


# Bring AI to your data

Dell Validated Designs for Generative AI  
April 2024



Filip Đurkin, Technical Sales  
Dell Technologies

# Generative AI is changing the game for enterprises

54%

of organizations are either already using GenAI or planning to in the next 12 months<sup>1</sup>

78%

say training on enterprise data is important<sup>2</sup>

60%

of the design effort of apps and websites will be from generative design AI, by 2026<sup>3</sup>



*“...the killer app for artificial intelligence is productivity gains...the real winners are those companies that have proprietary data, and lots of it, and the best pools of high-quality data...”<sup>4</sup>*

1-54% ESG Report - Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns, August 2023 (N= 790)

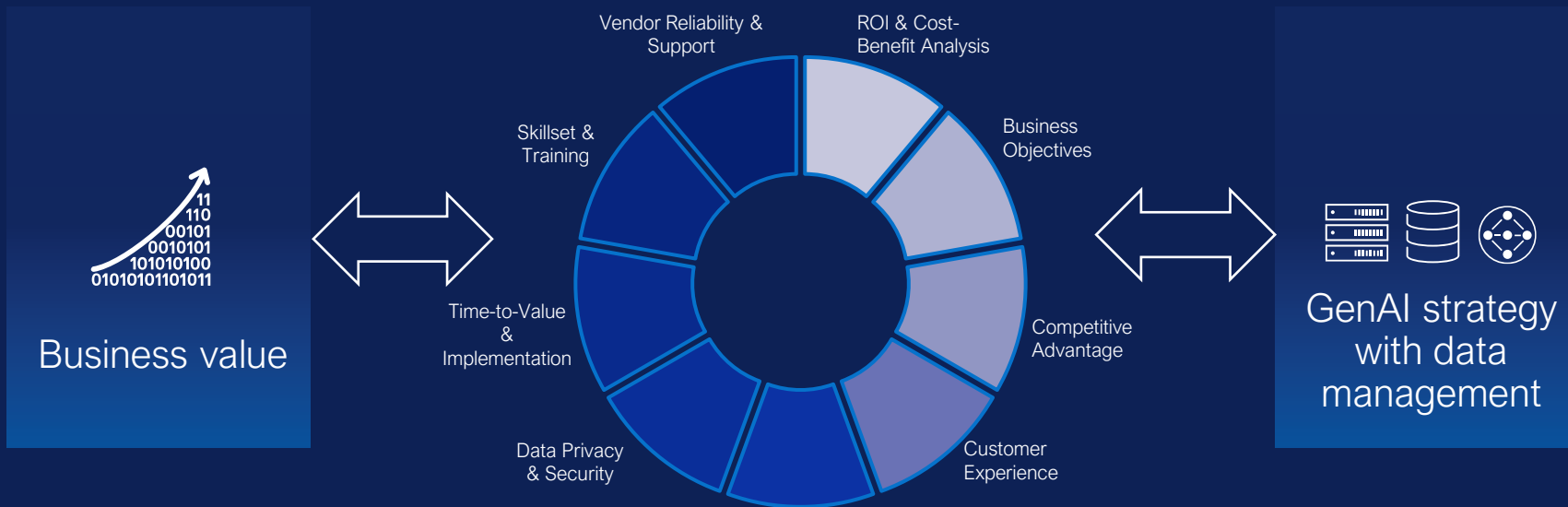
2-78% A New Beginning: Generative AI in the Enterprise – TECHanalysis Research Survey Report, May 2023 (N=1000)

3-60%: Fireflies.AI blog, The Generative AI Landscape: Where We Stand and Where We're Headed, 19 January 2023 –URL

4-4-ARK , BloombergTV interview, February 2023, <https://www.bloomberg.com/news/videos/2023-02-02/cathie-wood-on-deflation-risk-tech-stocks-and-bitcoin>

