WHITE PAPER

Pure Storage GenAI Solution for Drug Discovery with NVIDIA BioNeMo

A scalable AI solution for drug discovery using NVIDIA BioNeMo with Pure Storage®

Contents

Executive Summary/Introduction

The journey to develop and bring new drugs to market for pharmaceutical companies historically has been arduous, costly, and risky at all stages of the research and development (R&D) process. Growing economic, social, and political pressures over drug prices have pushed the industry into embracing fundamental changes to its approach to overcome these challenges. High research costs have plagued the industry for years. In 2022 the World Health Organization estimated that the average cost to develop a new drug ranges from US 43.4 million to 4.2 billion and continues to grow. Artificial intelligence (AI) and machine learning (ML) offer an opportunity to revolutionize the drug research and development journey from lab to market launch in a scalable way while driving clinical innovation and providing value to patients.

Designed to accelerate the building and training of models to use their own data and scale the deployment of models for drug discovery applications, NVIDIA BioNeMo is an AI platform that is specifically designed for the life sciences and healthcare industries. It leverages the power of large language models (LLMs) and the NVIDIA GPU architecture to advance research and applications in bioinformatics, genomics, drug discovery, and other areas of biomedical science.

This white paper introduces a scalable AI solution for drug discovery that highlights the functional, performance, and scalability advantages using NVIDIA BioNeMo with Pure Storage® for drug researchers at pharma and biotech companies. It simplifies and accelerates the training of models using your proprietary data, making it easier to develop and deploy AI models for biomolecular applications.

In addition, the paper also documents the scale-out architecture, deployment guidance, and benchmarking results for NVIDIA BioNeMo achieved with a NVIDIA DGX A100 system and Pure Storage FlashBlade//S500. We will outline a production-ready framework to use BioNeMo with shared scalable storage that accelerates time to science by enabling faster training using pre-trained models, a framework to implement scalable parallelism, and a holistic GPU accelerated framework which optimizes the use of NVIDIA GPUs.

Solution Overview

NVIDIA BioNeMo provides customizable AI models for various biomolecular tasks. These include predicting 3D protein structures, generating novel molecules, and optimizing molecular properties by using bioinformatics to help analyze genetic sequences, understand protein structures, and interpret biological data to accelerate drug discovery with cutting-edge models. By simulating molecular interactions and predicting compound behavior, BioNeMo accelerates the drug discovery process, leading to more efficient identification of potential therapeutic agents. Examples of some of the more popular models include protein and DNA sequence embeddings like ESM1nv, ESM2nv, ProtT5nv and, DNABERT; protein structures like OpenFold; novel small molecules (MegaMolBART); and molecular docking simulations like DiffDock and EquiDock.

BioNeMo training streamlines model development with seamless scaling that allows you to build enterprise-grade generative AI workflow, with your own data. It includes data loaders, validation loops, and enterprise support and is integrated with training frameworks for pre-trained models for protein sequences and models that understand chemistry and can be used in cheminformatics applications in drug discovery (MegaMolBart). NVIDIA BioNeMo enhances hyperparameter and checkpoint management, along with data and model parallelism, which enables near-linear scaling efficiency across hundreds of GPUs.

Pre-trained models for protein sequences (ESM-1nv), ProtT5nv (use cases) and MegaMolBart are models that understand chemistry and can be used in cheminformatics applications in drug discovery and are pre trained. NVIDIA BioNeMo enhances hyperparameter and checkpoint management, along with data and model parallelism. This enables near-linear scaling efficiency across hundreds of GPUs further speeding the time to science.

The Pure Storage solution for NVIDIA BioNeMo provides a scalable infrastructure for drug discovery with optimized performance and scalable infrastructure: Pure Storage FlashBlade® systems host the datasets (open and proprietary) and also provide a repository for the versioned models as well as output logs. FlashBlade is able to scale up performance to run multiple inference jobs in parallel, with a mix of training and inference workloads all in one GPU cluster. As training cycles get longer, the chances of disruption are higher. Pure Storage provides zero disruption storage infrastructure which will greatly reduce the cost of downtime for customers.

Technology Overview

The all-QLC architecture on FlashBlade//S™ systems deliver an optimized and cost-effective caching solution, crucial for the NVIDIA BioNeMo Framework. This architecture enhances performance while ensuring industry-leading efficiency per rack unit (RU), watt, and terabyte (TB). By providing a scalable and power-efficient solution, it empowers drug researchers to accelerate their time to science, ensuring they achieve breakthroughs faster.

The platform's unique modular architecture allows for independent scaling of compute and storage, which is vital in tailoring configurations to specific drug discovery workloads. This flexibility ensures consistent, high performance and adaptability to the ever-evolving demands of data-driven research. The platform's scalability is designed to accommodate the growing data needs inherent in drug discovery, maintaining performance integrity as data volumes expand.

Moreover, the multi-dimensional performance capabilities of the platform are indispensable for handling diverse and complex workloads, such as those in drug discovery. This versatility allows researchers to leverage the storage solution across various applications, including AI, machine learning, and high-performance computing, all crucial in modern drug research. The enhanced capacity and reliability offered by global media management on DirectFlash™ modules, which extracts up to 20% more capacity from NAND compared to competitors, further contribute to consistent performance and higher media endurance—key factors in supporting the demanding environments of drug discovery without relying on massive storage class memory (SCM) caches.

