

Dell Technologies AI Fabric with Dell PowerSwitch, Dell PowerEdge XE9680, AMD, and Broadcom stack

Design Guide

Abstract

This design guide describes the hardware and software components required for a small, medium to large GenAI fabric deployment with the Dell PowerSwitch, Dell PowerEdge, Dell PowerStorage, and Dell Enterprise SONiC product family.

Dell Technologies 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.

Chapter 1: Introduction.....	5
Business challenge.....	5
Solution introduction.....	5
Overview.....	5
Document purpose.....	6
Target audience.....	6
Disclaimer.....	6
Terminology.....	6
Chapter 2: Solution components.....	7
Overview.....	7
Software Component: Dell Enterprise SONiC Network Operating System.....	7
Hardware Components.....	8
Dell PowerSwitch.....	8
Dell PowerEdge XE9680.....	10
Dell PowerScale and ECS.....	11
Cables.....	13
Optics.....	13
Chapter 3: Solution Design Guide and Requirements.....	15
Solution requirements.....	15
Interoperable.....	15
High-performance.....	15
Scalable.....	15
Efficient.....	15
Solution topology.....	15
Small GenAI GPU cluster	16
Medium to large GenAI GPU cluster.....	16
Common GenAI fabrics overview.....	17
Out-of-Band management fabric.....	17
Back-end GPU fabric.....	18
Front-end GPU storage, application, and in-band management.....	19
Dell Technologies GenAI solution fabric design.....	20
Small GenAI cluster fabric design.....	21
OOB management fabric.....	22
Back-end GPU fabric.....	23
Front-end storage, application, in-band management fabrics.....	25
Infrastructure – Cabling and power.....	27
Medium or large GenAI cluster fabric design.....	29
OOB management fabric.....	30
Back-end GPU fabric.....	31
Front-end storage and application fabric.....	32
Infrastructure - Cabling and power.....	33

Chapter 4: Dell GenAI Environment — Orchestration and environment.....	38
Introduction.....	38
Infrastructure orchestration and monitoring.....	38
BeyondEdge Verity.....	38
Augtera.....	39
AI workload orchestration and monitoring.....	41
Appendix A: References.....	43
Dell Technologies documentation.....	43
External documentation.....	43
Appendix B: Feedback and technical support.....	44

Introduction

Business challenge

The landscape of Artificial Intelligence (AI) is becoming the hottest segment in the technology industry. It is undergoing rapid evolution and is becoming increasingly indispensable for businesses looking to gain a competitive foothold and stay competitive.

Deploying an AI solution is a complex endeavor. The intricacies of developing and overseeing the sophisticated architectures that support AI algorithms and models pose challenges, especially for organizations lacking the necessary expertise and time to craft, implement, and manage a comprehensive AI solution stack.

The scalability of AI applications is crucial for handling vast datasets and accommodating increased user demands. Maintaining optimal performance poses difficulties for enterprises without the appropriate infrastructure and surrounding ecosystem.

AI workloads tend to be resource-intensive, high-performance, and depend on lossless data transmission-which contributes to elevated infrastructure and operational costs. Therefore, it becomes crucial to streamline, manage, and scale resource utilization to minimize cost while ensuring operational efficiency.

Having clear visibility into a proven AI solution stack becomes the first hurdle for any organization looking to implement an AI solution. Dell Technologies provides several AI solution stacks consisting of networking, compute, storage, and application components.

Solution introduction

Overview

This design guide describes the software and hardware components as well as the processes required to build a Dell GenAI environment based on the Dell PowerEdge XE9680 compute node paired with the Dell PowerSwitch 400GbE data center product family. Dell Enterprise Software for Open Networking in the Cloud (SONiC) is the networking operating system that is used to deploy the two different topologies: stand-alone and leaf and spine.

These two different topologies show the different fabrics that support the deployment of a typical GenAI workload:

- Back-end fabric** This fabric supports the actual GenAI graphics processing unit (GPU) interconnect network.
- Front-end fabric** This fabric supports the deployment of the storage, application, and in-band cluster management components of the GenAI solution.
- Management fabric** This fabric supports the overall environment management network for the solution.

The GPU cluster is the Dell PowerEdge XE9680 with the AMD Mi300x GPU and the Broadcom Thor 2 400GbE network interface card.

This design guide uses the Dell PowerSwitch Z-series product family for the fabric elements. The Z9664F-ON and Z9432F-ON are 400GbE based platforms that run Dell Enterprise SONiC.

The storage element of the solution is provided by the Dell PowerScale or ECS platforms. Both platforms are extremely scalable and perfectly suited for AI and machine learning (ML) applications.

Document purpose

This document provides fabric design guidance for greenfield deployment of a Dell PowerEdge XE9680 GenAI single and multi-cluster GPU using the Dell Technologies product portfolio (PowerSwitch, PowerEdge, PowerScale, Dell Enterprise SONiC 4.2.1) and AMD Mi300x with a Broadcom Thor 2 NIC.

Target audience

This design guide is intended for AI solution architects, data engineers, IT infrastructure managers, and IT personnel who are interested in, or considering implementing, AI deployments.

Disclaimer

This document provides design guidance on deploying a typical small and medium to large GenAI GPU cluster, using basic networking concepts and leveraging key software features within Dell's Enterprise SONiC networking operating system to deliver an interoperable, high-performance, scalable, and efficient fabric.

The design guide does not provide performance data or sample infrastructure configurations. The Dell Validated Designs (DVDs) provide the specific infrastructure configurations that can be deployed at scale with Zero-Touch Provisioning (ZTP), Ansible, or a third-party partner orchestration tool

Terminology

The following table provides definitions for some terms used in this document.

Table 1. Terminology

Term	Definition
GPU	Graphics Processing Unit
CPU	Central Processing Unit
AI	Artificial Intelligence
SONiC	Software for Open Networking in the Cloud
Dell Enterprise SONiC	Dell Open-Source SONiC implementation
RoCE	RDMA over Converged Ethernet
RDMA	Remote Direct Memory Access
DHCP	Dynamic Host Configuration Protocol
GbE	Gigabit Ethernet
VLAN	Virtual Local Area Network
DAC	Direct Attach Copper
BGP	Border Gateway Protocol

Solution components

Overview

The following solution components make up this design guide.

Software Component: Dell Enterprise SONiC Network Operating System

Dell Technologies has been an innovator in the open-source and disaggregation arena for many years, starting with the integration of running third-party networking operating systems on Dell branded networking equipment to running open-source code such as SONiC.

Dell Technologies has created a world-class, enterprise-grade version of the open-source SONiC stack. The Dell Enterprise SONiC stack provides a full suite of enterprise, cloud, edge, and campus features.

This stack delivers scalable, high-performance, multi-tenancy connections for Hyper-Converged (HCI), Converged Infrastructure (CI), and AI workloads.

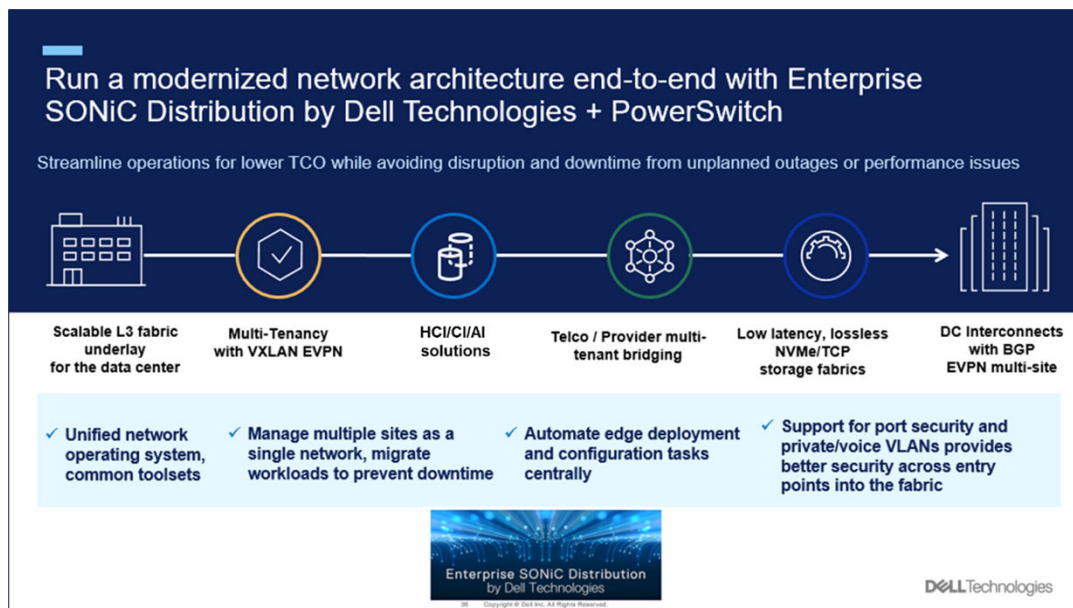


Figure 1. Dell Enterprise SONiC

Dell Enterprise SONiC, starting with version 4.2.0 on the Z9664F-ON, has implemented a series of networking features that facilitate the deployment of a single to multi-rack GPU cluster.

Figure 2 shows the initial Dell Enterprise SONiC for GenAI networking feature set. Each feature aims to address performance, quality-of-service, and ease of management for a GenAI deployment.

Dell Enterprise SONiC

Enabling the new world of Artificial Intelligence workloads

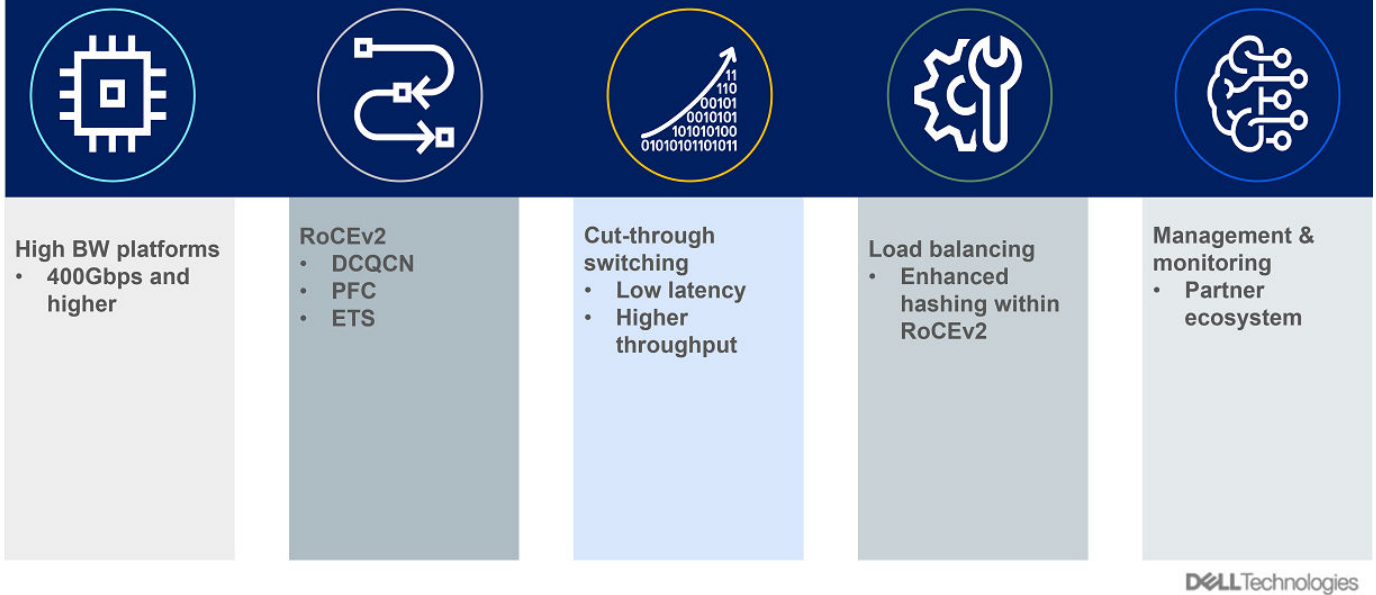


Figure 2. Dell Enterprise SONiC GenAI feature set

Besides offering a purpose-built AI software feature set, a robust set of management and monitoring capabilities are integrated into the Dell Enterprise SONiC network operating system.

Hardware Components

The hardware components of the Dell GenAI fabric solution design have been purpose built for the AI world. The hardware product family is light, high-performing, efficient, and based on open-standards. These characteristics make a Dell end-to-end solution design a good fit for unique workloads that require high-performance, efficiency, and easy-to-manage deployments.

Dell PowerSwitch

The typical GenAI environment consists of back-end, front-end, and management fabrics. The Dell PowerSwitch Z-series makes up the back-end fabric and supports the GPU interconnects.

The Dell PowerSwitch portfolio is the infrastructure that delivers the GenAI solution. In this design guide, the Z9664F-ON and Z9432F-ON are the primary 400GbE switches. With up to 51.2 Terabits per second (Tbps) switching capacity, these two switches are purpose-built for any kind of GenAI workload.

Customers looking for a high-performance, scalable, and lossless fabric for their GenAI environment should consider the Dell Z-Series platform.

Z9664F-ON 400GbE super spine switch

Powering up the next-generation IP fabric with 400GbE open networking

- State-of-the-art, high density 100/400GbE switch for demanding environments with 4X the throughput, 2X the price/performance and double the power efficiency of 100GbE platforms:
 - 64 ports x 400GbE
 - 128 ports x 200GbE (via breakout)
 - 256 ports x 10/50/100GbE ports (via breakout)
- 51.2Tbps switching capacity in 1RU (full duplex)
- Based on Broadcom Tomahawk 4 chipset

Purpose


- Built for Web 2.0, enterprise, and Tier1/Tier2 cloud service provider data center networks with intensive compute and storage traffic, cloud IoT, AI and streaming video requirements

Dell Technologies innovation

- Supports Open Networking (ONIE) and select 3rd party OS
- Flexible & multi-rate (10/25/40/50/100/200/400GbE) for cost-effective 100GbE connectivity and to help simplify migration to 100/200/400GbE
- Runs Dell SmartFabric OS10 or Enterprise SONiC* Distribution by Dell Technologies
- QSFP56-DD 400G form factor with low power, cost & space

Enterprise SONiC Distribution by Dell Technologies*

Dell PowerSwitch Z9664F-ON



2X

Switching throughput in 2RU form factor**

2X

Density of 400GbE switching ports***

* SONiC supported as of v4.1
 ** As compared to existing 400GbE Z-series switches
 *** As compared to existing 32-port 400GbE Z-series switches

DELLTechnologies

Figure 3. Dell PowerSwitch Z9664F-ON

Z9432F-ON 400GbE super spine switch

Powering up the next-generation IP fabric with 400GbE open networking

- State-of-the-art, high density 100/400GbE switch for demanding environments with 4X the throughput, 2X the price/performance and double the power efficiency of 100GbE platforms:
 - 32 ports x 400GbE
 - 64 ports x 200GbE (via breakout)
 - 128 ports x 10/50/100GbE ports (via breakout)
- 25.6Tbps switching capacity in 1RU (full duplex)
- Based on Broadcom Trident 4 chipset

Purpose

- Built for Web 2.0, enterprise, and Tier1/Tier2 cloud service provider data center networks with intensive compute and storage traffic, cloud IoT, AI and streaming video requirements


Dell Technologies innovation

- Supports Open Networking (ONIE) and select 3rd party OS
- Flexible & multi-rate (10/25/40/50/100/200/400GbE) for cost-effective 100GbE connectivity and to help simplify migration to 100/200/400GbE
- Runs Dell SmartFabric OS10 or Enterprise SONiC Distribution by Dell Technologies
- QSFP56-DD 400G form factor with low power, cost & space

ONIE

Enterprise SONiC Distribution by Dell Technologies

Dell PowerSwitch Z9432F-ON



4X

Switching throughput in 1RU form factor*

2X

Density of 100GbE switching ports**

* As compared to existing 100GbE switching platforms
 ** As compared to existing 64 port 100GbE switch

DELLTechnologies

Figure 4. Dell PowerSwitch Z9432F-ON

The front-end fabric is the GPU storage and application fabric. This fabric is supported by the Dell PowerSwitch S5200F-ON series.

The Dell PowerSwitch S5200 series delivers 100GbE connectivity from the storage and application layer to the GPU fabric. The S5200 series switchports range from 12 to 48 25GbE ports and four to 32 100GbE ports.