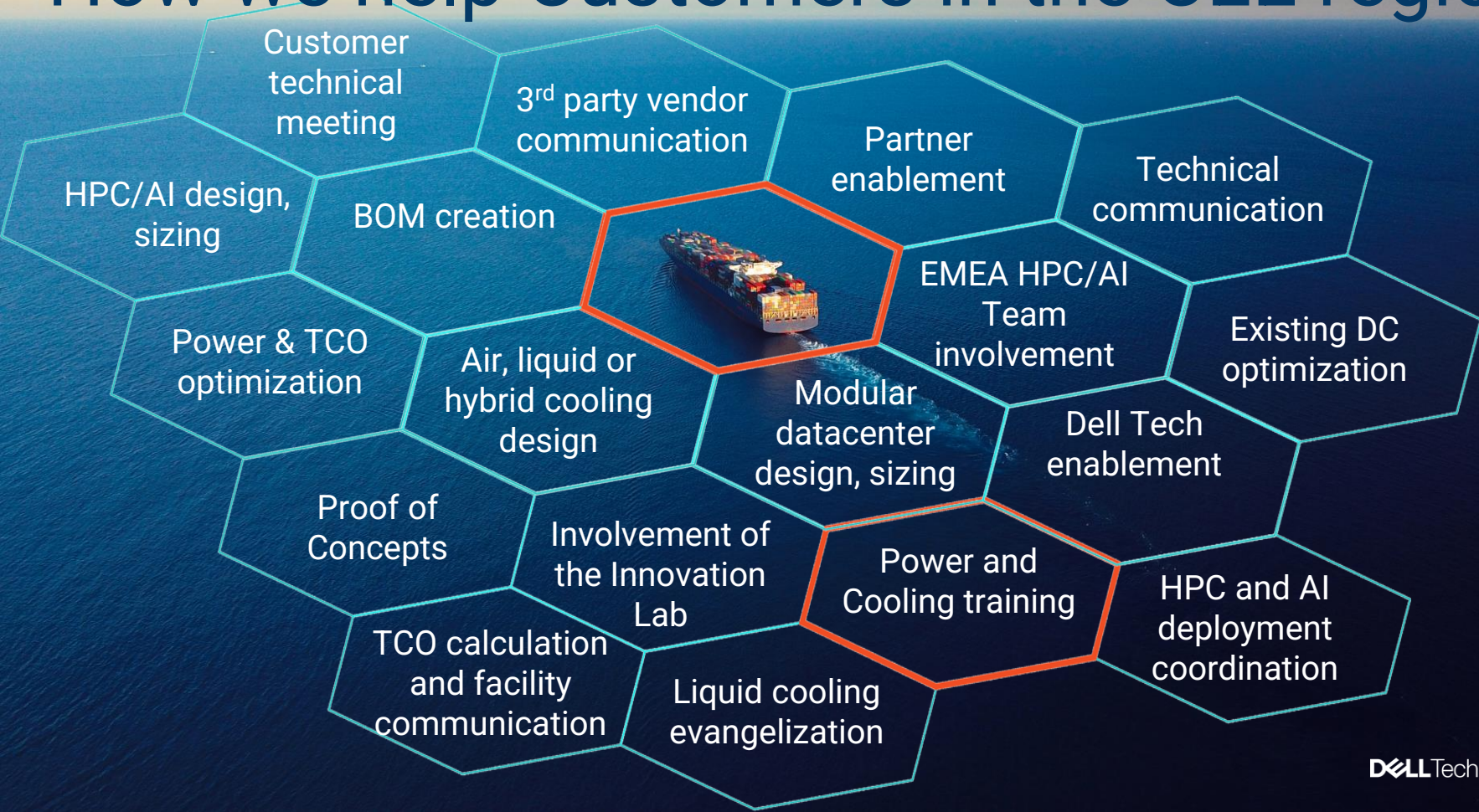
# Unlock the value of data with Generative AI

Aligning the business with your AI initiatives starts with questions



- What is the projected return on investment (ROI) for this AI solution?
- Is the AI solution scalable to handle our growing data and business demands?
- How long will it take to implement and deploy the AI solution?
- How will this AI solution enhance the customer experience?
- How will this AI solution give us a competitive edge in the market?