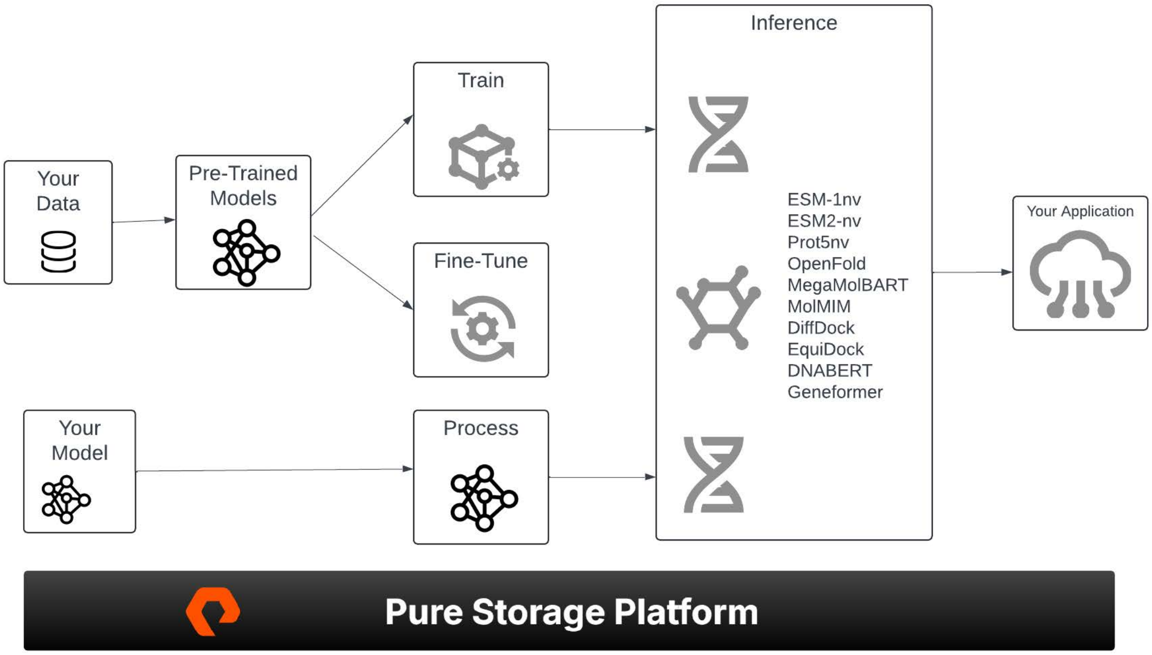


FIGURE 1 The Pure Storage platform

Technical Components of BioNeMo

The BioNeMo framework is a collection of programming tools and APIs used to access pre-trained biomolecular models and workflows, for building and customizing models for training and fine-tuning.

BioNeMo Framework

The NVIDIA BioNeMo framework is a comprehensive platform designed to advance AI-driven drug discovery by offering robust tools for model training and fine-tuning. It supports a range of neural network architectures, including transformers and graph neural networks (GNNs), which are specifically optimized for modeling complex molecular structures and interactions. The framework also excels in data handling, featuring advanced preprocessing and augmentation techniques that ensure biomolecular datasets that are of the highest quality and representativeness, thus enabling the training of highly accurate, and effective models. This cohesive integration of technical components positions BioNeMo as a powerful solution for accelerating drug discovery through AI.

Pretrained Models

The framework allows for incorporating a suite of pretrained models that are integral to various aspects of drug discovery. For protein structure prediction, BioNeMo integrates models like AlphaFold and RoseTTAFold, which enable precise 3D protein structure predictions directly from amino acid sequences. In the realm of generative models, the framework includes tailored versions of variational autoencoders (VAEs) and generative adversarial networks (GANs) designed to create novel molecular structures with specific desired properties. Additionally, BioNeMo features models for sequence generation, leveraging advanced techniques from natural language processing (NLP) to produce protein sequences capable of folding into precise structures or performing targeted functions. This integration of cutting-edge pretrained models makes BioNeMo a powerful tool for accelerating drug discovery.

NVIDIA BioNeMo NIMs (Neural Information Models)

The NVIDIA BioNeMo Neural Information Models (NIMs) are specialized microservices designed for high-performance inference in AI-driven drug discovery. Optimized specifically for NVIDIA GPUs, NIMs deliver efficient computation and scalability, ensuring that AI models run at peak performance. These microservices are accessible via RESTful APIs, allowing seamless integration into existing drug discovery pipelines and workflows, thereby enhancing the efficiency and effectiveness of the overall research process.

Containerization

This platform utilizes Docker containers to guarantee reproducibility and simplify deployment across various environments, whether on-premises or in the cloud. This containerization ensures that the platform can be consistently and reliably deployed, regardless of the underlying infrastructure, making it easier for researchers to implement and scale their AI-driven drug discovery workflows.

Advanced Applications

Virtual Screening and Docking

BioNeMo offers powerful tools for virtual screening and molecular docking, these are essential components in the drug discovery process. The platform includes advanced molecular docking capabilities, allowing researchers to predict how small molecules, such as potential drugs, bind to target proteins. This process involves scoring and ranking compounds based on their predicted binding affinity, helping to identify the most promising candidates. Additionally, BioNeMo can perform virtual screening of large compound libraries, rapidly identifying potential drug candidates for further testing and significantly accelerating the early stages of drug discovery.

De Novo Drug Design

By harnessing generative models in the field of generative chemistry to create novel molecules with specific properties, such as high binding affinity, low toxicity, and ideal pharmacokinetics, NVIDIA BioNeMo utilizes advanced optimization algorithms, including reinforcement learning, to iteratively enhance these molecules, ensuring they align with predefined goals. This potent combination of generative modeling and optimization enables BioNeMo to accelerate the discovery and development of drug candidates with precisely tailored attributes. The platform utilizes advanced optimization algorithms, including reinforcement learning, to iteratively enhance these molecules, ensuring they align with predefined goals.

Property Prediction

NVIDIA BioNeMo includes advanced models for predicting ADMET (absorption, Distribution, metabolism, excretion, and toxicity) properties, which are essential for evaluating the drug-likeness of compounds. Additionally, the platform offers models that predict key physicochemical properties, such as solubility, permeability, and stability. These predictions play a crucial role in the selection and optimization of viable drug candidates, ensuring that only the most promising compounds advance through the drug discovery pipeline.

Integration and Customization

NVIDIA BioNeMo provides the flexibility for researchers to develop custom models using their own data, which can then be seamlessly integrated into the BioNeMo framework. This capability allows for the creation of tailored solutions that address specific challenges in drug discovery. Furthermore, BioNeMo offers robust interoperability, supporting integration with other bioinformatics tools and databases. This ensures seamless data exchange and workflow automation, enhancing the efficiency and effectiveness of the drug discovery process.

Performance and Optimization

Optimize all computational tasks for NVIDIA GPUs, ensuring high performance and efficiency, including support for mixed-precision training and inference, which maximizes throughput. The platform is also built for scalability, enabling horizontal scaling across multiple GPUs and nodes, allowing researchers to conduct large-scale experiments with ease.

