PURESTORAGE®

NVIDIA ELITE PARTNER

TECHNICAL WHITE PAPER

Enterprise Medical Imaging with NVIDIA MONAI NIM and Pure Storage FlashBlade

Accelerate patient care and streamline clinical workflows with AI

Contents

Executive Summary

Medical imaging is the cornerstone of modern healthcare clinical AI use, and the demands for faster, more accurate diagnosis are pushing the limits of traditional IT infrastructures. Healthcare organizations contend with vast amounts of imaging data, privacy concerns, and the need for AI-driven analysis. The challenge now lies in building scalable, high-performance systems that can keep pace with these demands.

To meet these challenges, NVIDIA's MONAI, VISTA-3D, and MAISI models, in combination with Pure Storage® FlashBlade//S500, offer a powerful AI solution specifically designed for enterprise-scale medical imaging. This architecture streamlines the deployment of advanced deep learning models and leverages synthetic data generation to enhance 3D computed tomography (CT) workflows—delivering both speed and precision in clinical settings.

This solution integrates NVIDIA's MONAI, VISTA-3D and MAISI frameworks, minimizing reliance on sensitive patient data and reducing privacy risks while ensuring AI models are trained on robust and diverse datasets. Pairing one of these frameworks with a Pure Storage FlashBlade®, which provides high throughput and low latency, enables healthcare organizations to experience faster image segmentation, improved model performance, and the ability to scale effortlessly to meet growing demands.

In addition to optimizing data access and model processing, this architecture seamlessly fits into existing IT infrastructures, making it easier for healthcare providers to harness AI for precision medicine and clinical practice. Its performance advantages over traditional storage systems allow institutions to accelerate AI workflows and deliver quicker, more accurate results for diagnostic purposes. Designed with scalability in mind, this architecture is ideal for organizations aiming to future-proof their AI initiatives in medical imaging.

Introduction

Superior Performance with Pure Storage FlashBlade

The high-throughput, low-latency storage from a FlashBlade® system allows for significantly faster processing of medical images compared to traditional network attached storage (NAS) solutions. This performance gain is crucial for healthcare institutions that rely on timely analysis for diagnosis and treatment in clinical practice. By minimizing bottlenecks in data access, FlashBlade also accelerates AI inference, which enables real-time clinical analysis, something that many competitors with slower storage systems struggle to deliver.

The combined use of Pure Storage FlashBlade and NVIDIA MONAI frameworks ensures that the solution scales effortlessly to accommodate large datasets and growing computational demands. This scalability, in both data processing and model performance, gives this architecture a significant edge over solutions that struggle to handle large volumes of medical images or require costly infrastructure changes to scale.

The integration of cutting-edge AI frameworks like NVIDIA MONAI—which is specifically designed for healthcare use cases—positions this solution as a future-proof investment. MONAI's ability to rapidly adapt to new medical imaging challenges ensures that this pipeline can evolve with advancements in medical AI, keeping organizations at the forefront of technology. Competitors using less specialized frameworks or those reliant on proprietary systems are likely to face higher costs and slower innovation cycles.

Enterprise Scale Inferencing with NVIDIA VISTA-3D

NVIDIA's VISTA-3D NIM™ provides out-of-the-box segmentation for 127 anatomical structures and allows for interactive editing, combining automation with clinician expertise. This hybrid capability increases the accuracy of segmentation tasks and reduces the manual workload, differentiating the solution from competitors that offer either rigid automatic systems or exclusively manual processes. Furthermore, its zero-shot learning capability for new anatomical structures offers adaptability that most competitor systems lack.

Synthetic Data Generation with NVIDIA MAISI for Privacy and Data Augmentation

The NVIDIA MAISI NIM offers a unique advantage by generating high-quality synthetic 3D CT images, allowing institutions to expand their training datasets without relying on sensitive patient data. This capability not only reduces privacy risks but also enhances the model's ability to generalize across diverse medical scenarios. Competing solutions often lack the ability to generate synthetic data at this level of fidelity and scalability, making MAISI a standout feature in augmenting datasets while ensuring compliance with privacy regulations.

Solution Overview

NVIDIA MONAI is a comprehensive AI framework designed to streamline the development and deployment of deep learning models for medical imaging tasks such as segmentation. Built to handle the unique challenges of 3D medical imaging, MONAI provides healthcare organizations with customizable models that can be fine-tuned for specific medical applications, such as organ segmentation, tumor detection, and disease analysis.

The framework's ability to integrate with high-performance storage solutions like Pure Storage FlashBlade enhances its scalability and performance, making it ideal for enterprise-grade medical imaging workflows. MONAI supports both inferencing and fine-tuning of models, offering robust tools for managing large-scale medical datasets and complex, high-dimensional data such as CT and MRI scans.

MONAI VISTA and MAISI NIMs were deployed on FlashBlade systems, which provided high throughput and low-latency data access, resulting in superior performance for model inferencing and fine-tuning tasks. This integration enabled seamless handling of large volumetric image sets and improved the overall reliability of the system, allowing for uninterrupted processing even in multi-GPU environments.

MONAI's deep integration with NVIDIA GPUDirect Storage (GDS) further accelerates workflows by minimizing the reliance on CPU caching. GDS allows for direct data transfer between the storage and GPUs, resulting in faster processing times for both training and inferencing. When combined with the parallelism capabilities of MONAI, GDS ensures near-linear scaling of performance, making the framework suitable for the most demanding medical imaging applications.

Key Components

The key components of the solution are:

- **MONAI**: A deep learning framework optimized for healthcare, offering flexibility and specialized tools for medical imaging workflows, particularly suited for training complex AI models.
- **VISTA-3D**: A 3D segmentation model offering precise segmentation for 127 anatomical structures and interactive editing capabilities for healthcare professionals.
- **MAISI diffusion model**: A cutting-edge diffusion model that generates high-resolution synthetic 3D CT images, augmenting data for training pipelines while reducing the need for patient data, addressing privacy concerns.
- **Pure Storage FlashBlade**: A high-performance, scalable network-attached storage (NAS) solution that provides superior throughput for AI workloads, enhancing inference times and enabling parallel data processing.

High-level Benefits

The key benefits of the solution are centered around five key areas:

- **Improved performance**: Pure Storage FlashBlade accelerates the performance of AI models, particularly in processing large CT datasets, enabling faster inference, more efficient workflows, and the benefits of an all-flash enterprise platform.
- **Scalability**: The solution can scale with increasing data volumes, making it ideal for large healthcare providers with expanding medical imaging demands.
- **Synthetic data generation**: The MAISI NIM model supports the creation of high-quality synthetic 3D images, reducing dependency on patient data and ensuring privacy is maintained without compromising model accuracy.
- **Cost efficiency**: By reducing processing times and mitigating privacy-related risks, the architecture delivers a cost-effective solution for healthcare institutions looking to adopt advanced AI technologies.
- **Enhanced flexibility**: VISTA-3D offers both automated and interactive segmentation capabilities, while MAISI allows for fine-tuning based on specific clinical needs, providing flexibility for various imaging tasks.

Key Advantages of MONAI and FlashBlade

The Pure Storage FlashBlade solution for MONAI provides scalable, high-performance infrastructure, specifically optimized for AI-driven medical imaging. FlashBlade systems efficiently host large medical datasets and act as a reliable repository for storing models, outputs, and logs. By scaling up performance to run multiple inferencing jobs in parallel, FlashBlade supports both training and inference workloads within a unified GPU cluster.

A FlashBlade system's zero-disruption architecture ensures uninterrupted access to data during long training cycles, significantly reducing the risk and cost of downtime. For healthcare organizations, this means faster time-to-deployment for AI models, improved patient outcomes, decreased physician workloads, and a more efficient use of computational resources.

Technology Overview

The architecture of Pure Storage FlashBlade//S™ systems deliver a highly optimized and cost-effective storage solution, essential for running the NVIDIA MONAI in large-scale medical imaging applications. This architecture is designed to enhance performance while maintaining industry-leading efficiency per rack unit (RU), watt, and terabyte (TB), ensuring that healthcare institutions can process vast amounts of medical imaging data with speed and precision. By providing a scalable, power-efficient solution, FlashBlade systems enable healthcare professionals to accelerate time-to-diagnosis and improve patient outcomes through faster and more accurate AI-driven imaging workflows.

The modular architecture of FlashBlade//S allows for independent scaling of compute and storage, which is critical in tailoring configurations for specific medical imaging tasks such as segmentation, classification, and model fine-tuning within the NVIDIA MONAI framework. This flexibility ensures consistent high performance, allowing MONAI to handle the growing demands of data-heavy healthcare environments without sacrificing speed or accuracy.

The multidimensional performance capabilities of FlashBlade are indispensable for managing the diverse workloads that MONAI requires, including training, inferencing, and data augmentation. MONAI's ability to leverage synthetic data generation (through models like MAISI) and large-scale 3D imaging tasks (such as VISTA-3D) requires a storage platform that can seamlessly handle both high-throughput and low-latency requirements. The architecture of FlashBlade is designed to maintain performance integrity even as data volumes grow, ensuring that healthcare organizations can scale their AI operations as needed without compromising on reliability.

Global media management on DirectFlash® modules further enhances capacity and endurance by extracting up to 20% more capacity from NAND compared to traditional solutions. This feature is vital for maintaining consistent performance across MONAI's high-demand medical imaging workloads, which often involve processing large 3D CT and MRI datasets. Unlike systems that rely on large storage class memory (SCM) caches, FlashBlade provides these performance and capacity benefits directly, making it a more cost-effective and reliable choice for enterprise-scale healthcare applications.