# How we help Customers in the CEE region?



# Generative AI use cases are maximizing value today



# Dell Validated Design for Generative AI

Deploy Generative AI with Dell Technologies and NVIDIA expertise across the organization

## Validated designs

### Best-in-class infrastructure



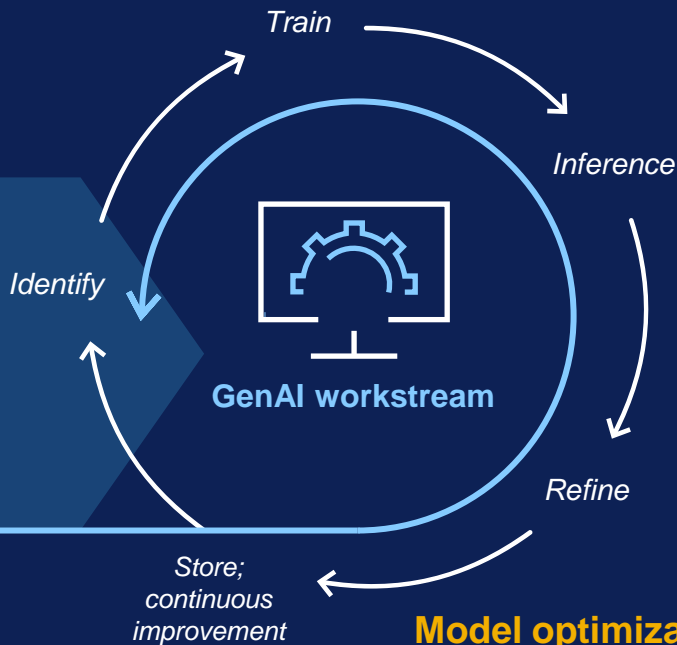
High performance platforms



Integrated GenAI stack



Expert advisors



## Enterprise-wide

→ **Trusted results.**  
Customized models with factual correctness, based on proprietary business data.

→ **Actionable decisions.**  
Democratize business-wide for step-improvements to drive faster transformation.

# Dell Validated Design for GenAI with NVIDIA

Delivering a better outcome by customizing the model

## GenAI Frameworks



NVIDIA NeMo Framework



Community Source Models

## AIOps & MLOps Platforms

NVIDIA Triton  
Inferencing Server

NVIDIA  
Merlin Framework

cnvrg.io

## Infrastructure Management

NVIDIA Base Command  
Manager Essentials

key solu

Kubernetes

o-value

Enterprise Linux

## Infrastructure



PowerEdge



PowerScale



ECS/ObjectScale



PowerSwitch



NVIDIA GPUs

Dell Professional Services



Tested and proven configurations  
reducing deployment times and risk



Scalable design approach with a choice  
of consumption models



Powerful acceleration-optimized compute  
with future-ready bandwidth



Support massive amounts of unstructured  
data with flexible networking speeds



Deploy clusters with 1000s of nodes  
using industry-leading latency and  
data throughput