Hardware Infrastructure

Pure Storage FlashBlade//S500 is fully compatible with NVIDIA DGX, NVIDIA OVX, and NVIDIA HGX systems, ensuring seamless integration with these advanced computing platforms. FlashBlade complements this setup by delivering the consistent, scale-out, multi-dimensional performance and capacity necessary for demanding AI applications like BioNeMo. In our testing, we used an NVIDIA SN3700 (NVIDIA Spectrum) ethernet switch, which provided robust networking capabilities. The BioNeMo Framework is optimized to run on NVIDIA GPUs, maximizing efficiency and performance by leveraging the CUDA ecosystem and GPU-accelerated libraries. This optimization ensures that BioNeMo operates at peak potential, fully utilizing the power of NVIDIA's hardware and software ecosystem.

NVIDIA BioNeMo Architecture with Pure Storage

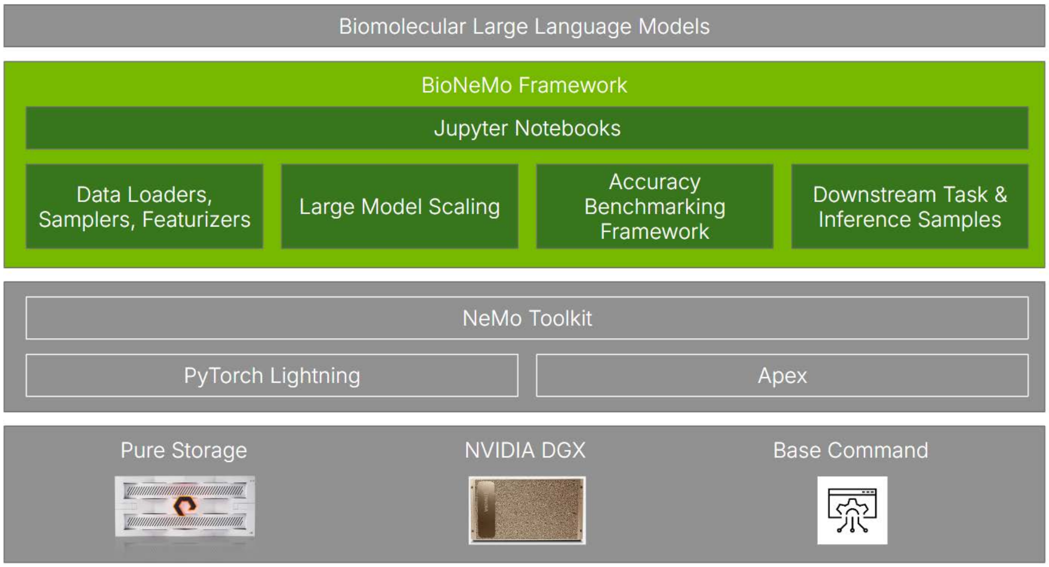

Biomolecular Large Language Models
BioNeMo Framework
Jupyter Notebooks
Data Loaders, Samplers, Featurizers
Large Model Scaling
Accuracy Benchmarking Framework
Downstream Task & Inference Samples
NeMo Toolkit
PyTorch Lightning
Apex
Pure Storage
NVIDIA DGX
Base Command

FIGURE 2 Integration of NVIDIA BioNeMo with a Pure Storage FlashBlade//S500 system

The above diagram illustrates the integration of NVIDIA BioNeMo with a Pure Storage FlashBlade//S500 system, demonstrating how these technologies collaborate to support AI-driven drug discovery and other computational tasks. The diagram effectively showcases the synergy between NVIDIA's powerful computing platforms and scalable storage solutions from Pure Storage. The integration enables a highly efficient, scalable, and high-performance environment for AI-driven tasks, such as those conducted in the BioNeMo framework for drug discovery. The use of NVIDIA's advanced networking hardware ensures that data flow between storage and compute resources is seamless, allowing the system to handle large-scale experiments and workloads with ease. The diagram illustrates a robust and scalable architecture optimized for AI applications, highlighting how NVIDIA and Pure Storage technologies combine to accelerate computational tasks in fields like drug discovery.

Technology Solution Design

The hardware stack consists of NVIDIA DGX servers with eight cores each, a FlashBlade//S500 in a three-chassis configuration and NVIDIA SN3700 network switch. The FlashBlade//S500 provides high throughput, low latency storage suitable for AI training and inference workloads. It provides industry leading PB/W performance for AI workloads, is able to scale and support large GPU clusters. BioNeMo and drug discovery are use cases where zero downtime is critical.

The NVIDIA DGX Servers are equipped with eight GPU cores each, delivering exceptional compute power tailored for AI and machine learning workloads. These servers leverage NVIDIA's NVLink technology, providing high-bandwidth, low-latency GPU-to-GPU communication, which is crucial for parallel processing tasks like training complex models in fields such as drug discovery. The DGX architecture is designed for scalability, allowing multiple servers to work together seamlessly, making it ideal for large-scale AI workloads that require substantial computational resources across numerous nodes.

The FlashBlade//S500 in the three-chassis configuration serves as the high-performance, scale-out storage solution that offers linear scalability, ensuring that data-intensive applications can keep pace with the computational demands of the DGX servers. Utilizing Pure Storage DirectFlash technology, the FlashBlade//S500 eliminates the performance bottlenecks typically associated with traditional SSDs, delivering higher throughput and lower latency. Its distributed metadata architecture ensures that metadata operations are spread across the storage system, preventing bottlenecks and maintaining high performance even as the system scales, which is particularly beneficial when multiple DGX servers access large datasets simultaneously.

The NVIDIA SN3700 network switch is designed to provide high-bandwidth, low-latency networking, which is essential for connecting NVIDIA DGX servers and FlashBlade storage in a unified fabric. Supporting 100GbE and 200GbE connections, the SN3700 prevents bottlenecks by ensuring that data moves quickly between compute and storage components. Its scalability and flexibility allow for easy expansion as AI workloads grow, ensuring high performance across an expanding network. Optimized for handling large data flows typical in AI environments, the SN3700 minimizes latency and maximizes throughput for both training and inference workloads, making it a critical component in maintaining the overall efficiency of the integrated solution.

The NVIDIA BioNeMo software stack is composed of three key components: the BioNeMo Framework, pre-trained models, and BioNeMo NIMs (neural information models), each playing a crucial role in AI-driven drug discovery. The BioNeMo Framework provides essential tools for training and fine-tuning AI models on biomolecular data and supports a variety of neural network architectures tailored to specific drug discovery tasks. It also includes robust data handling capabilities, with advanced preprocessing and augmentation techniques that ensure high-quality, representative data for model training.

