Scale AI Training and Fine-Tuning with Dell PowerScale and PowerEdge Servers

Run AI Model Training and Fine-Tuning Operations with Dell PowerScale File Storage and PowerEdge GPU Powered Servers

May 2024

H20044

Technical White Paper

Abstract

This white paper discusses the benefits of the Dell PowerScale scale-out file platform for AI. The Dell Reference Design describes a Generative AI solution architecture combining NVIDIA AI Enterprise and NEMO with Dell storage, server, and network infrastructure to successfully stage, run, and scale Generative AI model training.

Dell Technologies AI Solutions

Dell
Reference Design

Copyright

Contents

Executive summary

Overview

Generative AI has been front and center in the current tech landscape, especially as it relates to the compute and GPU aspect of the workflows. Regardless of the phase in the GenAI lifecycle—customization and fine-tuning, training, or inference—data is a constant. These operations simply cannot exist without data, from existing data being used to generate intelligent business insights and outcomes to new data created as a result, all of which must be supported by the hosting storage system.

Storage is a critical component within the architecture and will need to be able to support the various workloads that are applied throughout the AI workflows. A proper storage system must be able to support the various requirements of AI workflows including streaming and random reads and writes, as well as scale linearly alongside the compute and GPU requirements. The storage software will need to support hundreds to thousands of concurrent connections as each GPU draws data from the storage system.

Fortunately, Dell Technologies and NVIDIA have teamed up to bring together the industry's leading AI platform, leading scale-out file platform, and award winning server platforms.

Audience

This document guide is intended for anyone interested in the implementation of solutions and infrastructure for generative AI, including professionals and stakeholders involved in the development, deployment, and management of generative AI systems. Key roles include IT executives and decision makers such as Chief Technology Officers (CTOs), Chief Information Officers (CIOs), and principal systems architects. Other audience members may include system administrators and IT operations personnel, AI engineers and developers, and data scientists and AI researchers.

Revisions

Date	Part number/ revision	Description
May 2024	H20044	Initial release

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by email.

Author: Darren Miller

Note: For links to other documentation for this topic, see the Artificial Intelligence Info Hub.

Dell PowerScale architecture

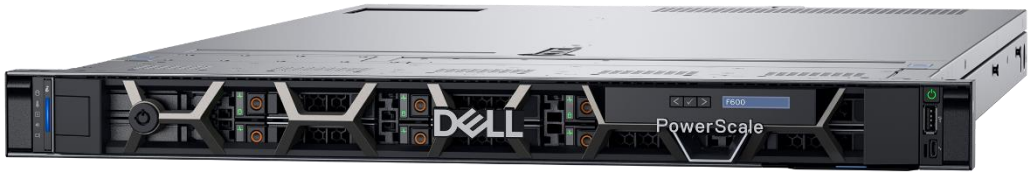


Figure 1. Dell PowerScale

Dell PowerScale is the industry's leading scale-out NAS platform. Dell PowerScale nodes cluster to form a single storage system that can scale up to 252 nodes. Internally, the powerful OneFS file system enables the delivery of the full cluster capacity and performance through a single namespace. The unique scaling capability of Dell PowerScale makes it an ideal storage solution for AI. As GPU clusters grow, storage performance and capacity can be increased non-disruptively by adding more Dell PowerScale nodes to the cluster.

Dell PowerScale is an AI-ready platform built to accelerate AI workloads wherever they are, from on-premises to the edge to the cloud. Powered by OneFS and NVMe disks, the all-flash platforms are ideal for AI architectures.

Performance and scale

When deploying GenAI platforms, a common question is how to size the infrastructure to meet business needs. This can be difficult to answer, especially for companies exploring GenAI for the first time. A scalable storage system provides the flexibility to begin with the correct initial investment and expand economically as needed. Enter Dell PowerScale, which provides predictable performance scaling that simplifies sizing systems and planning for future AI needs.