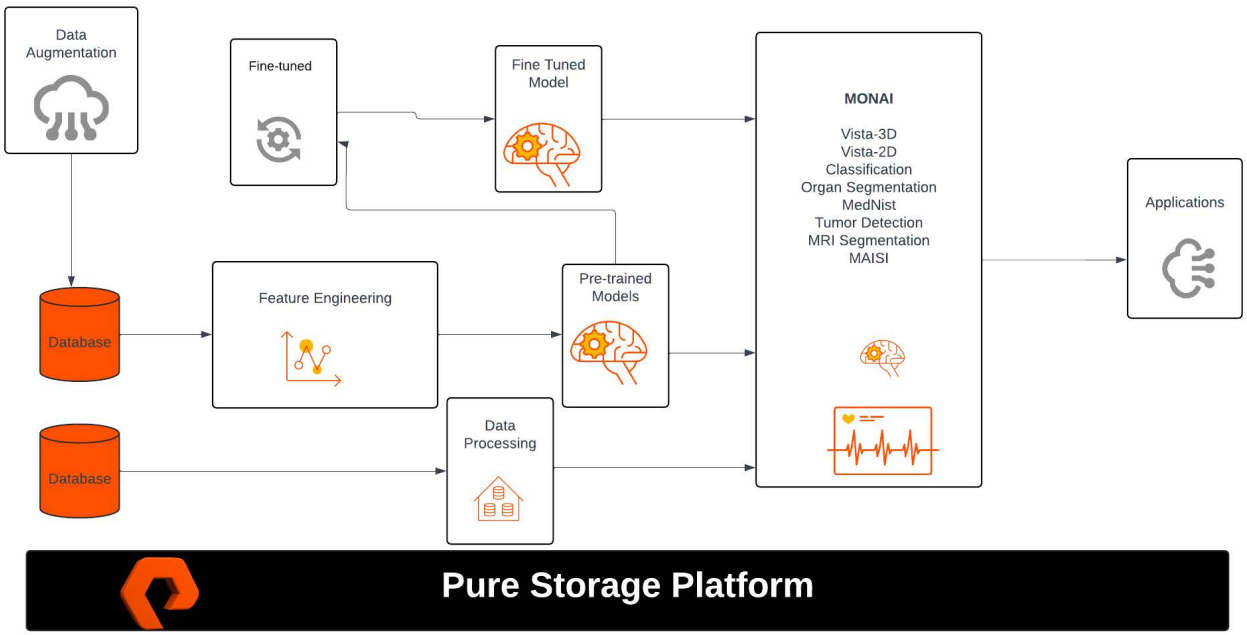

FIGURE 1 Pure Storage platform workflow for medical imaging AI

Technical Components of NVIDIA MONAI

MONAI Framework

The MONAI framework is a suite of programming tools and APIs developed specifically for AI-driven medical imaging. It provides a robust platform for building, training, and fine-tuning deep learning models for healthcare applications. By integrating state-of-the-art architectures, advanced data handling techniques, and a focus on medical-specific workflows, MONAI accelerates the adoption of AI in medical imaging.

VISTA Framework

The Versatile Imaging SegmenTation and Annotation (VISTA) framework is an advanced platform for high-precision segmentation in 3D medical imaging. Tailored for medical professionals and researchers, VISTA combines semantic segmentation with interactivity, allowing users to navigate complex anatomical structures with precision. VISTA supports a range of core workflows designed to optimize disease analysis, organ mapping, and the creation of accurate ground-truth data for AI models. The framework's ability to segment entire bodies, specific classes, or user-defined points makes it an adaptable solution for tackling diverse medical imaging needs. With VISTA, healthcare institutions can explore whole-body imaging, isolate specific organs for focused analysis, and create detailed annotations that fuel AI-driven diagnostics. Core workflows of VISTA include:

- **Segment everything:** This workflow enables a comprehensive exploration of the entire body, making it a vital tool for understanding complex diseases that affect multiple organs. It supports holistic treatment planning by providing a complete view of the body's anatomical structures.
- **Segment using class:** Offering focused, sectional views of specific anatomical areas, this workflow is essential for targeted disease analysis, such as identifying tumors in critical organs. It allows medical professionals to focus on specific regions while maintaining the accuracy of full-body scans.
- **Interactive point-prompt segmentation:** This feature enhances segmentation precision by enabling user-directed, click-based selection. By interacting directly with the model, users can create highly accurate ground-truth data quickly, a crucial step in medical imaging analysis and AI model training.

Figures 2 and 3 (below) highlight the difference between traditional CT and AI-augmented imaging. Figure 2 shows an un-augmented CT scan, while Figure 4 is the output of the AI model, demonstrating how segmenting multiple organs quickly can highlight areas of concern for medical staff to investigate.

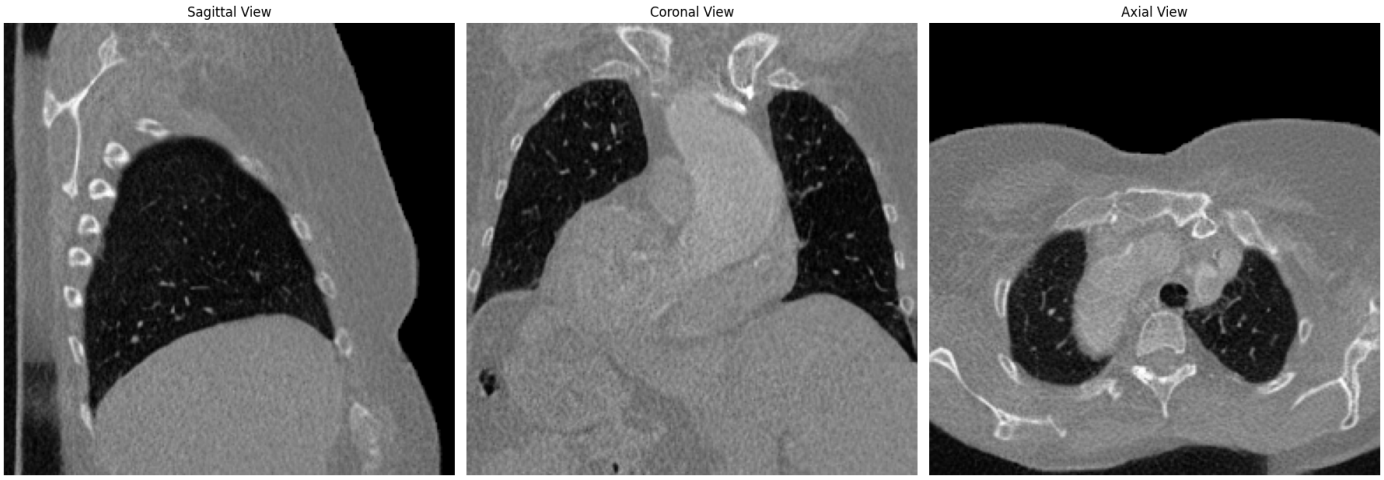

FIGURE 2 Original CT image

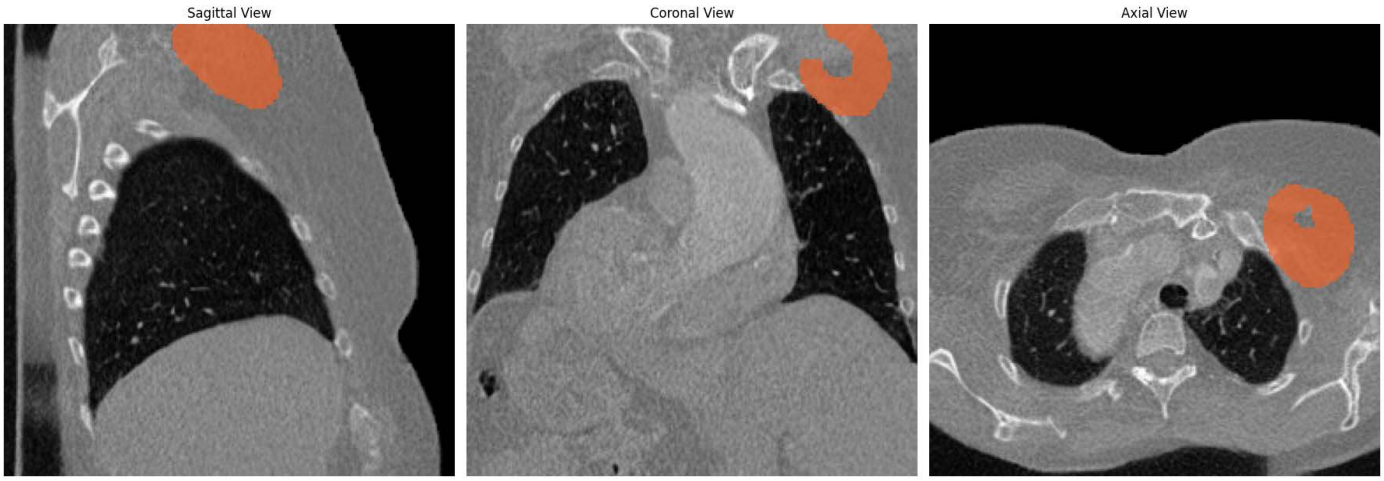

FIGURE 3 VISTA-3D CT output with kidney segmentation

MAISI Framework

The Medical AI for Synthetic Imaging (MAISI) framework is a cutting-edge 3D Latent Diffusion Model developed to generate high-quality synthetic CT images, addressing the challenges of data scarcity and privacy concerns in medical imaging. MAISI is particularly useful for augmenting datasets, allowing researchers to generate diverse, high-resolution 3D CT images, which are essential for training and improving the performance of other medical imaging AI models. Its ability to create synthetic images with or without anatomical annotations gives it a unique advantage in healthcare, where privacy and data availability are constant challenges. By producing paired segmentation masks and supporting a range of voxel sizes, MAISI empowers researchers to fine-tune their models for enhanced diagnostic accuracy and generalization across various conditions. Core features of MAISI include:

- **High-resolution CT image generation**: MAISI can generate 3D CT images with resolutions up to 512 × 512 × 768 voxels, supporting voxel sizes from 0.5mm to 5.0mm. This flexibility allows for a wide range of medical imaging applications, from routine scans to complex anatomical analyses.
- **127 anatomical classes**: Capable of annotating up to 127 anatomical structures, including organs and tumors, MAISI provides detailed segmentation for training medical imaging models. It also allows users to control the anatomy size of 10 specific classes, enhancing the customization of training datasets.
- **Paired segmentation masks**: MAISI's ability to produce paired segmentation masks makes it a powerful tool for creating annotated synthetic datasets, which are critical for training and validating AI models without exposing sensitive patient data.

Synthetic CT images (Figure 4) created with MAISI allow researchers to create large datasets, filling in gaps in data while reducing HIPAA concerns, as they are not derived from any one specific patient.