S5200-ON 25GbE & 100GbE in-rack switches

Enterprise SONiC Distribution by Dell Technologies

Dell PowerSwitch S5200-ON

Latest generation 25GbE & 100GbE open networking switches

- Low-cost fixed form factor top-of-rack switches offering multiple options of 25GbE SFP28 ports for in-rack server and storage connections and 100GbE QSFP28 & QSPDD-28 ports for uplink and clustering
- Based on Broadcom Trident3 chipset
- Enhanced buffering, higher forwarding tables and data plane support for VXLAN Routing (RIOT, Routing In and Out of Tunnels)
 - S5212F-ON – 1RU, half-width, 12 x 25GbE ports and 3 x 100GbE ports, 2.5X the throughput at 1/3 the size
 - S5224F-ON – 1RU, 24 x 25GbE ports and 4 x 100GbE ports
 - S5248F-ON – 1RU, 48 x 25GbE ports and 8 x 100GbE ports (4xQSFP28 100GbE and 2xQSPDD-28 2x100GbE ports)
 - S5296F-ON – 2RU, 96 x 25GbE ports and 8 x 100GbE ports
 - S5232F-ON – 1RU, 32 x 100GbE ports

Purpose

- Built and optimized for combinations of 25GbE connections in-rack with 100G to fabric and highly scalable and cost-effective 100GbE leaf/spine fabric between data center racks
- Ideal for Web 2.0, Enterprise, mid-market and Cloud Service Provider data center networks

Dell Technologies innovation

- High Density (96-port) for ToR/MoR/EoR
- QSFPDD-28 ports for higher density 100GbE uplink (S5248F)
- Open Networking running OS10 & ONIE or Enterprise SONiC Distribution (S5232F-ON, S5248F-ON, S5296F-ON)
- Fully tested and validated with 3rd party operating systems

S5212F-ON
S5224F-ON
S5248F-ON
S5296F-ON
S5232F-ON

2.5X Throughput of traffic
32 100GbE ports in 1RU

* Ports: Comparing S5296F (96) with S5048F (48) 25GbE

DELLTechnologies

Figure 5. Dell PowerSwitch S5200-ON Series

The last GenAI fabric is the Out-of-Band (OOB) management network. This network is the management access of the entire GenAI environment, and it includes the workload cluster devices.

The OOB fabric connections are all 1GbE copper-based RJ45 connections.

N3200-ON L2 & L3 Advanced 1G & 10G MultiGig Access

Enterprise SONiC Distribution by Dell Technologies

Dell PowerSwitch N3200-ON

Latest generation 1G and 10G MultiGig Campus Access Switches

- Cost optimized fixed form factor switches, with wide range of port density options for 1G and 10G MultiGig speeds and 802.3bt Type-4 (90W) and 802.3at (30W) PoE ports
- x86 platform based on Broadcom Trident t3 chipset (N3208PX-ON based on Broadcom Hurricane 3 MG chipset)
 - N3208PX-ON - Compact 4x5G 90W PoE and 4x1G 90W PoE ports
 - N3224T-ON - 1RU, 24x1G RJ-45 ports
 - N3224P-ON - 1RU, 24x1G 802.3at (30W) PoE ports
 - N3224F-ON - 1RU, 24x1G SFP ports
 - N3224PX-ON - 1RU, 24x1/2.5/5/10G RJ-45 802.3bt Type-4 (90W) PoE ports
 - N3248TE-ON - 1RU, 48x1G RJ45-ports
 - N3248P-ON - 1RU, 48x1G 802.3at (30W) PoE ports
 - N3248X-ON - 1RU, 48x1/2.5/5/10G RJ-45 ports
 - N3248PXE-ON - 1RU, 48x1/2.5/5/10G RJ-45 802.3bt Type-4 (90W) PoE ports

Purpose-built for

- 802.11ax WLAN deployments and 802.3bt Type-4 (90W) high power PoE applications
- Ideal for large enterprise campus networks and large retail deployments

Dell Technologies innovation

- Open Networking
- Runs OS6 or Enterprise SONiC for edge (N3248TE-ON, N3248PXE/X-ON)
- High density (48-ports) 10G MultiGig and 802.3bt Type-4 (90W) PoE
- 400G Stacking with up to 12 members
- 25G Uplinks to aggregation

48x 10G MultiGig
48x 90W PoE

DELLTechnologies

Figure 6. Dell PowerSwitch N3200-ON Series

Dell PowerEdge XE9680

The Dell XE9680 is a 6U server, and it is Dell's first 8 x GPU platform. This server is engineered to significantly enhance application performance by driving the most complex GenAI, ML, deep learning (ML/DL), and high-performance computing workloads (HPC).

The XE9680 server features up to 56-core 4th generation Intel Xeon processors and offers the highest GPU memory capacity and bandwidth currently available, making it capable of managing large and complex models and datasets.

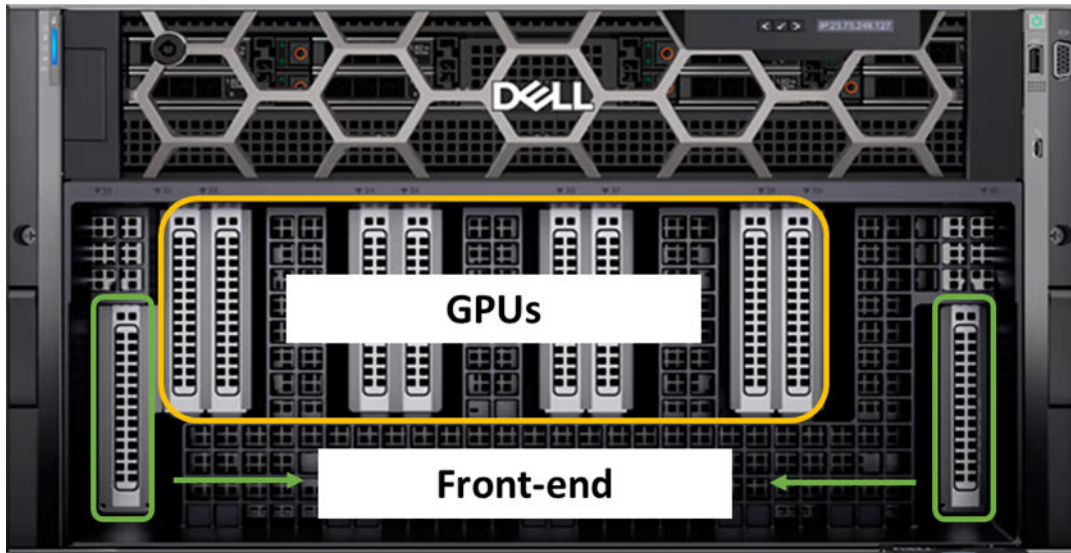


Figure 7. Dell PowerEdge XE9680

Two of its PCIe slots (located on the right and left) are used for front-end fabric connections. These slots are configured as 2 x 200GbE in this design guide.

Dell PowerScale and ECS

A GenAI environment requires large storage capabilities as it builds the large language models (LLM) dataset. GenAI data generation tends to be unstructured, random, and overall chaotic at its inception. To provide storage capabilities for this data, Dell Technologies offers two separate product lines.

Dell ECS Enterprise Storage provides three models for different scalability needs. The network connectivity of the ECS appliances is all 25GbE-based.

The architecture of ECS provides flexibility and scalability on-demand. With up to 24 disk drives per node and 16 node cluster size, the ECS storage is more than capable of supporting any AI or GenAI environment.

Figure 8 shows a detailed breakdown of the ECS Enterprise storage product family.




Object Storage			
ECS EX500	ECS EX5000	ECS EXF900	
			
Features	EX500	EX5000	EXF900
Node architecture	<ul style="list-style-type: none"> Intel x86 servers Integrated storage 12 or 24 disk drives per node 	<ul style="list-style-type: none"> Intel x86 servers Integrated storage Up to 100 disk drives per node 	<ul style="list-style-type: none"> Intel x86 servers Integrated storage 12 or 24 disk drives per node
Network connectivity	<ul style="list-style-type: none"> 25GbE FrontEnd 25GbE BackEnd 	<ul style="list-style-type: none"> 25GbE FrontEnd 25GbE BackEnd 	<ul style="list-style-type: none"> 25GbE FrontEnd 25GbE BackEnd
Rack configurations	<ul style="list-style-type: none"> 1, through 16 node configurations (5 node minimum initial rack) HA power 	<ul style="list-style-type: none"> EX5000S: 1, through 7 node configurations (5 node minimum initial rack) EX5000D: 2, through 14 node configurations (8 node minimum initial rack) HA power 	<ul style="list-style-type: none"> 1, through 16 node configurations (5 node minimum initial rack) HA power
Storage configurations	<ul style="list-style-type: none"> Unstructured storage up to 7680TB per rack 	<ul style="list-style-type: none"> Unstructured storage up to 14,000TB per rack 	<ul style="list-style-type: none"> Unstructured storage up to 5898TB per rack

Figure 8. Dell ECS Enterprise GenAI object storage




Object Storage			
PowerScale F900	PowerScale F6000	PowerScale F200	
			
Features	PowerScale F9000	PowerScale F6000	PowerScale F200
Node architecture	<ul style="list-style-type: none"> Intel x86 servers Integrated storage 24 (SSD) disk drives per node 	<ul style="list-style-type: none"> Intel x86 servers Integrated storage 8 (SSD) disk drives per node 	<ul style="list-style-type: none"> Intel x86 servers Integrated storage 4 (SSD) disk drives per node
Network connectivity	<ul style="list-style-type: none"> 100GbE front-end 100GbE back-end 	<ul style="list-style-type: none"> 100GbE front-end 100GbE back-end 	<ul style="list-style-type: none"> 100GbE front-end 100GbE back-end
Rack configurations	<ul style="list-style-type: none"> 1, through 252 node configurations (3 node minimum initial rack) HA power 	<ul style="list-style-type: none"> 1, through 252 node configurations (3 node minimum initial rack) HA power 	<ul style="list-style-type: none"> 1, through 252 node configurations (3 node minimum initial rack) HA power
Storage configurations	<ul style="list-style-type: none"> All-flash unstructured storage up to 736 TB per node 	<ul style="list-style-type: none"> All-flash unstructured storage up to 245 TB per node 	<ul style="list-style-type: none"> All-flash unstructured storage up to 30.72 TB per node

Figure 9. Dell PowerScale GenAI unstructured storage

For customers looking to increase their GenAI storage and application network connectivity, the Dell PowerScale product family provides 100GbE connectivity, plus increased cluster size.

With PowerScale, a cluster can range from three to 252 nodes, providing up to 186 Petabytes (PB) of storage capacity.

Both the ECS and PowerScale product lines provide comprehensive object storage capability for GenAI data storage and delivery.

For details about these two storage options, see the Solution Design chapter.

Cables

Both copper and fiber cabling options are offered as part of the Dell GenAI infrastructure solution.

For a copper option, Dell Technologies offers a Direct Attach Cable (DAC) and an Active Electrical Cable (AEC). These copper cables have a directly attached transceiver at both ends that spans 4 meters, and an AEC that spans 5 meters.

For a fiber option, Dell Technologies offers a single mode fiber (SMF) with respective optics.

Figure 10 shows the different cable options offered as part of the Dell GenAI fabric solution.



Figure 10. Dell GenAI 400GbE cable product line

Optics

The Dell GenAI infrastructure design uses both passive and active connections. Fiber connectivity is an active connection. Fiber is easier to deploy due to its thinner form factor and longer distance coverage span of 200 meters.

400G QSFP56-DD (400GbE) optical transceiver with MPO12 optical interface

400G QSFP112 (400GbE) optical transceiver with MPO12 optical interface

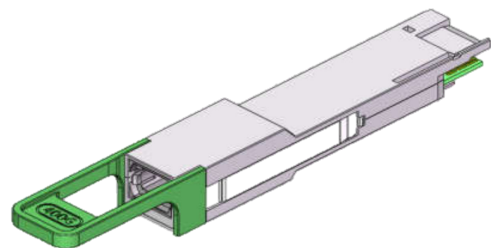
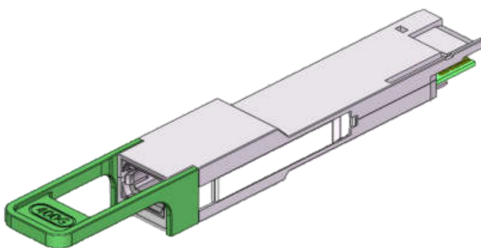


Figure 11. Dell GenAI infrastructure 400GbE optics

In this design guide, fiber connections will be used for connections that require a longer connection than 4 meters. These types of connections will take place at the end of row connections. For details, see Infrastructure - Cabling, Power.

Solution Design Guide and Requirements

Solution requirements

This section describes the requirements for the Dell GenAI fabric solution design.

Interoperable

The Dell GenAI fabric solution needs to operate at the highest level and leverage a well-established network ecosystem based on proven open standards such as Ethernet.

Several vendors are used to implement the different fabric topologies while delivering consistent performance and ease of management.

With an Ethernet-based approach, a better flexible design guide can be achieved.

High-performance

GenAI workloads are unique in that they require specific network or fabric requirements to perform optimally.

With Dell Enterprise SONiC as the networking operating system, a GenAI-specific feature set is included such as cut-through switching, RDMA over Converged Ethernet (RoCE) version 2, dynamic load balancing (DLB), and enhanced hashing to deliver the necessary network performance for GenAI workloads to perform. For details about these requirements, see Figure 2.

In addition to the software feature set by Dell Enterprise SONiC, the PowerSwitch Z-series provides the necessary non-blocking 400GbE switching fabric capacity.

Scalable

GenAI workloads can range from a single to a multi-cluster GPU environment. This design guide explores and provides two different fabric topologies: stand-alone fabric, and leaf and spine fabric.

The stand-alone fabric allows for a single GPU cluster from 32 to 64 GPUs. This fabric provides intra-GPU communication.

The leaf and spine fabric allows for a multi-GPU cluster ranging from 64 to 2,048 GPUs. This fabric provides inter-GPU communication.

Efficient

All intensive workload environments require a detailed, well-organized cable management and power allocation infrastructure.

The type of hardware that is being proposed for a GenAI environment requires specific power and cooling requirements.

The infrastructure or fabric needs to be designed such that end of row or middle of row hardware placement is considered to best accommodate cable length and optic specifications.

Solution topology

Figure 10 and Figure 11 demonstrate two physical architectures of the solution used to implement the Dell GenAI fabric options:

- Small GenAI GPU cluster
- Medium to Large GenAI GPU cluster

Small GenAI GPU cluster

The small GenAI cluster topology (shown in Figure 10) consists of:

- 1 PowerSwitch Z9664F-ON
- 1 PowerSwitch N3248TE
- 1 PowerSwitch Z-series, S5200-series, or S5448FF-ON
- 4 or 8 Dell PowerEdge XE9680s, each XE9680 has 8 Mi300X GPU
- 3-node Dell PowerScale cluster or 5-node ECS cluster
- Dell Enterprise SONiC 4.2.1

NOTE: Figure 10 shows a small GenAI cluster with an ECS 5-node cluster, where a 5-node cluster is the minimum size.

NOTE: For a smaller GenAI environment, the Z9664F-ON stand-alone PowerSwitch can be replaced with the Z9432F-ON, which would result in 4 XE9680s, or 32 GPUs.