Complementing the framework, the pre-trained models offer out-of-the-box capabilities for critical tasks like protein structure prediction, generative chemistry, and ADMET prediction. These models serve as a foundation that allow researchers to quickly apply advanced AI techniques to their projects.

BioNeMo NIMs further enhance the platform with optimized microservices designed for high-performance inference on NVIDIA GPUs. These microservices ensure that AI models run efficiently and effectively, leveraging the full computational power of the underlying hardware. Additionally, BioNeMo NIMs are equipped with RESTful APIs, making it easy to integrate these powerful tools into existing drug discovery pipelines and workflows. This seamless integration enables researchers to incorporate BioNeMo's capabilities into their current systems, accelerating the drug discovery process and driving innovation in related fields.

NVIDIA BioNeMo offers a robust suite of pre-trained models designed to address a variety of biomolecular tasks, making it an invaluable resource for drug discovery and related fields. The available models include:

- ESM1-nv: A transformer-based protein language model specifically optimized for tasks involving protein sequences, ESM1-nv enables advanced analysis and predictions in protein-related research.
- ProtT5-nv: Another powerful protein language model, ProtT5-nv is versatile and can be applied to a wide range of protein-related tasks, further enhancing research capabilities.
- MegaMolBART: Designed for molecule generation and representation learning, MegaMolBART excels in tasks involving small molecules, particularly those using SMILES representations. This model is especially useful for generating novel compounds in drug discovery.
- ESM2: An updated version of the ESM1 model, ESM2 offers improved performance for protein sequence tasks, providing even more accurate and reliable results for researchers.

These models are highly versatile and can be fine-tuned for specific tasks, allowing researchers to tailor them to their unique needs in drug discovery and other biomolecular applications. This flexibility makes the BioNeMo pre-trained models powerful tools for advancing research and accelerating the development of new therapeutics.

NVIDIA BioNeMo offers a user-friendly experience designed to simplify and accelerate drug discovery processes. Here are some key aspects of the user experience:

- **On-premises NIMs deployment**: Users can access BioNeMo through easy-to-use NIMs containers, allowing for interactive inference, visualization, and experimentation.
- **Turnkey Solutions**: The platform provides turnkey solutions for tasks like 3D protein structure prediction, de novo protein and small molecule generation, property predictions, and molecular docking.
- **Comprehensive Documentation**: Users have access to extensive documentation, tutorials, and community support, making it easier to get started and troubleshoot issues.
- **Scalable and Flexible**: The platform offers scalable, managed infrastructure, allowing users to experiment and build enterprise-grade generative AI workflows.

NVIDIA BioNeMo is designed to offer a user-friendly and powerful platform that simplifies and accelerates drug discovery processes. Below are some expanded details with examples to illustrate the key aspects of the user experience:

On-premises NIMs Deployment

Interactive inference and visualization: BioNeMo's on-premises NIMs deployment allows researchers to run high-performance inference tasks on local infrastructure. For example, a research team working on identifying potential inhibitors for a viral protein can use BioNeMo's NIMs containers to quickly test various molecules against the protein structure. The interactive visualization tools within the platform enable the team to see in real-time how different molecules bind to the protein, allowing for rapid iteration and refinement of potential drug candidates.

Turnkey Solutions

3D protein structure prediction: BioNeMo provides ready-to-use models like AlphaFold integrated with the platform, allowing users to predict the 3D structures of proteins from amino acid sequences without needing to configure complex environments. For instance, a biotech company focused on developing new antibiotics can use this turnkey solution to predict the structure of bacterial proteins and identify potential binding sites for drug development.

De novo protein and small molecule generation: BioNeMo includes tools like MegaMolBART for generating entirely new proteins or small molecules with desired properties. A pharmaceutical company could use this feature to design novel compounds with high affinity for a specific receptor implicated in cancer, significantly speeding up the lead optimization phase of drug development.

Property predictions and molecular docking: The platform offers solutions for predicting key properties of molecules, such as solubility and toxicity, as well as for performing molecular docking simulations. For example, a research group might use BioNeMo to screen a library of chemical compounds, predict their ADMET properties, and perform docking simulations to determine which compounds are most likely to be effective and safe drugs.

Comprehensive Documentation

Tutorials and examples: BioNeMo provides extensive documentation and tutorials that guide users through various tasks, such as setting up the environment, running models, and interpreting results. For example, a new user in academia might access step-by-step tutorials on how to use BioNeMo for protein-ligand docking studies, helping them get up to speed quickly and apply the platform to their research projects.

Community support: Users have access to forums and communities where they can ask questions, share insights, and collaborate with others using BioNeMo. A user encountering a specific issue with model convergence might find solutions in the community forums or collaborate with other users who have faced similar challenges.

Scalable and Flexible Infrastructure

Horizontal scaling: BioNeMo is built to scale across multiple GPUs and nodes, making it possible to handle large-scale experiments. For instance, a company conducting high-throughput virtual screening of millions of compounds against a target protein can leverage BioNeMo's scalable infrastructure to distribute the workload across several nodes, drastically reducing computation time and accelerating the discovery process.

Enterprise-grade AI workflows: The platform supports the development of robust generative AI workflows that can be scaled from research prototypes to full-scale production systems. A large pharmaceutical company might use BioNeMo to prototype a new AI-driven drug design pipeline, then scale it up to run thousands of simulations in parallel, supporting their global drug discovery efforts with a consistent and powerful computational backbone.

These features and examples demonstrate how NVIDIA BioNeMo is not only a powerful platform for AI-driven drug discovery but also one that is accessible, flexible, and designed to meet the diverse needs of researchers and organizations in the life sciences. By offering a user-friendly experience, BioNeMo enables teams to focus on what matters most: accelerating the discovery of new, life-saving drugs.

Design Validation

We validated the BioNeMo framework using the FlashBlade//S500 system to ensure scalability, efficiency, and ease of use for training models in drug discovery and running inferences. Our validation setup included a three-chassis FlashBlade//S500 system, paired with an NVIDIA DGX Server and SN3700 networking. We utilized between 1 to 8 GPU cores during testing, depending on the specific workload. The amount of input data ranged from 30 to 100 GB, allowing us to simulate a variety of real-world scenarios. Additionally, we tested the system's capability to handle multiple simultaneous jobs, further demonstrating the robust performance and seamless integration of BioNeMo with the FlashBlade//S500 for demanding AI-driven drug discovery workloads. This configuration enabled us to train models from scratch, fine-tune existing models, and perform inference tasks efficiently.