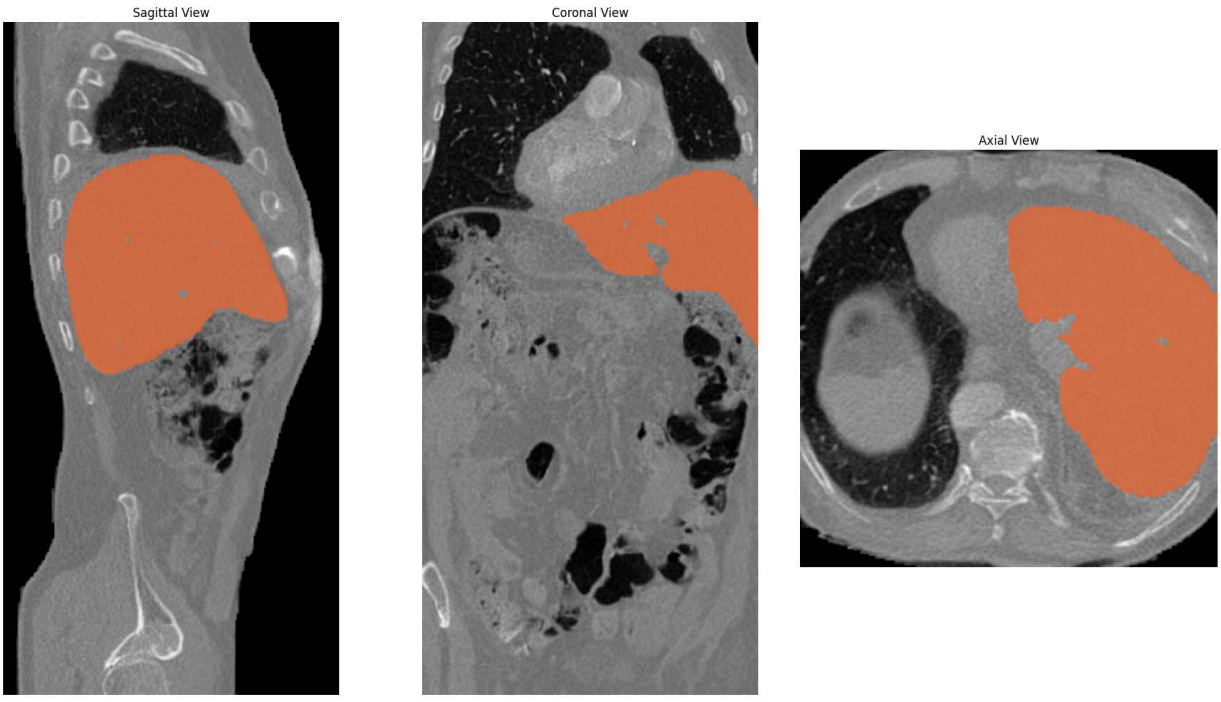

FIGURE 4 Synthetic image with liver segmentation

Integration with AI Tools and Frameworks

The architecture of FlashBlade is built to seamlessly integrate with a wide range of AI tools and frameworks, making it highly compatible within the MONAI framework and the broader AI and medical imaging ecosystem. Its ability to support multiple protocols, including Network File System (NFS) and object storage, allows it to function efficiently as a network-attached storage (NAS) solution while maintaining high performance across diverse AI workflows.

The NFS setup on FlashBlade ensures compatibility with a variety of machine learning frameworks, such as PyTorch, TensorFlow, and NVIDIA Clara, all of which are commonly used within MONAI for medical imaging tasks. A FlashBlade system's multiprotocol access allows AI models to seamlessly pull data for training and inferencing, regardless of the toolset being used. Whether researchers are working with real-world datasets from clinical sources or synthetic data generated through MAISI models, FlashBlade ensures that data is readily available for processing without bottlenecks.

This integration is further enhanced by the ability of FlashBlade to handle GPUDirect Storage, which is a key component of high-performance, multi-GPU configurations. The support of FlashBlade for GDS ensures that data is transferred directly from storage to GPU memory, bypassing CPU involvement and optimizing data transfer rates, particularly in multi-GPU setups. This feature works seamlessly with frameworks like PyTorch, which are optimized for GPU-based processing, making FlashBlade an ideal solution for large-scale AI workloads where throughput and low-latency data access are critical.

Additionally, the multiprotocol support of FlashBlade means it can integrate with other storage technologies and data frameworks within a healthcare or research environment. For example, a FlashBlade system is compatible with object storage systems and can interface with frameworks like Apache Spark for distributed data processing. This flexibility allows organizations to create a unified data platform that supports not only AI-driven medical imaging workloads but also other data-intensive applications, such as data analytics and clinical decision support systems.

By supporting a wide array of tools and frameworks, FlashBlade ensures that AI teams can adopt best-in-class tools for their specific workflows without worrying about storage compatibility or performance bottlenecks. Whether it's deep learning frameworks for AI model development, or data processing tools for handling complex imaging datasets, the architecture of FlashBlade ensures smooth interoperability, making it a critical component in the broader AI and medical imaging ecosystem.

NVIDIA VISTA-3D and MAISI NVIDIA Inference Microservices (NIM)

The NVIDIA VISTA-3D and MAISI NVIDIA Inference Microservices (NIM) are specialized microservices designed to enhance AI model performance in medical imaging. Optimized for NVIDIA GPU architectures, NIM provides high-speed inferencing and scalability, ensuring that large-scale medical imaging tasks, such as segmentation and synthetic data generation, are handled efficiently. Accessible through RESTful APIs, VISTA-3D and MAISI NIM integrates seamlessly into existing healthcare workflows, enabling institutions to quickly deploy AI models for diagnostic and research purposes while maintaining operational efficiency.

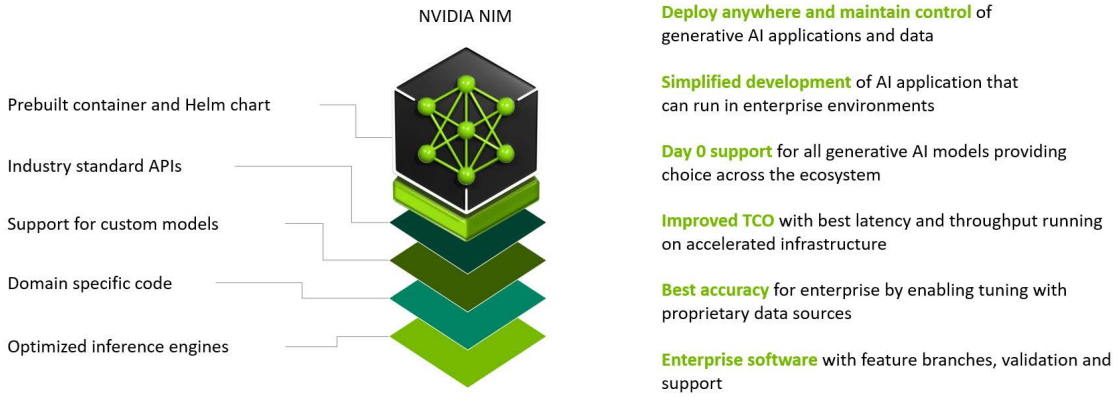

FIGURE 5 NVIDIA NIM platform features and benefits

Using NIM offers healthcare organization several advantages:

- **Scalability and performance**: NIM is designed to handle large-scale deployments, meaning it can manage multiple instances of models across various nodes, which is crucial for enterprises in the health sector, where processing large volumes of data and providing real-time responses are fundamental requirements. By leveraging NVIDIA's powerful GPUs, NIM ensures that the inference process is highly optimized. This optimization results in faster response times and more efficient processing of extensive datasets, which is vital for health analysts and medical doctors who rely on the data for decision-making.
- **Ease of deployment**: NIM utilizes container technology such as Docker, simplifying deployment across diverse healthcare environments, from on-premises data centers to cloud-based platforms. This containerized approach ensures that models are portable and can be deployed consistently in development, testing, and clinical production environments. For healthcare organizations, this means reduced time to operationalize AI models and greater flexibility in deploying new tools in clinical settings. Furthermore, NIM integration with Kubernetes allows for automated deployment, scaling, and management of containerized AI models, reducing operational overhead and ensuring high availability—both crucial for healthcare environments where continuous uptime is essential to patient care.
- **Operational efficiency**: NIM includes tools for monitoring and managing the performance of deployed models. Healthcare organizations benefit from the ability to track key performance metrics like latency, throughput, and resource utilization, all critical for maintaining optimal performance of medical AI models. Additionally, NIM's auto-scaling feature allows the system to automatically adjust the number of model instances based on demand. This ensures that resources are used efficiently and cost-effectively, enabling healthcare institutions to handle varying workloads without over-provisioning, which is particularly important for AI-driven medical imaging workloads.
- **Integration with existing systems**: NIM provides RESTful APIs, enabling seamless integration with existing healthcare IT systems such as electronic health record (EHR) platforms, picture archiving and communication systems (PACS), and other clinical data repositories. Clinicians and medical professionals can easily access AI capabilities through familiar interfaces without needing deep technical knowledge of AI. This API-based access facilitates the integration of advanced AI models into existing clinical workflows, ensuring that healthcare providers can quickly adopt AI without major infrastructure changes. Additionally, NIM supports various AI frameworks and integrates smoothly with existing data pipelines for medical imaging and diagnostics, promoting interoperability within the healthcare ecosystem.
- **Security and compliance**: NIM ensures that patient data is processed securely, adhering to healthcare-specific security standards such as HIPAA (Health Insurance Portability and Accountability Act) and other relevant regulations. Given the highly sensitive nature of healthcare data, NIM includes robust encryption and data protection mechanisms to ensure compliance with stringent security protocols. By supporting industry regulations and offering secure data handling practices, NIM helps healthcare institutions mitigate risks, protect patient privacy, and maintain trust with patients and regulatory bodies.

Key Applications of MONAI in Medical Imaging

NVIDIA's MONAI framework, in conjunction with advanced models like VISTA-3D and MAISI, offers healthcare providers powerful AI-driven tools to enhance medical imaging workflows. By enabling precise segmentation, synthetic data generation, and AI-powered diagnostics, MONAI addresses some of the most pressing challenges in modern healthcare, from privacy concerns to the need for scalable, high-performance infrastructures.

The application of MONAI extends across a wide range of medical use cases, with customization and fine-tuning for specific tasks, such as whole-body segmentation, drug effect tracking, and rapid anatomical mapping. MONAI's ability to scale through GPUDirect Storage on Pure Storage FlashBlade allows for the seamless handling of large datasets, faster inferencing, and superior reliability. Below is a breakdown of key applications.

Whole-body Segmentation for Drug Tracking (VISTA-3D)

Pharmaceutical companies are increasingly relying on medical imaging to track the effects of drugs on multiple organs across the human body. VISTA-3D's "Segment Everything" workflow allows for comprehensive body exploration, which is crucial when assessing drug efficacy and side effects in a holistic manner. By leveraging this feature, researchers can monitor how treatments affect different organs over time, providing deeper insights into the systemic effects of drugs. The ability to apply this to animal studies further enhances its utility, allowing biopharma companies to segment and analyze animal imaging data quickly, providing a faster path to understanding drug efficacy before moving to human trials.

AI-enhanced Imaging Workflows

Providers seek faster, more accurate segmentation in clinical workflows. VISTA-3D's interactive point-prompt segmentation accelerates the creation of ground-truth data, enabling clinicians to refine AI-generated images with precision. This is especially useful for radiology departments needing quick and accurate segmentations of organs or tumors for diagnostic purposes. By integrating AI into imaging workflows, healthcare providers can reduce the time required to process large datasets, allowing faster diagnostics and treatment planning while decreasing physician workload.