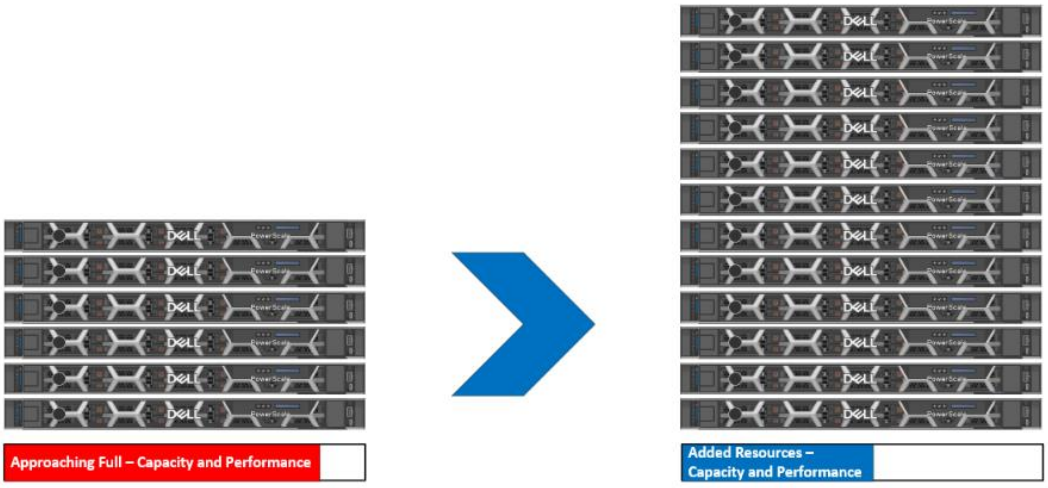


Figure 2. Balancing capacity and performance by adding PowerScale nodes to a cluster

As PowerScale nodes are added to a PowerScale OneFS cluster, the cluster increases in disk capacity and performance—including memory, CPU, and added network throughput. Dell PowerScale clusters can scale up to 252 nodes, 186PB of capacity, and over 2.5TB read/write throughput within a single namespace.

Concurrency

The AI development lifecycle spans preparing data, training, evaluating, retraining, and inferencing in production. Across all lifecycle phases is the need for concurrent IO. Throughout the training/retraining phases, GPUs read data from storage and write checkpoints back out. In large GPU clusters, this requires hundreds and sometimes thousands of simultaneous reads and writes to the storage system. To meet this level of demand, the storage platform must be able to handle sustained and transient concurrency peaks with ease.

Dell PowerScale was originally designed and optimized for industries with workloads that require extreme read and write concurrency. This inherent in-market experience has positioned Dell PowerScale ideally to meet emerging AI infrastructure market needs.