Dell Technologies small GenAI 32/64 GPU cluster – reference diagram

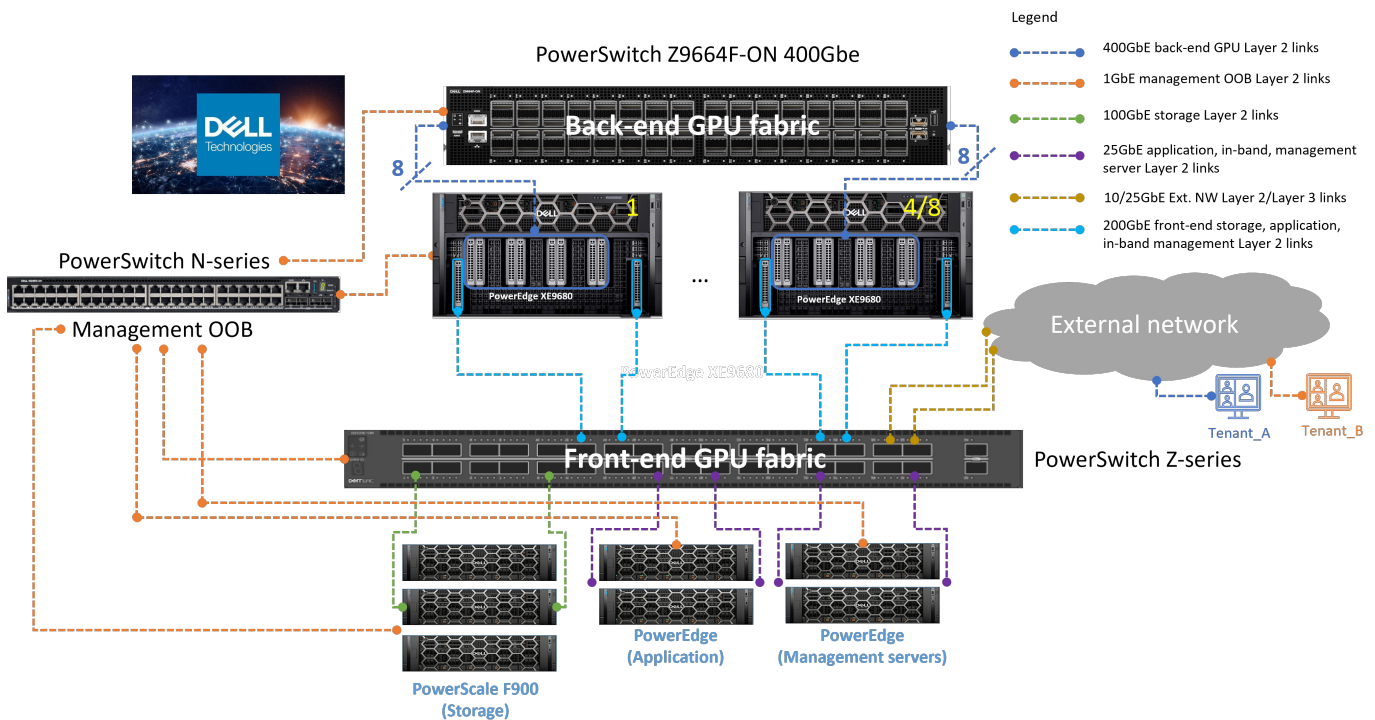


Figure 12. Small GenAI XE9680 GPU cluster

Medium to large GenAI GPU cluster

The medium to large GenAI architecture (Figure 11) consists of:

- Leaf and spine architecture
- 4 to 32 Dell PowerSwitch Z9664F-ON PowerSwitch as spine switches
- 8 to 64 Dell PowerSwitch Z9664F-ON PowerSwitch as leaf switches
- 4 to 10 Dell PowerSwitch N3248TE
- Dell PowerEdge XE9680, 32/256 node cluster, or 256/2,048 GPUs
- 3-node cluster Dell PowerScale, where a 3-node cluster is the minimum size, or a 5-node ECS cluster
- Dell Enterprise SONiC 4.2.1

Dell Technologies medium to large GenAI 256/2048 GPU cluster - reference diagram

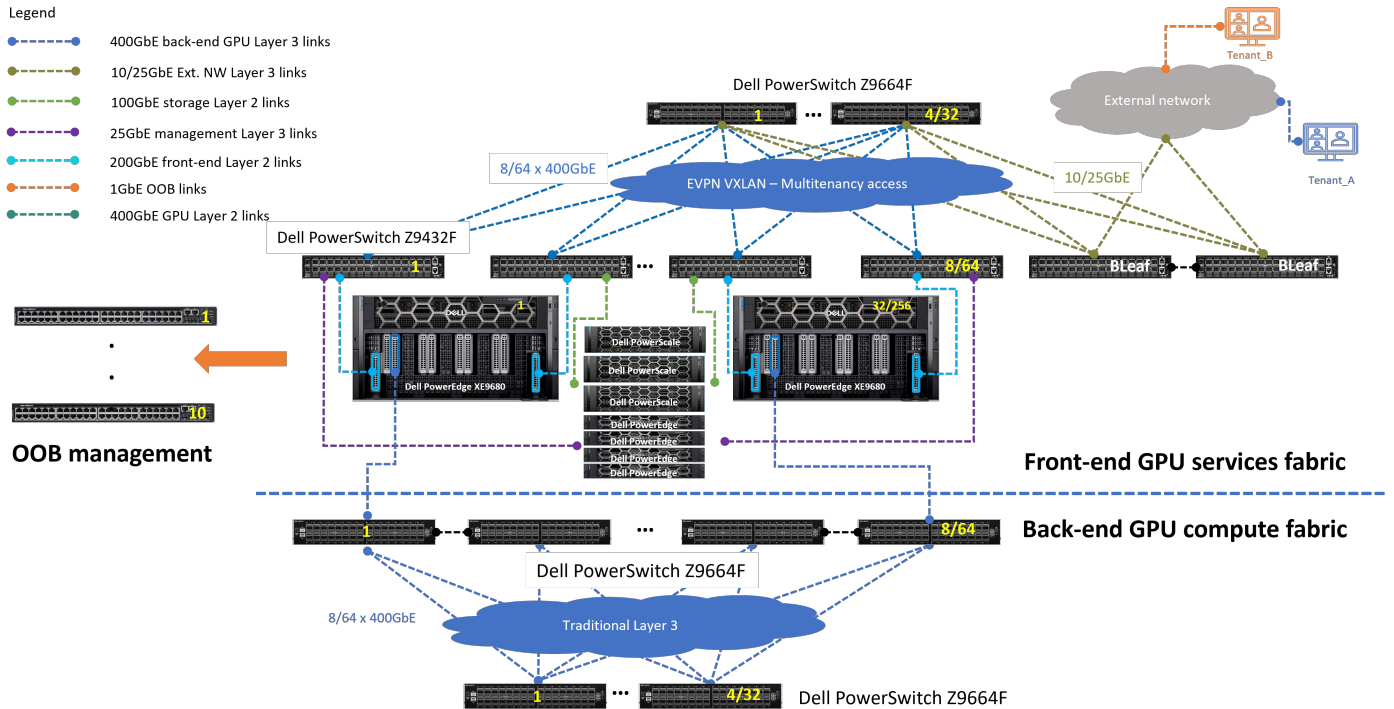


Figure 13. Medium to Large GenAI XE9680 GPU cluster

Common GenAI fabrics overview

This section of the design guide introduces the different GenAI fabrics that make up a typical GenAI infrastructure.

Out-of-Band management fabric

The OOB fabric provides management connectivity, whether local or remote, into the GenAI infrastructure. The OOB management fabric bandwidth is set to 1GbE.

The OOB connections range from simple iDRAC connections from the PowerEdge, PowerScale, or ECS appliances to RJ45 Ethernet management ports on the PowerSwitches.

The OOB fabric is assigned to a single subnet or VLAN. The default VLAN is 1; however, most customers use a different VLAN ID. This VLAN is then routed across the network, so it is reachable from anywhere.

The GenAI OOB fabric uses the PowerSwitch N3248TE. This switch supports Layer 2 and Layer 3 switching, and the switchport speeds are 1GbE. The Dell PowerSwitch N3248TE has four 10GbE SFP ports that can be used to uplink onto the legacy or existing network.

Figure 14 shows the OOB connections from all the devices in the Dell GenAI environment. Each PowerScale or ECS appliance has a single OOB connection to the Dell PowerSwitch N3248TE.

Dell Technologies AI fabric with AMD Mi300x GPU & BCM Thor2 stack – OOB management

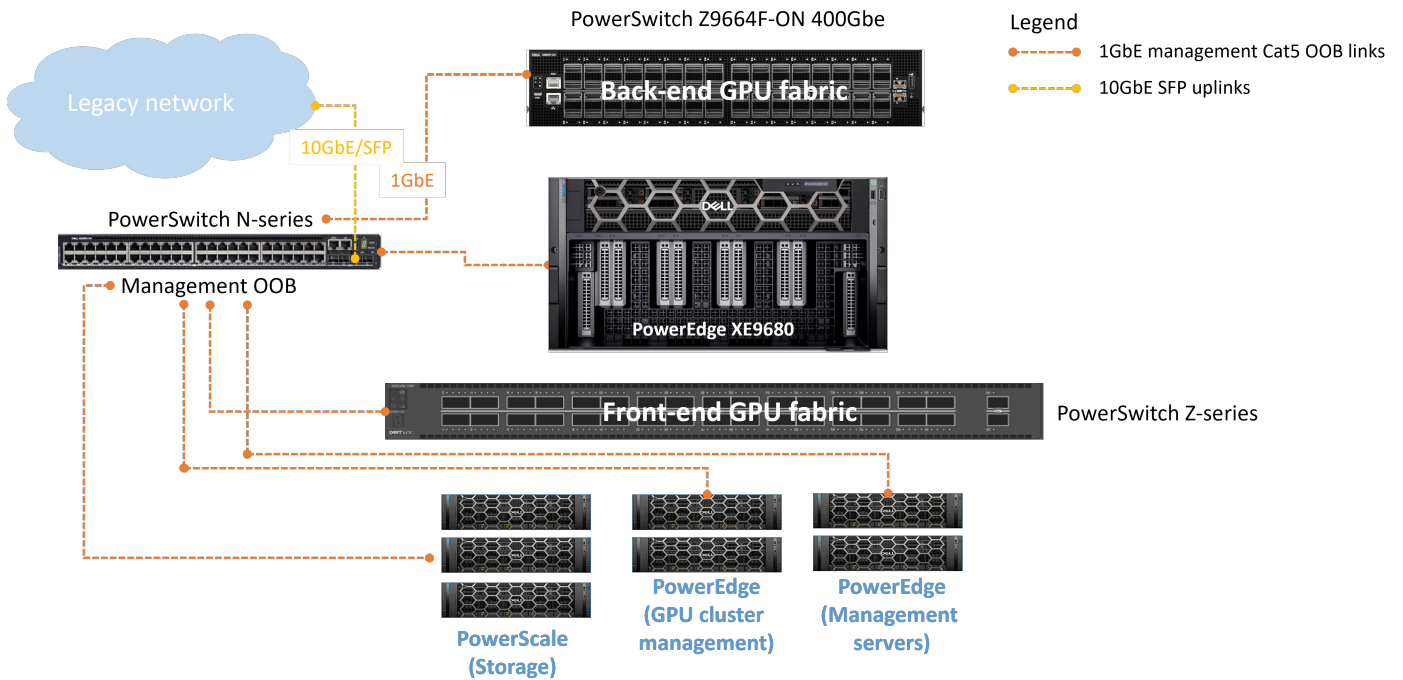


Figure 14. Out-of-Band Management GenAI reference topology

Back-end GPU fabric

The back-end GPU fabric is called the inter-GPU fabric. Of the three different fabrics, the back-end GPU fabric is the most bandwidth-intensive and requires careful planning.

Figure 15 shows the connections from the Dell PowerEdge XE9680 to the Dell PowerSwitch Z9664F-ON or Z9432F-ON.

Dell Technologies AI fabric with AMD Mi300x GPU & BRCM Thor2 stack – GPU fabric

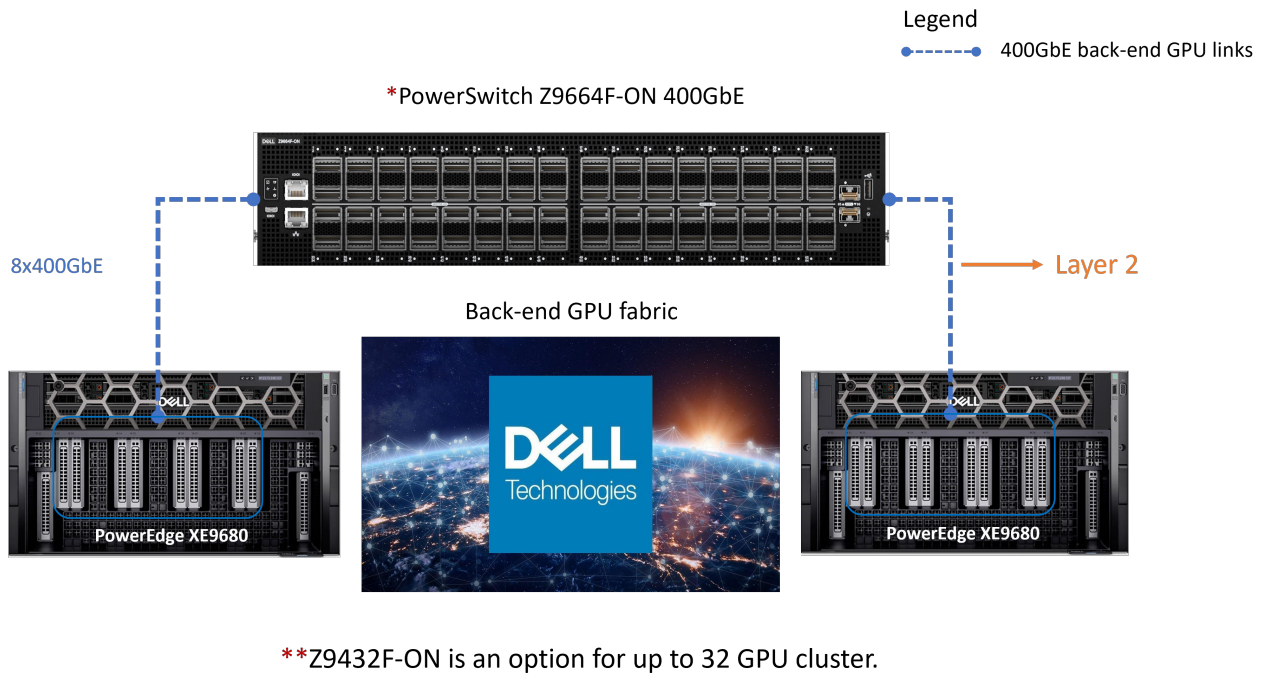


Figure 15. Back-end GPU GenAI reference topology

The connections in this topology are 400GbE, and they are not oversubscribed--that is, the connections are fully non-blocking 400GbE. These connections are Layer 2, and they are assigned to a specific VLAN ID. This VLAN ID is reserved for all inter-GPU communication and cannot be used elsewhere in the network.

Each GPU on the Dell PowerEdge XE9680 has a respective NIC port. In this case, each Broadcom Mi300x GPU connects to a respective Thor 2 NIC. The Thor 2 NIC has two 200GbE ports that can be virtually bundled into a single 400GbE port using a breakout cable (see Figure 10).

The Dell PowerSwitch can be configured as a simple Layer 2 switch with several VLAN IDs, or a single VLAN with all the connections belonging to the same VLAN ID.

Figure 15 shows a stand-alone or a single-rack GenAI GPU cluster; however, in a multi-rack GenAI GPU cluster, a leaf and spine architecture is recommended.

In a leaf and spine architecture, Layer 2 is configured from the GPU to the leaf, and Layer 3 connections are configured from the leaf to spine (BGP or BGP EVPN).

The Multi-rack GenAI cluster solution fabric design chapter will discuss the overall fabric design in more detail.

Front-end GPU storage, application, and in-band management

The front-end GPU fabric is dedicated for GenAI storage, application, and in-band GPU resources management support.

GenAI storage is needed to ingest, fine-tune, train, and infer large language models.

The GPU cluster environment service (such as Slurm or Kubernetes) is the application of the GenAI design solution. This GPU management cluster software stack provides a single pane of glass for GPU cluster management.

This fabric is not as bandwidth-intensive as the back-end inter-GPU fabric. At 200GbE or 100GbE, it is sufficient to provide storage, application, and in-band management access to and from the GPU fabric cluster.

Note: Unlike the back-end fabric where 400GbE is deployed, the front-end fabric can use 200GbE, 100GbE, 40GbE, 50GbE, or 25GbE. For GenAI however, 200GbE or 100GbE is recommended to accommodate the amount of traffic between the GPU cluster and storage needs.

Note: Although Dell ECS is mentioned as a storage solution for GenAI, this design guide will concentrate on the Dell PowerScale option.