Fine-tuning for Animal Studies and Research Clinics

Fine-tuning the VISTA-3D model is particularly important in the context of animal studies, such as creating a detailed animal atlas. For instance, a microCT mouse dataset is critical for research clinics that need faster segmentation tools to map anatomical structures accurately. VISTA-3D's fine-tuning capabilities, combined with MONAI's adaptable framework, allow for the rapid adaptation of models to specific anatomical needs in preclinical research, speeding up the research-to-clinic pipeline. This is particularly useful for translational research where animal data is being used to predict human outcomes.

Synthetic Data Generation and Augmentation (MAISI)

MAISI (Medical AI for Synthetic Imaging) enables the generation of high-quality synthetic CT images to address data scarcity, privacy concerns, and the need for diverse training datasets. Many customers face difficulties running MAISI due to its significant GPU requirements; however, once deployed on enterprise-level infrastructure like Pure Storage FlashBlade, MAISI delivers exceptional results. By generating synthetic images annotated with up to 127 anatomical classes, researchers can overcome the limitations of small or proprietary datasets, making it easier to train AI models for more accurate predictions and diagnostics.

Image-to-image Translation (CT to MRI)

One of the growing demands in medical imaging is translating data from one modality to another, particularly from CT to MRI, which provides more detailed soft-tissue information. Conditioning MAISI on age and other patient-specific factors can generate synthetic MRI images from existing CT data. This approach opens up possibilities for using CT data to predict MRI results, providing a non-invasive, low-cost way to acquire additional diagnostic information without needing further imaging procedures.

Business Impact and Market Analysis

The healthcare industry is rapidly adopting AI-driven solutions, particularly in the area of medical imaging. The growing demand for precision medicine, faster diagnostic capabilities, and AI-augmented workflows has driven the adoption of advanced AI frameworks like MONAI, VISTA-3D, and MAISI. The global medical imaging market is projected to exceed $45 billion by 2027, with AI-driven imaging systems representing the fastest-growing segment. The combination of deep learning models for 3D medical image analysis and automated workflows to reduce clinician workload is becoming a strategic priority for healthcare organizations.

FlashBlade offers a clear advantage for organizations looking to scale their AI workloads. Its ability to handle complex datasets provides a scalable, reliable, and cost-efficient infrastructure solution that aligns with the demands of the growing medical imaging AI market.

By integrating with NVIDIA's MONAI frameworks, FlashBlade enables healthcare organizations to achieve real-time processing, synthetic data generation, and AI model fine-tuning without sacrificing performance.

Market Needs

As healthcare organizations increasingly require AI systems that deliver faster inferencing, greater scalability, and enhanced data privacy, FlashBlade's ability to efficiently support enterprise-level workloads is a clear differentiator:

- **Faster inference times**: Healthcare providers need systems capable of handling large 3D medical datasets and delivering real-time inferencing to support diagnostic workflows. FlashBlade, with its high throughput and scalability, enables faster image processing times, meeting the demand for time-critical diagnostics.
- **Scalability and flexibility**: As medical imaging needs grow, organizations require systems that scale with increasing dataset sizes and model complexity. FlashBlade offers a scalable architecture capable of supporting both single-GPU and multi-GPU configurations, ensuring organizations can grow without losing performance.
- **Data privacy and synthetic data generation**: With stricter data privacy regulations, organizations are looking for solutions that generate synthetic data to reduce reliance on patient data. The high-performance architecture in FlashBlade systems complement AI frameworks like MAISI, which excel in synthetic data generation, allowing for privacy-preserving AI models without compromising model performance.
- **Cost efficiency and operational simplicity**: Healthcare organizations operate under budget constraints and require systems that maximize throughput while minimizing overhead. FlashBlade supports cost-effective scaling and simplified management and reduces infrastructure costs, enabling organizations to optimize their AI workflows efficiently.
- **Seamless integration**: FlashBlade's support for multiprotocol access, including NFS and object storage, allows it to integrate with legacy systems while maintaining compatibility with cutting-edge AI tools. This ensures organizations can modernize their AI workflows without overhauling existing infrastructure.

Technical Requirements

VISTA-3D

Hardware

- Minimum GPU memory (GB): 48GB
- A single NVIDIA GPU of Ampere or Hopper architecture.
- CPU x86-64 >= 8 core (Recommended)
- Memory >= 16GB (Recommended)
- Minimum Storage: 20GB (8GB container size)

Software

- Minimum NVIDIA Driver Version: 470
- Docker 24.0.7
- NVIDIA Container Toolkit

MAISI

Hardware Requirements

- Minimum GPU memory (GB): 60GB for generating images of size 512^3 or larger
- CPU x86-64 >= 8 core (Recommended)
- Memory >= 32GB (Recommended)
- Minimum Storage: 50GB (28GB container size)

Software Requirements

- Minimum NVIDIA Driver Version: 535
- Docker 24.0.7
- NVIDIA Container Toolkit

Containerization

Both VISTA-3D and MAISI NIM are containerized using Docker, ensuring that they can be reliably deployed across different environments, whether on-premises or in the cloud. This containerization guarantees reproducibility and simplifies the deployment process, making it easier for healthcare organizations to scale their AI-driven medical imaging workflows while ensuring consistent performance across diverse infrastructure setups.

Enterprise Stack

Table 1 outlines the components and functionality of the reference architecture.

Component	Functionality
NVIDIA AI Enterprise	End-to-end AI platform that accelerates data science pipelines and streamlines development and deployment of production-grade co-pilots and other generative AI applications.
NVIDIA Inference Microservice	Designed to bridge the gap between the complex world of AI development and the operational needs of enterprise environments, enabling 10-100X more enterprise application developers to contribute to AI transformations of their companies.
NVIDIA MONAI	The MONAI framework is a suite of programming tools and APIs developed specifically for AI-driven medical imaging. It provides a robust platform for building, training, and fine-tuning deep learning models for healthcare applications.
NVIDIA GPU Operator	Lifecycle management of software required to use GPUs with Kubernetes.
Docker	Container platform
NVIDIA DGX™ A100 / NVIDIA DGX H100 * if MAISI is not needed OVX servers with L40 GPUs will also work	Compute server(s)
Pure Storage FlashBlade//S500, 1-10 chassis	Storage
NVIDIA Spectrum™ Series Switches SN3700	Network

TABLE 1 Enterprise stack components.

Recommended Hardware Architecture and Design

We recommend the following for architecting the solution:

- **Computing resources**: NVIDIA DGX systems and NVIDIA OVX™ systems (if not needing MAISI) deliver the necessary computational power for AI workloads, supporting analytics, training, and inference operations.
- **FlashBlade storage layers**: FlashBlade//S500 provides a unified storage platform for NFS, S3, and SMB protocols, supporting large-scale vector databases and facilitating high-speed data retrieval. On a FlashBlade system, objects can be natively accessed via the S3 storage protocol, and simultaneously files can be accessed natively via the NFS or SMB storage protocols.

Tested Configuration and Scaling

Proper sizing and tuning of GPU and storage resources are critical for optimizing performance. In our testing and design of this Pure Storage and NVIDIA integrated solution, we used a singleDGX system (NVIDIA OVX systems are also supported with NVIDIA MONAI), a 3 chassis FlashBlade//S500 storage system, and an NVIDIA Spectrum™-2 SN3700 switch. Additional DGX or OVX systems can be added and up to 10 chassis are supported in a FlashBlade//S500 for greater scale.

- **Server—NVIDIA DGX or NVIDIA OVX systems**: NVIDIA DGX systems and NVIDIA OVX systems are powerful and versatile systems purpose-built for all AI infrastructure and workloads, ranging from analytics to training and inference.
- **Storage—1 to 10 chassis FlashBlade//S500**: FlashBlade//S 500 excels at energy efficiency, scalability, and multi-modal performance and is designed to handle unstructured data efficiently. The modular architecture of FlashBlade//S allows you to independently scale capacity and performance. This flexibility ensures greater efficiency and minimizes waste. You can adjust it to meet your growing needs and projections.
- **Network switch—NVIDIA Spectrum-2 SN3700**: The SN3700 enables connectivity to endpoints at different speeds and carries a throughput of 6.4Tb/s, with a landmark 8.33Bpps processing capacity. As an ideal spine solution, the SN3700 allows maximum flexibility, with port speeds spanning from 10GbE to 200GbE per port.

Software Stack with NIMs

We recommend the following software stack:

- NVIDIA MONAI
- NVIDIA VISTA-3D NIM
- NVIDIA MAISI NIM
- Pytorch 2.3.0 for deployment of models and GPU synchronization.
- Python 3.12.1 for running the scripts.
- Ubuntu 22.10 for the Linux environment.

Deployment Requirements

Below are the architecture's deployment requirements:

- **Compute**: NVIDIA DGX A100, DGX H100, DGX H200 or DGX B200 or OVX systems
- **Storage**: FlashBlade//S500, scalable from one to ten chassis
- **Network**: NVIDIA Spectrum SN3700 switches for high-throughput connectivity

Deployment Steps

For a successful deployment, the following steps need to be carried out:

- **Infrastructure setup**: Deploying and configuring servers, storage, and network components
- **Docker deployment**: Setting up a Docker container with NVIDIA GPU Operator
- **Software installation**: Installing KX Vector DB, NVIDIA NeMo™ Microservices, and other necessary software components
- **Configuration and tuning**: Fine-tuning settings to optimize performance for specific workloads

Network Setup and Configuration

In our testing, the Pure Storage FlashBlade system was connected to the NVIDIA DGX A100 via a high-speed 200GB Ethernet network. The system utilized NFS v3 as the file-sharing protocol, which is optimized for Linux environments. This setup allowed for a highly efficient transfer of data between the GPUs and the FlashBlade system, minimizing latency and maximizing throughput. The use of NFS over 200GB Ethernet ensured that FlashBlade could seamlessly handle the high bandwidth demands of AI workloads, particularly in multi-GPU configurations, where data transfer bottlenecks can significantly affect performance. This network-attached storage (NAS) configuration not only provided superior performance over local disk but also delivered the scalability and flexibility needed for enterprise AI environments.

Architecture Objectives

Technical Goals

- **Scalability**: Support the growing volume of 3D medical imaging data and increasingly complex models in a healthcare environment.
- **Reliability**: Ensure consistent performance and uptime across both model inferencing and synthetic data generation.
- **Performance**: Maximize throughput, particularly for large volumetric image sets, leveraging GDS to minimize bottlenecks.
- **Data privacy**: Generate synthetic data to reduce reliance on patient data, addressing privacy concerns and regulatory compliance.
- **Integration**: Seamlessly integrate into existing infrastructure while supporting future scaling and advanced AI workflows.