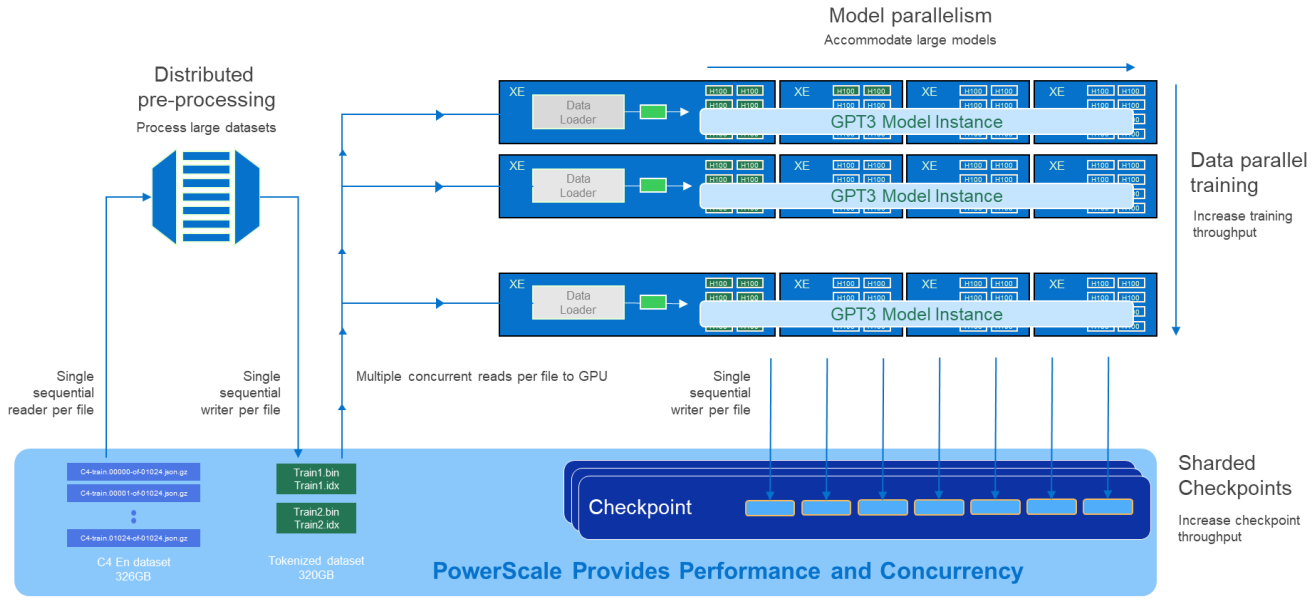


Figure 3. Architecture of workload characteristics

NFSoRDMA

Dell PowerScale was among the first storage platforms to support NFSoRDMA and NVIDIA's GPU Direct Storage (GDS). The low-latency direct transfer of data between storage and GPU memory that GDS enables ensures that Dell PowerScale can keep GPUs utilized while running large, demanding AI models.

PowerScale OneFS supports Remote Direct Memory Access (RDMA) over converged Ethernet (RoCEv2). When paired with NFS to enable NFSoRDMA, data is sent directly from the PowerScale network adapter to the compute or application memory, bypassing CPU and operating systems. Combining NFSoRDMA with NVIDIA's GPU Direct Storage and MagnumIO software modules, data can be transferred directly between NVIDIA GPU memory and PowerScale storage over a low latency, highly efficient network.

PowerEdge XE9680 server

The PowerEdge XE9680 server is a high-performance application server made for demanding AI, machine learning, and deep learning workloads that enable you to rapidly develop, train, and deploy large machine learning models.

The PowerEdge XE9680 server is the industry's first server to ship with eight NVIDIA H100 GPUs and NVIDIA AI software. It is designed to maximize AI throughput, providing enterprises with a highly-refined, systemized, and scalable platform to help achieve breakthroughs in NLP, recommender systems, data analytics, and more. Its 6U air-cooled design chassis supports the highest wattage next-generation technologies up to 35°C ambient. It features nine times more performance and two times faster networking with NVIDIA ConnectX-7 smart network interface cards (SmartNICs), alongside high-speed scalability for extreme scale AI architectures.

NVIDIA H100 Tensor Core GPU

The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With the NVIDIA fourth generation NVLInk Switch System, the NVIDIA H100 GPU accelerates AI workloads with a dedicated Transformer Engine for trillion-parameter language models. The NVIDIA H100 GPU uses breakthrough innovations in the NVIDIA Hopper architecture to deliver industry-leading conversational AI, speeding up large language models by 30 times over the previous generation.

Solution overview

This reference design should be reviewed alongside the GenAI Model Training design. The test methodologies and architecture are identical in both use cases. The solution described in this document covers a training workflow and examples of the performance impact on storage throughout the training workflow.

Solution approach

The solution approach for this design is to train a popular LLM to show both GPU and storage scaling. To keep the design applicable to today's GenAI workflows, a LLAMA 2 model architecture was chosen with 7B and 70B parameter counts.

