

Generative AI in the Enterprise with NVIDIA GPUs, NVIDIA Networking, and NVIDIA Software Stack


Dell AI Platform with NVIDIA

Abstract

This design guide describes the architecture and design of the Dell AI Platform with NVIDIA, based on Dell PowerEdge XE9680 and R760xa servers with NVIDIA GPU accelerators, Spectrum-X Networking, and AI Enterprise software, with Dell PowerScale storage.

Dell Technologies AI Solutions

Notes, cautions, and warnings

 **NOTE:** A NOTE indicates important information that helps you make better use of your product.

 **CAUTION:** A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.

 **WARNING:** A WARNING indicates a potential for property damage, personal injury, or death.

© 2024 - 2025 Dell Inc. or its subsidiaries. All rights reserved. Dell Technologies, Dell, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

The information in this publication is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose, and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages, or other liability, whether in an action of contract, tort, or otherwise, arising from, out of, or in connection with the publication or the use or other dealings in the publication.

The use, copying, and distribution of any software described in this publication requires an applicable software license.

THIRD-PARTY PRODUCTS DISCLAIMER. This solution be used with products, services, or other items that are provided by a third-party manufacturer or supplier and are not "Dell" or "Dell EMC" branded (collectively, "Third-Party Products"). Notwithstanding any other provisions: (1) such Third-Party Products are subject to the standard license, services, warranty, indemnity, support and other terms of the third-party manufacturer/supplier/community (or an applicable direct agreement between you and such manufacturer/supplier), which shall take priority; and (2) any claims we have acknowledged against Dell in relation to such Third-Party Products are expressly disclaimed and excluded.

Chapter 1: Introduction.....	5
Business challenge.....	5
Overview.....	5
Document purpose.....	6
Audience.....	6
Revision history.....	6
Chapter 2: Solution Architecture.....	7
Overview.....	7
Compute infrastructure.....	8
Network infrastructure.....	8
Cluster management.....	8
Storage infrastructure.....	8
Foundation model.....	8
NVIDIA Enterprise reference architecture.....	9
Software components.....	9
NVIDIA AI Enterprise and associated components.....	9
NVIDIA Run:ai.....	10
Security considerations.....	10
Chapter 3: Management and Compute Infrastructure.....	12
Overview.....	12
Management server configuration.....	12
GPU worker node configuration.....	13
Chapter 4: Networking design.....	15
Overview.....	15
Fabrics overview.....	17
NVIDIA Spectrum-X networking platform.....	17
Network architecture for a PowerEdge XE9680 cluster.....	18
Backend (east-west/GPU) network fabric.....	20
Frontend (north-south) network fabric.....	20
Network architecture for a PowerEdge R760xa cluster.....	20
Cables and optics.....	21
Network topologies for large clusters.....	22
Chapter 5: Rack and Power Design.....	23
Chapter 6: Solution validation.....	25
Overview.....	25
System configuration.....	25
Inference validation.....	26
Fine-tuning validation.....	27
Run:ai validation.....	27

Chapter 7: Inference Performance Characterization.....29
 Overview..... 29
 Use cases..... 29
 Key performance metrics and factors impacting performance..... 32
 Methodology.....33
 Results for the PowerEdge R760xa server with the NVIDIA H100 NVL GPU..... 34
 Results for the PowerEdge XE9680 server with the NVIDIA H200 SXM GPU.....38

Chapter 8: Conclusion..... 40
 Summary.....40
 We value your feedback..... 40

Chapter 9: References..... 41

Introduction

Topics:

- [Business challenge](#)
- [Overview](#)
- [Document purpose](#)
- [Audience](#)
- [Revision history](#)

Business challenge

A growing number of organizations are exploring private artificial intelligence (AI) and generative AI solutions. However, designing and implementing the advanced infrastructure that is required for AI applications presents many challenges.

AI tasks are resource-demanding, require high performance, and rely on error-free data transmissions. Organizations' expertise and timelines for designing and running a working AI solution also bring about more challenges.

Using traditional Ethernet, with its normal congestion, latency, and bandwidth issues, makes it difficult for businesses to achieve the peak network performance required for AI. NVIDIA Spectrum-X, however, provides better throughput and increased performance over traditional Ethernet, and offers the software and hardware component stack that is required to create the necessary infrastructure for the AI environment. This component stack includes the addition of a new and crucial fabric for highly efficient connectivity between GPUs in a GPU cluster to properly handle their tasks.

This Dell AI Platform with NVIDIA, a configuration of the Dell AI Factory with NVIDIA, provides organizations with the building blocks for seamless integration of AI models and frameworks into their operations, enabling them to turn their ideas into practical applications. From digital assistants to new code generation and natural language search applications, the AI Factory framework allows organizations greater control over their proprietary data and scales efficiently, providing a more affordable alternative to many public cloud solutions.

Overview

In the dynamic field of artificial intelligence (AI) and natural language processing, large language models (LLMs) have become essential for tasks such as generating text, answering questions, and analyzing sentiment. The Meta Llama 3 model, used in this design, is known for its exceptional ability to comprehend context and produce coherent responses and exemplifies a powerful conversational agent. Optimizing the performance of these models necessitates a strategic focus on a well-designed architecture that includes high-performance servers, high-speed networking, and scalable storage.

As organizations strive to use the full potential of AI, hardware acceleration emerges as a pivotal element for achieving superior performance. Recognizing this imperative, Dell Technologies and NVIDIA have joined forces to enhance the Dell portfolio for Generative AI in the Enterprise, offering businesses expanded options to support their unique AI initiatives.

This Dell AI Platform with NVIDIA GPUs, NVIDIA Networking, and the NVIDIA Software Stack is a configuration of the Dell AI Factory with NVIDIA, designed to help enterprises accelerate and optimize their AI initiatives. This validated design is based on Dell PowerEdge XE9680 servers with NVIDIA GPUs, NVIDIA Spectrum-X networking, and NVIDIA AI Enterprise software.

High-speed interconnects play a pivotal role in distributed Large Language Model (LLM) training, fine-tuning, and multinode inferencing. They enable efficient data and model parallelism, facilitating rapid communication between multiple GPUs or nodes. With the explosion in the number of parameters in LLM models, typically in the billions, the need for these interconnects is more pronounced. This trend is expected to continue, especially with the introduction of multimodal models. Training, fine-tuning, and inferencing of such massive LLMs necessitate low latency to reduce communication overhead and enhance the performance of the underlying network. These interconnects are crucial for scalability, facilitating the expansion of training across more resources as models and datasets grow.

The NVIDIA Spectrum-X networking platform enhances AI infrastructure deployed with Ethernet. This purpose-built platform improves performance, power efficiency, and predictability for Ethernet-based AI clusters. It outperforms traditional Ethernet solutions, especially for large AI workloads like language model training and fine-tuning. By tightly integrating NVIDIA

Spectrum-4 5600 Series Ethernet switches with the NVIDIA BlueField-3 SuperNIC, Spectrum-X delivers end-to-end network capabilities, reducing run times for massive transformer-based generative AI models. Network engineers, data scientists, and cloud service providers benefit from faster results and informed decision-making.

This validated design addresses generative AI for enterprises using Ethernet as the inter-GPU fabric. In this document, we discuss this design and its use cases. We describe the components, architecture, and other characteristics and provide guidance for designing high-performing generative AI environments for fine-tuning and inferencing.

Document purpose

This design guide describes a configuration of the Dell AI Platform with NVIDIA, a solution that is based on the PowerEdge XE9680 and R760xa servers with NVIDIA GPUs, NVIDIA Networking, and NVIDIA AI Enterprise software.

It provides an overview of the key concepts in the design, a detailed description of the solution architecture, including the hardware and software components, and how they are interconnected by the networking design. It also provides a description of our validation environment, methodology, and results.

Audience

This design guide is intended for enterprise practitioners and experts interested in the implementation of infrastructure and software solutions for generative AI, including professionals and stakeholders involved in the development, deployment, and management of generative AI systems.

Key roles include AI architects, IT infrastructure architects, and designers. Other audience members may include system administrators and IT operations personnel, AI engineers and developers, and data scientists and AI researchers. Some knowledge of AI model development and life cycle, generative AI principles, and terminology is assumed.

Revision history

The following table lists the revision history.

Table 1. Revision history

Date	Version	Change summary
November 2024	1.0	Initial release
January 2025	1.1	<ul style="list-style-type: none">• Addition of the NVIDIA H200 SXM GPU on the Dell PowerEdge XE9680 and the NVIDIA H100 NVL GPU on the Dell PowerEdge R760xa server.• Update to NVAIE 5.2 software.
May 2025	1.2	<ul style="list-style-type: none">• Support for NVIDIA AI Enterprise 6.2 and Ubuntu 24.04• Support for NVIDIA Enterprise Reference Architecture for PowerEdge R76xa-based cluster
July 2025	1.3	Addition of Run:ai for GPU orchestration

Solution Architecture

Topics:

- [Overview](#)
- [Compute infrastructure](#)
- [Network infrastructure](#)
- [Cluster management](#)
- [Storage infrastructure](#)
- [Foundation model](#)
- [NVIDIA Enterprise reference architecture](#)
- [Software components](#)
- [Security considerations](#)

Overview

This solution uses the NVIDIA Spectrum-X networking platform to provide high-speed networking, offering organizations guidelines and best practices to design and implement scalable, efficient, and reliable infrastructure that is specifically tailored for generative AI.

The following figure shows the key components of the reference architecture:

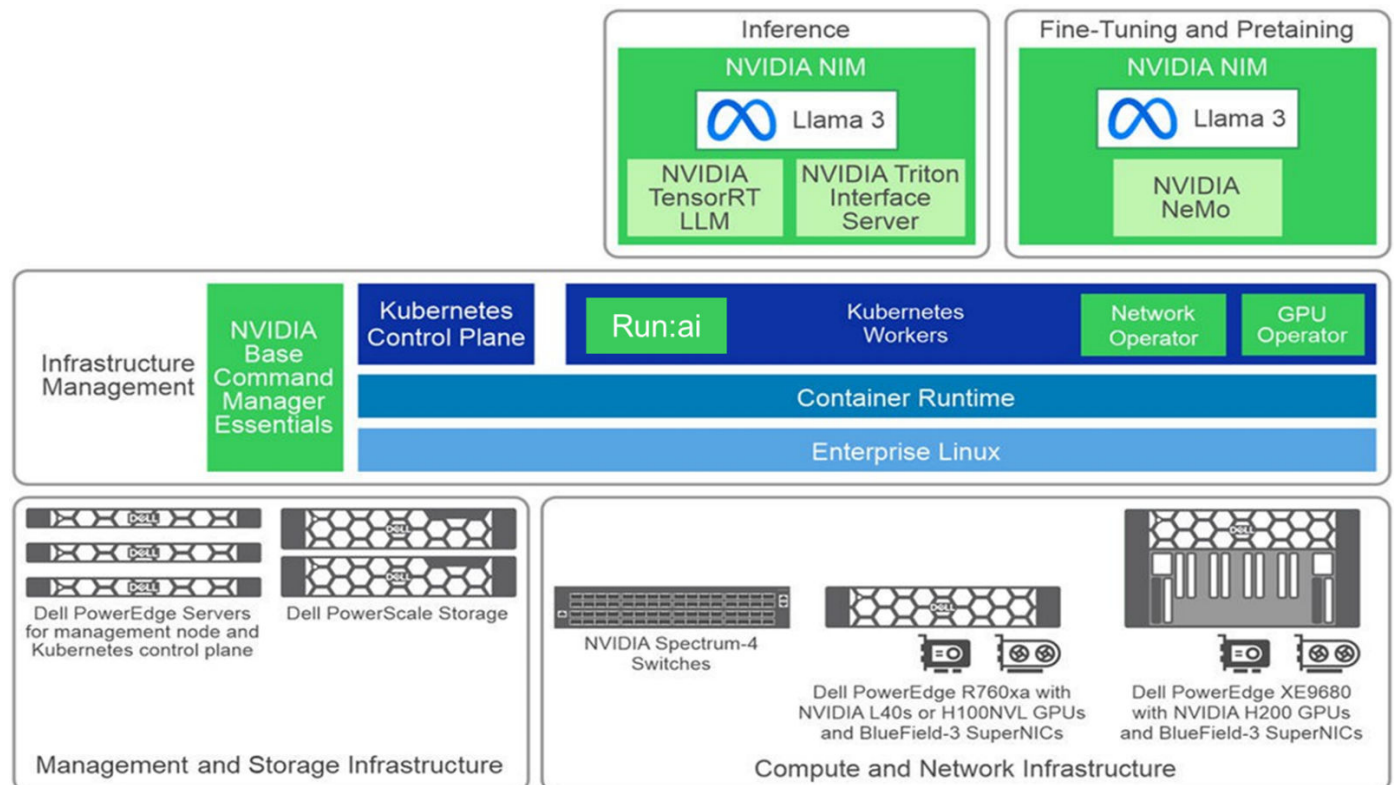


Figure 1. Reference architecture for enterprise generative AI infrastructure with NVIDIA Spectrum-X networking platform

The following sections describe the key components of the reference architecture.

Compute infrastructure

Dell Technologies offers a range of acceleration-optimized servers, which are equipped with NVIDIA GPUs, to handle the intense compute demands of LLMs. The following server models with GPU options and networking options are available based on this solution as compute resources for deploying LLM models in production:

- PowerEdge XE9680 server equipped with:
 - 8 x NVIDIA H100 SXM GPUs or 8 x NVIDIA H200 SXM GPUs with NVSwitch
 - 8 x NVIDIA Bluefield-3 Single Port 400 GbE QSFP112 PCIe FH (B3140H)
 - 1 x NVIDIA Bluefield-3 Dual Port 200 GbE QSFP112 PCIe FH (B3220L)
- PowerEdge R760xa servers equipped with:
 - 4 x NVIDIA H100 NVL GPUs or 4 x NVIDIA L40S GPUs
 - 2 x NVIDIA Bluefield-3 Single Port 400 GbE QSFP112 PCIe FH (B3140H)
 - 1 x NVIDIA Bluefield-3 Dual Port 200 GbE QSFP112 PCIe FH (B3220L)

Network infrastructure

This design incorporates three networks:

- A frontend network for management, storage, client/server traffic (sometimes referred to as north/south traffic)
- A backend network for internode GPU communication (sometimes referred to as east/west traffic) that is used for distributed training
- An out-of-band network for server management

NVIDIA Spectrum switches that run the Cumulus network operating system power these physical networks. Additionally, the storage infrastructure has its backend network that is powered by Dell PowerSwitch Networking. For more information, see the Networking Design [Overview](#).

Cluster management

NVIDIA Base Command Manager Essentials streamlines infrastructure provisioning, workload management, and resource monitoring. It provides all the tools you need to deploy and manage an AI data center. Base Command Manager can deploy Kubernetes or all the software required for running AI workloads on GPUs. It allows customers to manage both Kubernetes and Slurm clusters seamlessly. Base Command Manager Essentials can be used to configure the PowerEdge server either as part of the Kubernetes cluster or the Slurm cluster. They can be reconfigured quickly by rebooting. This method allows administrators to allocate resources quickly to either of the clusters on demand and with minimal overhead.