Key Requirements

- **High-performance storage**: Use Pure Storage FlashBlade for fast, scalable data access during model inferencing and training.
- **AI-driven workflows**: Enable deep learning workflows for model inferencing (VISTA-3D) and synthetic data generation (MAISI).
- **GPU acceleration**: Maximize the use of NVIDIA DGX GPUs, including multi-GPU setups, to ensure high efficiency in both single-node and distributed training scenarios.
- **Compliance**: Ensure the solution complies with healthcare industry regulations, particularly around data security and privacy.

High-level Architecture Diagram

This architecture integrates NVIDIA DGX accelerated computing systems and advanced network switches with Pure Storage FlashBlade platform to ensure seamless data flow between storage and compute resources. The diagram highlights how data is moved efficiently between these components, supporting tasks like AI model training, real-time inferencing, and synthetic data generation.

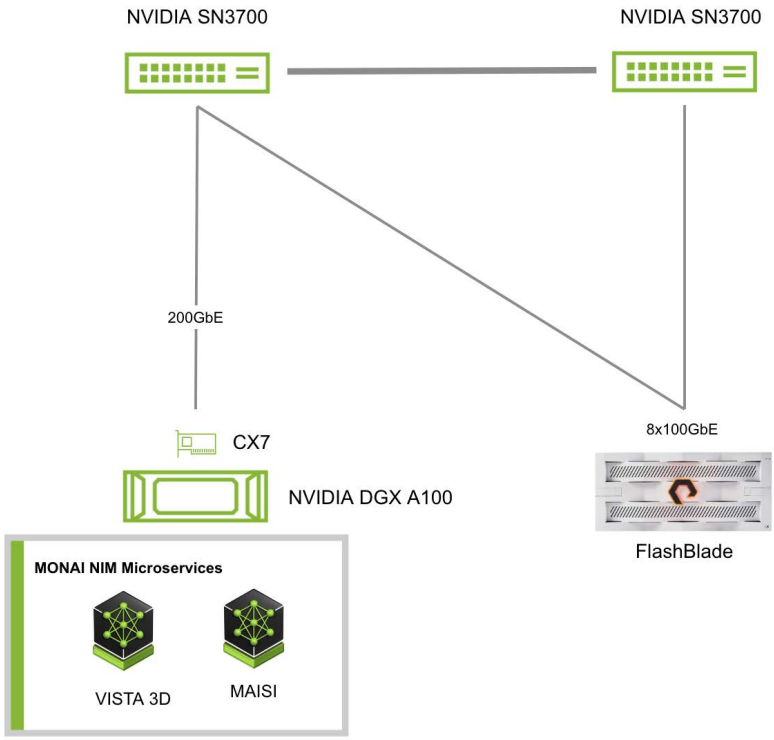

FIGURE 6 High-performance AI infrastructure with NVIDIA and Pure Storage

At the heart of the system is the NVIDIA DGX, which handles intensive inferencing and training workloads. Connected through a CX7 200GbE link to NVIDIA SN3700 switches, the DGX ensures rapid data transfers for compute tasks. The 200GbE bandwidth minimizes latency and maximizes throughput, ensuring that the GPUs inside the DGX can continuously access the data they need without interruption. This design is optimized for high-performance AI applications, where real-time access to data is crucial for maximizing GPU utilization and minimizing idle time.

Data is stored in the Pure Storage FlashBlade, a high-performance, scalable storage solution connected to the system via 8×100GbE links. FlashBlade acts as the central repository for vast amounts of structured and unstructured data, including high-resolution 3D medical images or synthetic datasets used in training. Its parallel I/O capabilities ensure that multiple data streams can flow simultaneously, feeding data into the GPUs in real time. This architecture ensures that data bottlenecks are eliminated, even as the volume of data scales. FlashBlade is also GPUDirect Storage certified, meaning it is optimized for the unique demands of AI workloads, allowing direct access between the GPUs and storage for even greater performance efficiency.

The NVIDIA SN3700 switches serve as the network backbone, orchestrating the data movement between compute and storage. These switches, equipped with low-latency and high-bandwidth capabilities, manage the flow of data from FlashBlade to the DGX system. Whether training large AI models or running real-time inferencing, the switches ensure that data is moved efficiently and without delay, further reducing latency and enabling distributed training when needed.

Design Validation

We validated the MONAI framework using the Pure Storage FlashBlade system to ensure scalability, performance, and ease of use for enterprise-scale medical imaging workloads. Our validation setup included a Pure Storage FlashBlade//S500 connected to an NVIDIA DGX A100 system. FlashBlade was connected to the DGX system via a high-speed 200GB Ethernet network, using NFS v3 as the file-sharing protocol. This setup minimized latency and maximized throughput, allowing the system to handle AI-driven medical imaging workloads efficiently, particularly in multi-GPU configurations.

We performed several tests across single and multi-GPU configurations. These tests focused on both inferencing and fine-tuning tasks in medical imaging, leveraging the VISTA-3D model for 3D anatomical segmentation and the MAISI diffusion model for synthetic data generation.

The datasets used for testing consisted of two key sources: the Medical Decathlon dataset and a suite of synthetic images generated using the MAISI NIM. The Medical Decathlon dataset is an industry-standard benchmark for evaluating the performance of organ segmentation algorithms. It includes a variety of medical imaging modalities, including CT scans of different organs, such as the liver, lungs, and brain, each varying in size, resolution, and complexity. This diversity allowed us to thoroughly assess the ability of FlashBlade to handle real-world medical imaging workloads with varying image sizes and formats. Given that the Medical Decathlon dataset is widely used for organ segmentation tasks, its inclusion ensured that our testing closely mirrored actual clinical conditions, providing a rigorous foundation for evaluating the performance of the MONAI framework.

In addition to real-world imaging data, we also utilized a set of synthetic images generated by the MAISI NIM, a latent diffusion model designed specifically for medical imaging tasks. MAISI produced high-resolution synthetic 3D CT images with and without annotations, covering various voxel sizes and anatomical classes. This synthetic data allowed us to test how FlashBlade performs when handling large, complex image sets, particularly in inferencing pipelines. The ability to augment datasets with synthetic images from MAISI added a layer of complexity and realism to our testing, further validating the capability of FlashBlade to scale seamlessly for enterprise-level medical imaging workflows.

Inferencing Capabilities

We evaluated the inferencing performance of the VISTA-3D NIM using volumetric image sizes ranging from 256×256×256 to 512×512×768 that were generated from the MAISI NIM. Each image was processed on a single GPU, comparing throughput on both Pure Storage FlashBlade and local disk. Additionally, to assess scalability, we spun up a total of eight containers to measure the system's ability to handle simultaneous inferencing tasks. FlashBlade outperformed or matched local disk performance in almost every test, providing superior throughput and faster inference times.

To further stress the system, we ran the entire Medical Decathlon dataset on a single GPU. The dataset's varying image sizes simulated a real-world scenario where medical imaging data is not standardized, and the FlashBlade system's throughput consistently outperformed local disk, even when processing larger or more complex images. This demonstrated the system's ability to efficiently handle the diverse demands typical in medical imaging workflows.

Fine-tuning Pipeline

Using the open source version of the VISTA-3D model, we conducted fine-tuning experiments to train models for 20 epochs on 240 labeled images from the Medical Decathlon dataset. Fine-tuning was performed on both single and multi-GPU configurations, with FlashBlade once again demonstrating superior performance over local disk during these tasks.

We explored three data-loading methods available within MONAI to optimize the training process:

- **Regular dataset loading:** This standard method uses the MONAI Dataset function to load images into the model. Performance was baseline, providing a reference for further optimizations.
- **Dataset Caching**: Using MONAI's Cache Dataset, we cached data in the CPU, significantly improving loading times for subsequent epochs.
- **GDS Caching**: The most performant option was NVIDIA Magnum IO™ GPUDirect® Storage caching using NVIDIA MONAI GDS. By bypassing the CPU and caching data directly on the GPU, we observed dramatically faster loading times, particularly in multi-GPU setups, and near-linear scaling with FlashBlade.

Additionally, we conducted experiments to measure checkpoint saving times during fine-tuning on both local disk and FlashBlade. Checkpointing models is critical for ensuring that progress is saved periodically, mitigating the risk of losing valuable training time due to system interruptions or for testing specific versions of a model. Our tests demonstrated that the ability of FlashBlade to handle high-frequency I/O made it significantly more efficient at managing these checkpointing tasks compared to local disk. Model states were saved quickly and reliably on FlashBlade without impacting overall performance, making it an ideal solution for environments where frequent model checkpointing is required.

Performance and Scalability

Our tests consistently demonstrated the scalability of the Pure Storage FlashBlade system, particularly in scenarios requiring parallel processing. When running multiple containers simultaneously for inferencing tasks, FlashBlade scaled effectively, delivering the same consistent throughput, keeping all containers busy and matching the same inference times as a single container, highlighting its ability to handle large-scale medical imaging workloads without bottlenecks.

FlashBlade consistently outperformed local disk during fine-tuning tasks, particularly in multi-GPU training and checkpointing jobs. Its high-performance capabilities ensured that even the most demanding AI workloads, such as large-scale model fine-tuning, were handled efficiently, leading to reduced training times and faster insights for clinicians and researchers.

Throughout our testing, FlashBlade seamlessly scaled across multiple GPUs and workloads, demonstrating robust and consistent performance across varying data sizes. This scalability, combined with the platform's ability to handle both fine-tuning and inferencing tasks in parallel, makes FlashBlade an optimal solution for AI-driven medical imaging at an enterprise scale.

Operational Simplicity and Efficiency

The combination of MONAI and FlashBlade not only delivers exceptional performance but also simplifies operational management. The modular architecture of FlashBlade allows for independent scaling of both storage capacity and performance, enabling healthcare organizations to tailor their infrastructure to their current and future needs. This scalability, paired with GDS optimizations, positions FlashBlade as a top-tier solution for handling the increasing demands of medical imaging and AI.

Additionally, the distributed metadata architecture in FlashBlade and DirectFlash modules contribute to its operational simplicity, reducing the complexity often associated with managing high-performance infrastructure. As a result, teams can focus more on refining their AI models and less on managing their storage and data pipelines.

Results

In our testing, the Pure Storage FlashBlade system, integrated with NVIDIA DGX A100 and MONAI, demonstrated exceptional performance across both inferencing and fine-tuning tasks. Across both single and multi-GPU tests, FlashBlade consistently outperformed local disk in terms of throughput, reliability, and scalability. The ability of FlashBlade to manage large datasets, including the Medical Decathlon and MAISI-generated synthetic images, resulted in faster processing times and lower performance variability. Its network-attached architecture allowed it to scale effectively across multi-GPU configurations, maintaining high throughput as workloads increased. In contrast, local disk exhibited higher performance variability, particularly with larger datasets and in multi-GPU setups.