Table 1. Solution system configuration

Component	Configuration 1	Configuration 2
Compute server for model customization	1 x PowerEdge XE9680 servers	6 x PowerEdge XE9680 servers
GPUs per server	8 x NVIDIA H100 SXM GPUs	8 x NVIDIA H100 SXM GPUs
Ethernet Network adapters	2 x NVIDIA ConnectX-6 DX Dual Port 100 GbE	2 x NVIDIA ConnectX-6 DX Dual Port 100 GbE
Ethernet Network switch	2 x PowerSwitch S5232F-ON	2 x PowerSwitch S5232F-ON
InfiniBand Network adapter	4 x NVIDIA ConnectX-7, Single Port NDR OSFP PCIe, No Crypto, Full Height	4 x NVIDIA ConnectX-7, Single Port NDR OSFP PCIe, No Crypto, Full Height
InfiniBand Network switch	QM9790	QM9790
PowerScale F600 Cluster	3 x PowerScale F600 Performance Optimized nodes	3 x PowerScale F600 Performance Optimized nodes

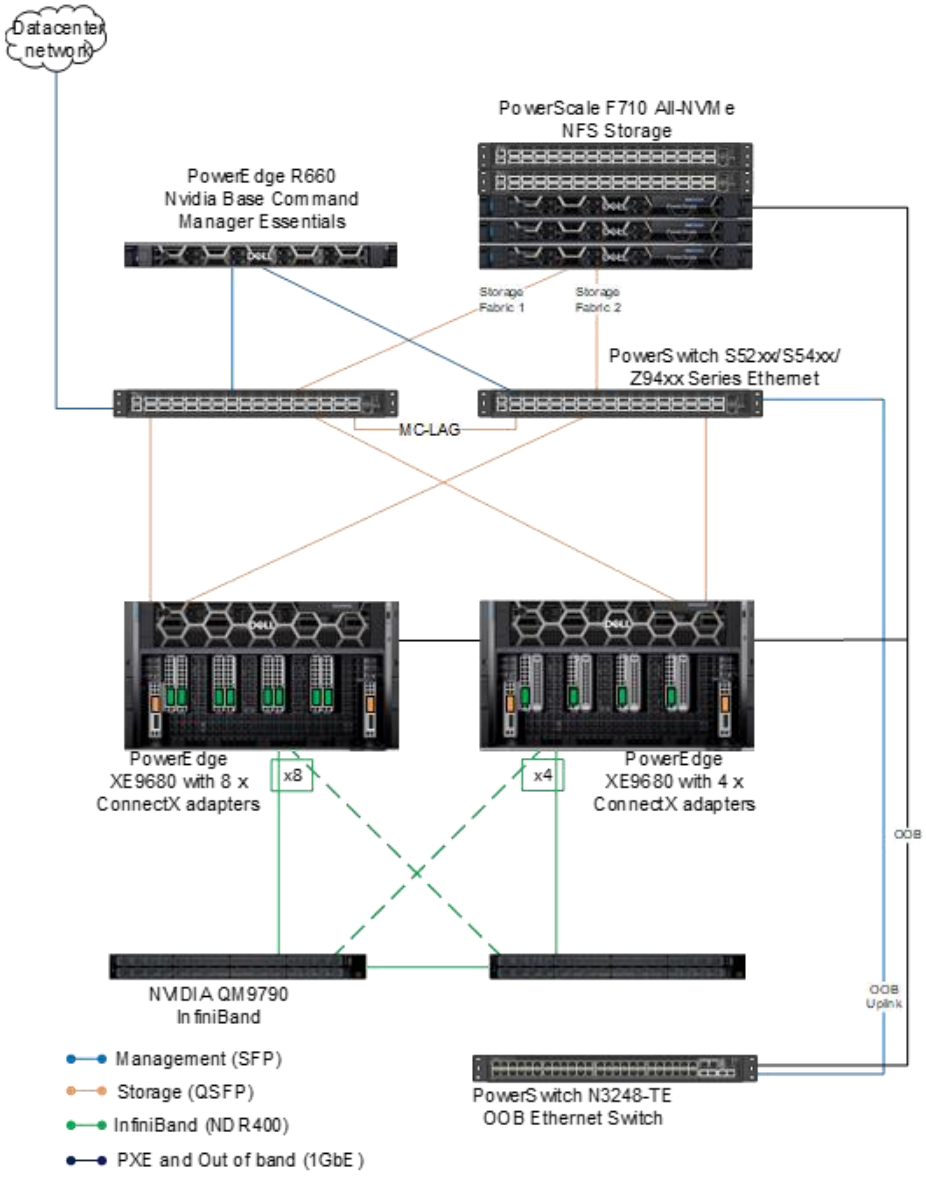


Figure 4. Network connectivity of PowerEdge training nodes, PowerScale storage, and control plane nodes

