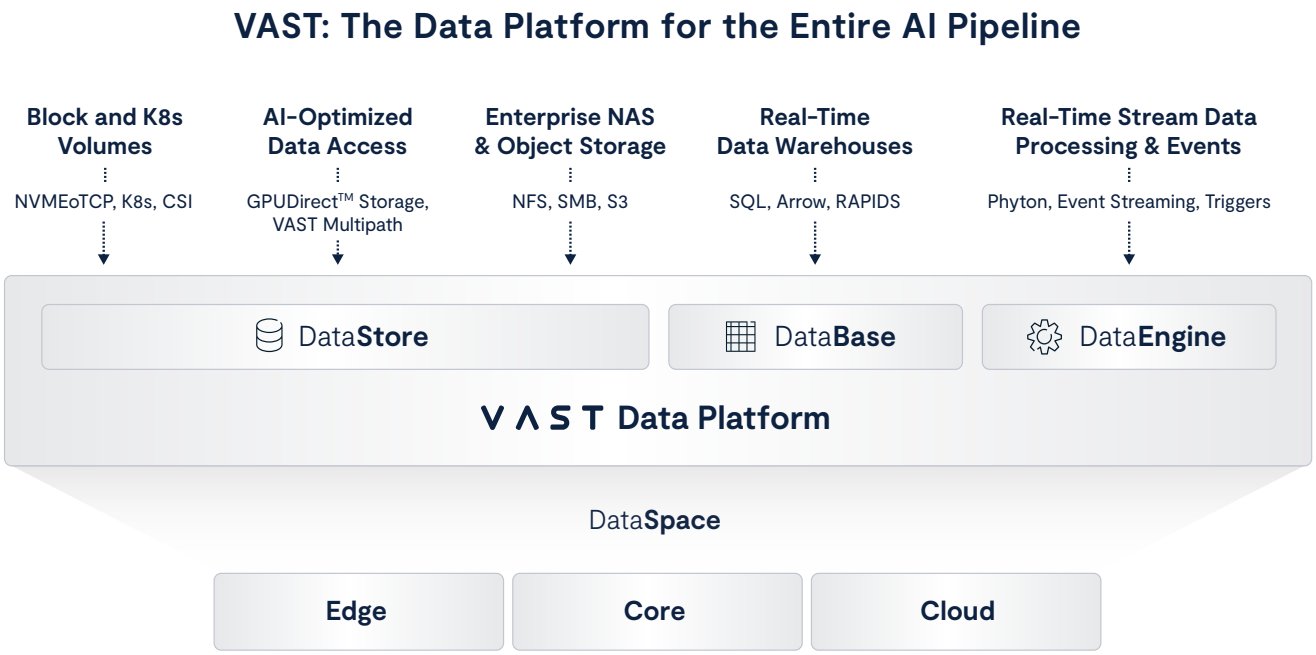
Audit and Transaction Logging

Audit and transaction logging are crucial for AI because they ensure transparency, accountability, and compliance in AI operations. These logs provide detailed records of AI system activities, enabling the monitoring of AI decisions, data usage, and system changes. The VAST DataBase is an ideal repository for this data, ensuring that organizations can respond to regulatory requests and improve on inference response. VAST's space-efficient snapshots can be used to curate the datasets used for training and inference to provide provenance.



VAST Data Platform Overview

To simplify AI pipelines VAST completely reimagined data management. Rather than continuing to add workarounds such as caches and tiers, VAST built an entirely new systems architecture designed for flash storage and high-performance, low-latency networking. The result is a data platform that unifies unstructured and structured data, and provides the scale and performance for the most demanding AI workloads, with efficiencies that make it affordable for organizations to store 100% of their data in a single AI-ready environment.