As the number of parallel jobs scaled, the FlashBlade//S500 system effectively handled the increased load, demonstrating impressive scalability. When we doubled the number of jobs running in parallel, the FlashBlade nearly doubled the bandwidth delivered to the GPUs. For instance, running two containers in parallel for a training job with 100 GB of data, the FlashBlade's bandwidth scaled from 1.9Gbps to 3.8Gbps, showcasing its ability to efficiently manage increased workloads.

Additionally, the storage throughput of FlashBlade significantly outperformed local disk during both inference and training workloads. For example, during an inference task with 100GB of data, the maximum read throughput on FlashBlade reached 2.9GB/s, which was nearly 2.5 times the 1.19GB/s read throughput observed with local disk. This highlights FlashBlade's superior performance in handling large-scale AI workloads.

We observed that the batch size was limited to 10 samples per batch per node, which constrained our ability to fully drive GPU utilization and maximize storage read throughput as data sizes scaled. Despite this limitation, FlashBlade seamlessly scaled across different GPU cores, the number of concurrent/parallel jobs, and the size of ingested data, proving its robustness and consistent performance for the BioNeMo framework.

Not only does this scalability enable faster time to innovation, it also helps teams easily meet future demands and provides Evergreen features that enable it to improve over time, non-disruptively.

When combined with NVIDIA GPUs, FlashBlade//S sets a new standard for energy and cost efficiency in both training and inference workloads. This synergy results in industry-leading performance metrics, such as the best PB/W in storage efficiency for AI workloads. This efficiency translates directly into a lower total cost of ownership (TCO), making the FlashBlade//S and NVIDIA combination as an optimal choice for organizations seeking high performance while maintaining sustainability and cost-effectiveness.

One of the standout features of the FlashBlade//S platform is its modular architecture, which allows for the independent scaling of both capacity and performance. This flexibility enables organizations to precisely match their infrastructure with their current needs and growth projections, ensuring maximum efficiency and minimizing waste. The platform's design, which includes a distributed metadata architecture, DirectFlash modules, and integrated networking, keeps complexity low and performance high. This modular approach means that FlashBlade can efficiently handle GPU loads across a wide range of configurations, from small, single-node clusters to expansive environments with several hundred nodes.

In addition to its performance, efficiency and scalability, FlashBlade//S excels in operational simplicity. The platform is designed to be easy to set up and monitor, reducing the time and expertise needed to manage the infrastructure. This ease of use allows teams to concentrate on their core objectives rather than getting bogged down in the complexities of infrastructure management, further enhancing the overall efficiency and effectiveness of the system.

Deployment

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

Infrastructure setup: Deploying and configuring systems servers, storage, and network components

- FlashBlade Setup
  - Login to FlashBlade. Create and export a file system (For Pure Storage customers, refer to the FlashBlade user guides for more details).
  - Configure Jumbo Frames on FlashBlade.
- DGX Server Setup
  - Ensure your DGX server is up and running with appropriate GPU drivers.
  - Mount the File System with NFSv3 protocol on `/biotest` mount point and create data & result folders for future use. NFS best practices can be found here.
  - Install Docker on DGX system (with GPU support, Docker Engine 19.03 or above).
  - Install NVIDIA Container Toolkit to allow Docker to access the GPUs.
- Network Setup
  - Configure MTU size on Linux Client

Software installation: Installing Bionemo framework

- Setup NGC account & API key configuration.
- Download the required model and start the container with bash prompt.

```
# Setting ngc config
$ ngc config set
# Downloading approrpriate model
$ python download_models.py --download_dir /workspace/bionemo/models megamolbart
# Starting the container with bionemo image
$ docker run -d -p <DGX IP>:6006:6006 -p <DGX IP>:8888:8888 --mount type=bind,source=/dev/shm,destination=/
dev/shm -v /biotest/data:/data -v /biotest/result:/result -it  --gpus all nvcr.io/nvidia/clara/bionemo-frame-
work:1.4.1 bash
```

- Execute the below to preprocess the dataset and start training.

```
# From container bash prompt, Navigate to specific model example folder

$ cd /workspace/bionemo/examples/molecule/megamolbart

# Downloading dataset and preprocessing it

$ python pretrain.py --config-path=conf --config-name=pretrain_xsmall_span_aug do_training=False model.data.links_file='${oc.env:BIONEMO_HOME}/examples/molecule/megamolbart/dataset/ZINC-downloader-sample.txt' model.data.dataset_path=$(pwd)/zinc_csv

# Executing training with single file in dataset

$ python pretrain.py --config-path=conf --config-name=pretrain_xsmall_span_aug do_training=True model.data.dataset_path=/data/zinc_csv model.data.dataset.train=x000 model.data.dataset.val=x000 model.data.dataset.test=x000 exp_manager.exp_dir=/results trainer.devices=1
```

The above trains with a single file in the dataset to provide a basic understanding of the process. To train on the entire dataset, remove the train, test, and validation parameters from the command. You can further fine-tune by modifying the `pretrain_xsmall_span_aug` file to suit your needs.

During execution you will be prompted for Weights & Biases (W&B) integration. If opted for W&B, results will be populated in W&B dashboards. Training related logs & outputs will be stored in results as per the training command execution.

Results

As the data set increases, the data set size and throughput also increase.