Figure 4 shows the network architecture and connectivity for the PowerEdge training nodes, PowerScale storage, and the three control plane nodes that incorporate NVIDIA Base Command Manager Essentials and other software components.

Software design

The NVIDIA AI software stack is the primary software used in this design. NVIDIA enterprise software solutions are designed to give IT admins, data scientists, architects, and designers access to the tools they need to easily manage and optimize their accelerated systems.

NVIDIA AI Enterprise

NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform, accelerates the data science pipeline and streamlines development and deployment of production AI, including

generative AI, computer vision, speech AI, and more. This secure, stable, cloud-native platform of AI software includes over 100 frameworks, pretrained models, and tools that accelerate data processing, simplify model training and optimization, and streamline deployment.

Table 2. Software components and versions

Component	Details
Operating system	Ubuntu 22.04.1 LTS
Cluster management	NVIDIA Base Command Manager Essentials 10.23.12
Slurm cluster	Slurm 23.02.4
AI framework	NVIDIA NeMo Framework v23.11

The solution design presented here is modular, and each of the components can be independently scaled depending on the customer's workflow and application requirements.

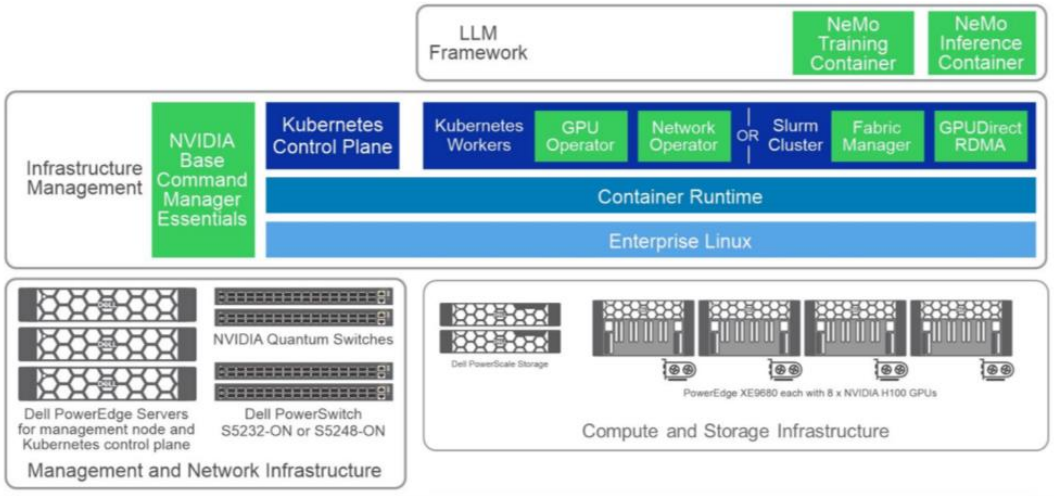


Figure 5. High-level software components used in testing

Results or findings

The goal in this validation was not to train a model to convergence and generate a complete foundational model, but rather to train for a defined number of steps.

The following list provides the details of our validation setup:

- **Model architectures –** 7B and 70B Llama 2 model architectures.
- **Cluster configuration –** Slurm for cluster management and job scheduling.
- **Dataset** – A Pile dataset was used for this validation. The Pile is an 825 GiB diverse, open-source language modeling dataset that consists of 22 smaller, high-quality datasets combined, derived primarily from academic and professional sources. Only 2 shards of the dataset (40G after tokenization) were used for this experiment.
- **Storage Impact** – Measure the impact of different training operations on the storage system. The initial data loading will read training data from the storage into GPUs while checkpointing during the training will show up as writes. These phases of the training process will be examined from a storage performance perspective.
- **Time for training** – The time to train will not be discussed in this document. Refer to GenAI in the Enterprise - Model Training for more on training time and a more thorough GPU examination.