Storage infrastructure

This design uses Dell PowerScale storage as a repository for datasets for model customization, models, model versioning and management, and model ensembles.

The flexible and robust storage capabilities of PowerScale offer the scale and speed necessary for training and operationalizing AI models, providing a foundational component for AI workflow. Its capacity to handle the vast data requirements of AI, combined with its reliability and high performance, cements the crucial role that external storage plays in successfully bringing AI models from conception to application.

There are several models used in generative AI solution architectures, such as PowerScale F210 and PowerScale F710, all powered by the PowerScale OneFS operating system and supporting inline data compression and deduplication. The minimum number of PowerScale nodes per cluster is three nodes. The maximum cluster size is 252 nodes, which supports up to 186 PB of raw capacity.

Foundation model

Pretraining, fine-tuning, and multi-node distributed inferencing large models such as Llama 3 require high performing interconnect to ensure that GPUs in different nodes can communicate efficiently. Based on open standards, the Dell

PowerSwitch Z series high-performance networking platform provides a scalable networking infrastructure for pretraining and fine-tuning such large models.

NVIDIA Enterprise reference architecture

NVIDIA's Enterprise Reference Architectures (Enterprise RAs) are designed to provide clear and consolidated recommendations for system partners and enterprise customers building AI Factories. This approach eliminates the guesswork and risk that is associated with deployment, ensuring optimal performance, utilization, uptime, total cost of ownership (TCO), and supportability. The goal is to help partners and customers achieve value sooner and maximize their return on investment. NVIDIA's extensive testing and best practices guide the configuration of systems to maximize performance and avoid common pitfalls, enabling faster and more confident delivery of AI solutions.

NVIDIA has a structured methodology for introducing reference architectures for new technologies, such as Hopper GPUs, Blackwell GPUs, Grace CPUs, Spectrum-X networking platform, and BlueField architecture-based offerings. Each new technology comes with configuration guides to assist partners in designing, building, and deploying optimized system configurations. The Enterprise RAs are tailored for enterprise-class deployments, supporting a diverse range of workloads and providing a versatile foundation for enterprise AI. Each reference architecture is designed around NVIDIA-certified servers, ensuring optimal performance and efficient deployment. For more information, see [NVIDIA Enterprise Reference Architecture Overview](#) white paper.

Dell AI Factories that are based on the Spectrum-X Networking Platform adhere to architectural principles, guidelines, deployment processes, and validated component versions. The PowerEdge R760xa cluster configuration aligns with NVIDIA Enterprise Reference Architecture 2-4-3. For future updates to Dell's design, the PowerEdge XE9680 server will align with NVIDIA Enterprise Reference Architecture 2-8-9.

Software components

NVIDIA AI Enterprise and associated components

NVIDIA AI Enterprise is a cloud-native platform that streamlines development and deployment of production-grade AI solutions including generative AI, computer vision, speech AI, and more. See the NVIDIA AI Enterprise documentation for more information about the components available with [NVIDIA AI Enterprise](#). The following networking components that are incorporated in this design are available as part of NVIDIA AI Enterprise:

- Base Command Manager (BCM) Essentials, which facilitates seamless operationalization of AI development at scale by providing features like operating system provisioning, firmware upgrades, network and storage configuration, multi-GPU and multinode job scheduling, and system monitoring.
- NVIDIA NeMo Framework is a scalable and cloud-native generative AI framework that is built for researchers and developers working on [Large Language Models](#), [Multimodal](#), and [Speech AI](#) (for example, [Automatic Speech Recognition](#) and [Text-to-Speech](#)). It enables users to efficiently create, customize, and deploy new generative AI models by leveraging existing code and pretrained model checkpoints.
- NVIDIA Network Operator, which manages networking-related components to enable RDMA and GPUDirect for workloads in a Kubernetes cluster. The goal of Network Operator is to manage all networking-related components to enable running workloads in a Kubernetes cluster.
- The NVIDIA GPU Operator simplifies the deployment and management of GPU resources in Kubernetes environments. By automating the installation and configuration of the necessary drivers, runtime libraries, and monitoring tools, it ensures optimal performance and reliability for GPU-accelerated workloads. This operator streamlines the integration of NVIDIA GPUs into containerized applications, enabling seamless scaling and efficient resource use. With its robust support for various Kubernetes distributions, the NVIDIA GPU Operator is an essential tool for organizations looking to use the power of GPUs in their cloud-native infrastructure.
- NVIDIA NIM is a set of optimized cloud-native microservices designed to simplify and accelerate the deployment of AI models in production environments. Part of NVIDIA AI Enterprise, NIM abstracts the complexities of AI model development and packaging, allowing developers to deploy generative AI models across various infrastructures, including cloud, data centers, and GPU-accelerated workstations. It supports industry-standard APIs, domain-specific models, and optimized inference engines, ensuring high performance and scalability. NIM also provides enterprise-grade support and security, making it a powerful tool for businesses looking to integrate AI into their operations efficiently.

NVIDIA Run:ai

NVIDIA Run:ai accelerates AI operations with dynamic orchestration across the AI life cycle. It maximizes GPU use, scales AI workloads efficiently, and integrates seamlessly into hybrid infrastructure with minimal overhead. The platform is built on a microservices architecture and provides a range of software components that work together to provide a consistent and efficient AI computing experience. Some of the key features of Run:ai include:

- GPU fractions—Run:ai GPU fractions allow you to split a single GPU into multiple virtual GPUs, enabling multiple workloads to share the same physical hardware. This feature is useful for large-scale AI workloads that require significant computing resources.
- Node pooling—The Run:ai node pool feature allows administrators to group and allocate or deallocate computing resources as needed. This feature is useful for workloads distribution and training that require certain types of compute resources.
- Job scheduler—The Run:ai job scheduler is a highly scalable and efficient scheduling system that allows you to manage and prioritize large numbers of AI workloads. The scheduler considers factors such as workload priority, resource availability, and job dependencies to ensure that workloads are run efficiently and effectively.
- GPU memory swap—The Run:ai GPU memory swap feature allows you to swap a model from GPU memory to CPU memory, enabling more efficient use of computing resources. This feature is useful for expanding the GPU physical memory to the CPU memory, which is typically an order of magnitude larger than that of the GPU.
- HuggingFace integration—Run:ai integration with Hugging Face allows you to use frameworks such as TGI and vLLM for natural language processing tasks. The integration provides a consistent and efficient way to deploy and manage Hugging Face models on the Run:ai platform.

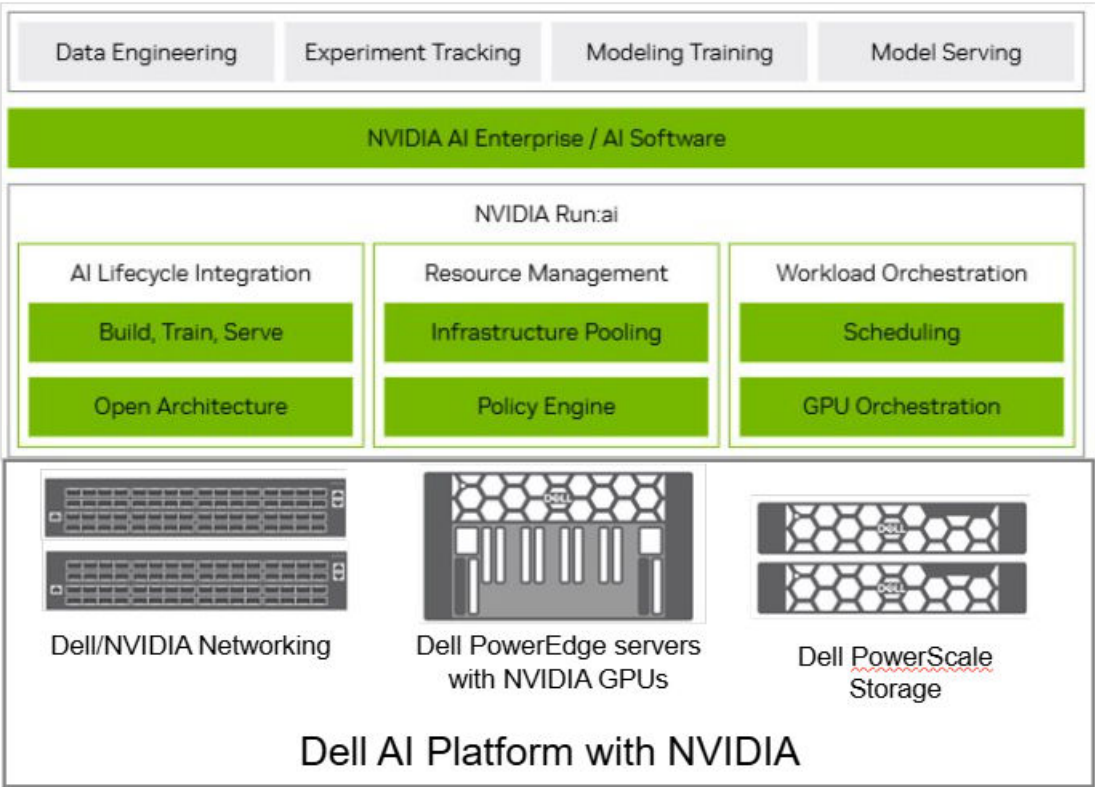


Figure 2. NVIDIA Run:ai on Dell AI Platform

Security considerations

AI platforms exist at the intersection of rapidly evolving AI technology and the ever-challenging security landscape. This unique position brings both incredible opportunities and significant risks. The security of any system exists in relation to the system's threat profile: the value of a system's assets, the likelihood of an attempted attack, and the nature of potential attackers in

terms of their resources, capabilities, and incentives. Furthermore, there is often a tension between security and convenience. Extra security measures might interfere with desirable use cases or lead to false-positive results. Therefore, Dell Technologies advises customers and architects to consider their specific operational contexts and conduct appropriate analysis to validate the applicability of the recommendations in this guide. Depending on the threat profile, additional services and security measures might be advisable.

The [Dell AI Platforms security best practices guide](#) describes some of the potential threats to an AI Platform and offers guidance on how to mitigate them. It also describes some common security configuration procedures, followed by guidance related to categories of potential threats and their respective mitigations.

Note that the recommendations in the security best practices guide are not a complete security evaluation of the Dell AI Platform in its entirety or its individual components, nor a complete threat and risk analysis. You can also find additional information at the [Dell Security and Trust Center](#).

Management and Compute Infrastructure

Topics:

- [Overview](#)
- [Management server configuration](#)
- [GPU worker node configuration](#)

Overview

Selecting the appropriate server and network configuration for generative AI model customization is crucial to ensure that adequate resources are allocated for model training. This section provides example configurations for both management and compute workloads and network architecture.

Management server configuration

The following table provides the recommended minimum configuration for the PowerEdge R660 head node and the control plane nodes:

Table 2. PowerEdge R660 head node and control plane configuration

Component	BCM head node and control plane nodes
Server model	4 x PowerEdge R660
CPU	1 x Intel Xeon Gold 6526Y 2.8G, 16C/32T, 20 GT/s, 37.5M Cache, Turbo, HT (195 W) DDR5-5200
Memory	4 x 16 GB RDIMM, 5600 MT/s, Single Rank
Operating system	BOSS-N1 controller card + with 2 M.2 960 GB (RAID 1)
RAID controller	PERC H755 with rear load Brackets
Storage	4 x 3.84 TB SSD SAS RI 24 Gbps 512e 2.5in Hot-Plug, AG Drive 1DWPD
OOB/PXE network (optional)	Broadcom 5720 Dual Port 1 GbE LOM
Frontend network	NVIDIA Mellanox ConnectX-6 DX Dual Port 100 GbE QSFP56 Network Adapter, Low Profile

Consider the following recommendations for the NVIDIA Base Command Manager Essentials and Kubernetes control plane configuration:

- We recommend four PowerEdge servers for management. Install NVIDIA Base Command Manager Essentials on one of the servers. If you require high availability, install NVIDIA Base Command Manager Essentials on two nodes as active-passive nodes. Install the Kubernetes control plane on the other three nodes.
- We recommend the same hardware configuration for all management servers for ease of configuration and maintenance.
- Because the node BCM and Kubernetes nodes do not require heavy computing, a single-processor server is sufficient.
- We recommend a storage-rich configuration to facilitate convenient storage of images and other essential tools. We recommend a minimum of four SSD drives. You can choose more drives or upgrade to NVMe for better performance.
- You can implement a PXE network using an isolated 1 GbE LOM or using VLANs with the frontend network.

GPU worker node configuration

Dell Technologies provides a selection of two GPU-optimized servers suitable for configuration as worker nodes for generative AI model customization: the PowerEdge R760xa and PowerEdge XE9680 servers. You have the flexibility to choose one of these PowerEdge servers that is based on the specific model size that you require. Larger models, characterized by a greater parameter size, require servers that are equipped with a higher GPU count and enhanced connectivity.

The GPU-optimized servers act as worker nodes in a Kubernetes cluster. The number of servers depends on the size of the model, the customization method, and the end-user requirements on training time. Larger models, characterized by a greater parameter size, require servers that are equipped with a higher GPU count and enhanced connectivity.

PowerEdge R760xa GPU worker node

The following table shows a recommended configuration for a PowerEdge R760xa GPU worker node:

Table 3. PowerEdge R760xa GPU worker node

Component	Details
Server model	PowerEdge R760xa
CPU	2 x Intel Xeon Gold 6548Y+ 2.5G, 32C/64T, 20 GT/s, 60M Cache, Turbo, HT (250 W) DDR5-5200
Memory	16 x 64 GB DDR5 5600 MT/s RDIMM
Operating system	BOSS-N1 controller card + with 2 M.2 960 GB (RAID 1)
Storage	2 x 3.84 TB Data Center NVMe Read Intensive AG Drive U2 Gen4
Frontend network	1 x NVIDIA BlueField-3 Dual Port 200 GbE QSFP112 PCIe Full Height (B3220L)
GPU	Either: <ul style="list-style-type: none">• 4 x NVIDIA H100 NVL, 94 GB PCIe GPUs• 4 x NVIDIA L40S, 48 GB PCIe GPUs
Backend network	2 x NVIDIA BlueField-3 Single Port 400 GbE QSFP112 PCIe Full Height (B3140H)

PowerEdge XE9680 GPU worker node

The PowerEdge XE9680 server is a two-socket, 6U server that is designed specifically for AI tasks. It supports eight accelerators, ideal for machine learning and deep learning training and inferencing workloads, especially for those workloads that are training LLMs.