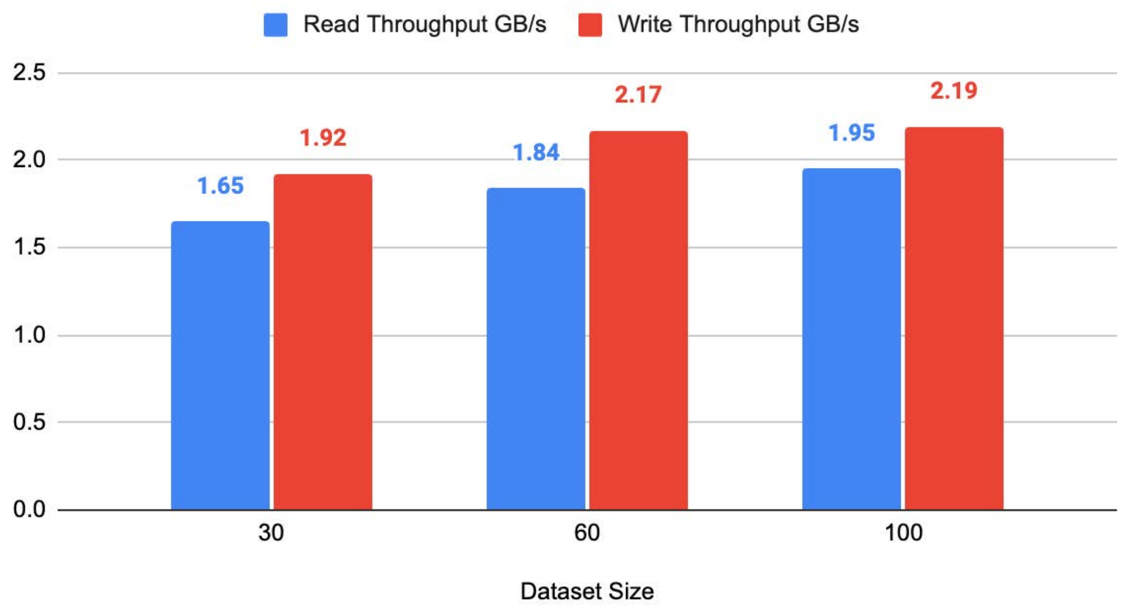


FIGURE 3 BioNeMo Training dataset size and throughput

The image above shows the comparative analysis of dataset sizes, read throughputs, and write throughputs during iterative runs in the context of training with BioNeMo. Dataset size plays a crucial role in determining both read and write throughputs during the testing. Larger datasets typically demand more from the storage system, requiring higher throughput to handle the data efficiently during both training and inference phases. Read throughput is critical during the inference and training processes, where data needs to be quickly accessed and processed sometimes iteratively. The image shows that as the dataset size increases, the read throughput must also scale to ensure that the GPUs are not starved for data. This is particularly important in iterative runs, where the same dataset may be accessed multiple times. Write throughput is essential when saving model states, checkpoints, or processed data back to storage during training. The result depicted compares how write throughputs vary with different dataset sizes and how efficiently the system handles these operations during iterative runs.

The comparisons highlight how well FlashBlade scales in terms of both read and write throughput as the dataset sizes increase. The FlashBlade proves it is adept at efficiently scaling and ensuring that the system can maintain high performance even as the workload grows. There are no bottlenecks, such as scenarios where write throughput cannot keep up with read throughput, potentially slowing down the iterative training process. This could indicate areas where optimization is needed, such as improving write speeds or balancing I/O operations more effectively. Iterative runs often involve repeated read and write operations, and the graph demonstrates how the system's performance holds up under such conditions. Higher read and write throughputs contribute to faster training cycles, allowing for more iterations within the same time frame, which is critical for model refinement.

Results of Scaling Training Datasets

The results we have shown underscore the importance of balanced and scalable read/write throughput to handle increasing dataset sizes efficiently, especially in the context of iterative runs in AI training with BioNeMo. This balance is key to maintaining high performance and ensuring that the GPUs and storage systems work harmoniously without causing delays or bottlenecks.

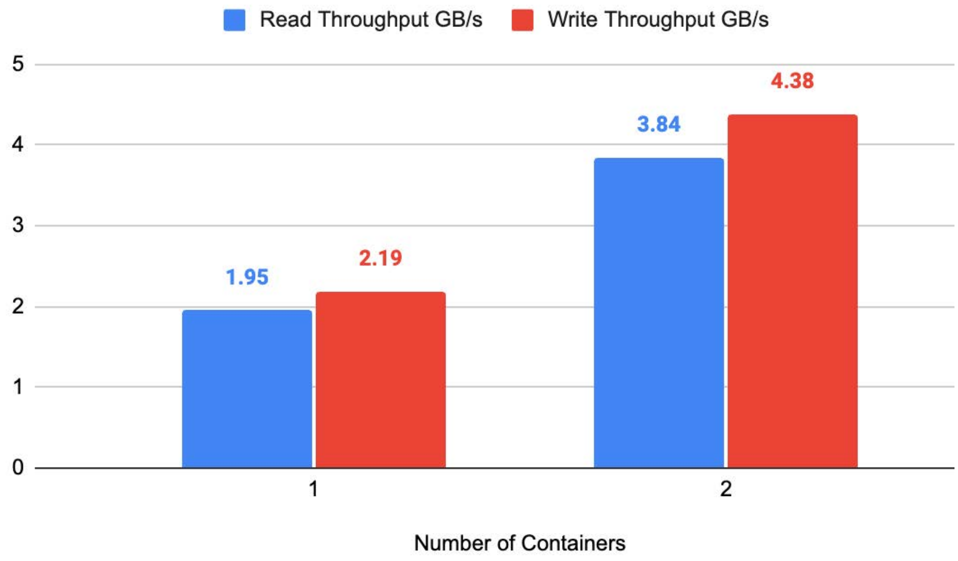


FIGURE 4 Parallel training throughput results

The above chart demonstrates that as the number of containers increases, the read throughput scales proportionally. This indicates that Pure Storage FlashBlade is capable of efficiently managing concurrent data access requests, ensuring that multiple containers can simultaneously read large datasets without performance degradation.

As an example, consider a scenario where a hospital is using AI models to analyze large volumes of medical imaging data, such as MRIs or CT scans, to detect early signs of diseases like cancer. These AI models need to access and process vast amounts of imaging data stored in a centralized system to provide accurate and timely diagnosis. Hospitals that have implemented a parallel training setup where multiple AI models (containers) are being trained simultaneously to improve detection algorithms for various types of cancer. As the number of containers increases, each container requires fast and efficient access to the stored imaging data to process it in real time.

If the storage system (e.g., FlashBlade) can effectively scale its read throughput with the increasing number of containers, it ensures that each AI model can quickly access the required imaging data without delays. This high read throughput is crucial because it directly impacts the speed and accuracy with which the AI models can analyze the data, leading to faster and more reliable diagnoses. For instance, if the system efficiently scales from handling 10 containers to 20 containers with minimal degradation in read throughput, the hospital can simultaneously train twice as many models.

This allows for parallel development of multiple specialized AI tools—one focusing on lung cancer, another on brain tumors, and so forth. This parallel processing accelerates the overall pace of innovation, enabling the hospital to deploy advanced diagnostic tools more quickly, ultimately improving patient outcomes by catching diseases earlier and starting treatment sooner.