The testing covered a range of volumetric image sizes, including the entire Medical Decathlon dataset, and involved various data loading techniques within MONAI, such as regular loading, CPU caching, and GPUDirect Storage (GDS). FlashBlade proved to be a robust, high-performance solution, handling diverse medical imaging workloads efficiently, with bottlenecks arising only from compute limitations, not storage. Scaling the computational resources in the infrastructure would further enhance the system's capabilities for enterprise-scale AI in medical imaging.

Inferencing

In our inferencing tests, we evaluated the performance of the VISTA-3D NIM across a variety of image sizes, comparing throughput between Pure Storage FlashBlade and local disk on both single and multi-GPU configurations. FlashBlade consistently delivered faster and more reliable inferencing, especially when handling large, complex datasets such as the Medical Decathlon images, demonstrating its ability to efficiently support enterprise-scale medical imaging workloads. The results highlight the performance benefits of using NAS fabric storage over local storage.

Single GPU

When running on a single GPU, Pure Storage FlashBlade consistently demonstrated faster and more reliable performance than local disk across various medical imaging workloads. The graphs (Figure 7) and table (Table 2) below illustrate the inference times for a set of MAISI-generated synthetic images, where the x, y values (sagittal and coronal axes) were held constant while varying the z (axial) dimension. FlashBlade maintained a more consistent throughput, as seen by the sharper peaks in the histograms, compared to local disk, which exhibited greater variability in inference times. In some instances, FlashBlade outperformed local disk by up to 40%.

The consistent performance of FlashBlade becomes especially important when handling larger image sizes and complex datasets. While the local disk struggled with variability in inference times, FlashBlade provided stable and predictable performance, highlighting its ability to handle real-world medical imaging workloads efficiently.

For larger image sizes like 512×512×512, FlashBlade demonstrated nearly 12% faster inference times than local disk, which becomes crucial when processing complex 3D medical datasets. This superior performance at scale makes FlashBlade an optimal solution for handling demanding medical imaging tasks.

In the accompanying histograms, the sharper peaks for FlashBlade compared to the broader spread for local disk illustrate the system's consistent throughput across various image sizes. This consistency becomes especially important in medical workflows, where predictable performance can significantly reduce processing time.

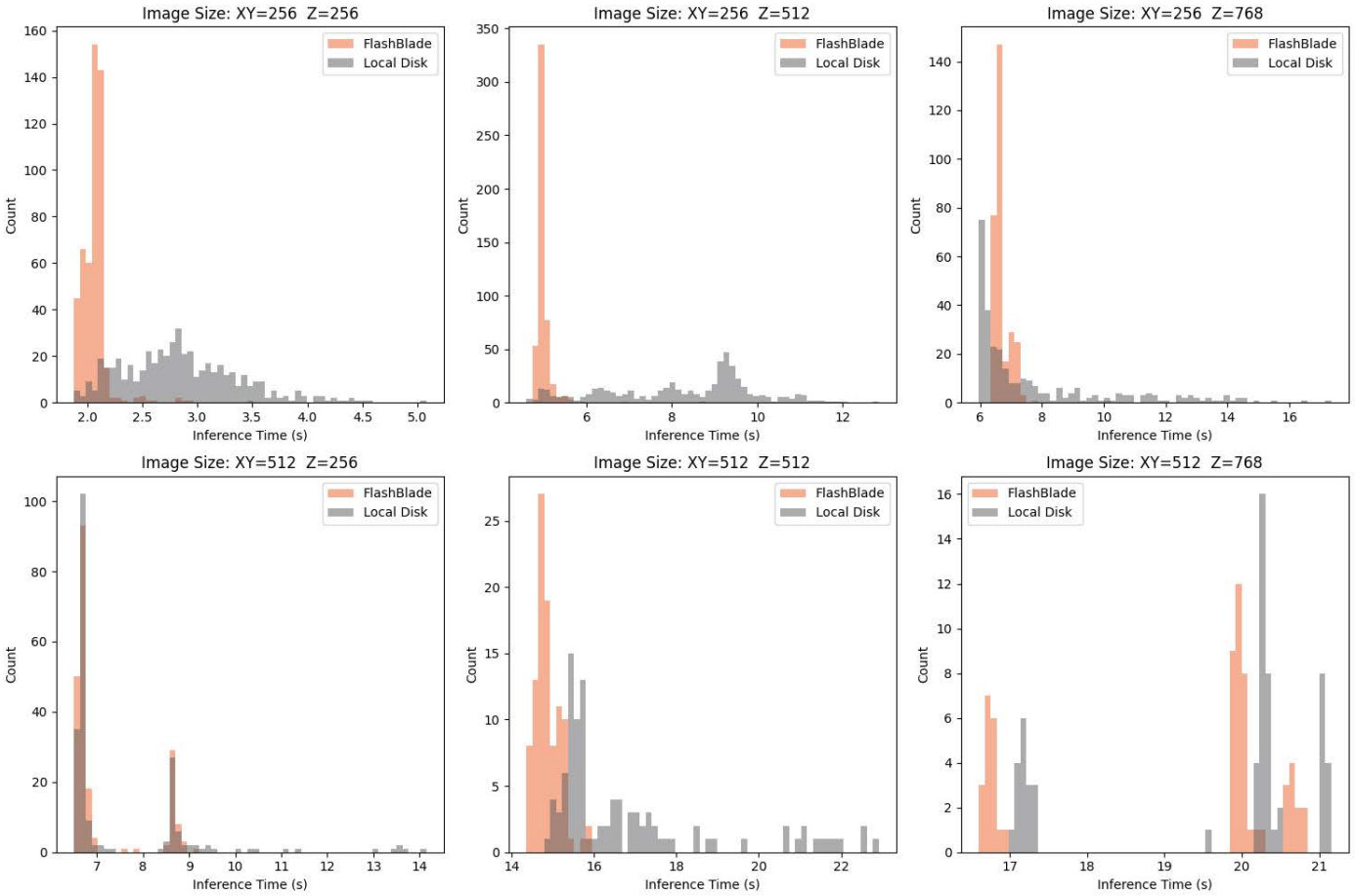

FIGURE 7 Inference times for a set of MAISI-generated synthetic images

Image Size	Images	Image Avg (MB)	FlashBlade Avg (s)	FlashBlade Std Dev(s)	Local Disk Avg (s)	Local Disk Std Dev (s)
256×256×256	500	25.51	2.07	0.13	2.85	0.54
256×256×512	500	55.91	4.99	0.39	8.25	1.70
256×256×768	300	70.99	6.68	0.70	7.09	2.41
512×512×256	212	45.14	7.09	0.82	7.44	1.47
512×512×512	100	93.00	14.88	0.31	16.85	2.08
512×512×768	61	97.43	19.15	1.58	19.57	1.52

TABLE 2 MAISI Image Stats and Inference Times

We also evaluated the performance of FlashBlade using the Medical Decathlon dataset (Figure 8, Table 3), which was designed to simulate the diversity of real-world medical imaging environments. Unlike the MAISI-generated images with standardized dimensions, the Medical Decathlon dataset includes images of varying sizes, mimicking clinical conditions where data is less structured. Once again, FlashBlade outperformed local disk, particularly in terms of consistency and scalability.

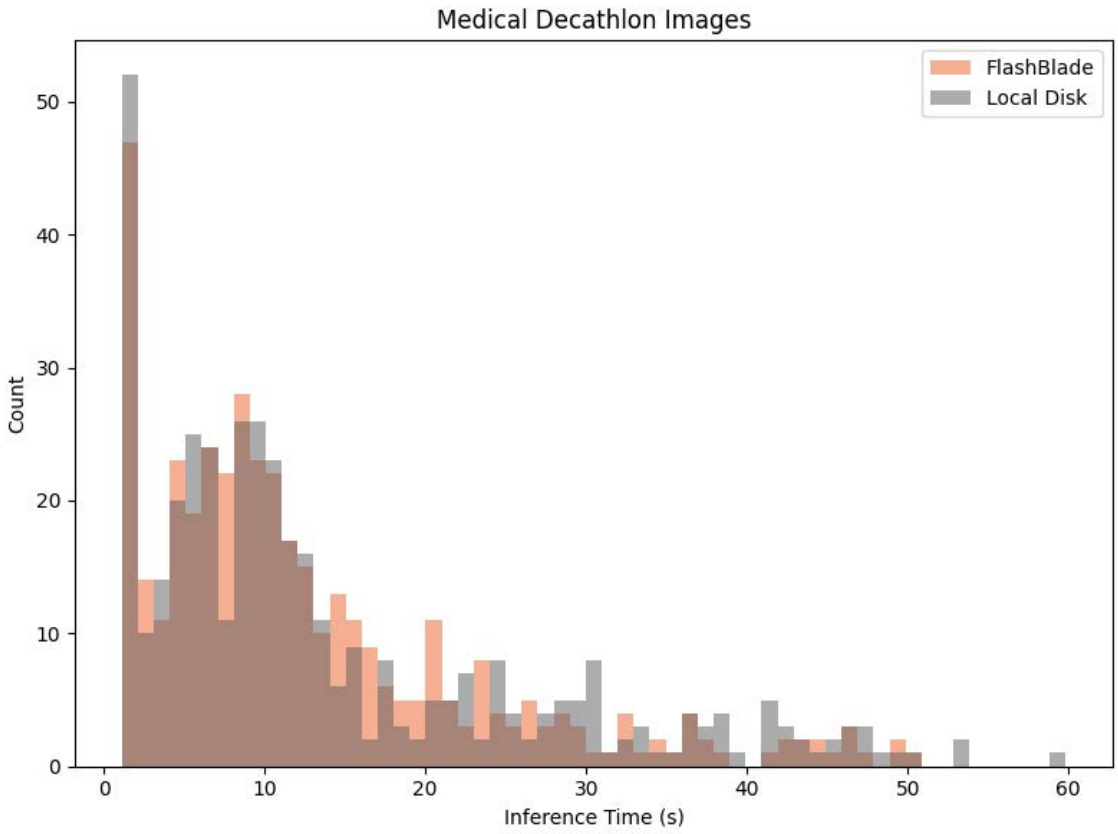

FIGURE 8 Medical Decathlon image counts and inference over time

Images	Image Size Avg (MB)	Min Image Size (MB)	Max Image Size (MB)	FlashBlade Avg (s)	FlashBlade Std Dev(s)	Local Disk Avg (s)	Local Disk Std Dev (s)
405	93.96	2.45	379.18	12.76	10.50	13.99	12.32

TABLE 3 Medical Decathlon Image Stats and Inference Times

Multi-GPU

To evaluate how FlashBlade scales in multi-GPU environments, we conducted similar tests using all available GPUs on the NVIDIA DGX system. Due to the VISTA-3D NIM lacking a helm chart for Kubernetes deployment at the time of testing, we manually spun up one container per GPU, splitting the inference workload across all GPUs to assess performance drop-offs, if any. The multi-GPU results showed that the FlashBlade system continued to provide consistent performance, standing up to the challenge of pushing more data through the network compared to local disk.