The PowerEdge XE9680 server with the NVIDIA H200 accelerator offers high-performance capabilities for enterprises seeking to unlock the value of their data and differentiate their business with customized LLMs. The following table provides a recommended configuration for a PowerEdge XE9680 GPU worker node:

Table 4. PowerEdge XE9680 GPU worker node

Component	Details
Server model	PowerEdge XE9680
CPU	2 x Intel Xeon Platinum 8562Y+ 2.8G, 32C/64T, 20 GT/s, 60M Cache
Memory	16 x 128 GB RDIMM, 5600 MT/s Dual Rank
Operating system	BOSS-N1 controller card + with 2 M.2 960 GB (RAID 1)
Storage	2 x 3.84 TB Data Center NVMe Read Intensive AG Drive U2 Gen4
Frontend network	1 x NVIDIA BlueField-3 Dual Port 200 GbE QSFP112 PCIe Full Height (B3220L)
GPU (accelerator)	Either: <ul style="list-style-type: none">• 8 x NVIDIA H100 SXM• 8 x NVIDIA H200 SXM

Table 4. PowerEdge XE9680 GPU worker node (continued)

Component	Details
Backend network	8 x NVIDIA BlueField-3 Single Port 400 GbE QSFP112 PCIe Full Height (B3140H)

The CPU memory allocation in the PowerEdge XE9680 GPU worker node configuration must be greater than 1.5 times the combined GPU memory footprint. Therefore, we recommend a minimum of 2 TB of total RAM space. While LLM tasks primarily rely on GPUs and do not significantly tax the CPU and memory, it is advisable to equip the system with high-performance CPUs and larger memory capacities. This provisioning ensures sufficient capacity for various data processing activities, machine learning operations, monitoring, and logging tasks. Our goal is to guarantee that the servers provide ample CPU and memory resources for these functions, preventing any potential disruptions to the critical AI operations on the GPUs.

Networking design

Topics:

- [Overview](#)
- [Fabrics overview](#)
- [NVIDIA Spectrum-X networking platform](#)
- [Network architecture for a PowerEdge XE9680 cluster](#)
- [Backend \(east-west/GPU\) network fabric](#)
- [Frontend \(north-south\) network fabric](#)
- [Network architecture for a PowerEdge R760xa cluster](#)
- [Cables and optics](#)
- [Network topologies for large clusters](#)

Overview

There are many types of applications that exist in a data center or network fabric. These applications can be loss-tolerant (that is, they can tolerate packet drops or are latency insensitive) or loss-intolerant (that is, they cannot tolerate packet drops or are latency sensitive).

Among these applications, few are as unique as AI whose data traffic pattern is characterized by:

- A large volume of data exchanged, particularly with LLMs and similar language models
- High-rate data exchange taking place during the initial training phase of any model
- Latency-sensitive applications exchanging vast amounts of data in a feedback loop fashion
- Diverse traffic patterns (for example, predictable and somehow ordered or unpredictable)
- Heterogeneous traffic size flows (for example, elephant and mice flows)
- Bursty dataset patterns

These workload characteristics differentiate AI because it must adhere to strict infrastructure requirements to be of any use to any organization.

Fabrics for generative AI compute clusters are challenged with delivering the highest bandwidth and lowest latency data transfer while avoiding packet loss or any kind of retransmission delays.

Due to the massive data volumes being pushed through the fabrics in support of AI workloads, these fabrics operate as closely as possible to saturation characterized by highly parallelized transmission of multiple elephant flows.

Finally, effective use of compute resources depends on minimizing delays due to the network transfers to allow the parallel compute jobs to progress in a synchronized fashion.

Unfortunately, traditional Ethernet is not sufficient for most AI infrastructure due to its inherent congestion, high latency, and unfair bandwidth. The Ethernet of yesterday cannot handle the tough requirements of AI traffic.

Ethernet fabrics for AI must deliver key features and use open standards to become a compelling fabric interconnect of choice for the AI world. The following sections describe these features.

Interoperable

The fabric must operate at the highest level and use a well-established network ecosystem that is based on proven open standards such as Ethernet. With an Ethernet-based approach, a flexible architecture can be achieved.

High performance

AI workloads are unique in that they require specific network or fabric properties to perform optimally.

Scalable

The requirements of AI workloads can range from a single GPU to a cluster of multiple GPUs. This design guide is relevant to the various AI fabric topologies: single switch, TOR-wired Clos topology, Pure Rail topology, and Rail Optimized topology.

Lossless

GPUDirect RDMA, which is specifically engineered for GPU acceleration, facilitates direct interaction between NVIDIA GPUs across different systems, circumventing system CPUs and removing the necessity for data buffer copies through the system memory. When run over RoCE, GPUDirect RDMA attains its best performance, particularly when implemented on a lossless network.

Load balancing

In a leaf and spine architecture, the fabric is Layer 3. It uses NVIDIA RoCE Adaptive Routing on equal cost multipath (ECMP) links, which results in a uniform traffic distribution across all links between the leaf and spine switches. NVIDIA Direct Data Placement (DDP) Technology augments RoCE Adaptive Routing by correcting the order of packets that are received in the receiving host/GPU memory.

Adaptive routing

The Spectrum-X platform also includes performance isolation measures to ensure that workloads do not impact each other's performance. With Spectrum-X's RoCE Adaptive Routing, performance isolation is attained using fine-grained data path balancing to avoid collision of flows across the leaf and spine.

Congestion control

NVIDIA Spectrum-X uses [Scaling Zero Touch RoCE Technology with Round-Trip Time Congestion Control](#) for end-to-end congestion control. This approach enables the SuperNIC to rate-limit transmissions based on telemetry data obtained from the switch. Congestion control provides AI environments with better throughput and increased performance over traditional Ethernet.

Multi-GPU communications

The NVIDIA Collective Communication Library (NCCL) optimizes communication between GPUs in multi-GPU and multinode setups. It provides high-performance collective communication primitives such as all-gather, all-reduce, broadcast, reduce, and reduce-scatter, which are crucial for distributed deep learning and other parallel computing tasks. NCCL is topology-aware, meaning it can automatically detect and use the most efficient communication paths across different interconnects like PCIe, NVLink, Ethernet, and InfiniBand. This capability helps in discovering the optimal network topology, ensuring high bandwidth and low latency communication, which is essential for scaling applications across multiple nodes. By using NCCL, tools such as NeMo Frameworks can achieve significant performance improvements in multinode clusters, making it easier to manage and optimize large-scale distributed computing environments.

Fabrics overview

A single GPU cluster can consist of up to four fabrics. The following figure shows an example of two AI servers (Dell PowerEdge XE9680 servers) connecting to all four network types:

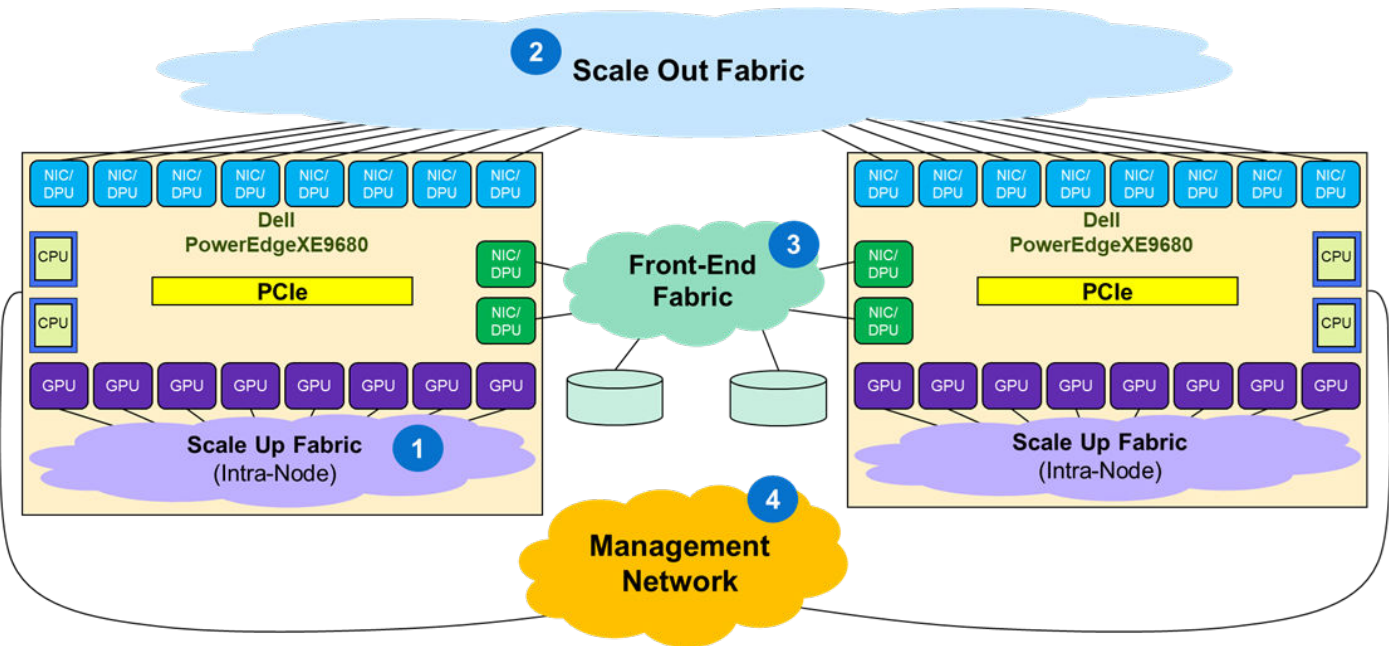


Figure 3. Example of a Dell PowerEdge XE9680-based GPU Cluster

This figure shows the different fabrics that support the deployment of a typical AI workload:

1. Scale Up Fabric–This fabric interconnects the eight NVIDIA GPUs to form a high-bandwidth domain through technologies such as NVIDIA NVLink and NVLink Switch.
2. Scale Out Fabric–Also known as a backend fabric or AI fabric, the set of GPUs composing a GPU cluster uses this fabric to exchange AI workload parameters.
3. Front-End Network–This is the traditional network fabric that supports the deployment of storage, application, and in-band cluster management components of the network.
4. Management Network–This fabric supports the overall environment management network for the solution.

Depending on the scale of the cluster, the front-end network, the management network and the backend network can be consolidated into a single physical network. In this design, AI clusters that are based on the PowerEdge R760xa server use the NVIDIA Spectrum SN5600 switch, which consolidates the above networks and aligns with NVIDIA Enterprise Reference Architecture.

NVIDIA Spectrum-X networking platform

This reference architecture incorporates NVIDIA Spectrum SN5600 Ethernet switches with 64 ports of 800 Gb/s. In harmony with the NVIDIA BlueField-3 SuperNIC, the Spectrum-X networking platform provides advanced feature sets vital for multitenant generative AI, optimizing performance in cloud and enterprise infrastructures.

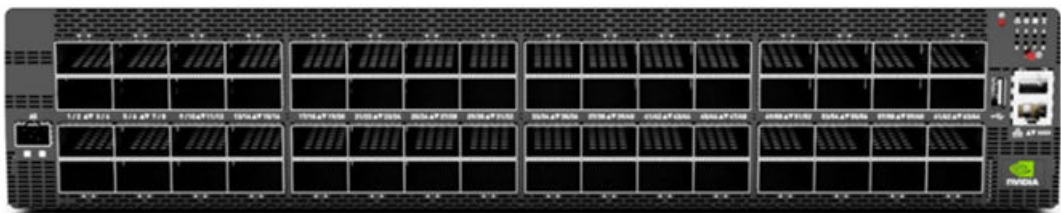


Figure 4. NVIDIA Spectrum SN5600 switch

The NVIDIA Spectrum SN5600 Ethernet switch, the world’s first Ethernet switch purpose-built for AI consists of:

- 64 x 800 GbE, 128 x 400 GbE, or 256 x 200 GbE ports for GPU-to-GPU traffic
- Extreme low latency; extreme effective bandwidth at scale
- NVIDIA RoCE extensions for scalable AI
- Standard Ethernet connectivity
- Full stack and end-to-end optimization
- Open operating system: Cumulus

NVIDIA B3140H BlueField-3 SuperNIC

As part of the Spectrum-X networking platform, the Bluefield-3 SuperNIC is installed in the PowerEdge AI servers. It is a new class of network accelerator purpose-built for connecting GPU servers over the Scale Out fabric, extending the robust performance of the AI fabric to the AI servers.



Figure 5. NVIDIA B3140H BlueField-3 SuperNIC

The NVIDIA B3140H BlueField-3 SuperNIC is a 400 GbE E-Series with eight Arm-Cores, single port QSFP112, PCIe Gen 5.0 x16, 16 GB DDR5 memory, and integrated BMC. The form factor is a single slot, half height, and half length.

NVIDIA B3220L BlueField-3 SuperNIC

The NVIDIA B3220L BlueField-3 SuperNIC is used for connecting storage and other devices over the frontend fabric.



Figure 6. NVIDIA B3220L BlueField-3 SuperNIC

The NVIDIA B3220L BlueField-3 SuperNIC is a 200 GbE E-Series with eight Arm-Cores, dual-port QSFP112, PCIe Gen 5.0 x16, 16 GB DDR5 memory, and integrated BMC. The form factor is single slot, full height, and half length.

For more information about the available BlueField options, see the [NVIDIA BlueField-3 Networking Platform User Guide](#).