Results of Scaling Containers with Parallel Training

The ability of the storage system to maintain high read throughput as the number of containers increases in a healthcare setting is vital for enabling fast, parallel processing of medical data. This capability supports the rapid development and deployment of AI-driven diagnostic tools, leading to more efficient healthcare delivery and better patient care

Write throughput is crucial during the training process, particularly for saving intermediate results, checkpoints, and final model states. The image might show that write throughput also scales with the number of containers but may not keep pace as closely as read throughput. This could indicate that while the system handles reads efficiently, writes could become a bottleneck as more containers are added.

If the image revealed that write throughput lagged behind read throughput as the number of containers increases, this would suggest that while data can be accessed quickly, saving or updating data may slow down, potentially leading to longer training times. This discrepancy can highlight the need for optimizing write operations or balancing the I/O workload.

The parallel, multi-dimensional architecture from Pure Storage is designed to minimize such bottlenecks by using advanced features like DirectFlash and distributed metadata. These technologies help maintain high write performance even under heavy parallel workloads, ensuring that iterative runs do not suffer from slow write operations.

The ability of the storage system to scale with the number of containers directly impacts the efficiency of iterative runs. If the results show that throughput remains high across all runs, it indicates that the system is well-suited for scaling AI workloads.

Conclusion

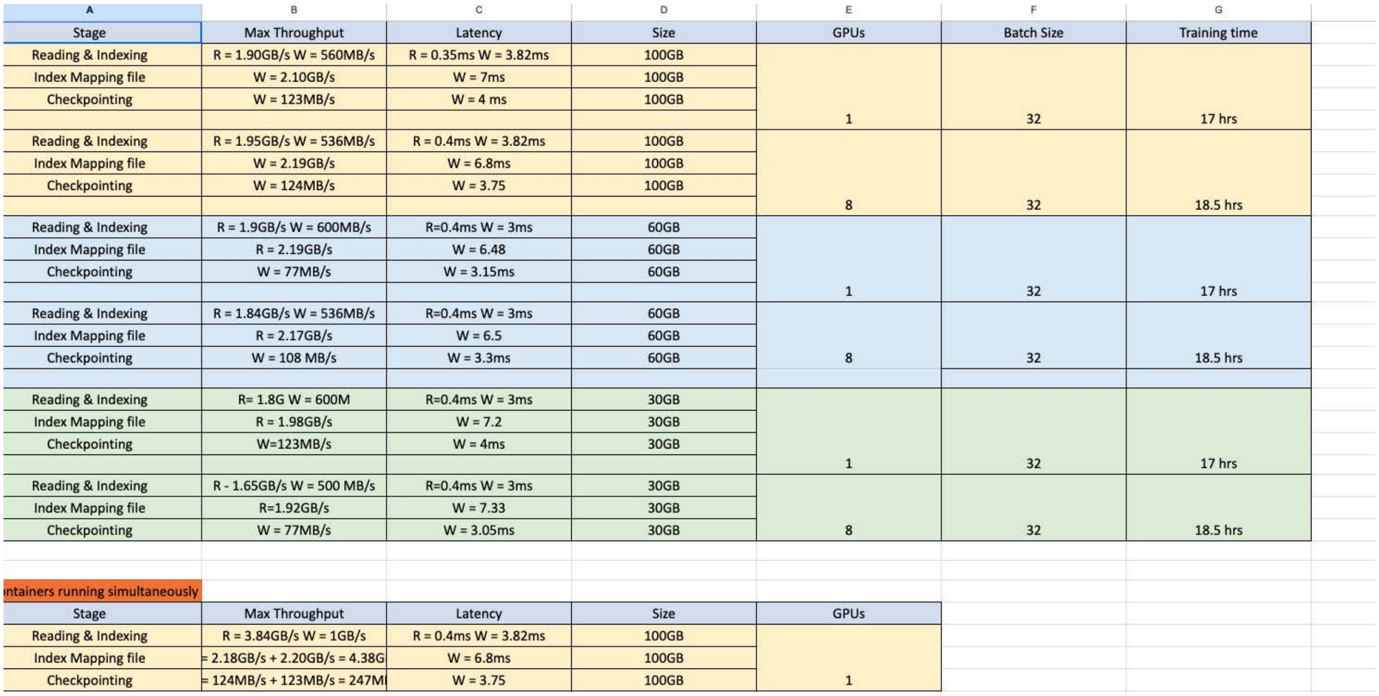
FlashBlade's modular architecture allows users to scale both capacity and performance independently. This means that as your BioNeMo workloads grow, you can add more containers or nodes without sacrificing performance, keeping your iterative training cycles fast and efficient.

The analysis of the parallel training data likely highlights the strong performance of FlashBlade in managing both read and write throughputs as the number of containers increases. While there may be slight discrepancies in scaling between read and write operations, the overall system performance demonstrates the capability of FlashBlade to support large-scale, iterative AI workloads like those in BioNeMo. By maintaining high throughput across parallel containers, FlashBlade ensures that organizations can scale their AI operations efficiently, leading to faster time-to-insight and innovation across various industries.

Additional Resources

- Visit our life sciences solution page to learn more about how we can help your organization.
- Learn more about the Pure Storage platform for AI and FlashBlade//S.
- Visit NVIDIA BioNeMo website.