Model architecture selection

Among the various LLMs available, the 7B and 70B parameters of Llama 2 architectures were selected for training based on several key factors:

- **Resource Usage**: Using these two models sizes helped us better understand the infrastructure resource usage and requirements for various training workloads for a range of model sizes.
- **Ease of Use**: The models have been readily available for consumption along with recipes and cookbook implementations, making modification to the codebase easier for customer use cases.

Parallelism

Following are the tensor and pipeline parallelism values we used for the models:

Table 3. Parallelism for Llama 2 architecture for training

Model	Configuration
Llama 2 7B	Tensor Parallelism = 2 Pipeline Parallelism = 1 Micro batch size = 1 Global batch size = 144 Sequence length = 4096
Llama 2 70B	Tensor Parallelism = 4 Pipeline Parallelism = 4 Micro batch size = 1 Global batch size = 144 Sequence length = 4096

Storage performance

The following table shows read and write performance during the initial epoch and the checkpoint operation for both the 7B and 70B parameter models. It includes the load, the dataset, checkpointing, and validation.

We evaluated two configurations:

- **Configuration 1**: LLAMA 2 7B parameter model with 1x PowerEdge XE9680 equipped with 8 NVIDIA H100 GPUs
- **Configuration 2**: LLAMA 2 70B parameter model with 6x PowerEdge XE9680 equipped with 48 NVIDIA H100 GPUs

Table 4. Time for training 40 steps of each model[1]

Models	# of XE9680 Servers	Time / mins	Checkpoint Size	Checkpoint Time (minutes)	Peak Read TP	Peak Write TP
Llama 2 7B	1	3.32	100 GB	2:30	827 KB/s	947 MB/s
Llama 2 70B	6	9.27	1.1 TB	3:28	3.8 MB/s	7.02 GB/s

The initial data load for both model examples had little performance impact on the storage. This is expected since most language- and text-based models have smaller dataset sizes and thus the model load portion of the training is minimal on the storage. This would account for the low read activity on the file system.

The checkpoint data is more interesting. The different parameter sizes of the two examples show the impact on the write throughput requirement on the OneFS file system during the checkpoint operation. The checkpoint during the 70B parameter model required significantly more write throughput than that of the 7B parameter model.

Note that benchmark results are highly dependent upon workload, specific application requirements, and system design and implementation. Relative system performance will vary due to these factors. Therefore, this workload should not be used as a substitute for a specific customer application benchmark when critical capacity planning and/or product evaluation decisions are contemplated. For benchmarking on Dell PowerEdge servers, refer to the MLPerf benchmarking page.

Image model training

Model types can have a major impact on storage performance, as shown in the previous exercise. The goal of this validation was to understand how the storage performance changes when training a dataset of images.

[1] All performance data contained in this report was obtained in a rigorously controlled environment. Results obtained in other operating environments may vary significantly. Dell Technologies does not warrant or represent that a user can or will achieve similar performance results.

Table 5. ResNet training system information

Configurations	ADD HEADER	ADD HEADER
Configuration 1	2x 8-way servers (16xH100 GPUs)	4-node PowerScale F600P cluster
Configuration 2	2x 8-way servers (16xH100 GPUs)	8-node PowerScale F600P cluster

The following list provides the details of our validation setup:

- **Storage Impact** – Measure the Dell PowerScale file system impact during training operations. Add PowerScale nodes and examine the change in file system performance.
- **Training performance** – Examine the training performance when scaling the PowerScale cluster with additional nodes.