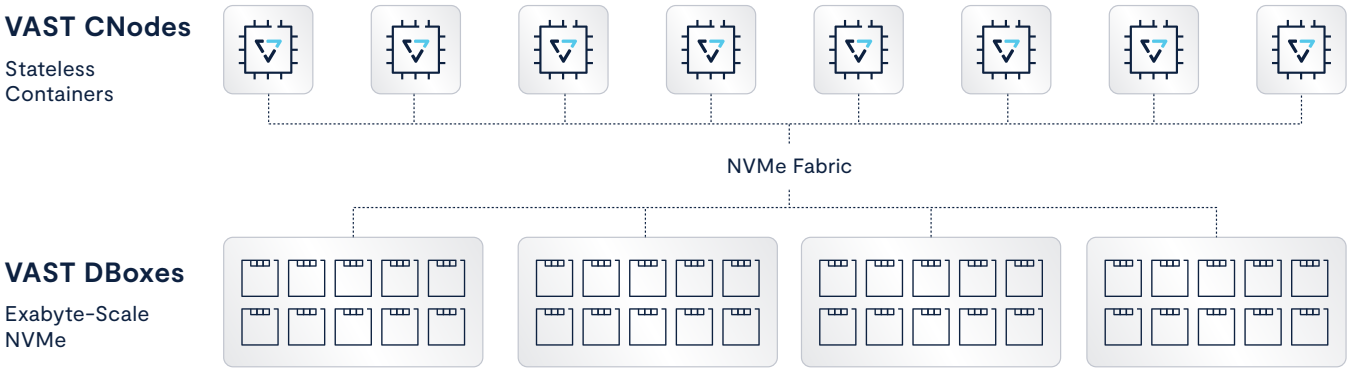


The VAST Data Platform is a breakthrough approach to data-intensive computing that serves as the comprehensive software infrastructure required to capture, catalog, refine, enrich, and preserve data through real-time deep data analysis and deep learning. This system is designed to provide seamless and universal data access and computing from edge-to-cloud, all from a platform that is designed for enterprises and cloud service providers to deploy on the infrastructure of their choosing.

Disaggregated, Shared-Everything Architecture (DASE)

By disaggregating the cluster's computational resources from its persistent data and system state VAST's architecture breaks the tradeoffs inherent in shared-nothing and shared-media architectures that have been the mainstay of IT for the last two decades.

VAST's shared everything model allows any VAST server or CNode to have direct access to all the data, metadata, and system state directly. In a DASE cluster, the system state is stored on NVMe SSDs in highly available NVMe enclosures known as DBoxes. By eliminating the need for CNodes to communicate with each other VAST clusters can independently scale performance and capacity well beyond the limitations of legacy architectures. Without the need to accommodate mechanical media, VAST implemented a new class of data protection algorithms that deliver superior resilience with radically lower overhead and fast disk rebuild. The result is a data platform that scales to exabytes of capacity, and delivers TB/s of performance, at a cost point that makes tiering and caching obsolete.



VAST DataStore

The DataStore is an unstructured data repository presenting popular file and object interfaces that VAST customers have been using since the inception of the company. Data ingested into the DataStore is stored not as files or objects but as data elements that can be presented over NFS, SMB, and S3 simultaneously with the flexibility to support future protocols and APIs. The DataStore manages metadata and data in a persistent and highly scalable architecture that eliminates the common causes of data corruption. Whether VAST is on-prem or in the cloud, the VAST Datastore serves, protects, and performs data reduction algorithms to optimally manage data.

Benefits of the VAST DataStore:

- Future-proof data access via the protocol of choice and eliminates the need to manage multiple storage platforms.
- AI / HPC class performance for all data simplifies pipelines
- Archive economics for exabyte-scale volumes of data on AI-ready flash at the cost of multi-cluster tiered solutions.

VAST DataBase

The VAST DataBase is a fully ACID-compliant high-speed data lake, that supports both transactional and analytical workloads at exabyte scale. Organizations can perform queries against not just real-time streaming data but also across the entirety of the historical data set. In the past this holistic view of data would only be possible by running separate databases, data warehouses, and data lake platforms all stitched together with complex ETL pipelines. The VAST DataBase provides support for query engines such as Trino and Spark with full push-down capabilities to accelerate query performance.

The VAST DataBase manages the VAST Catalog - for every file and object ingested to the DataStore metadata is automatically recorded to the VAST DataBase. This functionality supports audit and reporting capabilities required for AI provenance and regulatory compliance.

Benefits of the VAST DataBase:

- Simplifies and accelerates structured data workloads
- Eliminates complex data engineering such as vacuuming and partitioning
- Reduces compute resources needed by query engines with advanced filtration

VAST DataSpace

The VAST DataSpace creates a global namespace over data centers, cloud, and edge with a new approach that addresses the challenges of replication, remote data access, and consistency that have plagued traditional systems.