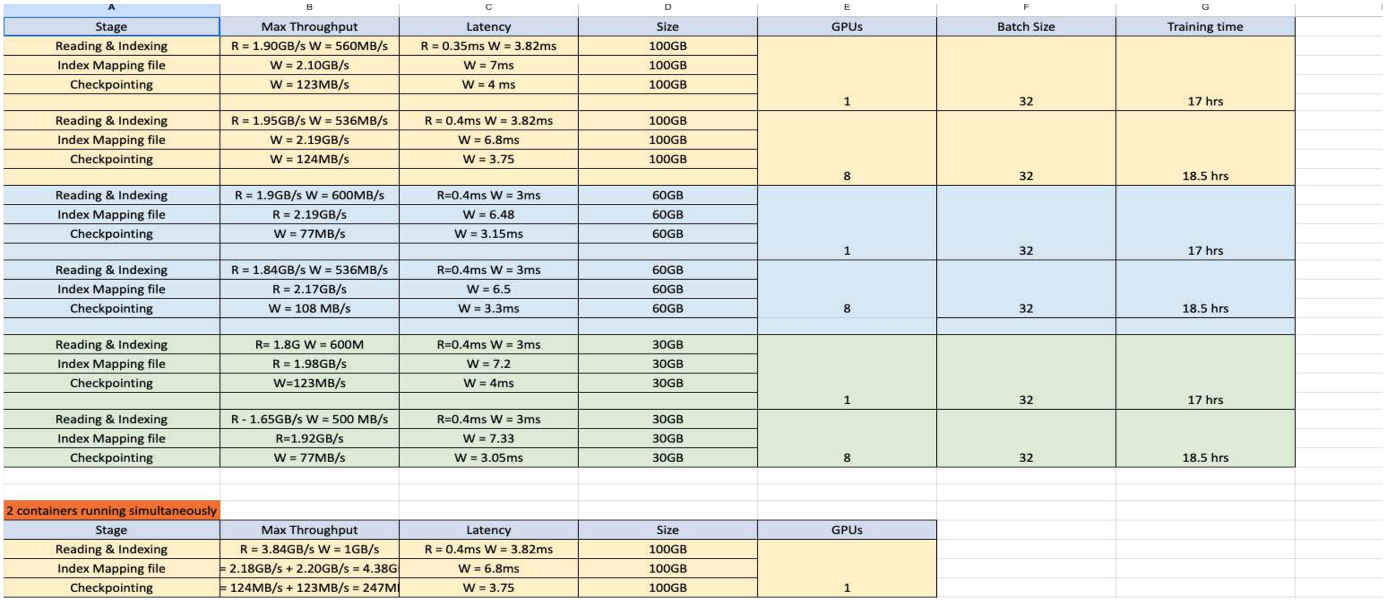
Appendix 1: Raw Data



A	B	C	D	E	F	G
Stage	Max Throughput	Latency	Size	GPUs	Batch Size	Training time
Reading & Indexing	R = 1.90GB/s W = 560MB/s	R = 0.35ms W = 3.82ms	100GB	1	32	17 hrs
Index Mapping file	W = 2.10GB/s	W = 7ms	100GB			
Checkpointing	W = 123MB/s	W = 4 ms	100GB			
Reading & Indexing	R = 1.95GB/s W = 536MB/s	R = 0.4ms W = 3.82ms	100GB	8	32	18.5 hrs
Index Mapping file	W = 2.19GB/s	W = 6.8ms	100GB			
Checkpointing	W = 124MB/s	W = 3.75	100GB			
Reading & Indexing	R = 1.9GB/s W = 600MB/s	R=0.4ms W = 3ms	60GB	1	32	17 hrs
Index Mapping file	R = 2.19GB/s	W = 6.48	60GB			
Checkpointing	W = 77MB/s	W = 3.15ms	60GB			
Reading & Indexing	R = 1.84GB/s W = 536MB/s	R=0.4ms W = 3ms	60GB	8	32	18.5 hrs
Index Mapping file	R = 2.17GB/s	W = 6.5	60GB			
Checkpointing	W = 108 MB/s	W = 3.3ms	60GB			
Reading & Indexing	R= 1.8G W = 600M	R=0.4ms W = 3ms	30GB	1	32	17 hrs
Index Mapping file	R = 1.98GB/s	W = 7.2	30GB			
Checkpointing	W=123MB/s	W = 4ms	30GB			
Reading & Indexing	R - 1.65GB/s W = 500 MB/s	R=0.4ms W = 3ms	30GB	8	32	18.5 hrs
Index Mapping file	R=1.92GB/s	W = 7.33	30GB			
Checkpointing	W = 77MB/s	W = 3.05ms	30GB			

ntainers running simultaneously

Stage	Max Throughput	Latency	Size	GPUs
Reading & Indexing	R = 3.84GB/s W = 1GB/s	R = 0.4ms W = 3.82ms	100GB	1
Index Mapping file	= 2.18GB/s + 2.20GB/s = 4.38G	W = 6.8ms	100GB	
Checkpointing	= 124MB/s + 123MB/s = 247MI	W = 3.75	100GB	



A	B	C	D	E	F	G
Stage	Max Throughput	Latency	Size	GPUs	Batch Size	Training time
Reading & Indexing	R = 1.90GB/s W = 560MB/s	R = 0.35ms W = 3.82ms	100GB	1	32	17 hrs
Index Mapping file	W = 2.10GB/s	W = 7ms	100GB			
Checkpointing	W = 123MB/s	W = 4 ms	100GB			
Reading & Indexing	R = 1.95GB/s W = 536MB/s	R = 0.4ms W = 3.82ms	100GB	8	32	18.5 hrs
Index Mapping file	W = 2.19GB/s	W = 6.8ms	100GB			
Checkpointing	W = 124MB/s	W = 3.75	100GB			
Reading & Indexing	R = 1.9GB/s W = 600MB/s	R=0.4ms W = 3ms	60GB	1	32	17 hrs
Index Mapping file	R = 2.19GB/s	W = 6.48	60GB			
Checkpointing	W = 77MB/s	W = 3.15ms	60GB			
Reading & Indexing	R = 1.84GB/s W = 536MB/s	R=0.4ms W = 3ms	60GB	8	32	18.5 hrs
Index Mapping file	R = 2.17GB/s	W = 6.5	60GB			
Checkpointing	W = 108 MB/s	W = 3.3ms	60GB			
Reading & Indexing	R= 1.8G W = 600M	R=0.4ms W = 3ms	30GB	1	32	17 hrs
Index Mapping file	R = 1.98GB/s	W = 7.2	30GB			
Checkpointing	W=123MB/s	W = 4ms	30GB			
Reading & Indexing	R - 1.65GB/s W = 500 MB/s	R=0.4ms W = 3ms	30GB	8	32	18.5 hrs
Index Mapping file	R=1.92GB/s	W = 7.33	30GB			
Checkpointing	W = 77MB/s	W = 3.05ms	30GB			

2 containers running simultaneously

Stage	Max Throughput	Latency	Size	GPUs
Reading & Indexing	R = 3.84GB/s W = 1GB/s	R = 0.4ms W = 3.82ms	100GB	1
Index Mapping file	= 2.18GB/s + 2.20GB/s = 4.38G	W = 6.8ms	100GB	
Checkpointing	= 124MB/s + 123MB/s = 247MI	W = 3.75	100GB	