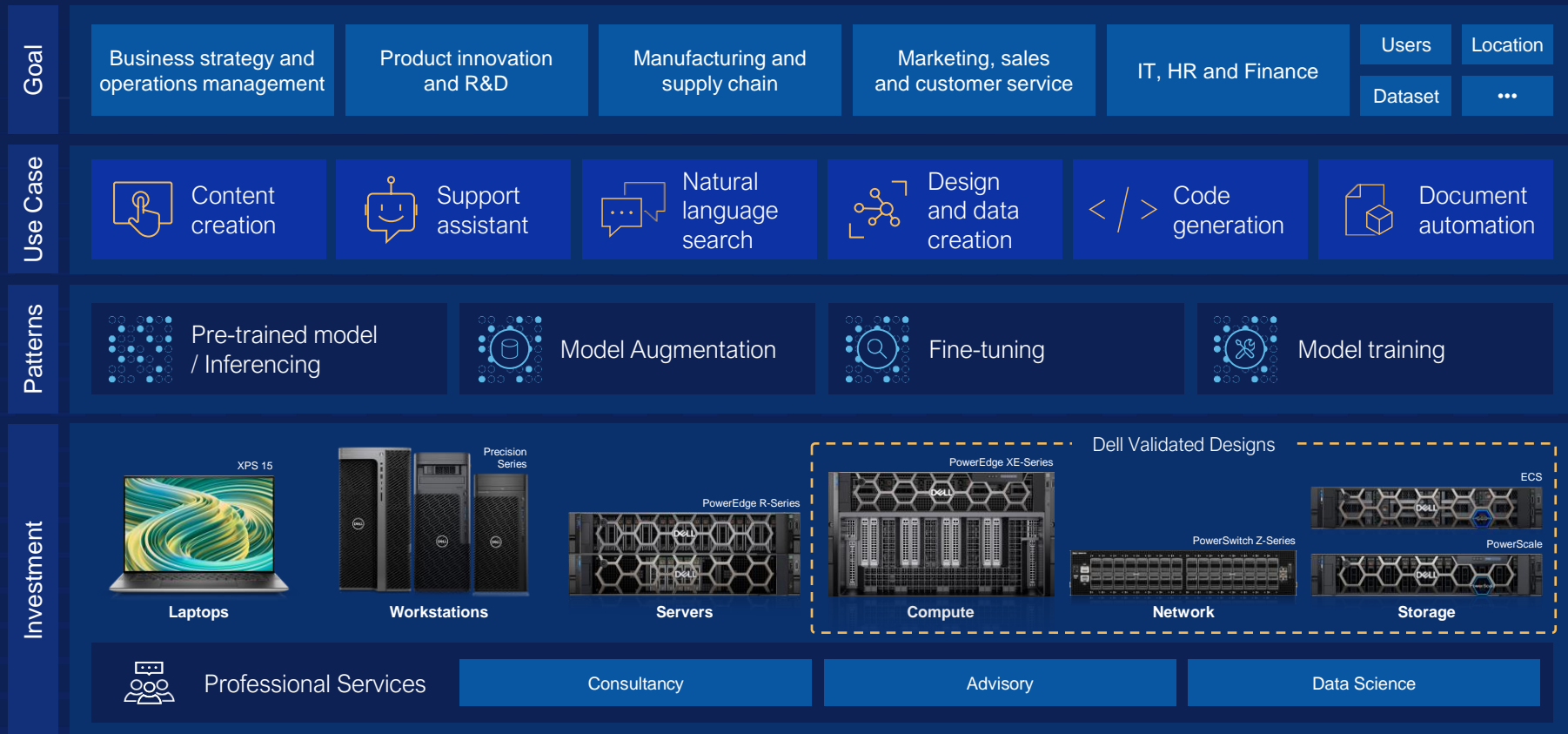
Dell Technologies

Which hardware is right for you?