While the single GPU tests highlighted the FlashBlade system's superiority in terms of throughput and consistency, the multi-GPU results confirm its ability to scale linearly. By using multiple GPUs in parallel, we can expect even greater performance gains in larger, multi-node setups, where FlashBlade's ability to handle large-scale, high-throughput workloads will further shine. The tests demonstrated that FlashBlade scales efficiently, providing a reliable and high-performance backbone for AI-driven medical imaging workloads, whether on a single or multi-GPU configuration. In multi-GPU environments, FlashBlade efficiently scaled as expected, ensuring data throughput kept pace with increased GPU demands. This capability is vital for imaging teams processing large volumes of scans in parallel, as it ensures that adding more compute power results in proportional reductions in processing time. As organizations grow and expand to larger, multi-node setups, FlashBlade's ability to manage high-throughput, data-intensive workloads will continue to deliver significant performance gains.

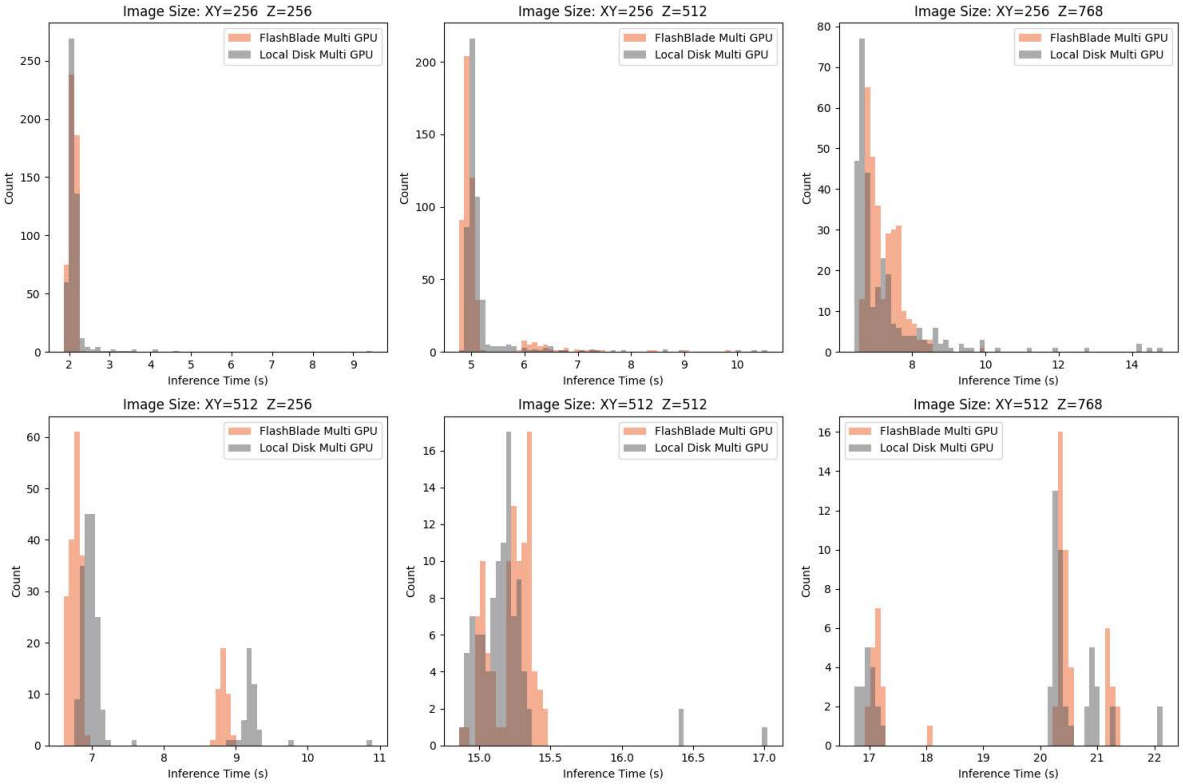

FIGURE 9 Multi-GPU inference times

Image Size	FlashBlade Avg (s)	FlashBlade Std Dev(s)	Local Disk Avg (s)	Local Disk Std Dev (s)
256×256×256	2.09	0.07	2.15	0.41
256×256×512	5.10	0.56	5.19	0.61
256×256×768	7.22	0.45	7.24	1.21
512×512×256	7.18	0.83	7.44	0.93
512×512×512	15.22	0.14	15.18	0.28
512×512×768	19.60	1.60	19.52	1.69

TABLE 4 MAISI image stats and inference times

Fine-tuning Workflows

In the fine-tuning workflows, we focused on training the VISTA-3D model using labeled data from the Medical Decathlon dataset, testing both single and multi-GPU configurations on Pure Storage FlashBlade and local disk. To optimize performance, we evaluated three different data loading methods available in MONAI: regular dataset loading, CPU caching, and GPUDirect Storage. FlashBlade consistently outperformed local disk across all workflows, demonstrating superior scalability and efficiency in handling large-scale fine-tuning tasks. This enabled faster training times and seamless model adaptation, crucial for high-performance AI-driven medical imaging pipelines.

While FlashBlade demonstrated significant performance improvements in both single and multi-GPU configurations, it's important to note that some bottlenecks we observed were compute-bound rather than storage-bound. In single GPU setups, the GPU's resources were taxed by handling both the model and data, which limited the potential benefits of GDS Dataset. However, in multi-GPU configurations, the ability of FlashBlade to bypass the CPU and deliver data directly to each GPU alleviated these bottlenecks, making it the ideal solution for scaling across multiple GPUs.

MONAI Dataset

The regular dataset loader, which uses no caching, is the standard method for loading data in most deep learning pipelines. It sequentially loads data from storage to the model, providing a straightforward and reliable process for training and inferencing tasks. While effective, it tends to be slower compared to more optimized methods, as it relies on fetching data from storage at each epoch without leveraging any intermediate caching techniques. Even when pulling data from ultra-performant storage systems like FlashBlade, the bottleneck remains in passing the data to the model on the GPU, limiting overall processing speed.

MONAI Cache Dataset

The MONAI Cache Dataset is built to speed up repetitive tasks by temporarily storing data that has already been processed. For example, if certain image adjustments or preprocessing steps are needed multiple times, this dataset will cache those changes so they don't have to be recalculated each time. This significantly speeds up the overall workflow, especially during training that spans many iterations, by reducing the need to re-process data with each run. It's a practical solution for ensuring faster training times when working with large medical imaging datasets.

MONAI GDS Dataset

The MONAI GDS Dataset is designed to improve the speed at which data moves between storage and the GPU, which is essential in large-scale imaging projects. It bypasses the CPU altogether, allowing data to move directly from storage to the GPU, minimizing delays and improving overall performance. This method is ideal for cases where speed is critical, such as when working with large datasets for real-time analysis or training AI models. This approach ensures that the system can handle heavy workloads without being slowed down by traditional data transfer methods.

Dataset Loader	Best for	Advantages	Drawbacks
Regular Dataset	Baseline	Standard data loading process; Simple to implement	Slow, reliant on CPU for data loading; Not optimized for multi-GPU setups
Cache Dataset	Single-GPU	Caches data in CPU memory, improving efficiency for smaller workloads	Bottlenecks in multi-GPU environments due to CPU-to-GPU data transfer
GDS Dataset	Multi-GPU	Direct data access from storage to GPU, bypassing CPU; Faster processing times in large-scale setups	GDS Dataset may be overkill for smaller datasets where the performance gains are less noticeable.

TABLE 5 Use cases, advantages, and drawbacks of specific datasets

Single-GPU Fine-tuning Results

The results from the single GPU fine-tuning runs highlight some interesting trends regarding the different dataset loading techniques. While the GDS Dataset is typically designed for faster data transfers by bypassing the CPU, the Cache Dataset method showed slightly better performance in this case. This is likely due to the nature of single GPU workloads, where the GPU itself is responsible for handling the model, parameters, and data simultaneously. In this scenario, the Cache Dataset, which caches data in CPU memory, allows the GPU to focus more on processing without being bottlenecked by storage I/O, as it benefits from a more stable and predictable data feed. The additional data handling by the num_workers configuration in the data loading likely improved the efficiency of the loading, further boosting the performance.

On the other hand, the GDS Dataset method, while still performing well, did not provide the same gains in this single GPU setup. The reason for this is that the GPU's resources are already stretched by handling the model and parameters, and so the additional speed advantage offered by direct GPU storage access is not fully utilized in this context. GDS Dataset shines more when scaling to multiple GPUs, where the ability to feed large amounts of data directly to each GPU in parallel provides a significant speedup, as demonstrated in the multi-GPU tests below.

When running on a single GPU, FlashBlade consistently matched the performance of local disk, particularly in terms of throughput consistency. In single GPU workloads, the Cache Dataset method allowed the GPU to focus on processing the model and parameters, avoiding bottlenecks caused by data transfer from storage. This method is well-suited for single GPU runs, particularly when working with datasets like the Medical Decathlon, where the ability of FlashBlade to provide stable and reliable throughput is critical as image sizes vary.

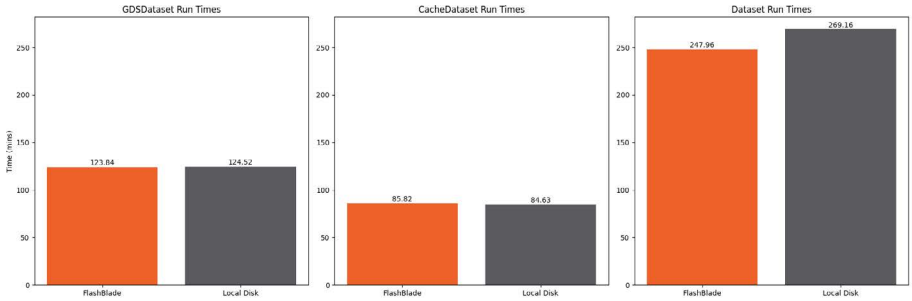

FIGURE 10 Data set runtimes compared in different environments with single GPU

Multi-GPU Fine-tuning Results

In the multi-GPU fine-tuning tests, GDS Dataset emerged as the clear leader in performance, significantly outperforming both Cache Dataset and the standard dataset loader. In these scenarios, the benefits of GDS Dataset became evident as it was able to take full advantage of GPUDirect Storage, directly streaming data from the FlashBlade storage to the GPU memory without relying on the CPU for intermediate data transfers. This efficiency is especially noticeable in multi-GPU setups, where the ability to provide direct, high-bandwidth access to each GPU leads to considerably faster data processing. The performance of FlashBlade in these GDS Dataset runs highlights its ability to support large-scale, data-intensive workloads, achieving speed and consistency that is critical in high-performance AI-driven medical imaging environments.