Dell Technologies AI fabric with AMD Mi300x GPU & BCM Thor2 stack – front end

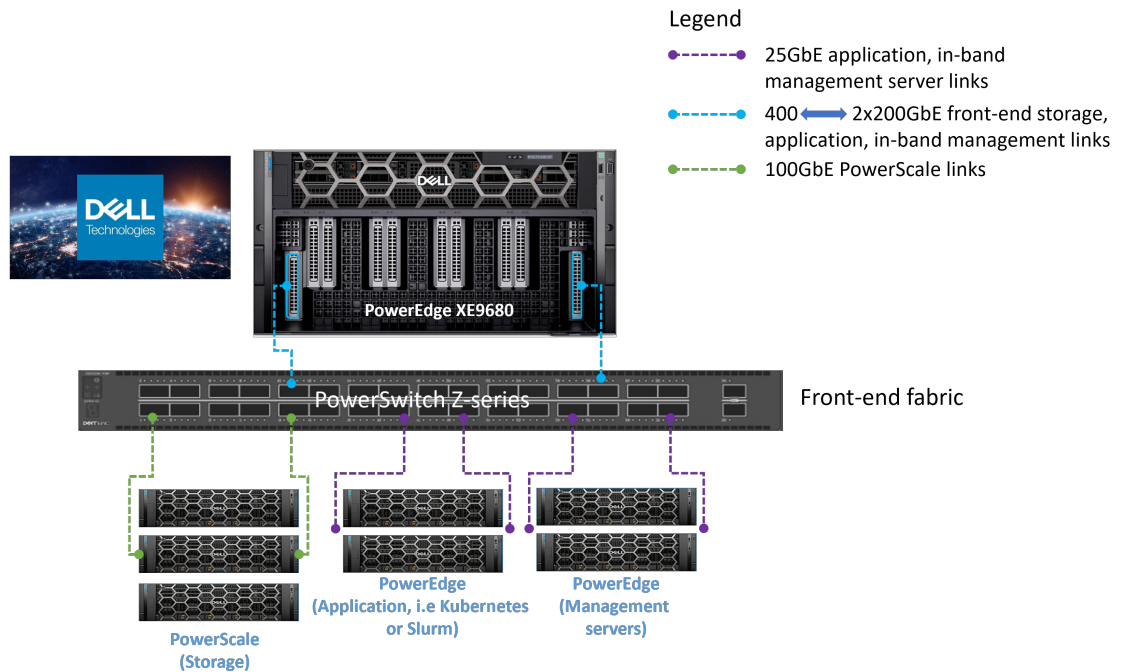


Figure 16. Front-end Dell GenAI GPU storage and application reference topology

The front-end fabric is a converged fabric where storage, application, and in-band GPU cluster management connections take place at different speeds.

GPU storage, application, and in-band GPU cluster management speeds are configured at 200GbE, 25GbE, and 25GbE respectively from the Dell PowerSwitch to the PowerScale, application, and in-band GPU cluster servers (see Figure 16).

The Dell PowerSwitch that provides the front-end fabric is configured as a Layer 2 switch with three different VLAN segments (storage, application, and in-band GPU management cluster).

The connections from the storage, application, and in-band cluster management to the Dell PowerSwitch are Layer 2, and these connections are configured as link-aggregation (LAG) bundles providing link redundancy upon a link failure.

The PowerScale cluster has two fabrics. The back-end fabric is for inter-PowerScale appliance communication, and the front-end fabric is for storage to GPU communication. This design guide only covers the front-end infrastructure design.

The back-end PowerScale fabric uses both Ethernet or InfiniBand, and it is not user-configurable. This fabric comes pre-configured or deployed from the manufacturing floor.

Using Dell PowerScale, the minimum cluster size is a three-node cluster. With Dell PowerScale, scalability is not an issue as the cluster can grow up to 255 nodes.

For GPU cluster management, Slurm or Kubernetes is deployed on a 2-node PowerEdge cluster. For in-band cluster management such as scheduling GPU jobs, another 2-node PowerEdge cluster is used.

Each of these services (storage, application, and in-band GPU cluster management) is assigned a unique VLAN ID or subnet. These VLANs are part of an EVPN VXLAN configuration on the front-end fabric switch.

Dell Technologies GenAI solution fabric design

This section describes a Dell Technologies GenAI small and medium to large GPU cluster solution design showing the specific optics, cables, network protocols, software feature set, and recommendations in the deployment of the solution.

NOTE: The two designs (small and medium to large) are a reference. They are not meant to be the only design approach for an AI environment.

Small GenAI cluster fabric design

Figures 17 and 18 show a small non-redundant and redundant GenAI GPU cluster. The components of the solution are:

- Dell Enterprise SONiC 4.2.1
- 1 Dell PowerSwitch Z9664F-ON
- 8 Dell PowerEdge XE9680 with eight GPUs per XE9680
- 1 Dell PowerSwitch N3248TE
- 1 Dell PowerSwitch Z, S5248F-ON or S5448F-ON
- 3-node PowerScale cluster, or 5-node ECS cluster
- 4 Dell PowerEdge R760 with dual 2x25GbE NIC ports

Dell Technologies small GenAI 32/64 GPU cluster

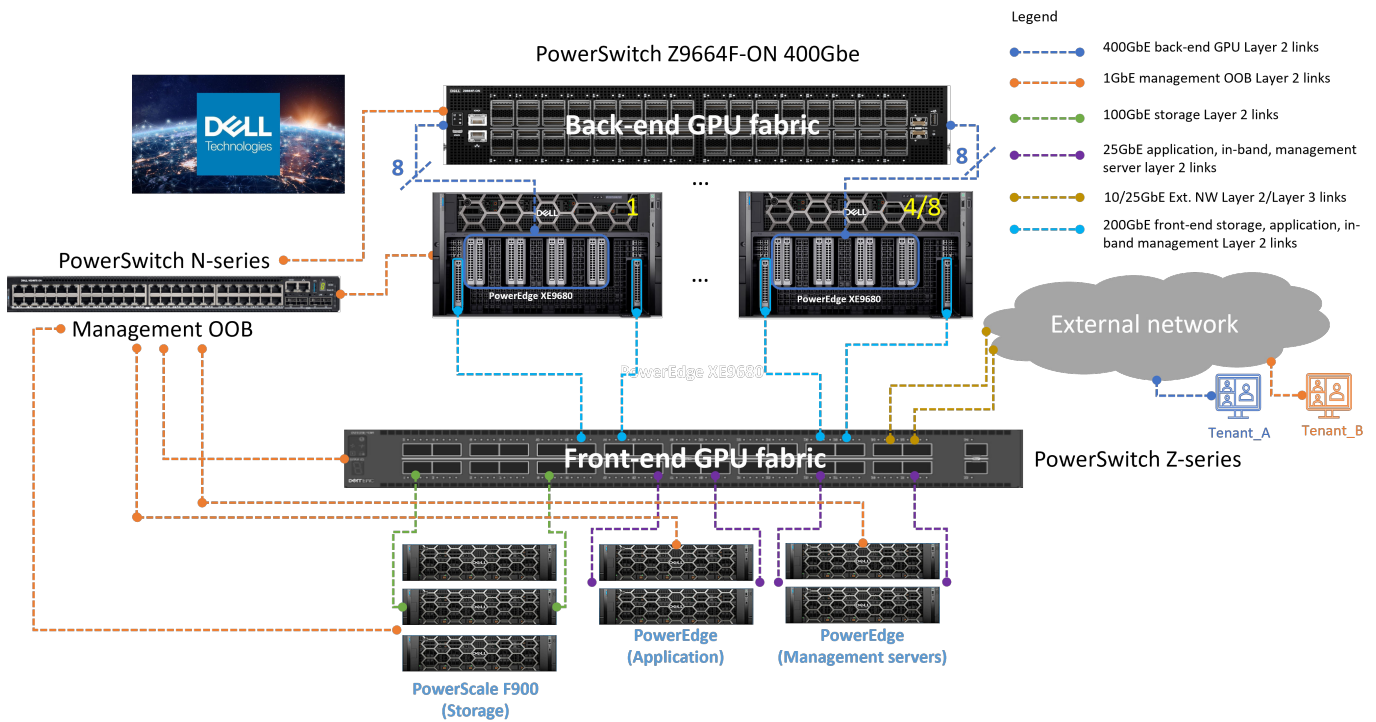


Figure 17. Dell Technologies small GenAI XE9680 32/64 non-redundant GPU cluster

The single rack is an entry-level GenAI cluster that starts at 32 up to 64 GPUs.

Dell Technologies small GenAI 32/64 GPU redundant mini-rail fabric cluster

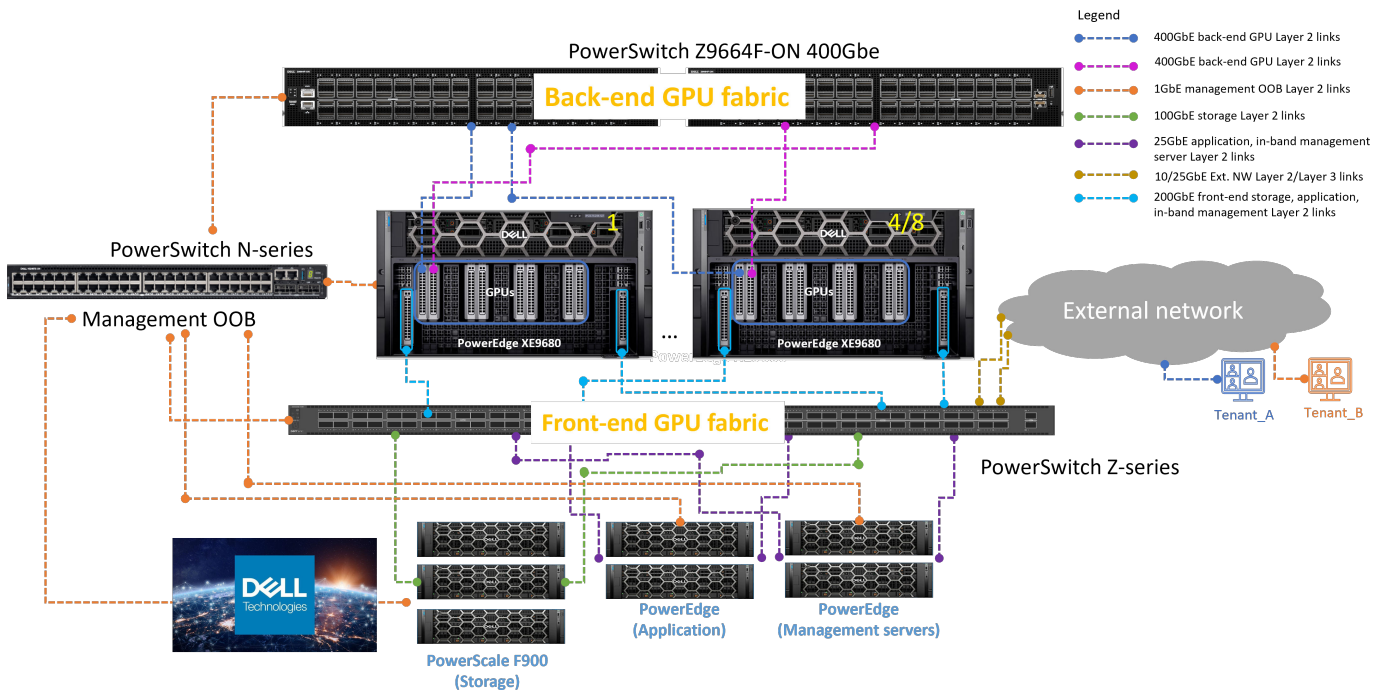


Figure 18. Dell Technologies small GenAI XE9680 32/64 redundant GPU cluster

OOB management fabric

The OOB management fabric provides access into the environment. In this design, the Dell PowerSwitch N3248TE provides 1GbE connectivity to all the devices in the solution.

When designing the OOB management fabric, the following networking guidelines should be followed:

- Assuming all the devices and applications in the environment are in the same management subnet, then all N3248TE switchports connected into the device management port or server iDRAC should be configured under the same VLAN ID. This design guide uses VLAN ID 99.

The traffic from this VLAN segment is either terminated on the same N3248TE switch or extended into the legacy network and routed by the legacy network infrastructure.

- For this design guide, management traffic (VLAN 99 for example) is extended into the legacy network through the 10GbE SFP uplinks.
- The 10GbE SFP fiber uplink ports (2 ports) on the N3248TE are configured as a LAG bundle. When configuring the LAG on the N3248TE, make sure the LAG mode (static or dynamic) matches on the N3248TE and the upstream device inside the legacy network cloud.

This LAG is configured as a trunk to carry VLAN 99 out into the legacy network. See Figure 18.

Dell Technologies AI fabric with AMD Mi300x GPU & BCM Thor2 stack – OOB management

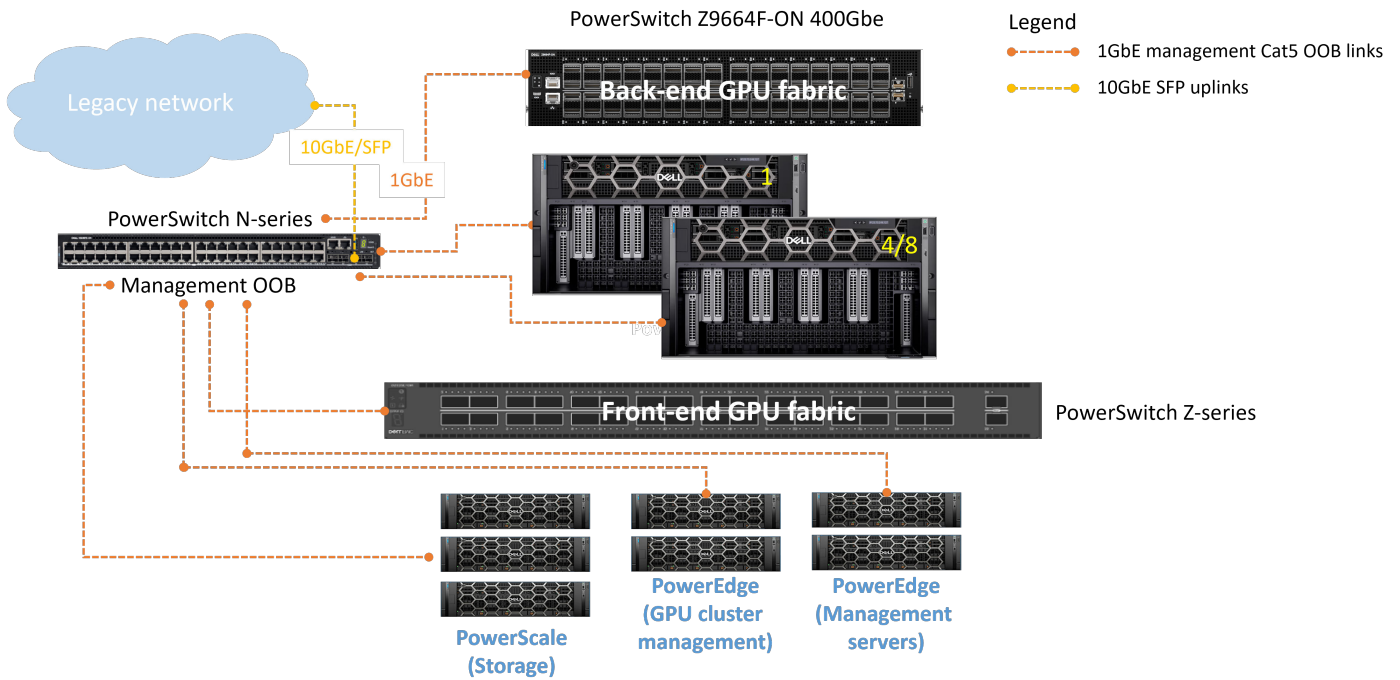


Figure 19. OOB management GPU cluster fabric

- The legacy network device terminates VLAN 99 as a default gateway. From the legacy network, VLAN 99 is routed to provide external management connectivity into the GenAI environment.
- The GenAI management subnet is divided into several subnets. For example, if the Dell PowerSwitches management subnet is VLAN 99 and the servers' iDRAC management is VLAN 100, then the OOB management switch (N3248TE) switchports need to be divided and configured into their respective VLANs.
- VLANs 99 and 100 are extended and terminated into the legacy network cloud. From the legacy network cloud, these two VLAN segments are routed.