Network architecture for a PowerEdge XE9680 cluster

This solution uses the NVIDIA Spectrum-X networking platform, and consists of three physical networks:

- Frontend network for management, storage, and client/server traffic that is powered by two Spectrum-4 SN5600 switches

- Backend network for internode GPU communication using one or more Spectrum-4 SN5600 switches
- Out-of-band management traffic that is powered by the Spectrum SN2201 switch

The following figure shows the network topology that is used for the solution:

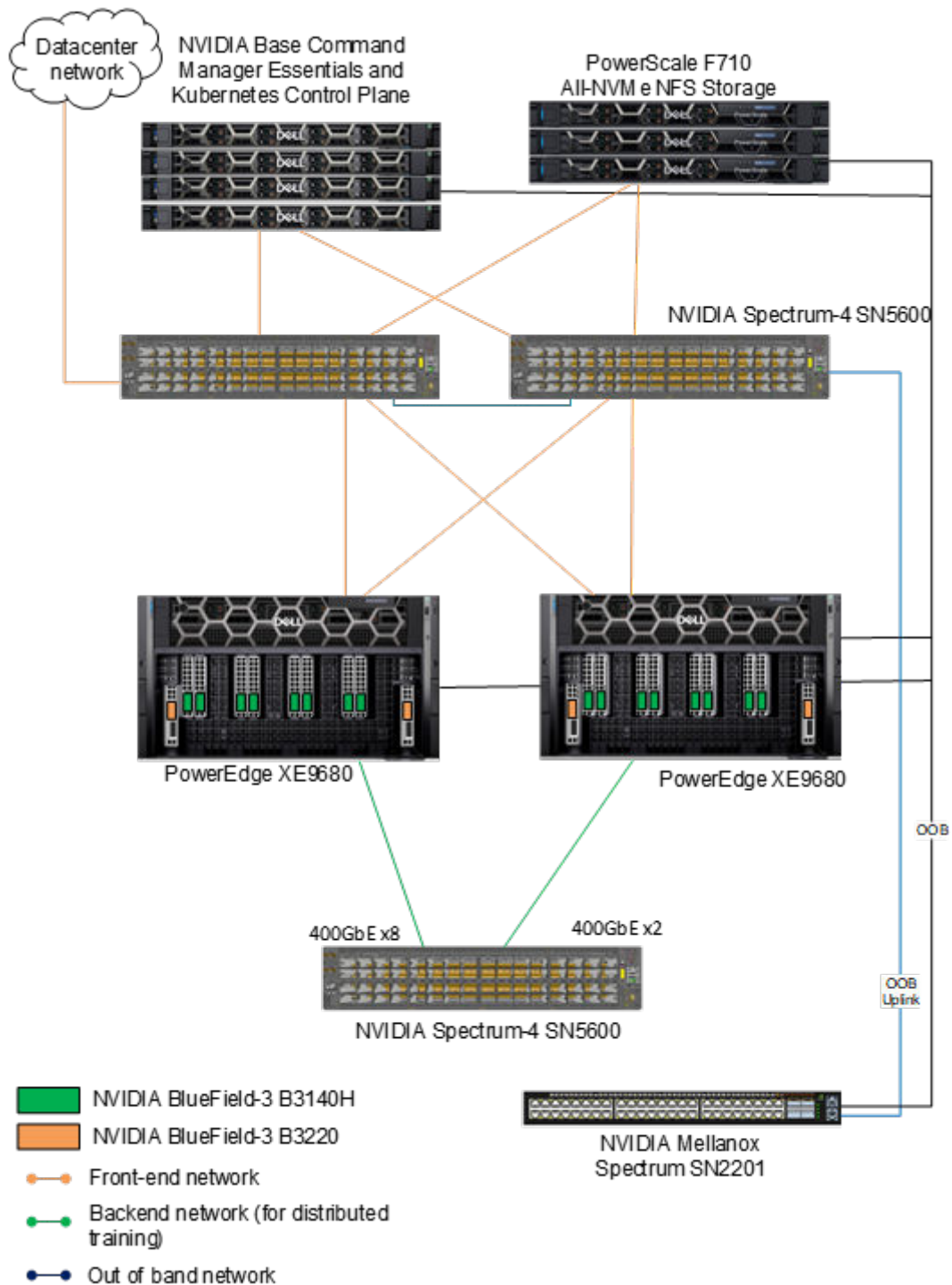


Figure 7. Network configuration for PowerEdge XE9680-based cluster

NOTE:

The figure shows only two PowerEdge XE9680 servers for the purpose of illustration.

Backend (east-west/GPU) network fabric

The backend network fabric, also referred to as the GPU, Scale Out, or east-west network fabric, is a 400 Gb Ethernet high-bandwidth low latency network fabric dedicated for inter-node GPU fabric communication. It is required for distributed fine-tuning of LLMs to facilitate rapid communication between different nodes in a distributed computing environment. Backend network fabric is typically not required for inference-only clusters, in which the model fits inside a single server. This design uses the Llama 3 model, which fits in a single PowerEdge XE9680 server for inferencing and a backend network fabric is not needed.

The backend network is powered by a Spectrum-4 SN5600 switch, which is a versatile network switch that serves as a smart-leaf, spine, and super-spine. It provides 64 ports of 800 GbE, all packed into a compact 2U form factor. This switch is suitable for an inter-GPU fabric. It supports both conventional leaf/spine designs with top-of-rack (ToR) switches and end-of-row (EoR) topologies. The SN5600 switch offers a wide range of connectivity options, from one to 800 GbE. It stands out in the industry with a leading total throughput of 51.2Tb/s.

Spectrum-X incorporates the following capabilities to support innovations to achieve the highest effective bandwidth under load and at scale:

- NVIDIA RDMA over Converged Ethernet (RoCE) Adaptive Routing on Spectrum-X switches
- NVIDIA Direct Data Placement (DDP) on Spectrum-X SuperNIC
- NVIDIA RoCE Congestion Control on both Spectrum-X switches and SuperNICs
- NVIDIA AI Acceleration Software
- End-to-end AI network visibility using NVIDIA NetQ

There are 10 PCIe Gen 5 slots on the PowerEdge XE9680 server, which are internally connected to the CPU and GPU using PCIe switches. Each of the 10 PCIe slots is configured with eight NVIDIA Bluefield-3 Single Port 400 GbE B3140H adapters that are connected to a Spectrum-4 5600 switch, which constitutes the backend fabric.

Frontend (north-south) network fabric

The frontend network fabric in this design refers to the standard Data Center Fabric providing connectivity to resources outside the AI cluster. As shown in Figure 6 , it is a consolidated fabric that supports management storage, external data center access, and Kubernetes management operations. This network fabric is often referred to as the north-south network.

A pair of Spectrum-4 SN5600 switches power the frontend network fabric. These switches connect to the PowerScale F710 storage arrays and the PowerEdge R660 Kubernetes and NVIDIA Base Command Manager. Also, the switches provide external access by connecting to the data center network. They also connect to the PowerEdge XE9680 server node using NVIDIA Bluefield-3 Dual Port 200 GbE.

Network architecture for a PowerEdge R760xa cluster

This solution uses the NVIDIA Spectrum-X networking platform. Dell Technologies aligns with NVIDIA Enterprise Reference Architecture and has one consolidated physical network that is supported by two Spectrum-4 SN5600 switches for frontend network (management, storage, and client/server traffic) and backend network. The Spectrum SN2201 switch powers out-of-band management traffic.

The following figure shows the network topology that is used for the solution:

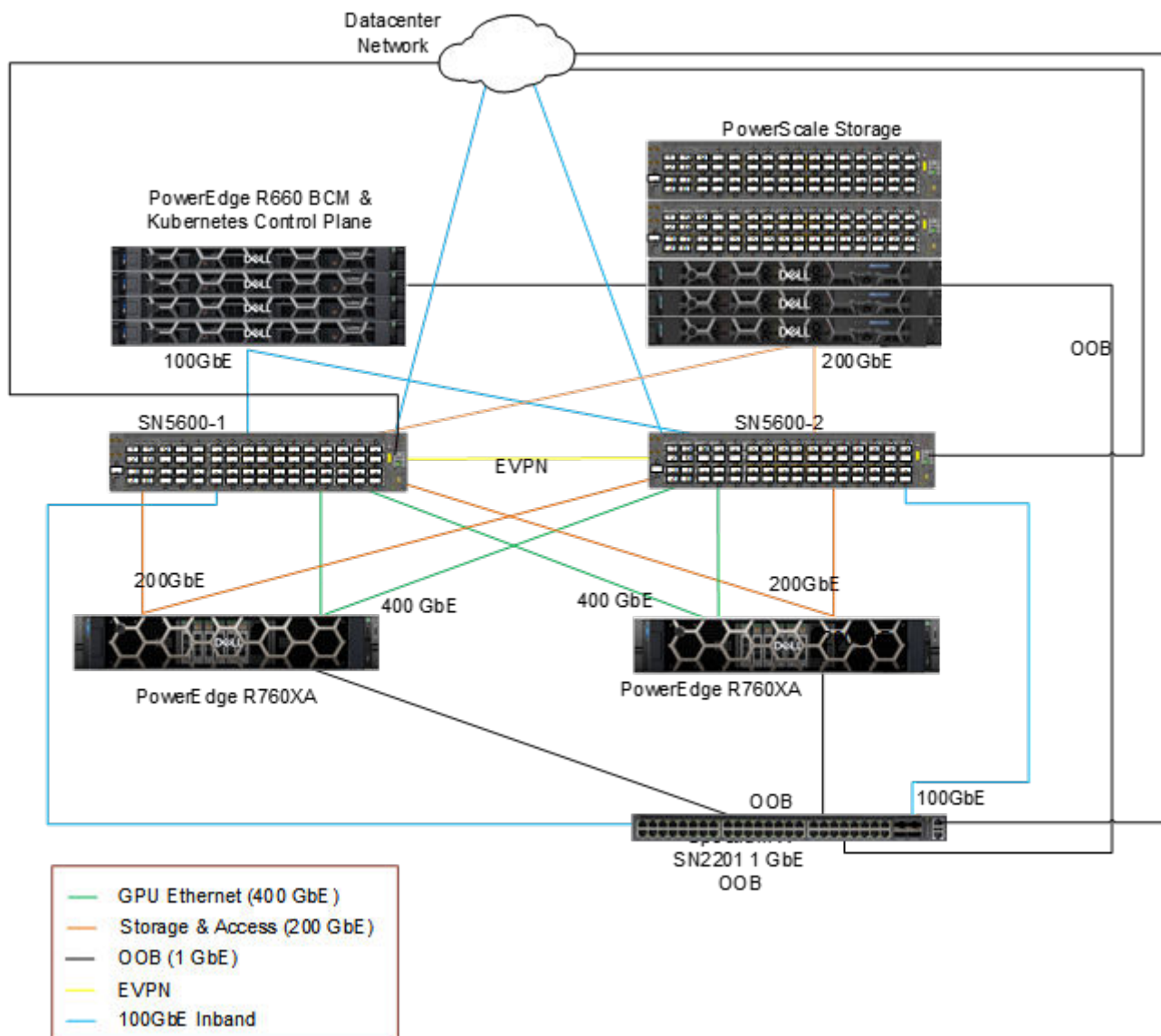


Figure 8. Network configuration for PowerEdge R760xa-based cluster

NOTE: The figure shows only a single PowerEdge XE9680 server and a single PowerEdge R760xa server for the purpose of illustration.

In the preceding figure, the consolidated network represents connectivity between PowerEdge R760xa servers using BlueField-3 B3220 (frontend) and BlueField-3 B3140H (backend) adapters, and an NVIDIA Spectrum SN5600 switch.

This converged network design uses VLAN aware VXLAN for network segmentation and scalability. It relies on Ethernet VPN (EVPN) for L2 network connectivity and uses multihoming for high availability.

Cables and optics

NVIDIA's [Ethernet Switch Configurator](#) can be used to help identify and order the appropriate cables and optics for your AI fabrics. After entering your topology information into the form, it lists the available product options for your environment, along with part numbers and other information.

For the network design that is shown in Figure 8, customers can either use Direct Attached Copper (DAC) or fiber.

- For the backend network:

- **DAC option**—800 Gb/s Twin-port OSFP to 2x 400G QSFP cables are required. The 800 Gb OSFP connects to the Spectrum-X SN5600 switch while the 400G QSFP112 connects to BlueField-3 B3140H. See an [example DAC Splitter cable product specification](#) for more information.
- **Fiber option**—NVIDIA 800 Gb/s OSFP twin port transceiver can be used on the switch side, NVIDIA 400 Gbps QSFP112 single port transceiver on the BlueField-3 B3140H and NVIDIA passive fiber cable (MPO12 APC to MPO12) can be used to connect them.
- For the frontend network:
 - **DAC option**—400 Gb/s QSFP-DD to QSFP112 cables are required. The 400G QSFP-DD connects to Spectrum-X SN5600, while the QSFP112 connects to BlueField-3 B3220. For more information, see the [list of supported transceivers and cables](#).
 - **Fiber option**—NVIDIA 400 Gbps QSFP-DD transceiver can be used on the switch side and 400 Gbps QSFP112 transceiver on the BlueField-3 B3220. NVIDIA passive fiber cable, MMF, MPO12 APC to 2xMPO12 APC can be used to connect them.

Network topologies for large clusters

In modern AI and high-performance computing environments, efficient data exchange between GPUs across multiple nodes is crucial. The scale-out fabric interconnects NICs coupled with GPUs in PowerEdge XE9680 and PowerEdge R760xa servers, forming a robust GPU cluster. This fabric facilitates the seamless exchange of AI workload parameters, ensuring optimal performance and scalability.

in modern ai and high-performance computing environments, efficient data exchange between gpus across multiple nodes is crucial. the scale out fabric interconnects nics coupled with GPUs in PowerEdge XE9680 and PowerEdge R760xa servers, forming a robust GPU cluster. This fabric facilitates the seamless exchange of AI workload parameters, ensuring optimal performance and scalability.

A scale-out fabric can be implemented in different topologies that include:

- Single switch topology
- TOR-Wired Clos topology
- Pure Rail topology
- Rail Optimized topology

The [Dell Technologies AI Fabrics Overview](#) white paper helps you explore various topologies and guidelines for deploying a scale-out fabric, highlighting the benefits and trade-offs of each approach.

Rack and Power Design

The following figure shows an example rack layout for this design:

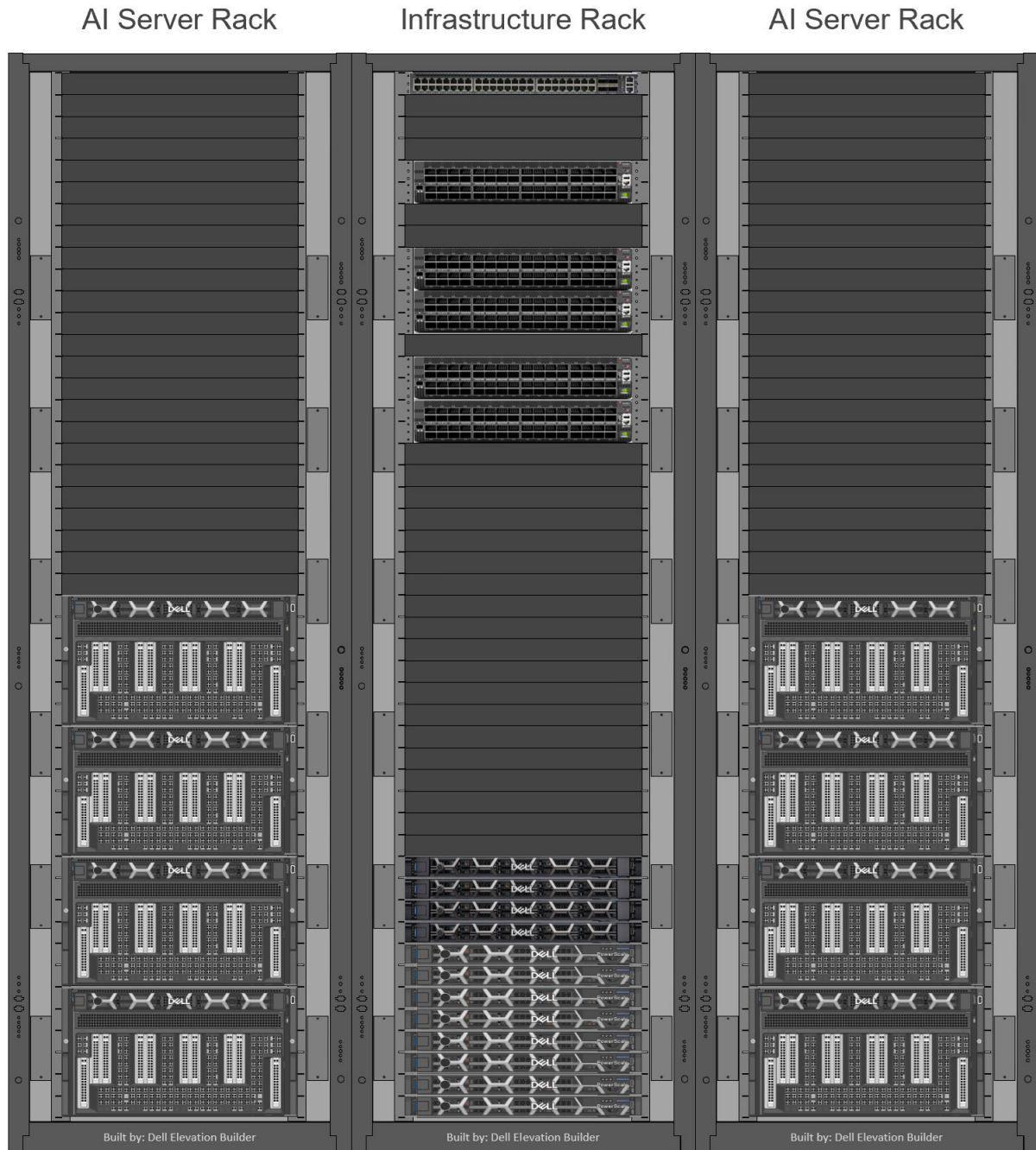


Figure 9. Example rack configuration

This rack design was created by using the Dell [Enterprise Infrastructure Planning Tool](#) (with the illustrations of the switches enhanced). Filler panels are not shown. You can use the tool to determine your solution and determine weight, power requirements, airflow, and other details.