Model architecture

ResNet-50 is a real-world image classification dataset that has become the standard benchmark to characterize the performance of a deep learning training workflow on storage and GPU compute platforms. This benchmark performs training of an image classification convolutional neural network (CNN) on labeled images using MXNet. Essentially, the system learns whether an image contains a cat, dog, car, train, and so on. The well-known ILSVRC2012 image dataset—often referred to as ImageNet—was used. This dataset contains 1,281,167 training images in 144.8 GB1. All images are grouped into 1000 categories or classes.

The individual JPEG images in the ImageNet dataset were converted to RecordIO format. The dataset was not resized nor normalized and no preprocessing was performed on the raw ImageNet JPEG images. It maintains the image compression offered by the JPEG format, and the total size of the dataset remained roughly the same (148 GB). The average image size was 115 KB.

Storage performance

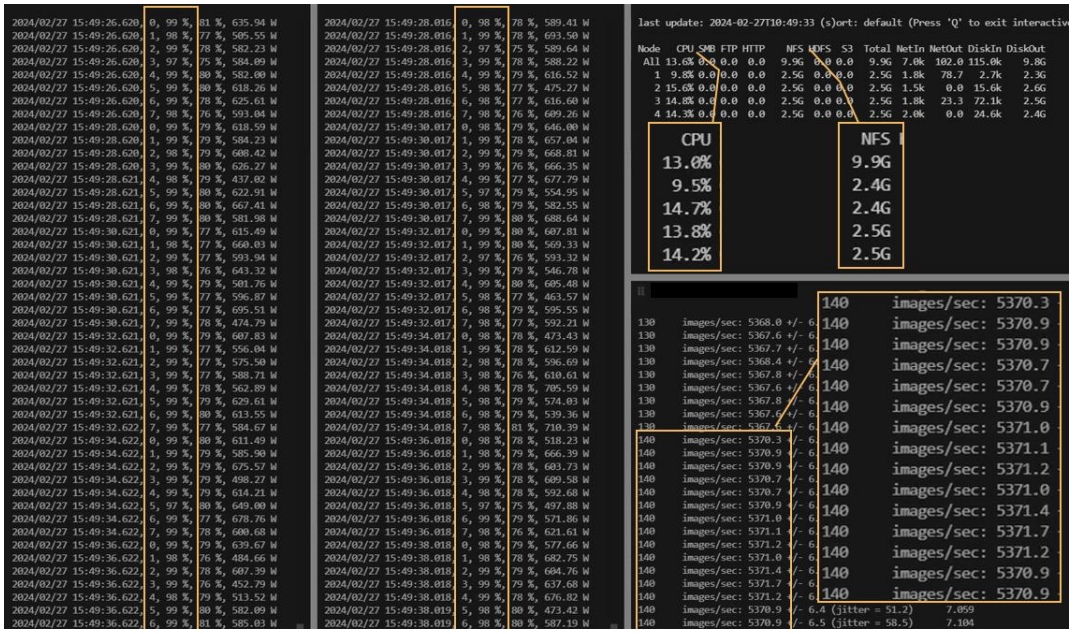
The following table summarizes the findings during the testing. When the PowerScale cluster scales from 4 to 8 nodes, there is a 41% reduction in CPU cycles and 50% reduction in NFS ops across the cluster nodes. The training performance remains consistent for both images/sec per GPU and GPU utilization.

Table 6. Results summary of PowerScale cluster performance.