Back-end GPU fabric

The back-end GPU fabric provides inter-GPU communication. In this design, the Dell PowerSwitch Z9664F-ON provides 400GbE non-blocking to all the GPUs in the cluster.

When designing the back-end GPU fabric, follow these networking design guidelines:

- The design of the back-end GPU fabric is based on a flat Layer 2 implementation. All the switchports on the Dell Z9664F-ON are configured to be part of the same VLAN ID or broadcast domain.
By placing all the GPUs on the same broadcast domain, GenAI workloads do not incur additional potential latency by having Layer 2 GPU traffic terminate on a default Layer 3 gateway to cross into another Layer 2 broadcast domain.
- Enable jumbo frames (9000 - 9216 bytes) on all the switchports.
- Enable RoCE version 2 on the switches. RDMA bypasses the CPU by accessing memory directly. This increases compute-to-compute communication.
- Enable cut-through switching on the switches.
- Enable dynamic load balancing.

Dell Technologies GenAI 32/64 GPU cluster – GPU non-redundant back-end fabric

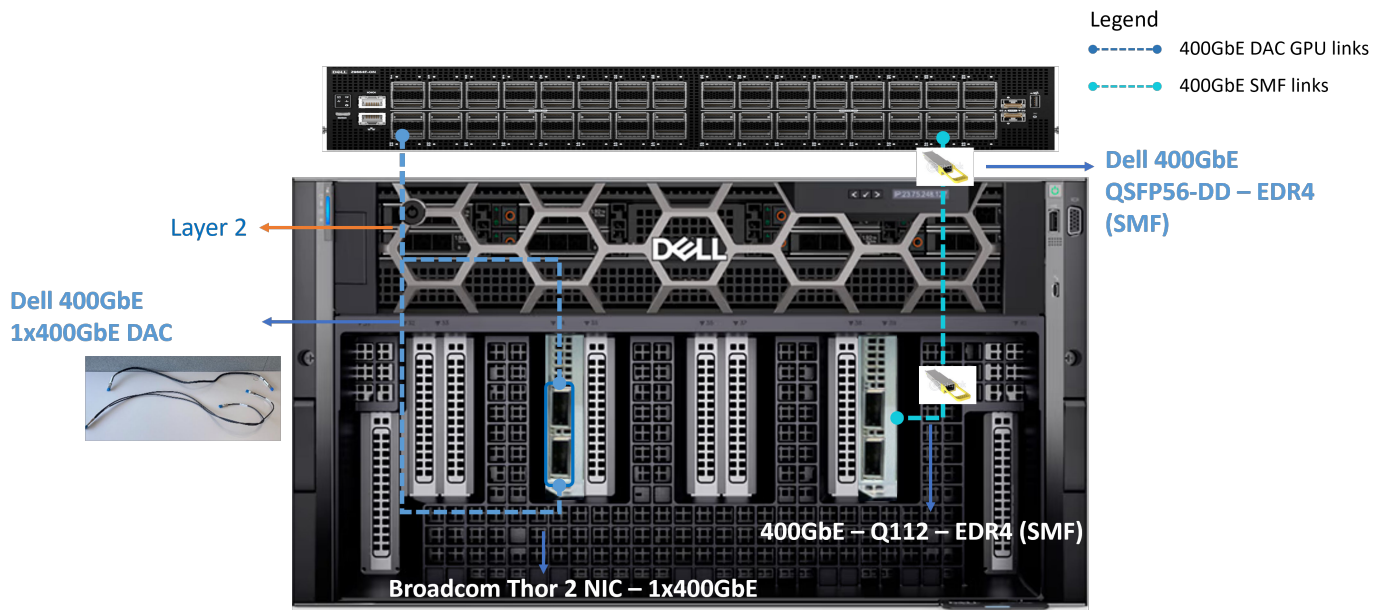


Figure 20. Back-end GPU cluster physical connection

Figure 20 shows a single XE9680 with its GPU physical connection to the Dell PowerSwitch Z9664F-ON using the Dell 400GbE to 1x400GbE DAC breakout cable or single mode fiber (SMF) with its optic (EDR4).

In the case of a redundant Z9664F, the connections would be repeated.

Repeat the same connections for all the XE9680s in the cluster.

The GPU DAC connection has two physical and functional aspects.

Figure 20 and 21 show a physical single DAC broken into two DAC connections onto their respective NIC ports on the Broadcom Thor 2. However, functionally this DAC connection is a single 1x400GbE connection.

Even though the Broadcom Thor 2 NIC shows two physical NIC ports, it operates and behaves as a single 400GbE NIC by default.

NOTE: The single 400GbE connection on the Thor 2 NIC can be configured as a 2x200GbE using the same physical DAC cable if the Z9664F-ON switchport is configured as a 2x200GbE as well. (See Figure 20.)

GPU fiber connectivity is simpler. When using a fiber connection option, only one NIC port is used on the Broadcom Thor 2 NIC to deliver 400GbE. The second NIC port cannot be used. (See Figure 21.)

Figure 21 shows the physical connections made when using a DAC or single mode fiber cable.

400GbE Solution – Thor2 NIC to Dell PowerSwitch Z-Series

Dell Switch Platforms: Z9664F-ON & Z9432F-ON

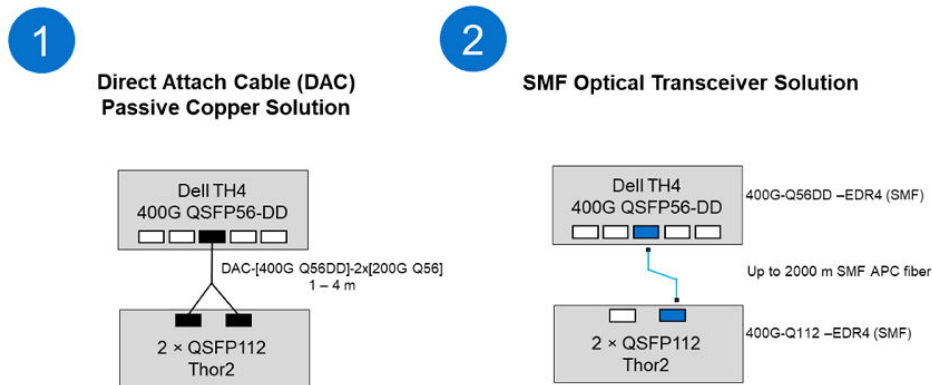


Figure 21. 400GbE Thor 2 to Dell PowerSwitch connections.

Front-end storage, application, in-band management fabrics

The front-end fabric, the Dell PowerSwitch Z9432F-ON, provides three different GPU cluster services: storage, application, and GPU cluster in-band management.

The XE9680 storage connections use the Thor 2 NICs on slots 31 and 40. These connections are configured as 2x200GbE. This is not the default speed configuration and must be configured on the Thor 2 NIC.

There are two deployment options for the GenAI front-end fabric.

- Non-redundant: A single Z9432F-ON is used as the front-end fabric. See Figure 22.
- Redundant: A pair of Z9432Fs is used to provide link and fabric redundancy for the front-end fabric. See Figure 23.

On the Dell Z9432F-ON, the switchports connected to the Thor 2 NICs are configured as 2x200GbE. (See Figure 22 or Figure 23.)

The physical cable used is a Dell DAC breakout cable (Figure 21).

Dell Technologies GenAI 32/64 GPU cluster – front-end non-redundant fabric

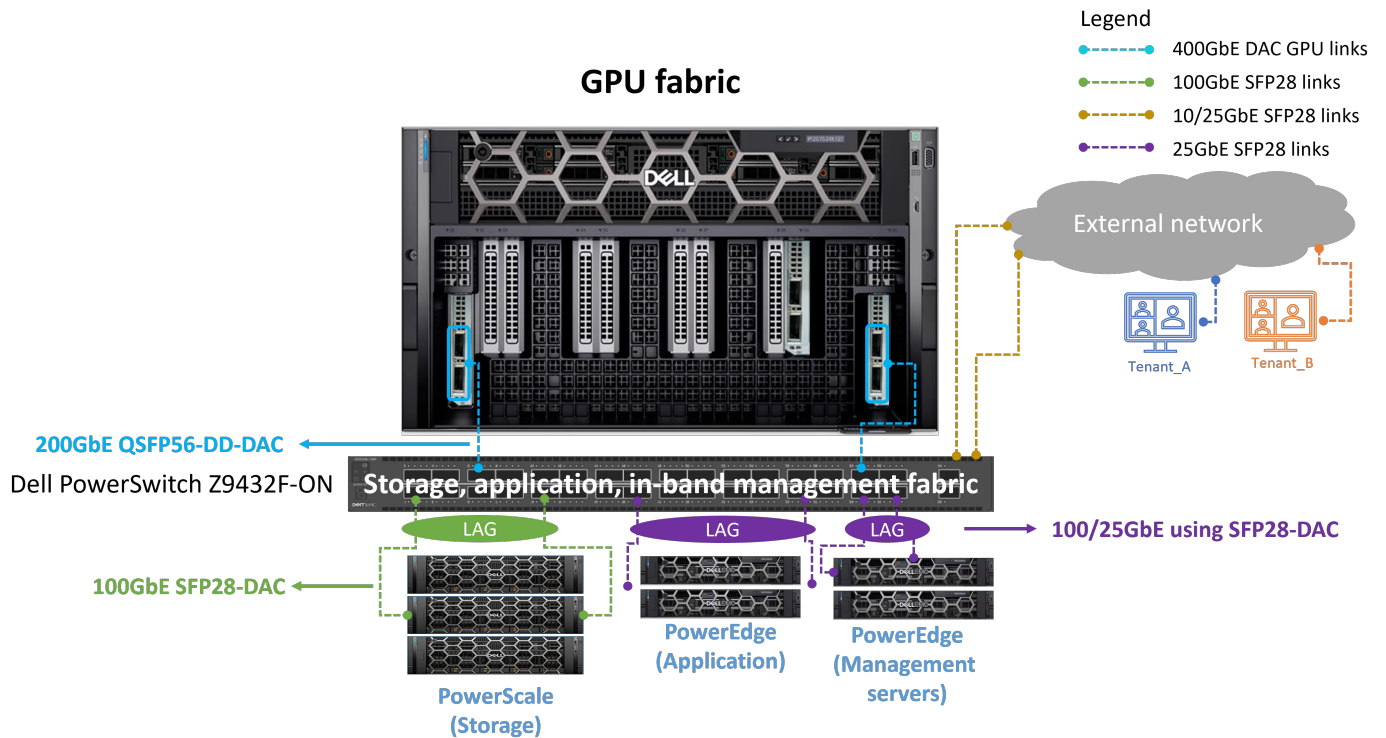


Figure 22. Front-end GPU cluster non-redundant fabric option 1

Dell Technologies GenAI 32/64 GPU cluster – front-end redundant fabric

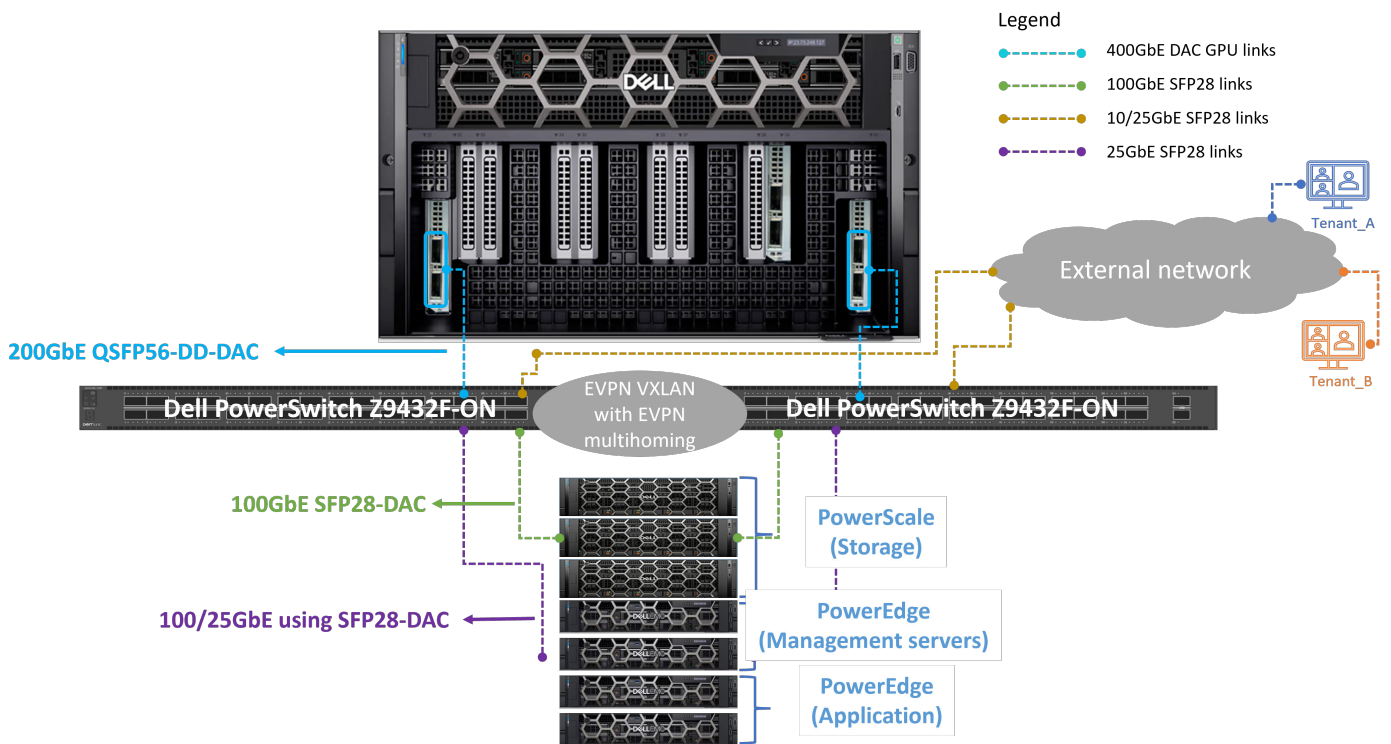


Figure 23. Front-end GPU cluster redundant fabric option 2

The connections from the Dell Z9432F-ON to storage, application, and GPU services are 100GbE and 25GbE. Figures 21 and 22 show a single XE9680 connection to the front-end fabric. These connections are repeated for all the XE9680s in the cluster.

The network features enabled on the Dell Z9432F-ON are:

- 400GbE into 2x200GbE breakout configuration
- 400GbE into 4x25GbE breakout configuration
- 400GbE into 4x100GbE breakout configuration
- EVPN multi-homing for workload redundancy
- EVPN VxLAN for multitenancy access into the GPU environment
- Jumbo frames (9000 - 9216 bytes) on the switchports connected to the PowerScale cluster
- Basic Layer 3 configuration to the external network

Infrastructure – Cabling and power

A 32/64 GenAI GPU cluster deployment requires three 48RU racks.

Racks 1 and 3 host 4 XE9680s each, while rack 2 hosts the switches, storage, and servers.