VAST introduces a concept of global lease management combined with intelligent data movement. The members of a VAST DataSpace have full visibility to all data (within the permission model) and can granularly take ownership or a "write lease", such that when a process at a remote location needs to update data only the precise required data is moved across the wire, and while the update is taking place leaseholder becomes the authoritative owner of the data to ensure consistency until the lease is released. This new concept allows for consistency to be assured across multiple geographies without wasteful and expensive data movement.

Benefits of the VAST DataSpace

- A single data namespace that spans multiple locations and clouds without the high cost of legacy replication policies
- Transparent access to data for analysis and ML processing
- A consistent view of data and elimination of copy sprawl

VAST DataEngine

The VAST DataEngine, is a distributed processing environment designed to power event-driven AI workflows. By natively integrating data processing and event notifications, the DataEngine lays the foundation for continuous AI training, inference, and discovery. In its initial iteration, the DataEngine introduces a data streaming interface that writes events directly into the VAST Database. This enables real-time evaluation or triggers, where functions can be invoked to initiate new processes and automate the critical task of AI discovery.

Benefits of the VAST DataEngine:

- Iterative event-based processing for deep learning breakthroughs
- Location \ resource-aware computing engine determines where to invoke processes for optimal data access or when to move to data to compute resources
- Integration with VAST DataBase to give structure to unstructured data

Zero Trust Architecture

VAST is uniquely positioned as an AI-class data management solution with enterprise-grade security. VAST is inherently multi-tenant with the ability to dedicate CNode Pools, create QoS policies, isolate networking, and uniquely encrypt each tenant's data with enterprise key management (EKM). This is a key differentiator that has made VAST the defacto software stack for service providers creating cloud-scale AI offerings.

"Previous technology in HPC was not designed with security in mind. This is where VAST provided us with performance and enterprise features like Zero Trust."

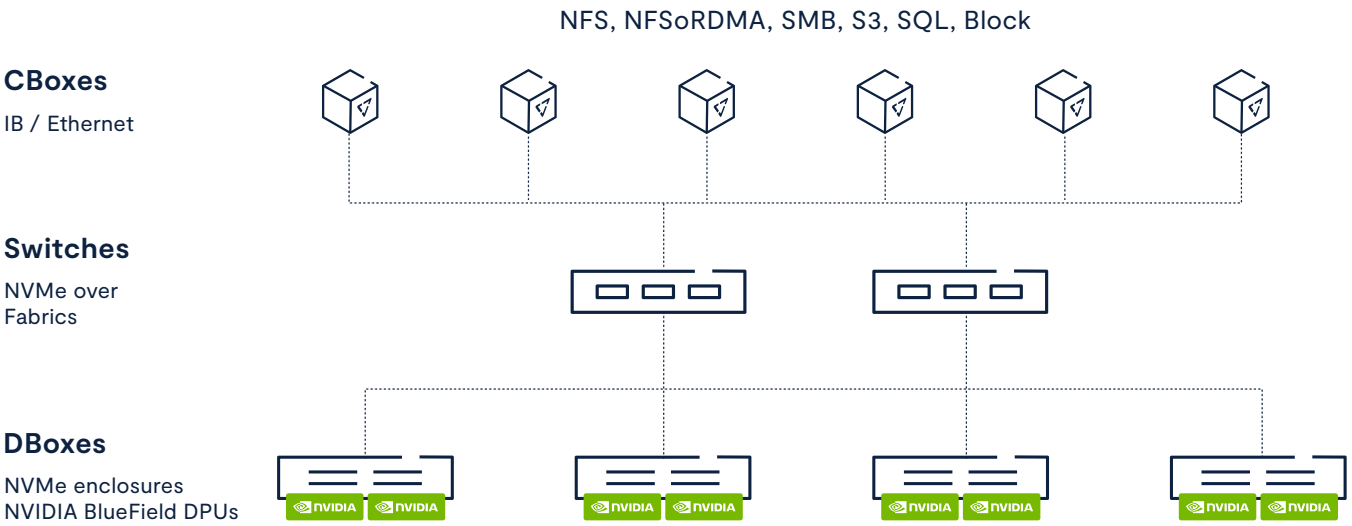
- Edmondo Orlotti,
Chief Strategy Officer