This example shows eight PowerEdge XE9680 servers in two racks. The four PowerEdge XE9680 servers in a single rack require four 17 kW Power Distribution Units (PDUs) per rack. However, carefully evaluate your own power and cooling requirements and your preference for rack layout, power distribution, airflow management, and cabling design.

If significant growth is anticipated in the size of the deployment, consider separate racks for compute, storage, and management nodes to allow sufficient capacity for that growth.

APC rack models AR3300 (42U), AR3350 (42U), AR3307 (48U), and AR3357 (48U) are racks that can be considered for this solution. APDU10450SW and APDU10452SW are recommended APC PDU models.

To understand the physical aspects of deploying a PowerEdge XE9680 server, see the [PowerEdge XE9680 Rack Integration](#) technical white paper.

Solution validation

Topics:

- [Overview](#)
- [System configuration](#)
- [Inference validation](#)
- [Fine-tuning validation](#)
- [Run:ai validation](#)

Overview

The Dell AI Platform for Generative AI simplifies and accelerates the deployment of complex infrastructure for generative AI by providing the reference architecture and configurations. It helps customers by reducing the guesswork and potential risks that are associated with designing and implementing initial custom solutions.

System configuration

The following tables list the system configurations and software stack that are used for the validation efforts in this design:

Table 5. System configuration and versions

Component	Configuration 1	Configuration 2
Compute server for model customization	4 x PowerEdge XE9680 servers	4 x PowerEdge R760xa servers
GPUs per server	8 x NVIDIA H200 SXM GPUs	4 x NVIDIA H100 NVL PCIe GPUs
Backend network adapters	8 x NVIDIA Bluefield-3 Single Port 400 GbE QSFP112 PCIe FH (B3140H) (Firmware version: 32.44.1036)	2 x NVIDIA Bluefield-3 Single Port 400 GbE QSFP112 PCIe FH (B3140H) (Firmware version: 32.44.1036)
Frontend network adapters	1 x NVIDIA Bluefield-3 Dual Port 200 GbE QSFP112 PCIe FH (B3220L) (Firmware version: 32.44.1036)	1 x NVIDIA Bluefield-3 Dual Port 200 GbE QSFP112 PCIe FH (B3220L) (Firmware version: 32.44.1036)
Storage	PowerScale F710	PowerScale F210
Network switches	<ul style="list-style-type: none"> • 2 x NVIDIA Spectrum 5600 switches for frontend (firmware: 5.12.1) • 2 x NVIDIA Spectrum 5600 switches for backend (firmware: 5.12.1) • 1 x NVIDIA Spectrum SN2201 (firmware:5.12.1) 	<ul style="list-style-type: none"> • 2 x NVIDIA Spectrum 5600 consolidated switches frontend (firmware: 5.12.1) • 1 x NVIDIA Spectrum SN2201 (firmware: 5.12.1)

Table 6. Software components and versions

Component	Details
Operating system	Ubuntu 24.04
AI software platform	NVIDIA AI Enterprise version 6.2
AI framework	NVIDIA NeMo Framework v25.02
Cluster management	NVIDIA Base Command Manager Essentials 10.25.03

Table 6. Software components and versions (continued)

Component	Details
Kubernetes cluster	Version 1.30.11 with: <ul style="list-style-type: none">• Network Operator: 25.01 (with OFED driver v25.1.0 (DOCA 2.10.0))• GPU Operator: 25.03 (with Data Center Driver 570.133.20)

Inference validation

Inference is the stage where AI generates actionable results, driving innovation across various industries. This section presents the performance test results of our AI solutions, highlighting key metrics such as model concurrency, scalability, and efficiency. These metrics demonstrate the strengths and capabilities of our integrated hardware and software platform.

Understanding latency and throughput are essential for determining when to scale workloads. While some business units or job functions can tolerate a few seconds of delay, others require sub-millisecond responses for optimal performance. This section describes our test results, offering insights into how different configurations perform under varying loads.

We performed inference validation on the following five use cases:

Use Case 1—Single model baseline test

The initial step in the scaling process establishes a baseline by deploying a single model on a single GPU to measure its performance. This baseline serves as a reference for evaluating the impact of scaling and different configurations. The tests are conducted on PowerEdge R760xa and XE9680 servers using Llama 3 models (8B and 70B). The primary goals are to assess the impact on latency and throughput across various concurrency scenarios (ranging from 1 to 1000). Specific configurations include running Llama 3 8B and 70B models on different GPU setups within the PowerEdge XE9680 and R760xa servers.

Use Case 2—Scaling up multiple instances of Llama 3 models

This use case focuses on running multiple instances of the Llama 3 8B model on PowerEdge XE9680 and R760xa servers. It gathers data about latency, throughput, and system metrics (CPU, GPU, memory, and network use). The goal is to understand the performance impact as the number of model instances increases, ensuring full GPU memory utilization and 100 percent reservation of GPU capacity.

Use Case 3—Cluster testing with multiple instances of Llama 3 models

The objective is to demonstrate the scalability of Llama 3 8B models across multiple nodes in a Kubernetes cluster. The test deploys the model across two PowerEdge R760xa servers, fully loading both servers, and using a frontend load balancer to distribute the load. Key metrics such as throughput, latency, and host system performance (GPU, CPU, memory) are gathered to evaluate performance.

Use Case 4—Running multiple different models on the PowerEdge XE9680 server

This scenario includes running multiple different models on a single PowerEdge XE9680 server, using all eight GPUs. One Llama 3 70B model uses four NVIDIA H200 GPUs, while four Llama 3 8B models each use a single NVIDIA H200 GPU. The aim is to gather throughput and latency metrics and compare them to the baseline results. This setup allows for efficient GPU resource utilization and is critical for applications requiring diverse model deployments.

Use Case 5—Impact of running models with different quantization on the PowerEdge XE9680 server

This use case examines the performance impact of running the Llama 3 70B model with different quantization (FP16 and FP8) on the PowerEdge XE9680 server. The goal is to evaluate and compare performance metrics such as throughput and latency to understand the effects of quantization on model performance.

The performance metrics and results of these use cases can be found at [Maximizing AI Performance: A Deep Dive into Scalable Inferencing on Dell with NVIDIA](#).

Fine-tuning validation

The following list provides the details of our validation setup:

- **Foundational models**—We validated 8 B and 70 B Llama 3.
- **Model customization techniques**—We used SFT and LoRA. The next sections show the results of these fine-tuning methods.
- **Cluster configuration**—We used Kubernetes clusters.
- **Dataset**—We used the [Dolly dataset from Databricks](#) (databricks-dolly-15k). It is an open-source dataset of instruction-following records that thousands of Databricks employees generated in several of the behavioral categories that are outlined in the [InstructGPT paper](#). The categories include brainstorming, classification, closed QA, generation, information extraction, open QA, and summarization.
- **Time for training**—Usually, data scientists train a model until it reaches convergence, a point influenced by factors like the dataset, model complexity, and chosen hyperparameters. Our aim was not to achieve convergence for every scenario, as it is specific to our chosen dataset and parameters, offering limited insight into a customer's needs. To maintain a consistent metric across all scenarios, we conducted training jobs for a minimum of 1000 steps.

Run:ai validation

We deployed and validated Run:ai on Dell AI Platform with NVIDIA. We then validated the following Run:ai version 2.21.19 capabilities on PowerEdge XE9680 servers:

GPU slicing—We conducted a series of tests using fractional GPU allocations across three workload types: development workspace, inference, and training. We created a workspace using 10 percent of an NVIDIA H200 GPU with Jupyter Notebook as the development environment. For inference, we deployed the Llama 3.1 8B model using 35 percent of a GPU, authenticated by using Hugging Face, and confirmed model responsiveness with curl. For training, we allocated 50 percent of a GPU and used JupyterLab for interactive development. We monitored each workload through the Run:ai dashboard and verified GPU use with nvidia-smi, confirming accurate resource allocation and efficient multitenant GPU sharing.

Memory swap—We enabled the feature on Dell infrastructure by configuring CPU memory limits (250Gi), reserving 4GiB of GPU memory for nonswappable allocations, and labeling nodes to accept swap-enabled workloads. We created a dynamic compute resource with 30 percent minimum GPU memory allocation and sufficient CPU resources to support multiple concurrent models. Using this setup, we deployed three large language models—Llama 3.1 8B, Gemma 2 9B, and Phi-3.5 MoE—each of which typically require a dedicated NVIDIA H200 GPU. Through Run:ai's memory swap, we successfully ran all three models on a single GPU, dynamically swapping them between GPU and CPU memory. A web UI was used to interact with each model, demonstrating that while the swap introduces a few seconds of latency, it is a viable solution for low-demand environments.

Node pool—We created a dedicated node pool to manage a group of nodes with shared characteristics—specifically, nodes equipped with NVIDIA GPUs. Using the Run:ai platform, we defined the node pool using the Resources tab and associated it with a project by editing the project settings and assigning the node pool, which enabled scheduling of up to eight GPUs. We then validated the setup by deploying an NVIDIA NIM-based inference workload. This task involved configuring NGC credentials, selecting the appropriate node pool during workload creation, and confirming successful workload scheduling and execution. This validation demonstrated the effectiveness of node pools in targeting specific hardware resources for optimized workload placement.

Scheduler—We tested the scheduler's ability to manage AI workloads with fairness, efficiency, and support for complex scheduling scenarios. We evaluated key features such as quota enforcement, overquota scheduling, preemption, and priority-based resource allocation. Our tests included deploying training and inference jobs under various quota conditions across multiple node pools. We confirmed that the scheduler correctly handled over-quota requests when idle resources were available, enforced nonpreemptive behavior for inference workloads, and prioritized high-importance jobs by preempting lower-priority

ones. The scheduler's queue fairness mechanism ensured equitable job distribution across projects, even when job submission volumes varied. These validations demonstrated the scheduler's effectiveness in optimizing GPU utilization, maintaining workload fairness, and supporting high availability for critical inference services.

Inference Performance Characterization

Topics:

- [Overview](#)
- [Use cases](#)
- [Key performance metrics and factors impacting performance](#)
- [Methodology](#)
- [Results for the PowerEdge R760xa server with the NVIDIA H100 NVL GPU](#)
- [Results for the PowerEdge XE9680 server with the NVIDIA H200 SXM GPU](#)

Overview

Before using an infrastructure for LLM tasks such as training and inferencing, benchmarking the infrastructure:

- Helps to understand the capacity of the infrastructure for handling the computational demands of these tasks. LLM tasks often require significant computational resources, and benchmarking can provide insights into whether the current infrastructure can meet these demands.
- Allows for the optimization of resource allocation to ensure that the infrastructure is used efficiently and cost-effectively.
- Highlights potential bottlenecks in the infrastructure that can hinder the performance of LLM tasks.

By identifying these issues in advance, you can take steps to mitigate them, ensuring smooth and efficient operation. Benchmarking provides a baseline against which future upgrades or changes to the infrastructure can be measured, aiding in continuous improvement efforts. Therefore, benchmarking is a vital step in preparing an infrastructure for LLM tasks.

The following sections show some of the end-to-end performance benchmarking that we performed as part of the validation of this design on the inferencing.

Use cases

Evaluating LLM inference performance requires distilling what technical requirements are associated with the use cases being pursued. Many modern LLMs such as GPT 4, Claude 3.5, Mistral, and Llama 3 are built for next-token inference as a decode only. The solutions (and their configurations) that can be built around these LLMs are impacted by the possible variability of the:

- Input Sequence Length (ISL)—The length of the text that is provided to the model as input
- Output Sequence Length (OSL)—The length of the text that is generated by the model as output

There are four common combinations of ISL and OSL pairings that we used to categorize many common LLM inference use cases:

1. Short input, short output
2. Short input, medium-length output
3. Medium-length input, short output
4. Long input, medium-length output

The following table shows the input/output lengths that we used for performance testing for each use case combination:

Table 7. Input/output lengths

Use case category	Input token-sequence length	Output token-sequence length	Primary use case	Critical performance metric and recommended value
Use case 1—Short input, short output	200	200	Chatbot interactions,	TTFT under 2000 ms

Table 7. Input/output lengths (continued)

Use case category	Input token-sequence length	Output token-sequence length	Primary use case	Critical performance metric and recommended value
			language translation	
Use case 2—Short Input, medium-length output	200	1000	Email drafting, creative writing, and code generation	TTFT under 2000 ms
Use case 3—Medium-length input, short output	2000	200	Summarization, meeting notes, customer support	Request Response Latency under 4 seconds
Use case 4—Long input, medium-length output	7000	1000	Legal or research document analysis, RAG	Throughput of 500 tps with Request Response latency < 1 minute

NOTE: These token-sequence lengths are chosen for benchmarking purposes. They are used to simulate use cases representative of certain realistic scenarios. They are not universally representative of all scenarios that might fall under the corresponding use cases.