Dell Technologies GenAI 64 GPU cluster – physical and functional view

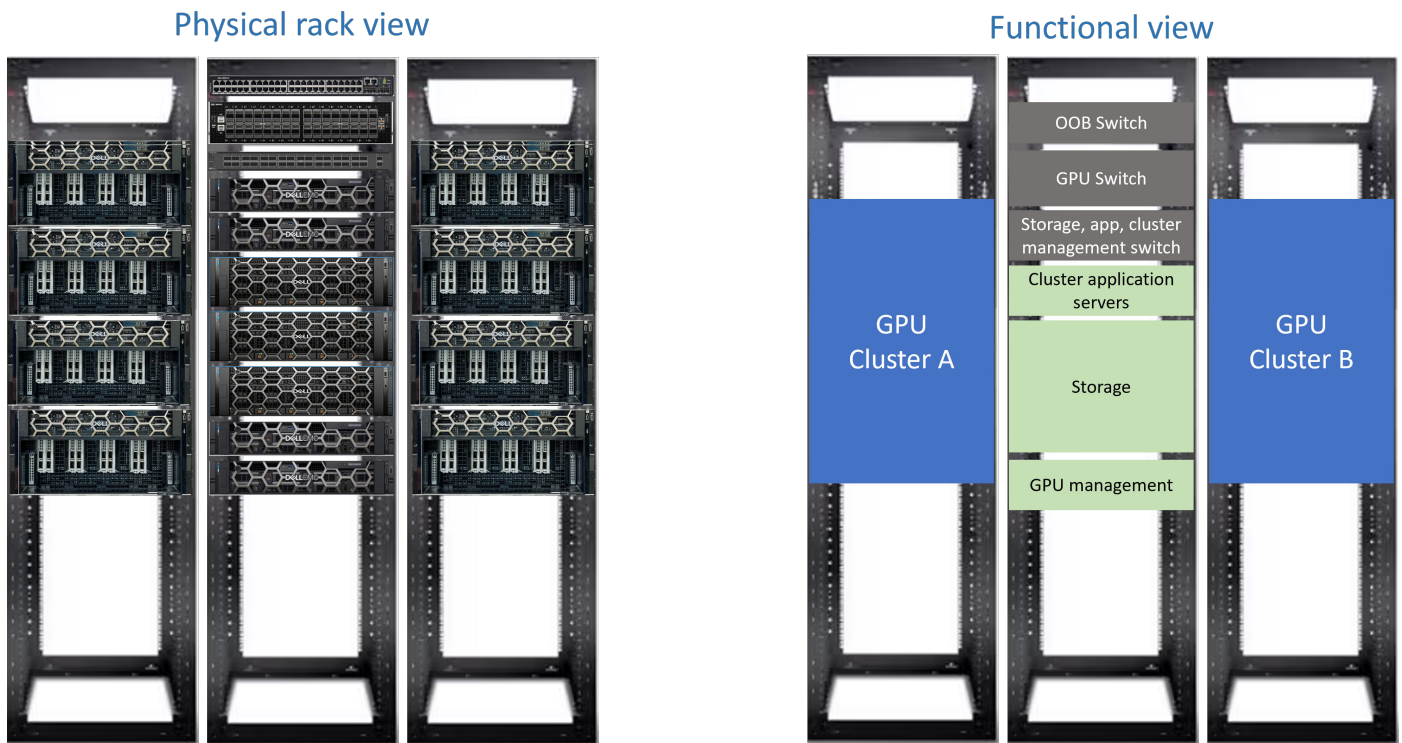


Figure 24. Dell Technologies GenAI 64 GPU cluster rack view

All connections converge on Rack 2 (the GPU cluster). The Dell cables (DACs and fiber) are long enough to span to both racks (Rack 1 and Rack 3).

This design guide assumes that each rack can accommodate 4 XE9680s. Every deployment will be different; however, there are some basic design guidelines that can be followed to ensure that the rack and power implementation is efficient.

- Whenever possible, infrastructure elements (routers and switches) should be flanked by the workloads.
- Racks hosting the infrastructure elements can be inserted in the middle or at the end of a row. Fiber can be used as an interconnection to address distance limitations that can be encountered if using copper DACs.

In terms of power planning:

- This design guide assumes 4 XE9680s per rack. A 32/64 GPU cluster solution requires up to 2 racks (8 total XE9680s, divided by 4 XE9680s per rack), plus another rack for the infrastructure devices.
- Redundant 30 amp circuits per rack other than XE9680 racks.
- Each XE9680 rack requires up to 16 Kilowatts, assuming all 6 power supplies are used. Each rack requires two circuits for redundancy.

Table 2. Small GenAI GPU cluster Bill of Material (BOM)

4/8 Node, 32/64 GPU cluster		
Item	Description	Qty
1	Compute nodes: Dell PowerEdge XE9680 6U chassis with 8 GPU 8 x 2.5 NVMe only DNP [G7N1GYC] [321-BKTN]	4/8
2	GPU: AMD Mi300X 8-GPU OAM DNP [GEUZ19M] [490-BJZD]	32/64
3	XE9680 network interface card: Broadcom 57608 Dual Port 200G Q112 PCIe Full Height DNP [250P7]	40/80
4	Ethernet high performance, switching fabric (back-end): Dell PowerSwitch Z9664F-ON, 64x400GbE QSFP56-DD	1
5	Ethernet high performance, switching fabric (front-end): Dell PowerSwitch Z9432F-ON, 32x400GbE QSFP56-DD	1
6	Ethernet front-end GPU fabric: Dell PowerSwitch Z9432F-ON, 32x400GbE QSFP56-DD	1 or 2, non-redundant or redundant
7	Ethernet OOB management network: Dell PowerSwitch N3248TE, 48x1GbE RJ45	1
8	Storage nodes: Dell PowerScale F900 - SKU 210-AXRS	3
9	Dell PowerScale F900 network interface card: F900 Network Interface Card [406-BBSL(CX6) QSFP28]	6
10	Management nodes: Dell PowerEdge R660	4
11	Rack infrastructure: Dell 4820W Server Rack Enclosure - 48U rack [A7096292]	3
12	Cables - Switch to GPU 400GbE connections:	32/64

Table 2. Small GenAI GPU cluster Bill of Material (BOM) (continued)

4/8 Node, 32/64 GPU cluster		
Item	Description	Qty
	Dell DAC- [400G Q56DD]-2x [200G Q56] - 4 meter	
13	Cables - Front-end storage connections: Dell DAC- [400G Q56DD]- 4x [100G ACC/AEC] - 5 meter	2
16	Cables - Front-end server 25GbE connections: Dell [DAC-Q28-4S28-25G-5M] - 5 meter	2
17	Cables - Ethernet: Dell RJ45 Ethernet 1GbE - 5 meter Cat5/6 RJ45 Ethernet	17

Medium or large GenAI cluster fabric design

The medium to large GenAI GPU cluster consists of:

- Dell Enterprise SONiC 4.2.1
- 8 or 64 Dell PowerSwitch Z9664F-ON as spine switches
- 8 or 64 Dell PowerSwitch Z9664F-ON as leaf switches
- 8 or 64 Dell PowerSwitch Z9432F-ON as leaf switches
- 32 or 256 Dell PowerEdge XE9680 with 8 GPUs per XE9680
- 10 Dell PowerSwitch N3248TE
- 2 Dell PowerSwitch Z, S5248F-ON, or S5448F-ON as border leaf switches
- 3-node PowerScale cluster
- 4 Dell PowerEdge R760 with dual 2x25GbE NIC ports

Dell Technologies medium to large GenAI 256/2048 GPU cluster - reference diagram

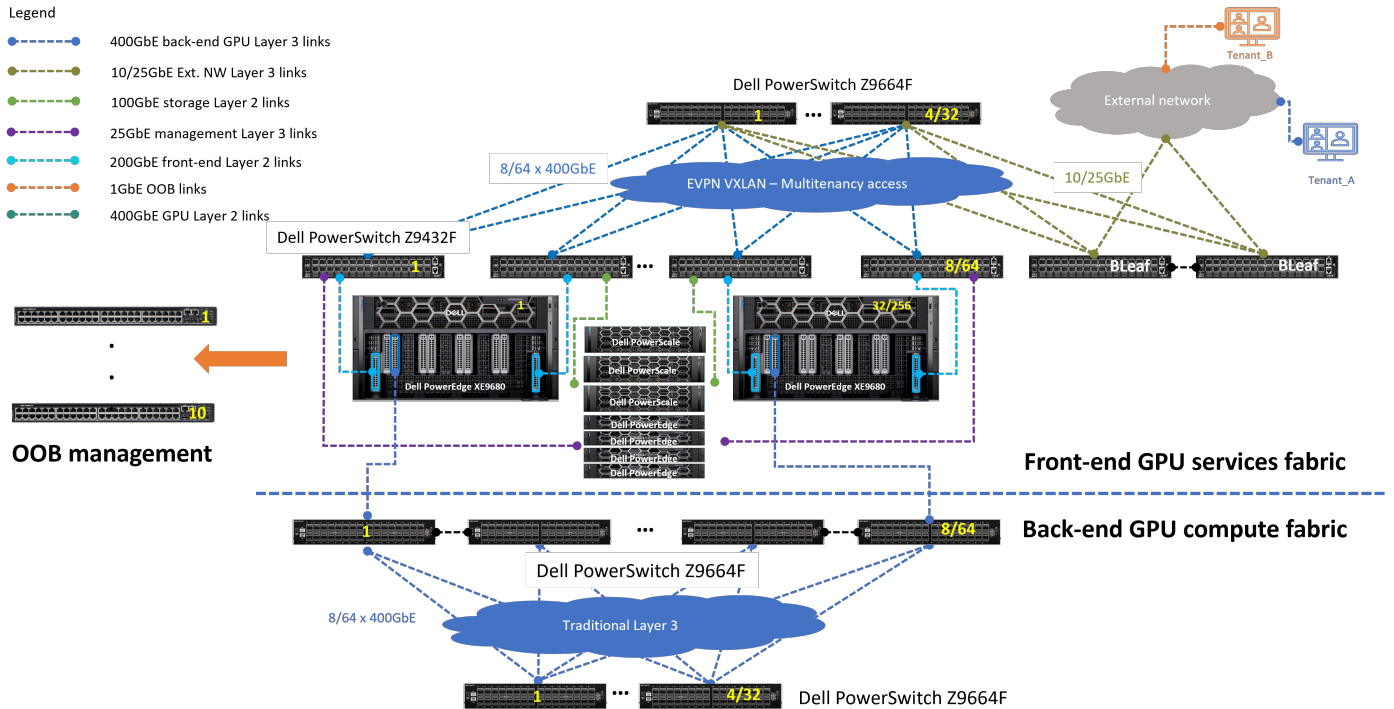


Figure 25. GenAI GPU cluster

The medium or large GenAI GPU cluster scales from 256 up to 2,048 GPUs.

OOB management fabric

Unlike the small GenAI GPU cluster, the OOB management fabric design follows a distinct set of guidelines considering the size of the cluster:

- The total number of devices for the large GPU cluster requires 457 OOB management connections.
- Assuming a class B subnet is used to assign OOB IP addresses, then a single subnet will be enough to assign all OOB IPs in the same subnet.
- If a class C subnet is used to assign all OOB management IP addresses, then two separate class C subnets will be required. Each subnet will be part of a unique VLAN ID. This tagged VLAN traffic will be carried on the 10GbE SFP uplinks of the N3248TE switches to the external network for routing purposes.
- Use all 48 1GbE ports on each N3248TE OOB switch whenever possible, and configure all 48 ports in the same VLAN ID.
- Trunk the OOB subnet using the external 10GbE SFP uplinks from each N3248TE onto the external network for routing.

A trunk interface is a Layer 2 interface that carries tagged Layer 2 traffic.

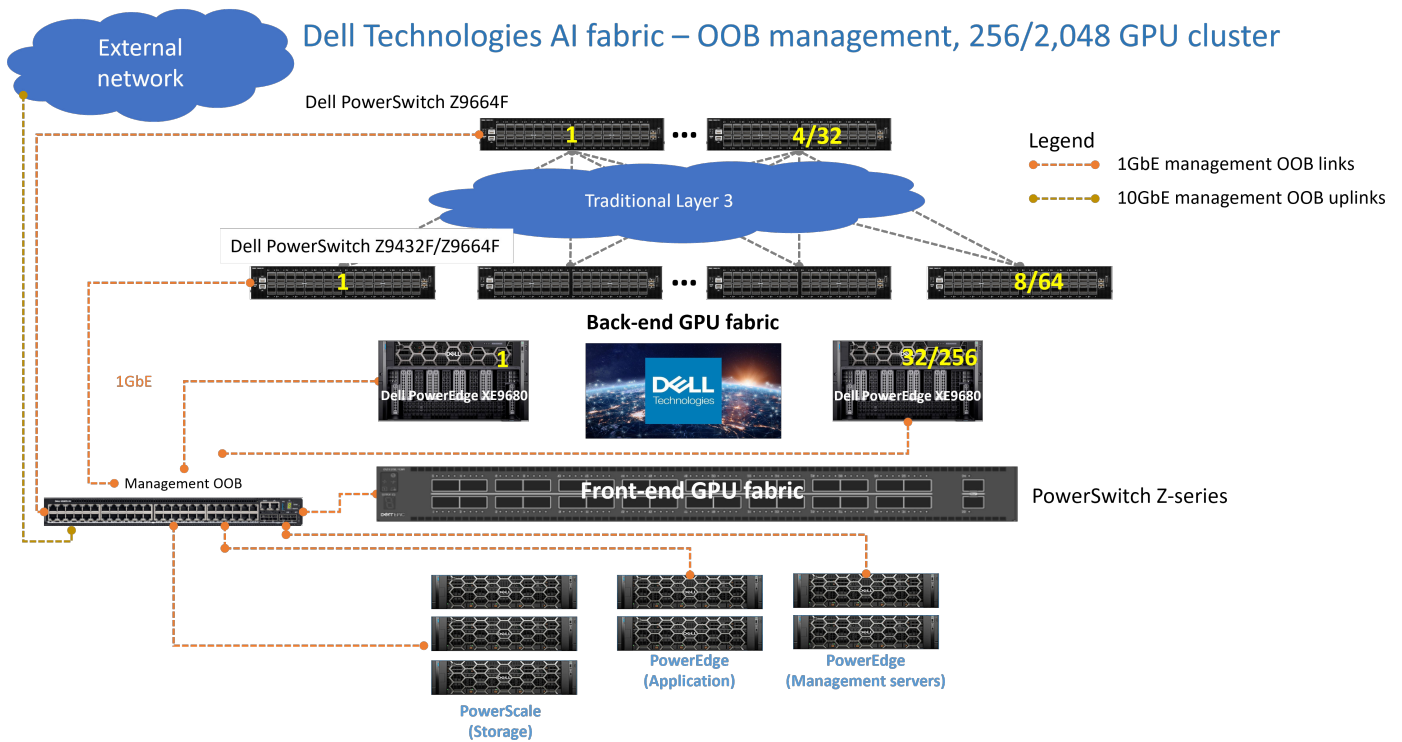


Figure 26. Medium-Large OOB management fabric

- The 10GbE SFP uplinks are configured as a LAG.
- If the OOB management design is divided into more than a single subnet, then the same approach as the small OOB management fabric can be followed (see OOB management fabric).

Back-end GPU fabric

The back-end GPU fabric provides inter-GPU communication. In this design, the Dell PowerSwitch Z9664F-ON provides 400GbE non-blocking to all the GPUs in the cluster.

When designing the back-end GPU fabric, follow these networking design guidelines:

- Assuming the fabric is Layer 2, base the back-end GPU fabric on a flat Layer 2 or Layer 3 implementation. For a 256 XE9680 cluster, the number of GPUs is 2,048. In a pure Layer 2 fabric, all the GPU connections can be assigned to the same VLAN ID, which creates a full mesh or common broadcast domain across all the GPUs.
- If the back-end fabric is based on traditional Layer 3 with BGP, the VLANs terminate on the leaf Z9664F-ON. The VLAN ID is assigned an IP address commonly known as an L3 VLAN.

The L3 VLANs act as L3 BGP interfaces, establishing a full BGP network connection in the fabric with each other.

NOTE: If Layer 3 is deployed, BGP unnumbered is enabled on the fabric.