core42

VAST Data Platform Deployment Options

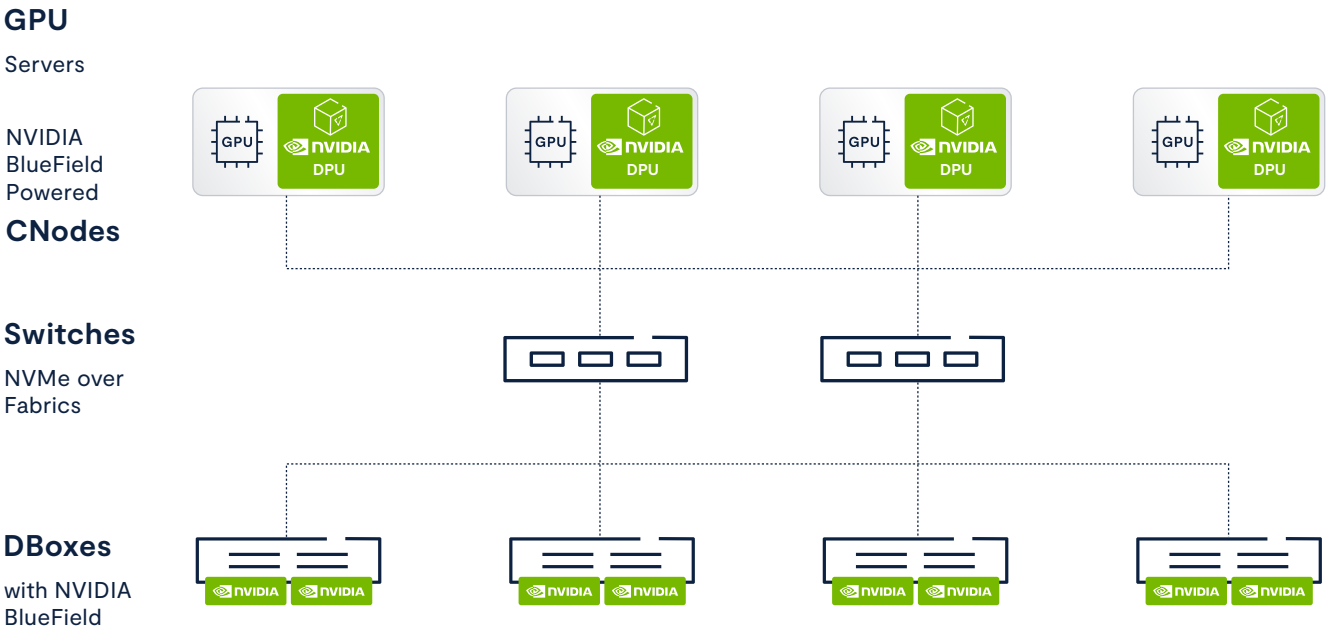
VAST's flexible and scalable architecture offers customizable deployment options to meet the diverse needs of data-driven organizations - from enterprise-scale on-premises deployments to seamless integration with hyperscale cloud environments and increasingly as the foundation for a new generation of GPU cloud service providers.

Workload Optimized



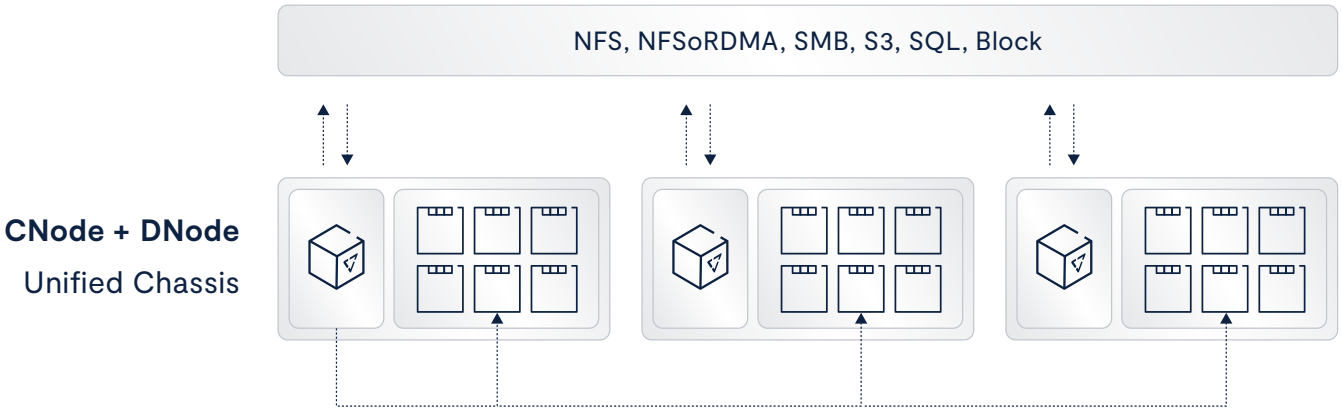
VAST for workload-optimized deployments gives organizations the ability to independently scale compute and capacity to meet their unique requirements. VAST CBoxes are x86 servers with VAST containerized CNodes that serve file, object, table, and block protocols over InfiniBand and Ethernet. VAST DBoxes are highly available NVMe enclosures housing Storage Class Memory for highly durable and persistent metadata and transactional layer in addition to ultra-dense Hyperscale Flash for cost-effective file, object, and table storage.