Short input, short output

This category addresses quick, concise tasks with both short inputs and outputs:

- **Chatbot Interactions**—LLMs can engage in conversations, maintaining context and providing relevant, concise responses to user queries. This application enables more natural and effective human-computer interactions in customer support, virtual assistants, or interactive learning environments. Additionally, in educational settings, chatbots can assist students by answering questions, providing explanations, and guiding them through learning materials, making education more accessible and interactive. Note that as the context length increases, the input size also increases.
- **Language translation**—LLMs can accurately translate short phrases or sentences between languages, maintaining context and nuance. This application is useful for quick translations in travel applications, multilingual customer support, or international business communications, enabling quick communication across language barriers.
- **Code refactoring**—Given a short code snippet, LLMs can suggest improvements or simplifications while preserving functionality, enhancing code readability and efficiency. This application assists developers to maintain clean, optimized code bases, aiding developers in maintaining clean codebases.
- **Keyword extraction**—LLMs can identify and summarize key terms from short documents, useful for search engine optimization, content tagging, or quick document classification. This application aids in content management and improves searchability of digital assets, useful for indexing and search optimization.
- **Microblogging**—LLMs can generate or optimize short posts for platforms such as Twitter, ensuring maximum impact within character limits. This application helps social media managers and marketers create engaging, concise content, enabling quick content creation.
- **Multimodal language translation**—LLMs can translate text while considering accompanying images or diagrams, ensuring that the translation preserves the overall meaning and context. This application is useful for translating infographics, memes, or other visual content with text, enhancing cross-language communication.
- **Multimodal code generation**—By analyzing both text descriptions and visual representations (such as flowcharts), LLMs can generate short, effective code snippets. This application bridges the gap between conceptual design and implementation in software development, aiding in complex development tasks.
- **Multimodal fraud detection**—LLMs can quickly analyze transactions or documents that include text, images, and signatures, identifying potential fraud or anomalies in a concise report. This application enhances security measures in financial transactions, document verification, or identity authentication processes.

Short input, medium-length output

This category generates longer content from brief inputs:

- **Creative writing**—LLMs can generate short stories, poems, or creative narratives based on brief prompts. For example, a prompt such as "A mysterious door in an abandoned house," the model can craft a 1000-word story exploring the concept. This application is useful for writers seeking inspiration or for generating content for creative writing workshops, allowing for creative exploration with minimal input.
- **Email drafting**—From a short summary or bullet points, LLMs can compose professional, well-structured email messages. This application is useful for busy professionals who must communicate complex ideas efficiently. The model can expand on key points, maintain a consistent tone, and ensure that all necessary information is in the email message, streamlining communication in corporate settings.
- **Code generation**—LLMs can create code snippets or entire functions based on concise descriptions of functionality. For example, "Create a Python function to sort a list of dictionaries by a specific key" can result in a fully implemented and documented function. This application can accelerate development processes and assist programmers to address complex coding tasks, aiding developers in rapid prototyping.
- **Marketing content**—Given a brief outline or key points about a product or service, LLMs can generate compelling marketing copy, including product descriptions, ad text, or social media posts. The model can adapt its writing style to suit different platforms and target audiences, ensuring consistent brand messaging across various channels and enhancing content creation workflows.
- **FAQ creation**—From a few user questions, LLMs can develop comprehensive FAQ sections, anticipating related questions and providing detailed answers. This application is invaluable for businesses looking to improve their customer support resources or creating educational materials, improving customer support resources.
- **Visual storytelling**—LLMs can generate detailed stories or descriptions that are based on a series of images, bridging the gap between visual and textual content. This application is useful in fields such as digital marketing, journalism, or eCommerce, where compelling narratives must be created around visual content, combining text and visuals to tell compelling stories.
- **Product recommendation**—By analyzing short text descriptions and images of products, LLMs can provide personalized product recommendations, complete with reasoning and comparisons. This application enhances the shopping experience by offering tailored suggestions that are based on user preferences and product features, improving eCommerce experiences.

Medium-length input, short output

This category addresses medium-length inputs with concise outputs:

- **Medium-length summarization**—LLMs can summarize articles or reports into brief, informative paragraphs, distilling key information for quick consumption. This application is ideal for creating abstracts, executive summaries, or brief overviews of longer content.
- **Chatbot interactions**—LLMs can engage in multiturn conversations, maintaining context and providing relevant, concise responses to user queries. This application enables more natural and effective human-computer interactions in customer support, virtual assistants, or interactive learning environments, enhancing user experience in customer support.
- **Contextual Q&A**—Given a medium-length context, LLMs can answer specific questions accurately and concisely, making them ideal for information retrieval systems. This application is useful in educational settings, customer support, or any scenario where quick, accurate answers are needed based on a given context, facilitating knowledge retrieval in conversational settings.
- **Meeting notes**—LLMs can process longer meeting transcripts and generate concise, actionable notes, highlighting key decisions and action items. This application saves time in postmeeting documentation and ensures that important points are not missed, improving postmeeting productivity.
- **Multimodal news summarization**—By analyzing news articles along with accompanying images and infographics, LLMs can generate concise summaries that capture the essence of the news story. This application is valuable for news aggregators, media monitoring services, or individuals wanting to stay informed about current events, providing a quick overview of current events.
- **Multimodal customer support**—LLMs can provide personalized support by understanding and responding to customer queries that include text, images, or screenshots, offering concise solutions. This application enhances customer support efficiency by quickly addressing issues that might require visual context, improving customer satisfaction.
- **Multimodal event recaps**—LLMs can generate brief summaries of events or conferences, combining key takeaways from presentations, speaker information, and relevant visual content. This application is useful for attendees, organizers, or people who did not attend but want a quick overview of the event, aiding event participants.

Long input, medium-length output

This category processes and summarizes large amounts of information:

- **Long-form article summarization**—LLMs can distill extensive articles or research papers into concise, informative overviews, capturing key points and main arguments. This application is useful for researchers, students, or professionals who must grasp the essence of lengthy documents quickly, saving time for readers.
- **Legal document analysis**—By processing lengthy legal documents, LLMs engage in multiturn discussions with legal professionals. They use chain-of-thought reasoning to clarify complex legal concepts, extract and summarize key points, identify potential issues, and provide a high-level overview of complex legal matters. This application can reduce the time that lawyers and paralegals spend on document review and analysis, aiding legal professionals in quick decision-making.
- **Research paper review**—LLMs can analyze comprehensive research papers, providing critiques, identifying strengths and weaknesses, and suggesting areas for further research. This application assists researchers in literature reviews and helps journal editors in the peer review process, aiding academics in their evaluations.
- **Retrieval-augmented generation (RAG) for educational content**—LLMs can use RAG to retrieve relevant information from a database of textbooks and articles. The model engages in multiturn conversations with students, answering questions based on retrieved content and providing detailed explanations.
- **Medical diagnosis assistance**—LLMs can analyze comprehensive patient records, including text notes and medical images. The model summarizes the patient's history and engages in a multiturn conversation with healthcare providers, using RAG to retrieve relevant medical literature for informed decision-making.

Key performance metrics and factors impacting performance

To understand the performance of an LLM, you can measure several metrics, each of which is relevant for assessing different aspects of the model's performance. The following list describes how each metric contributes to understanding LLM performance:

- **Time Per Output Token (TPOT)**—This metric, also known as Inter-Token Latency (ITL), indicates the time to generate each output token for a user query. It reflects the perceived “speed” of the model. A lower TPOT means that the model can generate tokens quickly, enhancing the overall responsiveness and efficiency of the system.
- **Throughput (TPS)**—This metric measures the throughput of the model, indicating how many tokens the model can generate per second on average. It is measured in tokens per second (TPS). It provides insight into the overall speed of the model and its ability to process input data efficiently.
- **Time To First Token (TTFT)**—This metric measures the time for the model to generate the first token of a response after receiving an input prompt. It reflects the initial processing time and is crucial for real-time applications for which low latency is essential.
- **Request Response Latency (RRL)**—This metric measures the total time that is required for the model to generate a complete response to a given input prompt. It includes the time for processing the entire input sequence and generating the output sequence, providing a comprehensive view of the end-to-end latency experienced by users.

The following factors impact the performance of the LLM model inference:

- **Model size (number of model parameters)**—This parameter is the size of a model, which is measured by the number of parameters, that directly influences its memory use and computational demands. Larger models with more parameters typically need more powerful hardware to perform inference efficiently.
- **Input token-sequence length and output token-sequence length**—These parameters influence both GPU memory use and computation time. Longer input and output token-sequences demand more memory. Processing and generating these longer token-sequences increase the computational load, resulting in longer inference times.
- **Concurrent requests**—This parameter specifies the number of input sequences that are processed simultaneously during inference. Larger concurrent requests require more memory to store each sequence and the corresponding computations. Certain LLM deployment configurations can potentially improve throughput by processing multiple input sequences in parallel, fully using the GPUs. However, while larger concurrent requests/batch sizes improve throughput, they might also increase latency when handling many requests in a single batch. This approach is useful for offline benchmarking purposes.
- **Tensor parallel size**—This parameter enables the distribution of model weights across multiple GPUs. For example, in a system such as the Dell PowerEdge XE9680 server equipped with eight GPUs, setting the tensor parallel size to 8 allows each GPU to handle a portion of the model weights. This configuration reduces the memory load on each GPU and permits larger Key-Value (KV) cache sizes, potentially improving inference efficiency.

Methodology

To measure the LLM inference performance, we use the GenAI-Perf tool. GenAI-Perf is a powerful command-line tool that is designed to measure the throughput and latency of generative AI models when served through an inference server. Tailored for LLMs, GenAI-Perf provides comprehensive metrics including output token throughput, time to first token, time to second token, intertoken latency, and request throughput.

You can specify the model's name, inference server URL, input types (either synthetic or from a predefined dataset), and the wanted load parameters (such as the number of concurrent requests and request rate). GenAI-Perf then generates the specified load, evaluates the performance of the inference server, and presents the metrics in a clear, tabular format on the console. All results are logged in CSV and JSON files, enabling further analysis and visualization.

We ran the tests by using Triton container version [24.09](#), which is available at the [NGC Catalog](#). To avoid resource contention, we recommend running the container on a different system than the system that is running the model inference.

The objective of the use cases and concurrencies is to help customers and partners estimate the behavior and performance of the model in production. However, they cannot be used as absolute numbers because different versions of drivers and inference software can influence the overall performance of the model.

The following table shows the infrastructure and configuration parameters that we used to test the performance of the Llama 3.1 8B and Llama 3.1 70B models:

Table 8. Infrastructure and configuration parameters

Parameter	Value
Large Language Model	Llama 3 8B Instruct and Llama 3 70B Instruct
NIMs container tags used	<ul style="list-style-type: none">nvcr.io/nim/meta/llama-3.1-8b-instruct:1.3.3nvcr.io/nim/meta/llama-3.1-70b-instruct:1.3.3
Servers	PowerEdge XE9680 with 8 x H200 SXM, PowerEdge R760xa with 4 x H100 NVL or 4 X L40S
Tensor parallel size	Equal to the number of GPUs
Iterations	10
Data type	float16

Results for the PowerEdge R760xa server with the NVIDIA H100 NVL GPU

The following figures show the latency, throughput, and TTFT of Llama 3 models running on the PowerEdge R760xa server with the NVIDIA H100 NVL GPU:

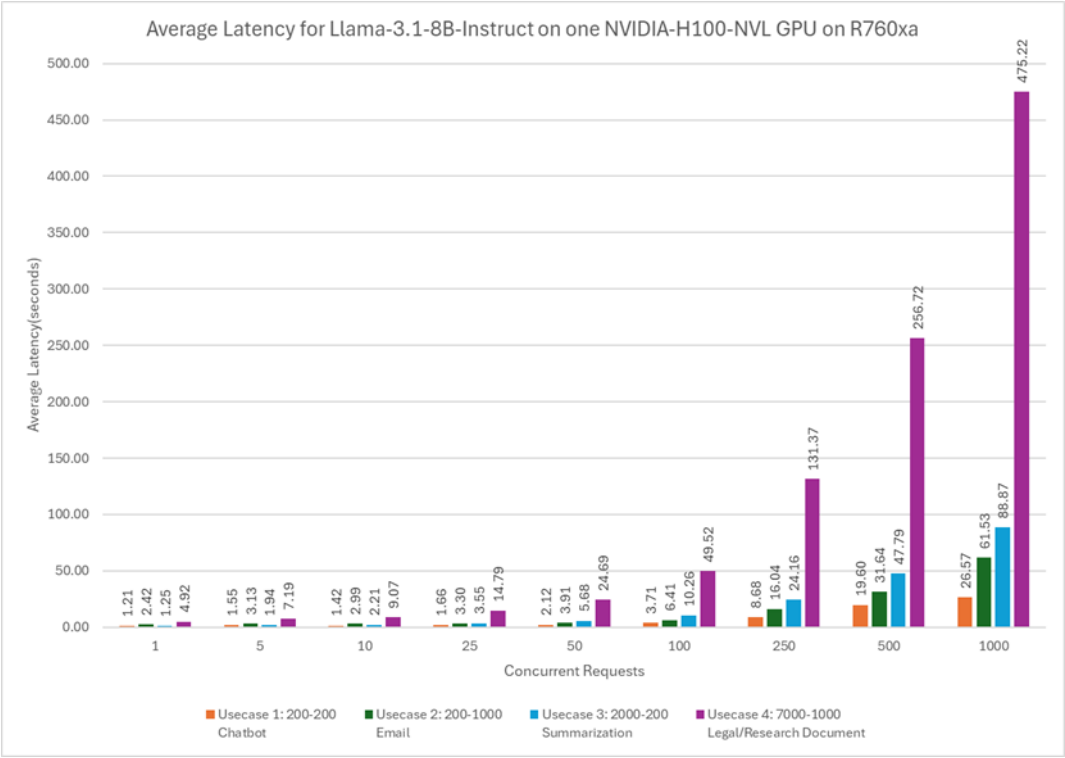


Figure 10. Average latency for Llama-3.1-8B-Instruct on one NVIDIA H100 NVL GPU

Latency measures the total time that is required for the model to generate a complete response to a given input. Analyzing the latency results for Llama 3.1 70B and Llama 3.1 8B provides a comprehensive comparison of average latency (in seconds) across various concurrent requests and four distinct use cases with different input and output lengths. Notably, Use Case 1 (Input Length = 200, Output Length = 200) and Use Case 2 (Input Length = 200, Output Length = 1000) exhibit lower latencies, likely due to their smaller input lengths.