Dell Technologies Gen AI 32/64 nodes, 256/2,048 back-end GPU fabric cluster

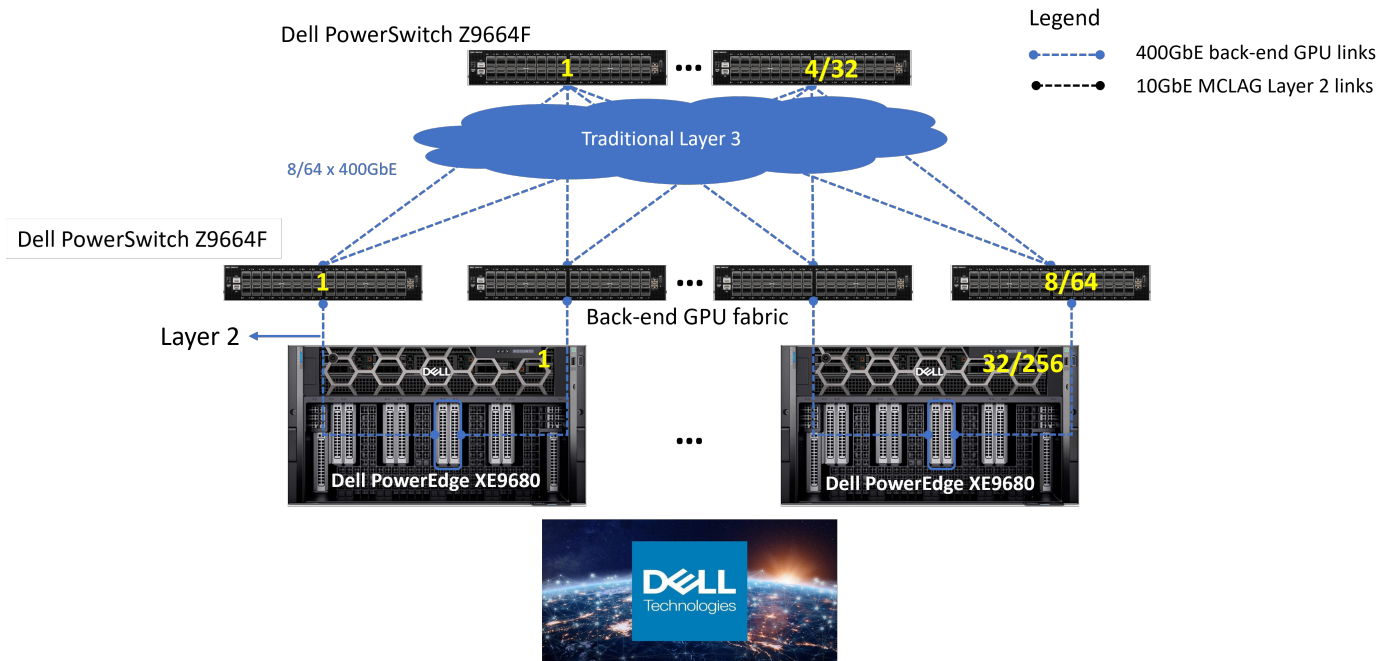


Figure 27. Multi-rack GenAI GPU fabric

Figure 28 shows a diagram of a leaf and spine physical GPU connection to the Dell PowerSwitch Z9664F-ON using the Dell 400GbE to 1x400GbE DAC breakout cables or single mode fiber (SMF) with their respective optic (EDR4).

It is important to consider the physical and functional aspects of the GPU DAC connection.

Figure 20 and Figure 21 show a physical single DAC broken into two DAC connections onto their respective NIC ports on the Broadcom Thor 2. However, functionally this DAC connection is a single 1x400GbE connection.

Even though the Broadcom Thor 2 NIC shows two physical NIC ports, it operates and behaves as a single 400GbE NIC by default.

NOTE: The single 400GbE connection on the Thor 2 NIC can be configured as a 2x200GbE using the same physical DAC cable if the Z9664F-ON switchport is configured as a 2x200GbE as well. For details, see Figure 21.

Front-end storage and application fabric

The front-end fabric design for the medium or large GPU cluster follows the same guidelines as the small GPU cluster.

The basic difference between these two models is the size of the fabric. The small GPU cluster uses a single switch for the front-end fabric, whereas the medium or large deployment uses a leaf and spine architecture.

Figure 27 shows the front-end design with a pair of border leaf switches. The border leaf switches provide external multi-tenancy access into the GenAI environment.

Dell Technologies GenAI 256/2,048 GPUs fabric – front-end storage fabric

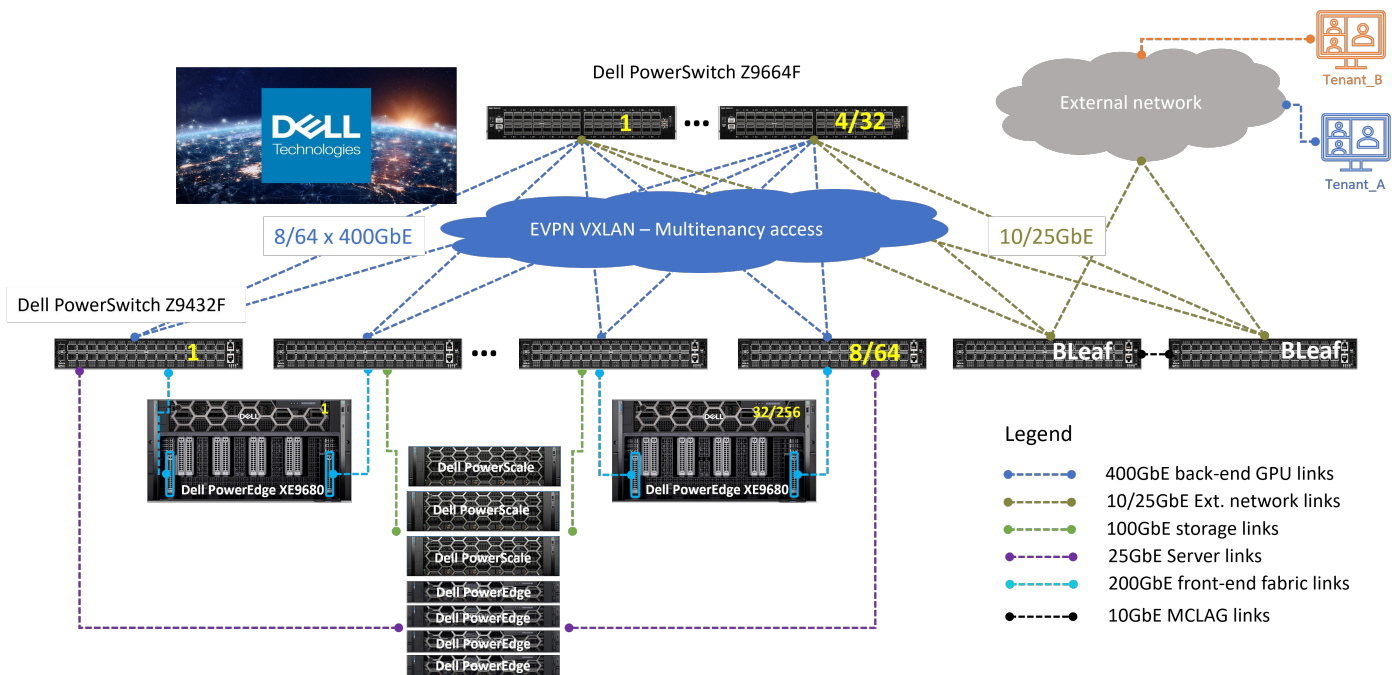


Figure 28. Multi-rack front-end GPU fabric

Follow these design guidelines:

- Enable jumbo (9000 - 9216 bytes) frames in the entire fabric.
- Enable EVPN VXLAN on the fabric.
- Enable EVPN multi-homing on the leaf switches to provide workload redundancy. EVPN multi-homing is similar to MC-LAG without the need to implement MC-LAG chassis.
- Assign different VLANs to storage, application, and in-band cluster management services.
- Integrate all VLANs into the EVPN VXLAN environment, and make them reachable from the outside world.
- EVPN VXLAN should support multi-tenancy connections from the outside world to manage the GPU cluster by the in-band service.

NOTE: LAGs are not implemented in the front-end fabric. This is because there is no MC-LAG configuration between the leaf switches.

Infrastructure - Cabling and power

The cabling and power infrastructure design for the medium and large GPU cluster has many options.

The following assumptions are made to create the rack elevation design of the medium to large GPU cluster:

- Whenever possible, infrastructure elements (routers, switches) should be flanked by the workloads.
- To address the distance limitations encountered if using DACs (Copper), racks hosting the infrastructure elements can be inserted at the middle or end of a row using fiber as interconnects.
- The border leaf switches part of the front-end fabric can be placed at the middle or end of a row, and they can be connected to the leaf and spine front-end fabric by single-mode fiber connections.

The approach described in this design is not meant to be exhaustive; it describes basic guidelines following cabling length, type, and power allocation per rack.

In terms of power planning:

- This design guide assumes 4 - XE9680s per rack. The large GPU cluster solution requires up to 64 racks (256 XE9680s, divided by 4 XE9680s per rack).
- The racks require redundant 30 amp circuits per rack, other than the XE9680 racks.
- Each XE9680 rack requires up to 16 Kilowatt assuming all 6 power supplies are used. Each rack requires two circuits for redundancy. The circuits for each XE9680 rack require proper planning as each rack would require 68 amps.

Dell Technologies GenAI 32/256 node – 256/2,048 GPUs cluster design

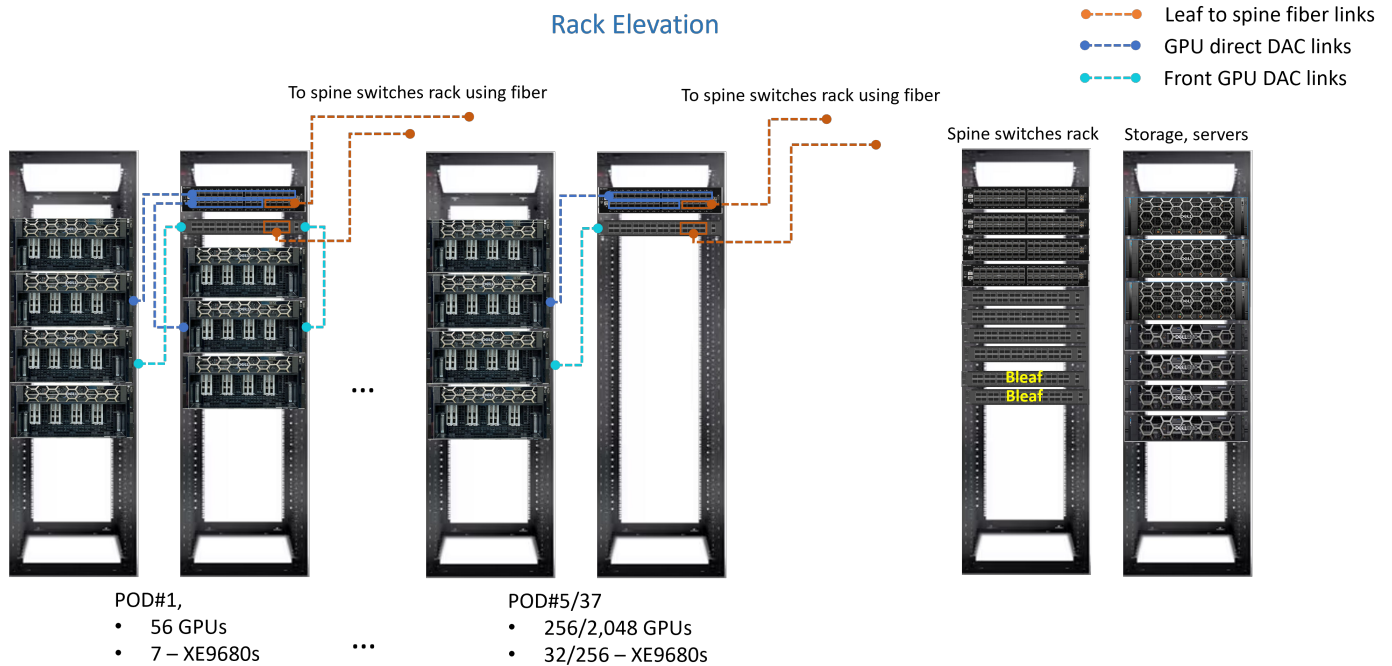


Figure 29. 256/2,048 GPU cluster rack view

Figure 28 shows a pod design of the medium to large GPU cluster. Each pod consists of:

- Two racks that are 52U tall.
- Rack 1 hosts 4 XE9680s, which equals 32 GPUs (8 GPUs per XE9680).
- Rack 2 hosts the leaf Z9664F switches that are part of the back-end and front-end fabrics.
- The racks are connected using two different connections (400GbE and 200GbE). They are connected with the 400GbE DAC cable, which spans 4 meters.
- From Rack 2, the leaf switches are connected to the spine rack using single mode fiber connections. These are 400GbE connections.
- Pod#5 and Pod#37 are the last GPU clusters for the medium and large design. The last two racks are used for spine, border leaf switches, storage, application, and management servers.

Table 3. Medium 256 GPU cluster bill of materials

32 node, 256 GPU cluster		
Item	Description	Qty
1	Compute nodes - Dell PowerEdge XE9680 chassis: XE9680 6U chassis with 8 GPU 8 x 2.5 NVMe only DNP [G7N1GYC] [321-BKTN]	32
2	GPU: AMD Mi300X 8-GPU OAM DNP [GEUZ19M] [490-BJZD]	256
3	XE9680 network interface card: Broadcom 57608 Dual Port 200G Q112 PCIe Full Height DNP [250P7]	320
4	Ethernet high performance, switching fabric (back-end):	16

Table 3. Medium 256 GPU cluster bill of materials (continued)

32 node, 256 GPU cluster		
Item	Description	Qty
	Dell PowerSwitch Z9664F-ON, 64x400GbE QSFP56-DD	
5	Ethernet high performance, switching fabric (front-end): Dell PowerSwitch Z9432F-ON, 32x400GbE QSFP56-DD	8
6	Ethernet OOB management network: Dell PowerSwitch N3248TE, 48x1GbE RJ45	2
7	Storage nodes: Dell PowerScale F900 - SKU 210-AXRS	3
8	Network interface card: Dell PowerScale F900 network interface card [406-BBSL(CX6) QSFP28]	6
9	Management nodes: Dell PowerEdge R660	4
10	Rack infrastructure: Dell 4820W server rack enclosure - 48U rack [A7096292]	12
11	Cables - Switch to GPU and front-end 400GbE Dell DAC- [400G Q56DD]-2x [200G Q56] - 4 meter	320
12	Dell Switch 400GbE optic [400G-Q56DD-EDR4]	64
13	Fiber cables: Dell single mode fiber MPO-12APC	64
14	Cables - Front-end storage 100GbE connections: Dell [DAC-Q28-4S28] - 4 meter	2
15	Cables - Front-end server 25GbE connections: Dell [DAC-Q28-4S28-25G-5M] - 5 meter	4
16	Dell RJ45 Ethernet 1GbE - 5-10 meter Cat5/6 RJ45 Ethernet	58

Table 4. Large 2,048 GPU cluster bill of materials

256 Node, 2,048 GPU cluster		
Item	Description	Qty
1	Compute nodes: Dell PowerEdge XE9680 6U chassis with 8 GPU 8 x 2.5 NVMe only DNP [G7N1GYC] [321-BKTN]	256
2	GPU: AMD Mi300X 8-GPU OAM DNP [GEUZ19M] [490-BJZD]	2,048
3	XE9680 network interface card Broadcom 57608 Dual Port 200G Q112 PCIe Full Height DNP [250P7]	2,560
4	Ethernet high performance, switching fabric (back-end): Dell PowerSwitch Z9664F-ON, 64x400GbE QSFP56-DD	128
5	Ethernet high performance, switching fabric (front-end): Dell PowerSwitch Z9432F-ON, 32x400GbE QSFP56-DD	64
6	Ethernet OOB management network: Dell PowerSwitch N3248TE, 48x1GbE RJ45	10
7	Storage nodes: Dell PowerScale F900 - SKU 210-AXRS	6
8	Network interface card Dell PowerScale F900 network interface card [406-BBSL(CX6) QSFP28]	6
9	Management nodes: Dell PowerEdge R660	4
10	Rack infrastructure: Dell 4820W server rack enclosure - 48U rack [A7096292]	76
11	Cables - Switch to GPU and front-end 400GbE Dell DAC- [400G Q56DD]-2x [200G Q56] - 4 meter	2,560
12	Dell Switch 400GbE optic [400G-Q56DD-EDR4]	512

Table 4. Large 2,048 GPU cluster bill of materials (continued)

256 Node, 2,048 GPU cluster		
Item	Description	Qty
13	Dell single mode fiber MPO-12APC	512
14	Cables - Front-end storage 100GbE connections Dell [DAC-Q28-4S28] - 4 meter	2
15	Cables - Front-end server 25GbE connections Dell [DAC-Q28-4S28-25G-5M] - 5 meter	4
16	Cables Dell RJ45 Ethernet 1GbE - 5-10 meter	457

The design options described in this guide are a reference for AI/GenAI workloads that can be deployed using different network architectures. The design options in this guide follow networking best practices or guidelines using current networking software and hardware features within Dell Technologies, to deliver an interoperable, high-performance, scalable, and efficient fabric.

These designs do not aim to address all intricacies that can potentially affect the overall design. Certain proprietary GPU-to-GPU communication architectures or GPU-to-NIC leveraging IPv6/4 offloading may require additional fabric fine-tuning, and this guide does not address these scenarios.