In comparison, Cache Dataset—while still showing solid performance—lags behind in multi-GPU environments. This is largely because caching data in CPU memory introduces a bottleneck when working with multiple GPUs, as it must manage data movement from the CPU to each GPU. Although Cache Dataset works well for single GPU setups by optimizing data transfer between CPU memory and the GPU, it becomes less efficient as the number of GPUs increases. The advantage of GDS Dataset in these multi-GPU runs showcases how the FlashBlade network-attached architecture effectively eliminates these bottlenecks, resulting in faster, more efficient fine-tuning across all available GPUs.

What is particularly important here is that FlashBlade consistently meets or exceeds the performance of local disk, while offering substantial enterprise-grade benefits that go beyond raw speed. FlashBlade provides scalability, reliability, and simplified management, ensuring that it can handle the growing data demands of large healthcare organizations without disruption.

In multi-GPU configurations, the performance advantage of FlashBlade became even more pronounced. Using GDS Dataset, FlashBlade could bypass traditional CPU bottlenecks, delivering data directly from storage to the GPUs, resulting in faster processing times and near-linear scaling. The ability of FlashBlade to handle multiple high-bandwidth data streams without losing performance ensures it's the superior choice for multi-GPU configurations, particularly in large-scale medical imaging tasks.

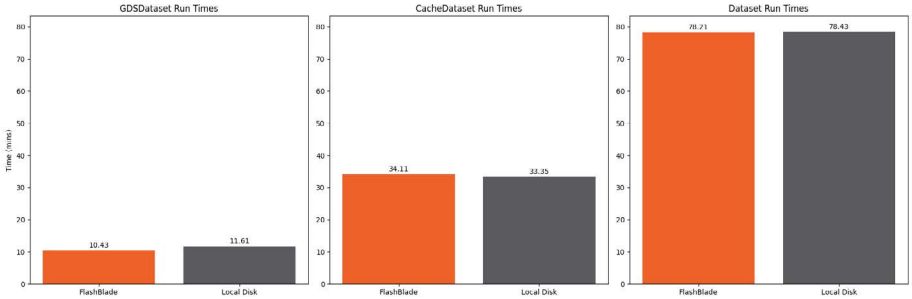

FIGURE 11 Data set runtimes compared in different environments with single GPU, with multi GPU

Checkpointing Times

The checkpoint saving times for both FlashBlade and local disk, as shown in Figure 12 below, indicate that while the average checkpoint times are roughly the same, FlashBlade offers significantly better consistency and reliability. The FlashBlade system's save times are more tightly clustered between 1.5 and 3.5 seconds, while local disk exhibits a much wider distribution, with more variability and longer save times, extending up to 4.5 seconds. This variability in local disk performance can result in less predictable training durations, particularly when frequent checkpointing is required.

Additionally, the longer checkpoint saves seen in both FlashBlade and local disk are primarily associated with multi-GPU runs, which require more coordination between GPUs to write model states to disk. Despite this added complexity, FlashBlade provides a more consistent and reliable checkpointing process with less variability in save times. This stability ensures that checkpointing is completed efficiently, reducing interruptions in the training workflow and offering a clear advantage in multi-GPU environments where consistent performance is crucial.

The average time to save a total of 500 model checkpoint for both FlashBlade and local disk was 2.9 seconds with FlashBlade having a standard deviation of 0.89 seconds and local disk having a standard deviation of 1.08 seconds. Each checkpoint was 830 MB in size.

Checkpointing is an essential process in long-running model training jobs, especially for medical imaging tasks where model fine-tuning can take several hours or even days. Frequent checkpointing helps safeguard progress and ensures that work isn't lost in the event of a system failure. The superior I/O performance of FlashBlade ensured that checkpointing was both faster and more consistent than with local disk, making it ideal for enterprise AI workloads that require reliability and efficiency. This advantage becomes even more critical in multi-GPU configurations, where coordination between GPUs to write model states can introduce variability in save times.

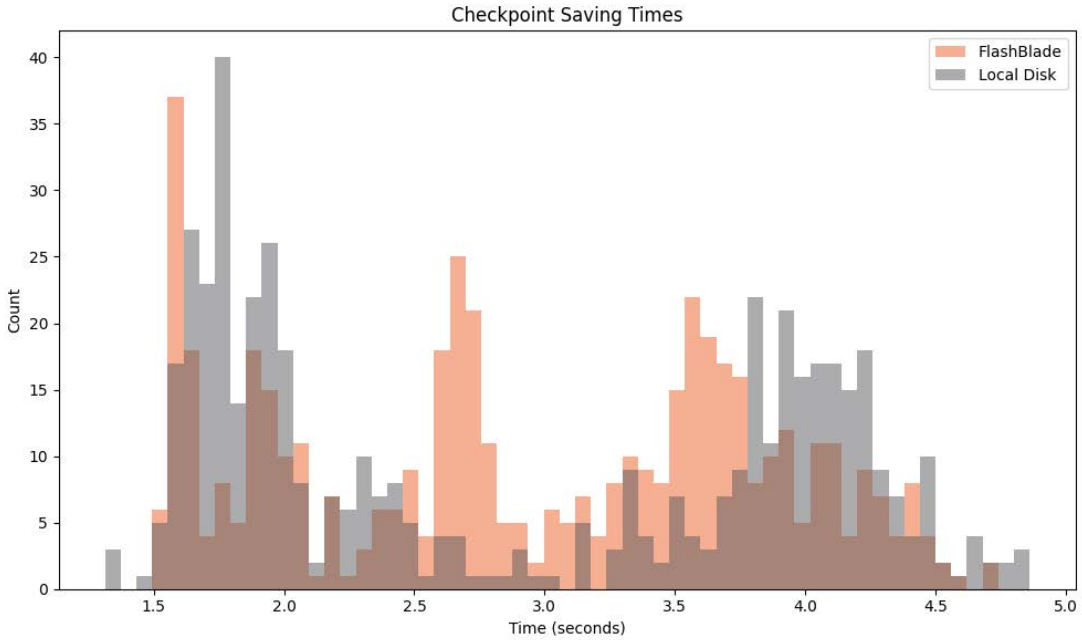

FIGURE 12 Checkpoint time savings with scaleout parallelism

Resource Utilization

One of the key benefits of using FlashBlade for AI-driven medical imaging is its ability to maximize resource utilization on the NVIDIA DGX system. By decreasing the inference times, we were able to increase the utilization of the GPUs by up to 5%. This efficiency allowed for more consistent data flow to the GPUs, reducing idle time and ensuring that computational resources were fully leveraged. As a result, the overall throughput of the system improved, making it possible to handle more workloads without additional hardware investments.

Compute-bound vs Storage-bound Workloads

It is important to note that many of the bottlenecks observed during our testing were compute-bound rather than storage-bound. This means that the limitations in performance were primarily due to the capacity of the GPUs to process the model and data, rather than FlashBlade's ability to deliver data quickly enough. FlashBlade's high throughput ensures that data is fed to the GPUs without delay, but in scenarios where the GPUs are already fully utilized, the storage system is not the limiting factor. For example, Flashblade has shown to deliver XYZ Gbps via MLPerf

For instance, in multi-GPU fine-tuning tasks, the GPUs were operating at their full capacity, and the performance gains from FlashBlade were most evident in situations where the data transfer rates could keep up with the processing power of the GPUs. However, further performance improvements in such cases would require additional compute resources (e.g., more GPUs or distributed training across multiple nodes), not improvements in storage speed. The results underscore that FlashBlade is more than capable of handling the data needs of AI-driven medical imaging workloads, and scaling up compute resources would unlock even greater overall performance.

In addition to raw performance improvements, FlashBlade offers enterprise-grade benefits that go beyond just speed. It provides scalability, reliability, and simplified management, making it a robust solution for organizations that need to scale their AI workloads without disruption. Features such as data protection, high availability, and seamless scaling make FlashBlade ideal for healthcare environments where data integrity and consistent performance are critical. While local disk may suffice in some scenarios, FlashBlade delivers a more comprehensive solution that can grow with the increasing demands of modern medical imaging pipelines.

Finally, in our checkpointing experiments, while the average save times were similar between FlashBlade and local disk, FlashBlade consistently demonstrated less variability in save times, ensuring more predictable checkpointing, particularly in multi-GPU environments. This reliability, coupled with FlashBlade's ability to efficiently manage high-frequency I/O tasks, makes it the preferred choice for workloads that require frequent checkpointing, where interruptions in training workflows can have costly consequences.

In summary, Pure Storage FlashBlade stands out as a critical enabler of high-performance, scalable, and reliable AI workloads in medical imaging. Its integration with multi-GPU configurations and ability to manage large, complex datasets make it an ideal solution for organizations aiming to optimize their AI-driven imaging workflows at enterprise scale.

Conclusion

The results of our tests demonstrate that Pure Storage FlashBlade provides substantial performance improvements and scalability for AI-driven medical imaging workloads, especially when compared to local disk. Integrated with NVIDIA DGX and the MONAI framework, FlashBlade consistently outperformed local disk across inferencing, fine-tuning, and data-loading tasks, regardless of whether single or multi-GPU configurations were used. The combination of FlashBlade's high throughput, scalability, and efficient handling of diverse datasets such as the Medical Decathlon and MAISI-generated synthetic images highlights its role as a critical infrastructure component for enterprise-scale AI workloads.

The superior performance of FlashBlade was especially notable in multi-GPU setups, where its ability to provide near-linear scalability in throughput helped ensure minimal performance drop-offs even as workloads increased. In inferencing tests, FlashBlade demonstrated faster, more reliable performance, particularly when handling larger and more complex medical datasets. The consistency of FlashBlade, as evidenced by the tighter peaks in inference time histograms, ensures that it can handle real-world medical imaging workflows with far greater predictability and reliability than local disk.

When scaling up to multi-GPU fine-tuning workflows, FlashBlade facilitated much faster data movement in multi-GPU configurations, enabling faster fine-tuning times. FlashBlade's network-attached architecture, coupled with GDS Dataset, ensures that data moves efficiently between storage and GPUs, making it the superior choice for fine-tuning large-scale MONAI models in multi-GPU setups.

Additional Resources

- Visit our Enterprise Imaging solution page to learn more about how we can help your organization.
- Learn more about the Pure Storage platform for AI
- Learn more about AI-ready Infrastructure (AIRI) and FlashBlade//S.

©2025 Pure Storage, the Pure Storage P Logo, DirectFlash, FlashBlade, FlashBlade//S, and the marks in the Pure Storage Trademark List are trademarks or registered trademarks of Pure Storage Inc. in the U.S. and/or other countries. The Trademark List can be found at purestorage.com/trademarks. Other names may be trademarks of their respective owners.

PS2699-02-en 02/25