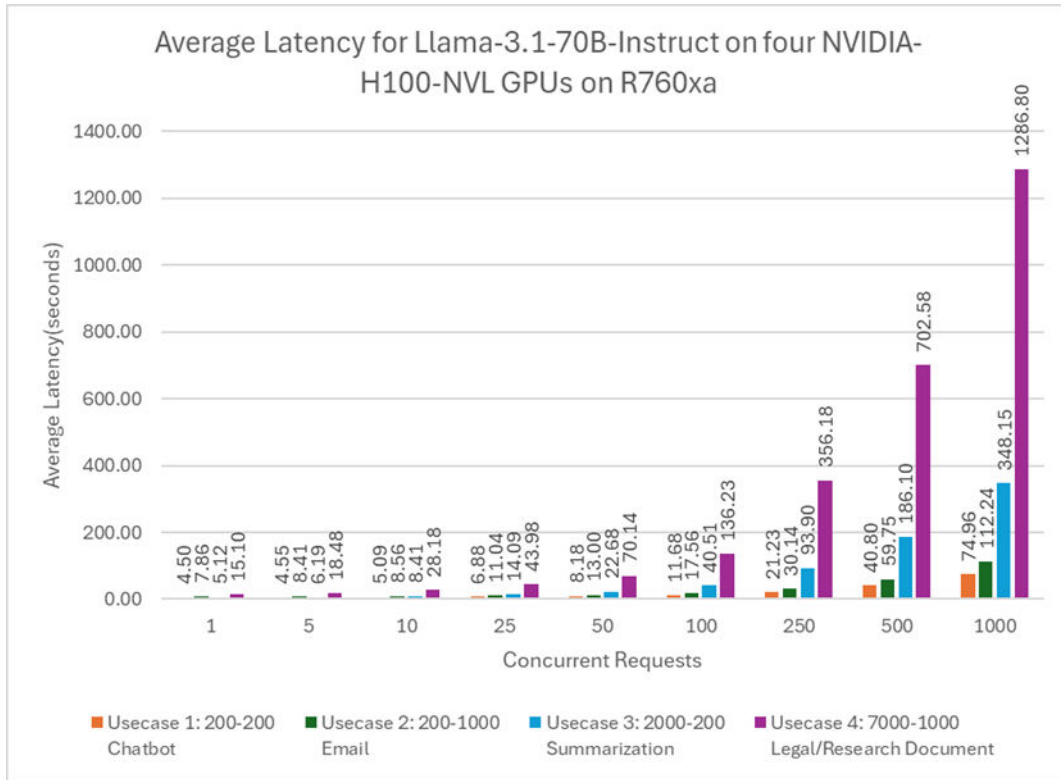


Figure 11. Average latency for Llama-3.1-70B-Instruct on four NVIDIA H100 NVL GPUs

Also, the impact of concurrent requests on performance is evident, with a significant increase in latency observed for larger concurrent requests, particularly in UseCase4 (Input Length = 7000, Output Length = 1000). These findings highlight the importance of optimizing concurrent requests for offline inference and understanding specific use case requirements to ensure efficient performance.