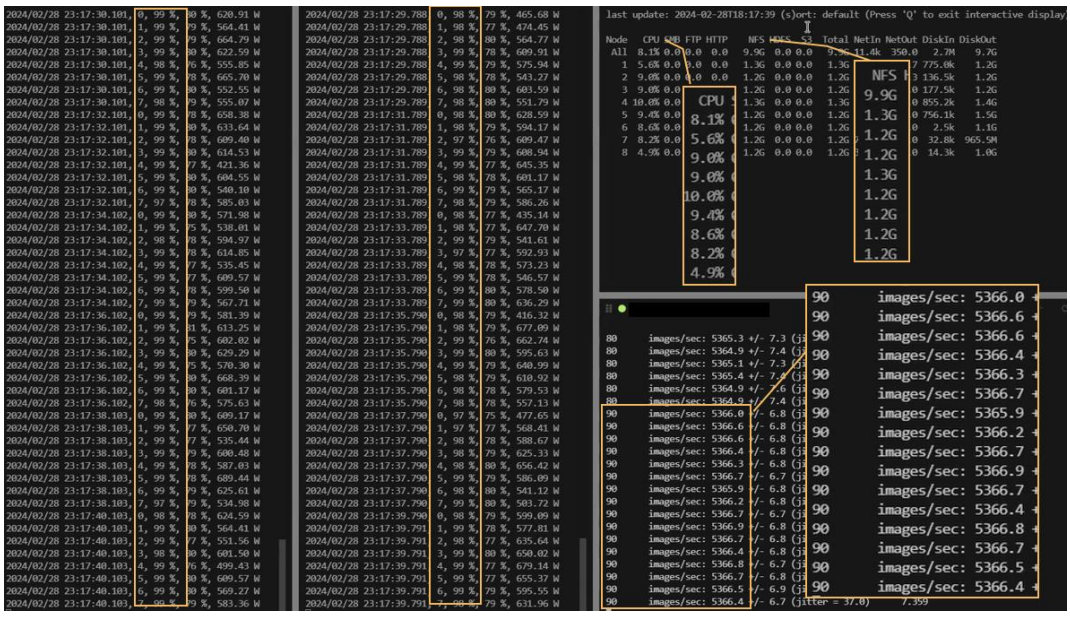
PowerScale Cluster size	CPU	NFS Ops	Images/sec/GPU	GPU %
4 Nodes	13.6%	2.5G/node	5370	99%
8 Nodes	8.1%	1.2G/node	5366	99%

The following figure shows the performance of the PowerScale 4 node cluster through isi statistics, NVIDIA SMI, and the ResNet training logs. The cluster performance in the upper right corner shows the CPU and NFS statistics while the left shows the GPU

utilization as reported by the compute nodes. The bottom-right display shows the images per second per GPU that the system is processing.



This following image shows the performance of the PowerScale cluster after adding 4 additional nodes. Here, we see the reduction in cluster CPU and NFS ops across the 8 PowerScale nodes while the GPU % remains at 99%, and images per second per GPU remains consistent with the 4 node training performance. In both instances, the GPUs were 99% active and therefore fully utilized for the duration of the training exercises.



Conclusion

The Dell Reference Design for Generative AI Model Training with PowerScale provides a comprehensive, scalable, and high-performance architecture for training LLMs. This

design addresses the challenges of LLM training, offering a modular solution that can be tailored to various enterprise use cases.

The design leverages the power of NVIDIA's AI software stack, including NVIDIA AI Enterprise and NVIDIA NeMo, to streamline the development and training of generative AI models. It also provides a robust Dell infrastructure for efficient model training, with considerations for network architecture, software design, and storage performance.

The validation of this design using Llama 2 model architectures demonstrates its effectiveness in delivering reliable and high-performance solutions for generative AI model training. The design offers flexibility in terms of model architectures, allowing organizations to choose the most suitable setup for their needs.

In conclusion, the Dell Reference Design for Generative AI Model Training with PowerScale serves as a valuable guide for organizations interested in understanding storage requirements for training with different model types. It answers questions about how parameter differences and data set differences change the performance requirements of the storage during different phases of training.

References

Dell Technologies documentation

This Dell Technologies documentation provides additional and relevant information related to this solution.

- Dell Technologies Info Hub for AI Solutions
- Dell Technologies Info Hub for PowerScale Analytics
- Dell Reference Design for Generative AI in the Enterprise Model Training

NVIDIA documentation

The following NVIDIA web sites and documentation provide additional and relevant information.

- NVIDIA AI Enterprise NVIDIA NeMo
- Nvidia Base Command Manager Essential
- NVIDIA Data Center GPUs
- NVIDIA Networking