NVIDIA BlueField Optimized



The flexibility of VAST's architecture allows for the adoption of new technologies as illustrated with VAST Cnodes running directly on NVIDIA BlueField-3 DPUs. NVIDIA BlueField DPUs are powerful networking, storage, and security processors designed to offload, accelerate, and isolate data center workloads. To support deployment on NVIDIA BlueField DPUs VAST engineering ported the VAST container to the ARM architecture to provide native file and object services to the GPU server without proprietary file systems clients. Now cloud service providers building GPUs aaS offerings can build fully disaggregated data centers while reducing their infrastructure footprint and enhancing their security posture.

Hyperscale Optimized



The world's largest cloud builders seek efficiency at every level of infrastructure. At "Hyperscale" there are significant operational advantages to using a common platform for multiple services. For hyperscale operations, VAST is introducing a new platform where both the CNode and DNodes are housed in a unified chassis. These simple units are building blocks of data center scale AI computing. While the CNode and DNodes now share a common enclosure the core tenants of DASE architecture are upheld with massive parallelism achieved by each CNode maintaining independent access to every NVMe device in any chassis. Furthermore the failure of any CNode or DNode has no impact on data availability or performance regardless of being collocated in a single chassis.

Cloud Deployments

VAST software can be deployed on Amazon Web Services (AWS) and Google Cloud, with Microsoft Azure support later in 2024. When deployed on the public cloud, VAST supports the full range of features available in the largest on-premises VAST clusters. Combined with the VAST DataSpace, customers can deploy VAST across their choice of hybrid and multi-cloud platforms for a consistent set of standard protocols and APIs to support and simplify AI workflows regardless of the cloud environment.

Partnering to accelerate insights

VAST Data's position as a leader in AI infrastructure is bolstered by its strategic partnerships and deep integrations with pioneering companies in the deep learning ecosystem.

NVIDIA DGX SuperPODand DGX BasePod Storage Certifications

The VAST Data Platform the is first enterprise NAS certified as a storage solution for both NVIDIA DGX SuperPOD and DGX BasePOD with the ability to seamlessly scale as customers expand their NVIDIA accelerated computing capabilities.

NVIDIA AI Enterprise

The VAST Data Platform seamlessly integrates with NVIDIA AI Enterprise, enhancing its end-to-end, cloud-native capabilities for accelerating data science pipelines and streamlining the development and deployment of generative AI applications. This collaboration provides enterprises with a robust foundation, leveraging VAST's data infrastructure alongside NVIDIA's advanced AI microservices, frameworks, and tools to ensure secure, efficient, and stable transitions from pilot projects to full-scale production, supporting the demand for high-performance, production-grade AI and generative AI applications.

VAST Data has partnered with Run:ai to bridge operational gaps in AI deployments, enhancing the efficiency and scalability of AI workflows. This collaboration integrates VAST's advanced data infrastructure with Run:ai's orchestration capabilities, enabling enterprises to optimize resource utilization, improve GPU utilization, and accelerate AI model development and deployment processes. The partnership aims to provide a seamless, scalable solution for managing complex AI workloads across diverse computing environments.

VAST Data and Supermicro have collaborated to leverage Supermicro's NVIDIA-Certified systems with VAST's innovative software platform, creating a unified solution for cloud service providers and AI driven enterprise. The combined strengths of Supermicro's hardware and VAST Data's software ensure advanced scalability, efficiency, and robust security, catering to the demanding requirements of large-scale AI deployments.

For more information on the VAST Data Platform and how it can help you solve your application problems, reach out to us at hello@vastdata.com.

03/2024