Dell GenAI Environment — Orchestration and environment

Introduction

The Dell GenAI environment consists of two major components:

- Infrastructure or fabric
- Workloads

Infrastructure orchestration and monitoring

Dell Technologies offers infrastructure automation and monitoring capabilities through Ansible and northbound interface integration with Dell Enterprise SONiC using OpenConfig.

The Ansible integration leverages modules and collections with predefined networking features arranged in an Ansible playbook that is deployed at scale from a single Ansible controller.

While Ansible leverages modules and collections, OpenConfig uses gRPC, an open-source Remote Procedure Call (RPC) originally developed by Google to configure and obtain information from the network infrastructure.

In addition to the embedded orchestration and monitoring capabilities with Dell Enterprise SONiC, two key partnerships with BeyondEdge and Augtera further complement this offer.

BeyondEdge Verity

BeyondEdge Verity is an intent-based network infrastructure orchestrator integrated with Dell Enterprise SONiC. BeyondEdge Verity uses simple text-based network configurations created by Dell Fabric Design Center (FDC).

Through Zero-Touch Provisioning (ZTP), BeyondEdge Verity discovers the presence of Dell PowerSwitches and deploys at scale the network configurations obtained through FDC.

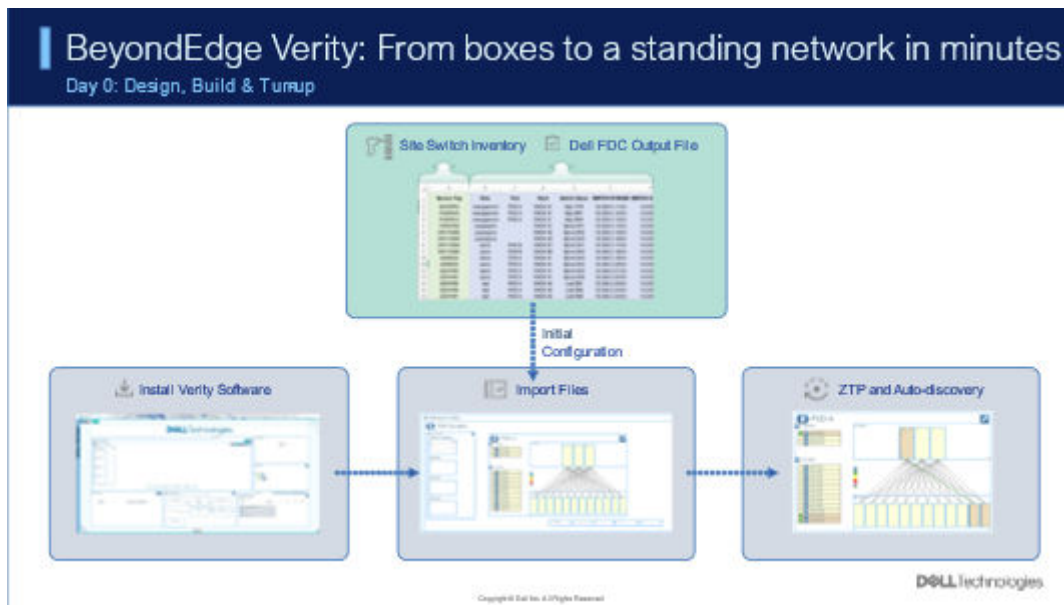


Figure 30. BeyondEdge Verity - Day 0 process

Figure 29 shows the basic process that BeyondEdge Verity performs during the automated orchestration of the fabric.

Besides BeyondEdge Verity, the Dell GenAI infrastructure orchestration can also be deployed by Ansible. Dell Enterprise SONiC supports the latest Ansible modules and collections used to create the relevant YAML playbooks needed to automate the infrastructure.

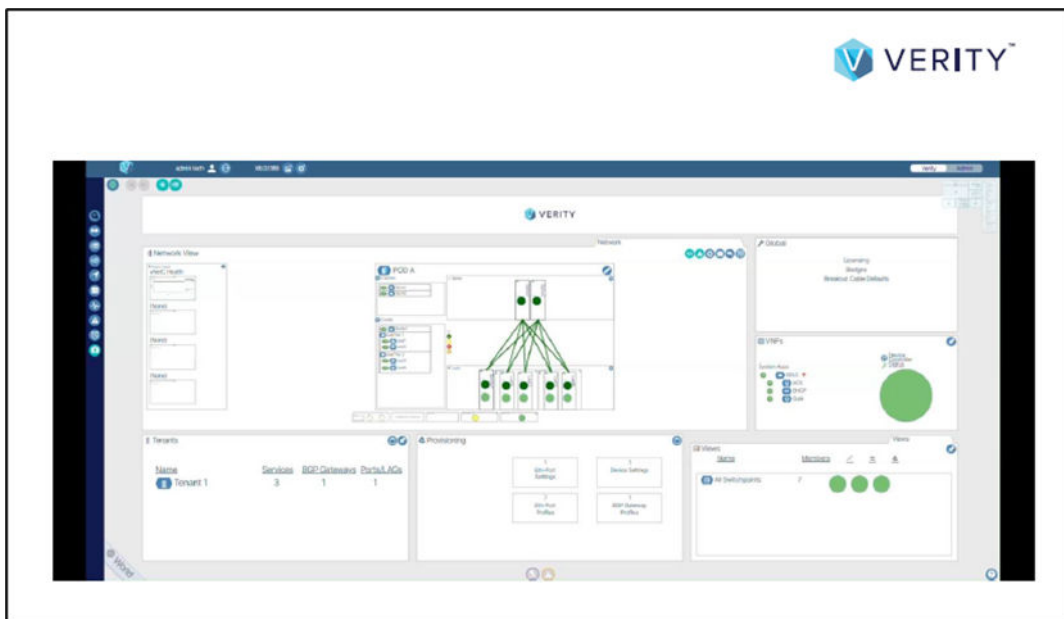


Figure 31. BeyondEdge Verity infrastructure zoom

Figure 30 shows the zooming capabilities of BeyondEdge Verity, which can amplify configurations and events in the fabric

Augtera

Provisioning the fabric is not enough; visibility into the solution is key. With Augtera, Dell Enterprise SONiC integrates monitoring and deep insight into the GenAI solution.

The integration uses open-standards traffic monitoring technology such as sFlow, everFlow, SNMP tracking, and SONiC gRPC, gNMI Telemetry.

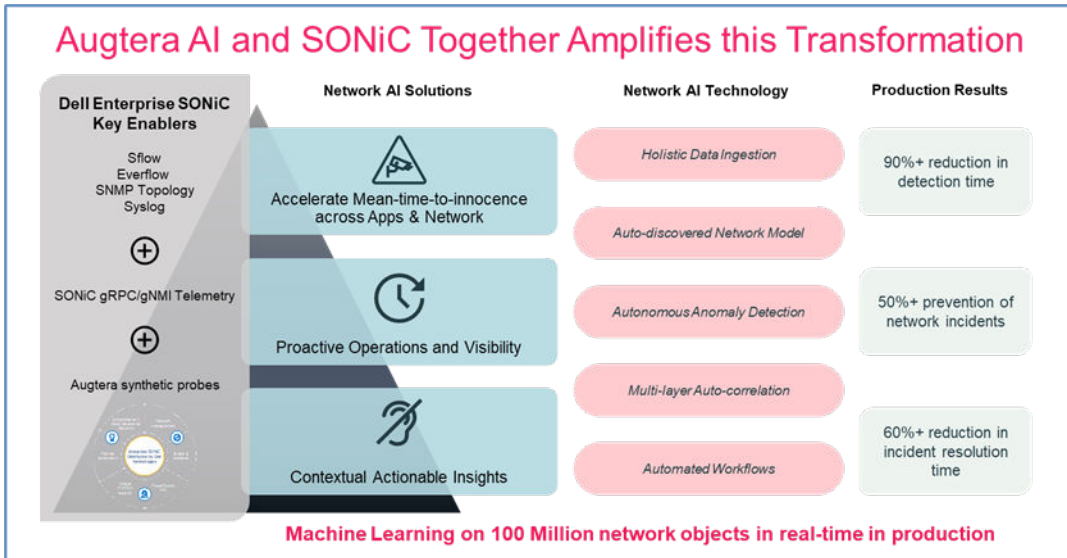


Figure 32. Augtera and Dell Enterprise SONiC integration

Augtera's dashboard can be configured to obtain different views of the environment.

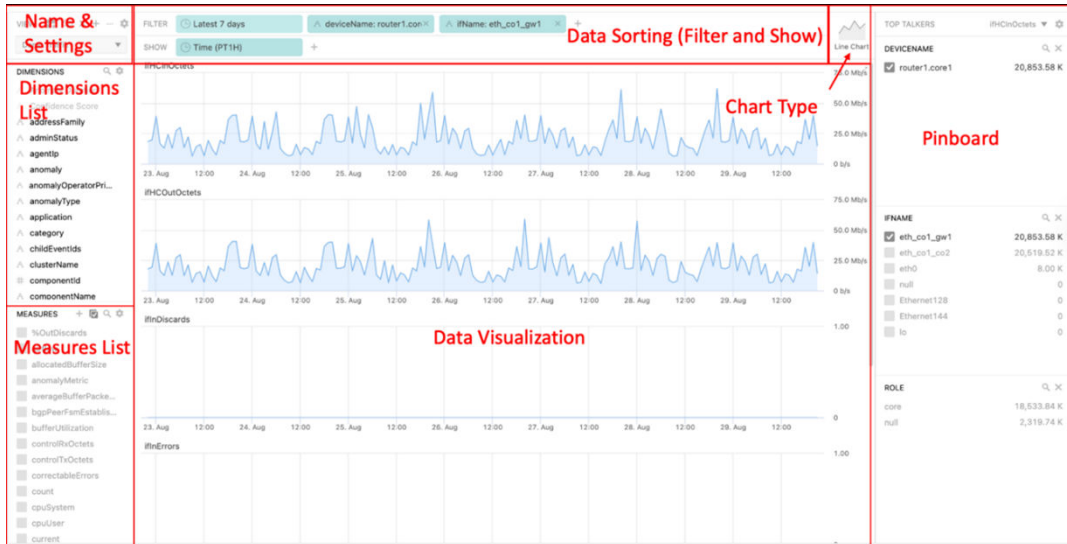


Figure 33. Augtera dashboard configuration



Figure 34. Augtera AI workload deep insight

Figure 33 shows actual AI workload performance obtained with the Augtera Network AI platform.

GenAI is a complex undertaking, and Dell Technologies is ready to become the technology partner of choice for GenAI deployments.

Its product portfolio and partnerships deliver the right solutions based on open standards, scalability, and high-performance benefits.

AI workload orchestration and monitoring

The challenges of orchestrating and monitoring the resources of a dynamic environment such as an AI cluster--where CPU, memory, storage, GPU, and other compute resources spread across hundreds if not thousands of nodes--create an urgent need towards a more efficient workload orchestration and monitoring approach.

The OMNIA project is an open-source initiative to make deploying consolidated workloads easy and painless. The project uses open-source and free-use software started by Dell Technologies, HPC, and AI Innovation Lab.

The OMNIA software stack deploys two types of workload management software: Slurm and/or Kubernetes.

Both Slurm and Kubernetes allow deployment at any scale, bringing all the compute resources into a single entity.

Figure 34 and Figure 35 describe the respective workload management stacks.

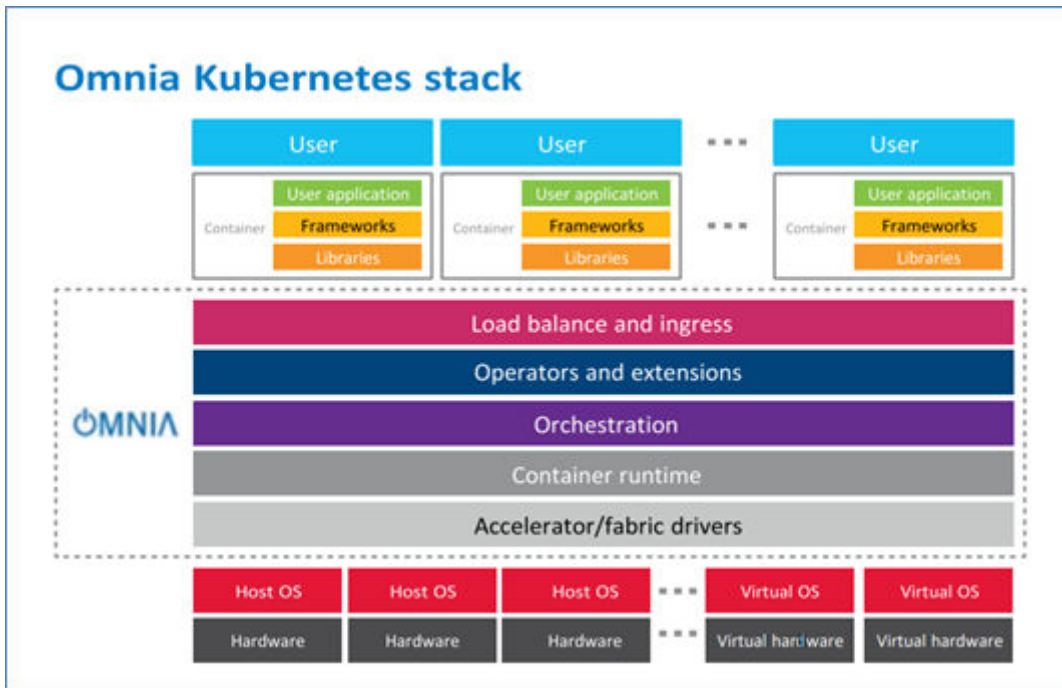


Figure 35. OMNIA Kubernetes stack

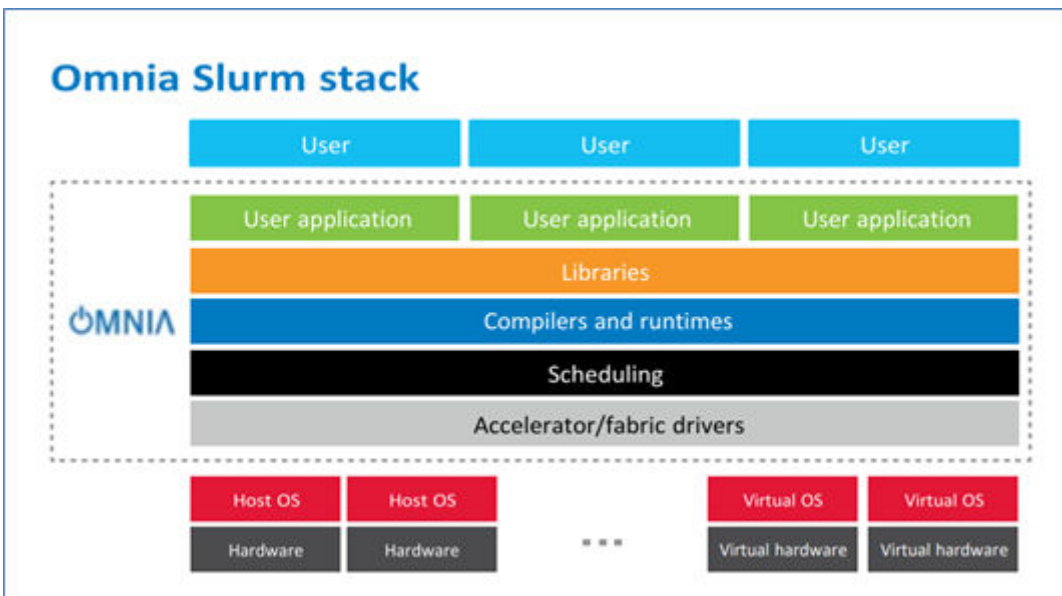


Figure 36. OMNIA Slurm stack

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 PowerSwitch](#)
- [Dell PowerEdge XE9680](#)
- [Dell PowerScale](#)
- [Dell Enterprise SONiC](#)
- [Dell Technologies OMNIA Overview](#)

External documentation

The following documentation provides information about partner technologies.

- [BeyondEdge](#)
- [Augtera](#)

Feedback and technical support

We encourage readers to provide feedback on the quality and usefulness of this publication by sending an email to Dell_Networking_Solutions.

For technical support, visit [the Dell Support site](#).