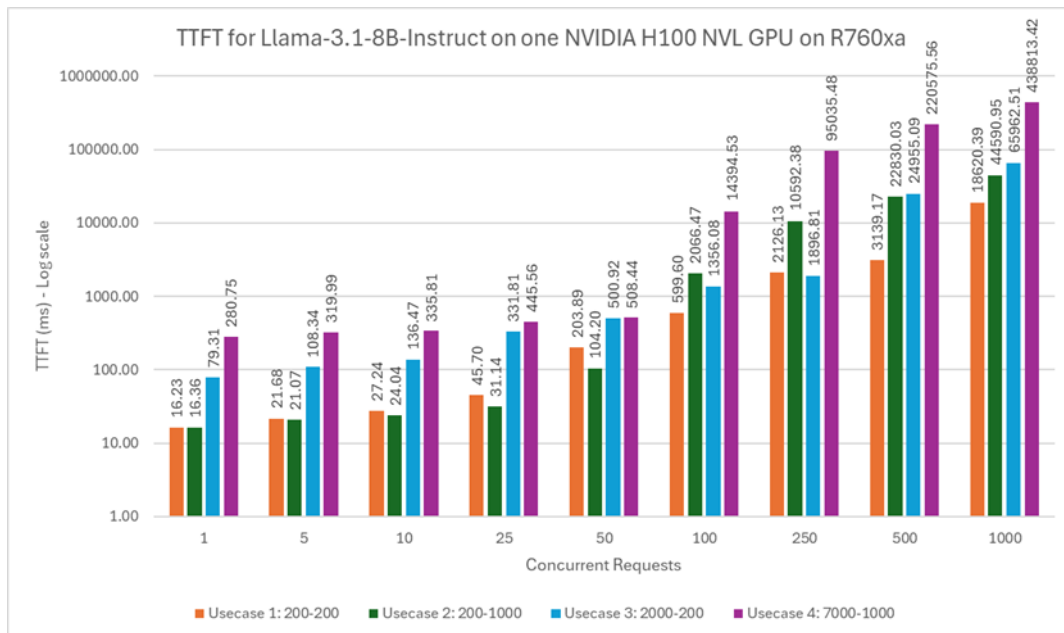


Figure 12. TTFT for Llama-3.1-8B-Instruct on one NVIDIA H100 NVL GPU

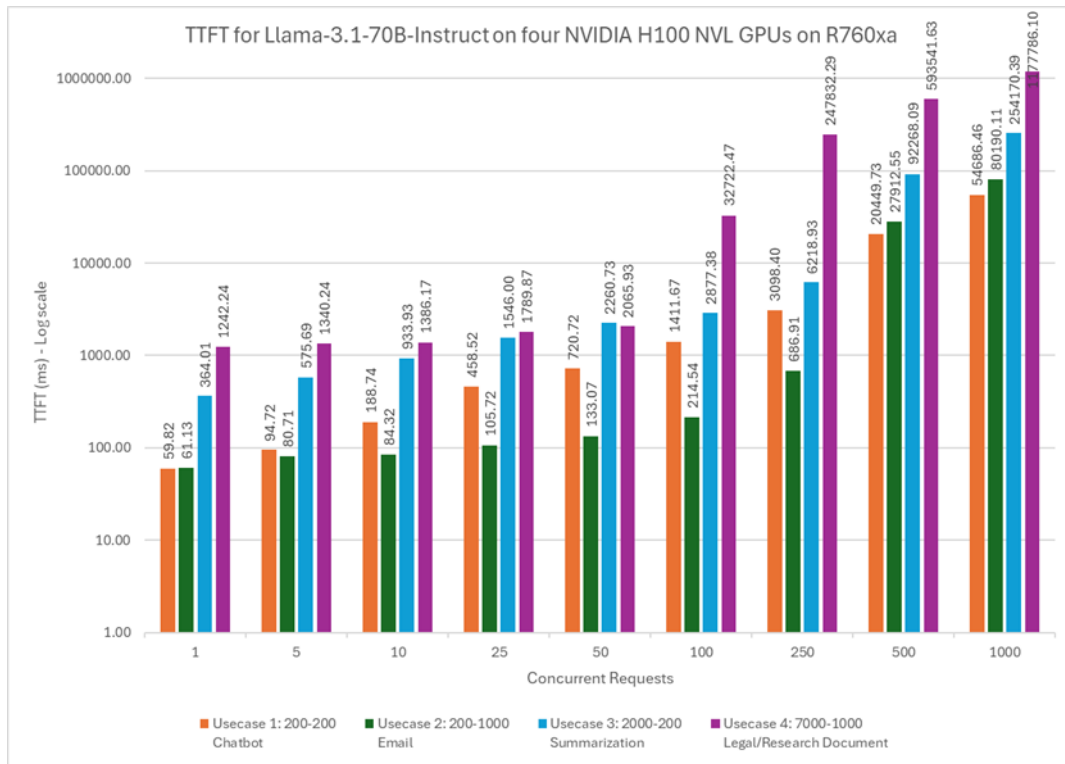


Figure 13. TTFT for Llama-3.1-70B-Instruct on four NVIDIA H100 NVL GPUs

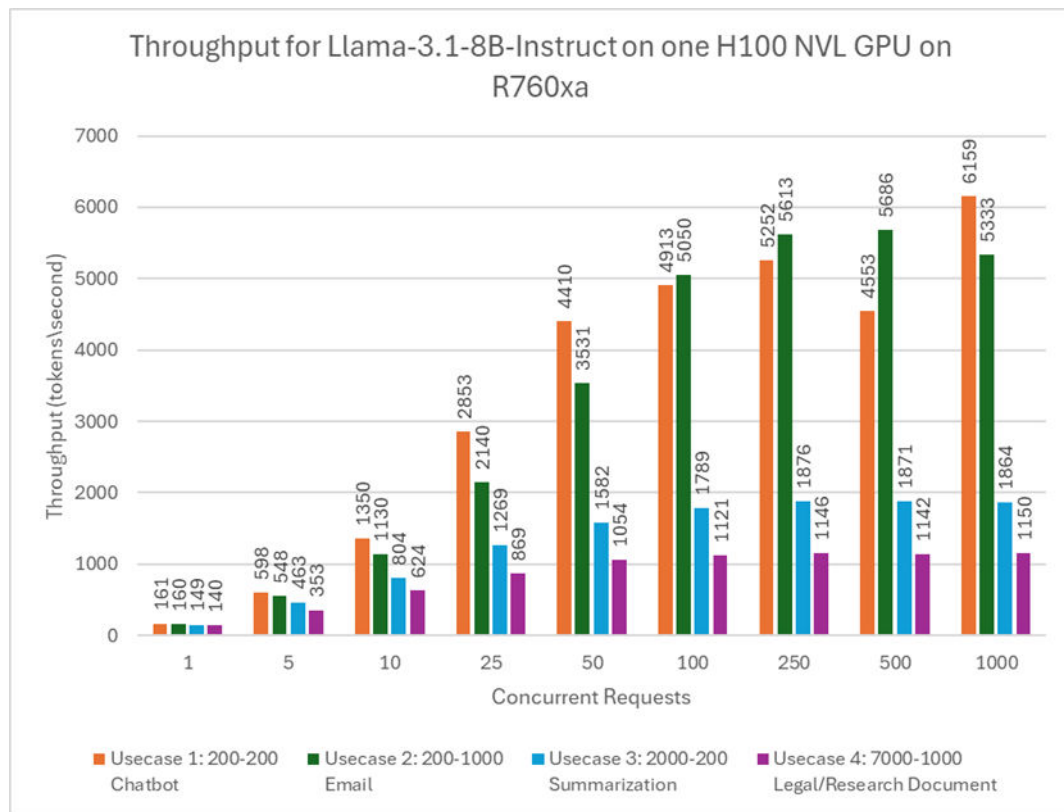


Figure 14. Throughput for Llama-3.1-8B-Instruct on one H100 NVL GPU

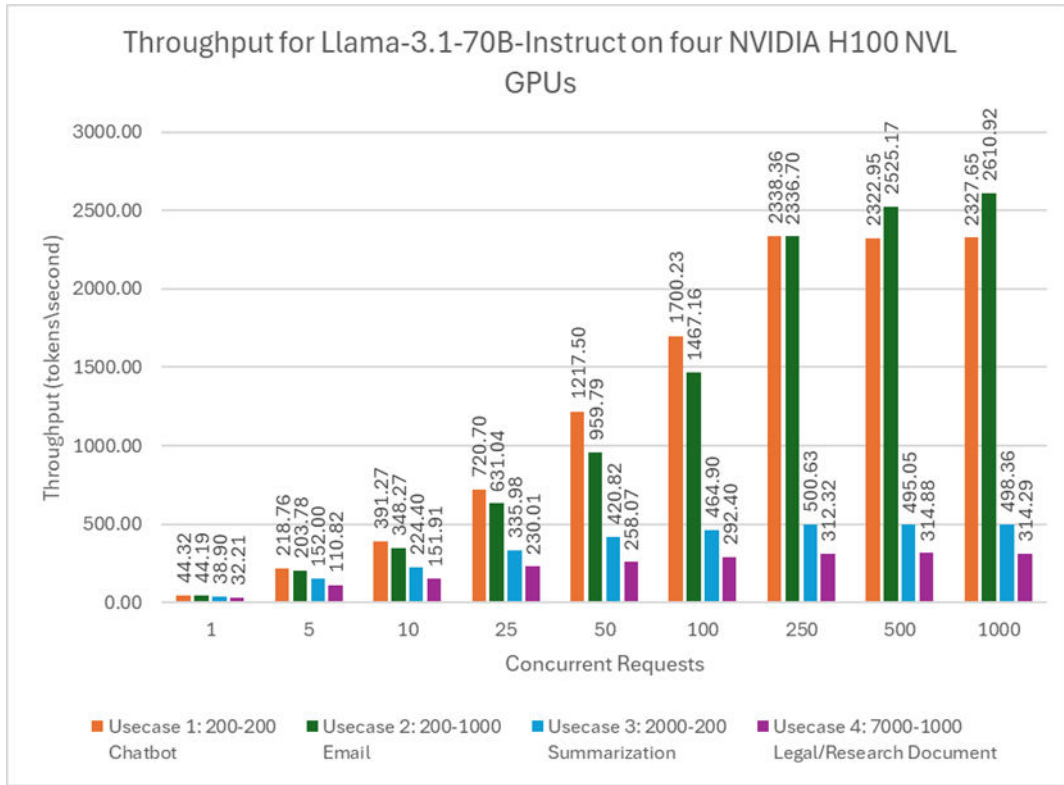


Figure 15. Throughput for Llama-3.1-70B-Instruct on four NVIDIA H100 NVL GPUs

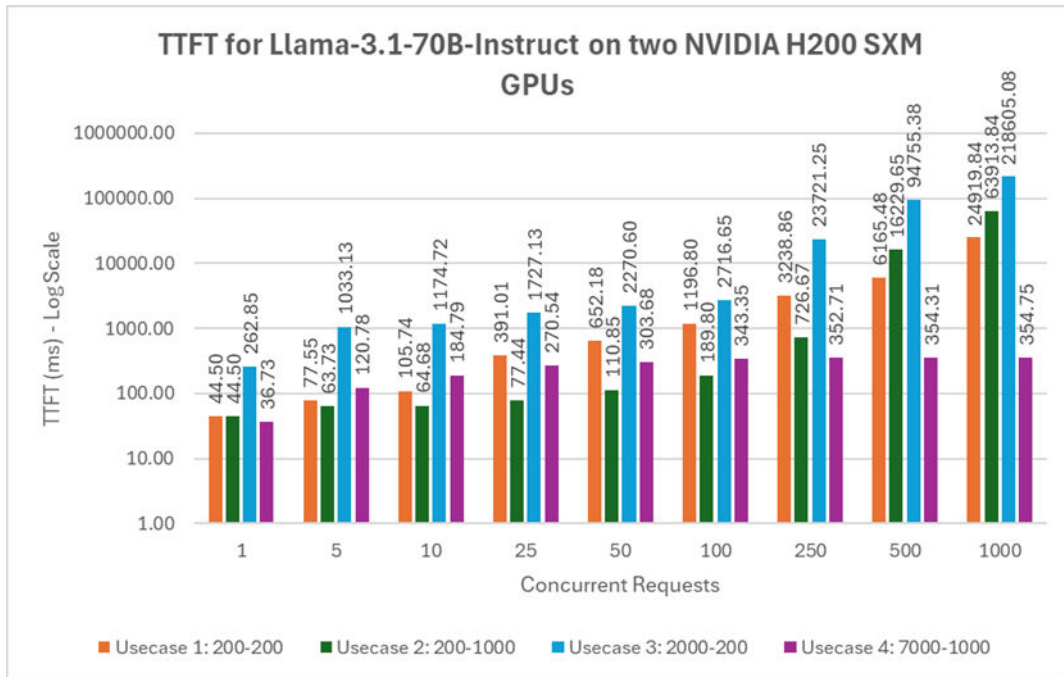


Figure 16. TTFT for Llama-3.1-70B-Instruct on two NVIDIA H200 SXM GPUs

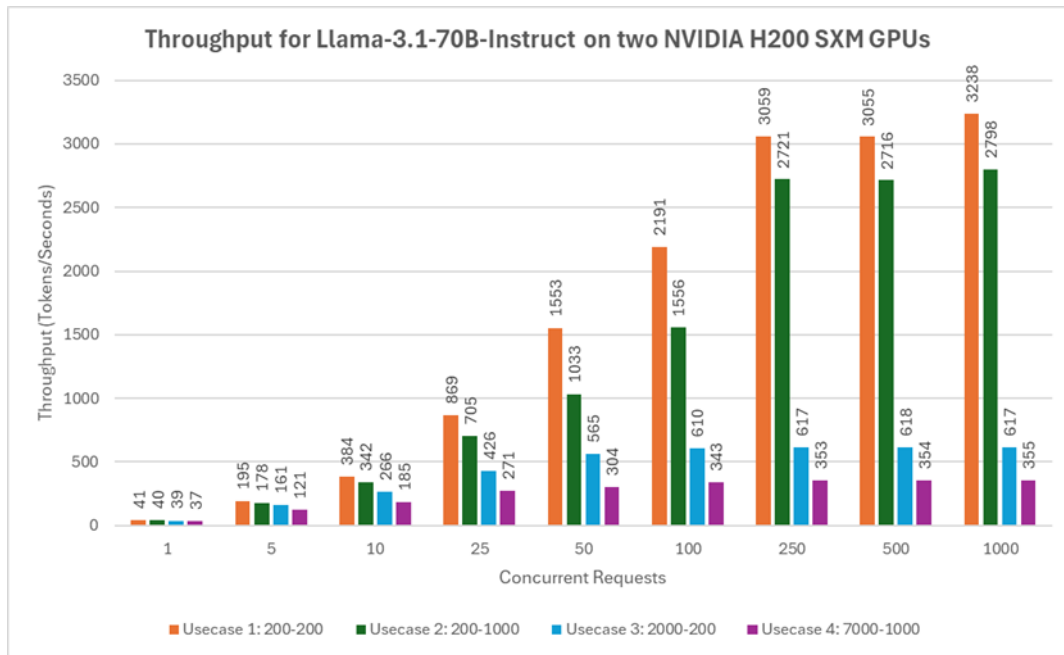


Figure 17. Throughput for Llama-3.1-70B-Instruct on eight NVIDIA H200 SXM GPUs

Results for the PowerEdge XE9680 server with the NVIDIA H200 SXM GPU

The following figures show the latency, throughput, and TTFT of Llama 3 models running on the PowerEdge XE9680 server with the NVIDIA H200 SXM GPU. We measured inference performance for the Llama 3.1 70B model. As of publication of this documentation, the Llama 3.1 8B profile is not available through NVIDIA AI Enterprise NIM.

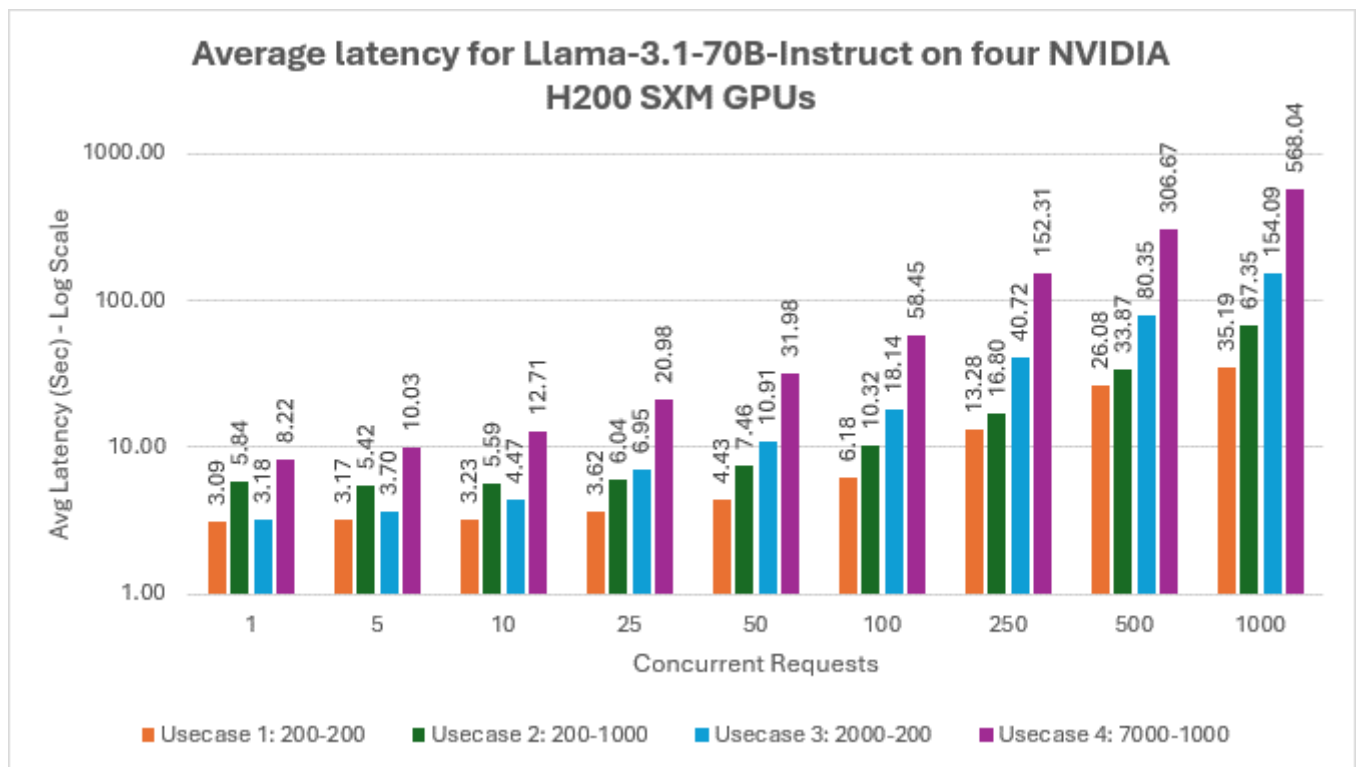


Figure 18. Average latency for Llama-3.1-70B-Instruct on four NVIDIA H200 SXM GPUs

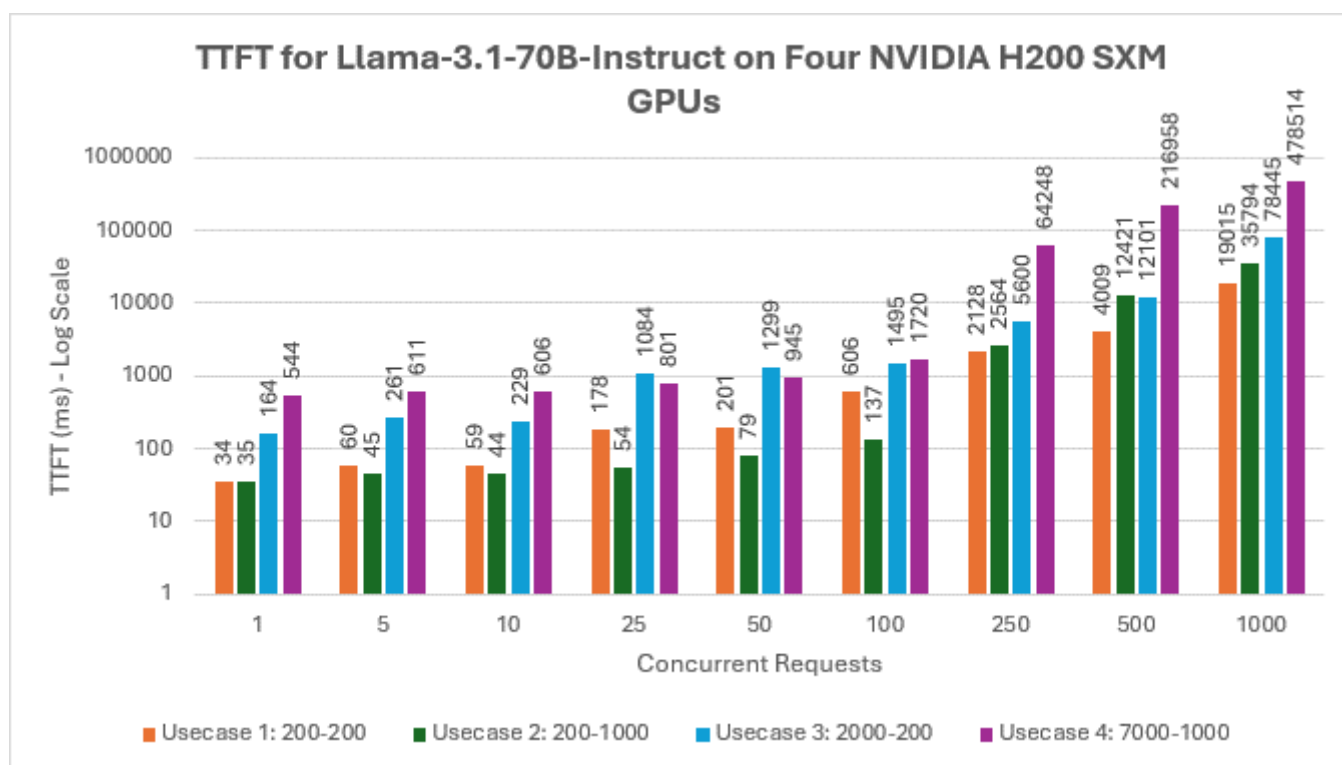


Figure 19. TTFT for Llama-3.1-70B-Instruct on four NVIDIA H200 SXM GPUs

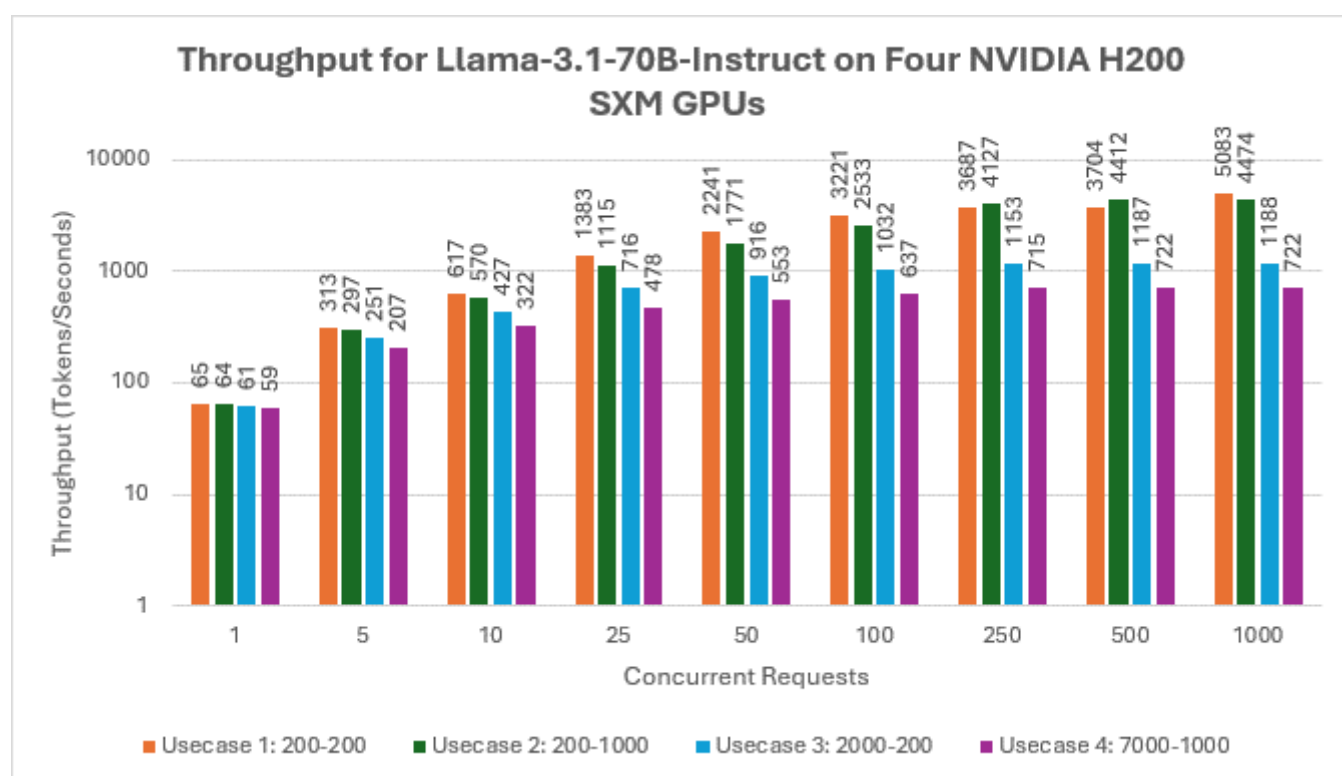


Figure 20. Throughput for Llama-3.1-70B-Instruct on four NVIDIA H200 SXM GPUs

Conclusion

Topics:

- [Summary](#)
- [We value your feedback](#)

Summary

The Dell AI Platform with NVIDIA, based on this validated design for Generative AI in the Enterprise with NVIDIA Spectrum-X Networking Platform, addresses the requirements of enterprises that must develop and run custom generative AI LLMs using domain-specific data that is relevant to their own organization.

Dell Technologies and NVIDIA have designed a scalable, modular, and high-performance architecture that enables enterprises to quickly design and deploy an AI inferencing solution that can be customized to their specific needs using fine-tuning and RAG methodologies. All these methodologies have been validated and performance-tested to accelerate the time to value and to reduce the risk and uncertainty by using a proven design.

The NVIDIA Spectrum-X networking platform with Dell PowerEdge servers exceeds all bandwidth and low latency requirements for AI. The synergy of coupling Dell Technologies and NVIDIA accommodates the most data-intensive requirements for AI applications.

A world-class solution is created for the data-intensive AI industry that delivers optimal performance for AI distributed fine-tuning and inferencing by exclusively joining network and compute hardware, both purpose-built for AI.

Dell Technologies and NVIDIA enable organizations to deliver full-stack generative AI solutions that are built on the best of Dell infrastructure and software, which is combined with the latest NVIDIA accelerators and AI software. This combination of components enables enterprises to use purpose-built generative AI on-premises to solve their business challenges. Together, we are leading the way in driving the next wave of innovation in the enterprise AI landscape.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document and the information that it contains. Contact the Dell Technologies Solutions team by [email](#).

References

Dell Technologies documentation

The following Dell Technologies documentation provides additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative:

- [Dell XE9680 Specification Sheet](#)
- [Dell R760xa Specification Sheet](#)
- [Dell PowerEdge Servers for AI](#)
- [The Dell AI Platform with NVIDIA](#)
- [Dell GenAI solutions](#)
- [Dell AI Services](#)
- [Dell Security and Trust Center](#)
- [Security Best Practices for Generative AI in the Enterprise](#)
- [Dell Info Hub](#)

NVIDIA documentation

The following NVIDIA documentation provides additional and relevant information:

- [NVIDIA AI Enterprise](#)
- [NVIDIA NeMo Framework](#)
- [NVIDIA Spectrum-X Networking Platform](#)
- [NVIDIA BlueField-3 Networking Platform](#)
- [NVIDIA Run:ai](#)