# Right-sizing your AI investment



# Use case example | Departmental document automation

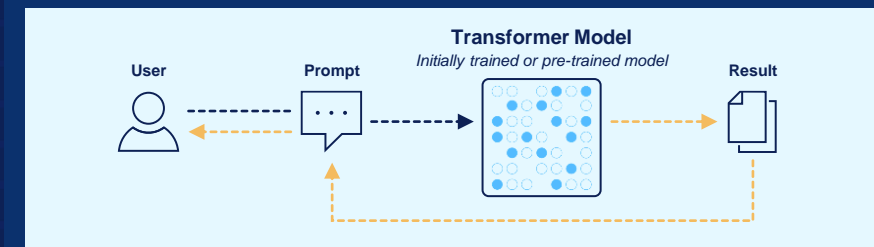
Use case	GenAI purpose	Users	Benefits
 <p>Document automation</p>	<ul style="list-style-type: none"> <li>Analyzes large amounts of information to find or articulate a result (interpretation and synthesis)</li> <li>Translates information for new audience</li> <li>Includes additional context</li> </ul>	<ul style="list-style-type: none"> <li>Sales / marketing</li> <li>Internal users</li> <li># of users</li> <li>Etc.</li> </ul>	<ul style="list-style-type: none"> <li>Faster education in a new topic or domain</li> <li>Can be used to refine data for new audiences</li> <li>Helps standardize, format, classify, review and/or organize documents</li> <li>Enhances productivity</li> </ul>

## Why inferencing?



Pre-trained model / Inferencing

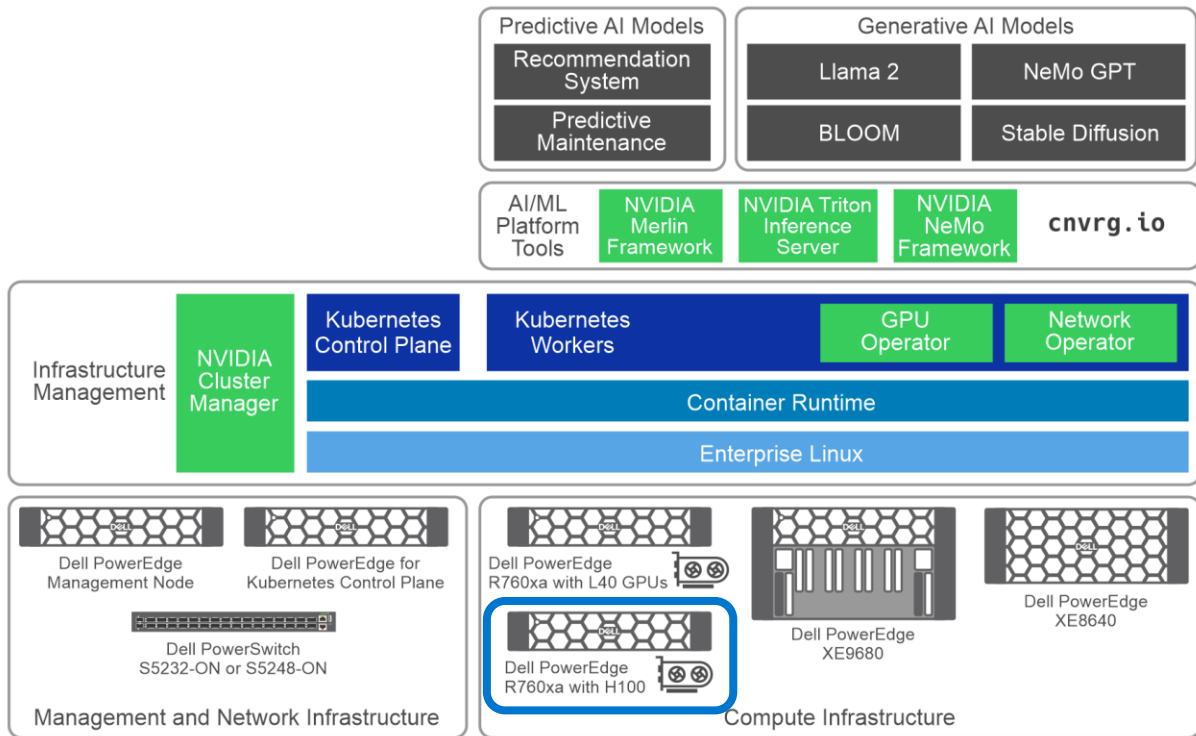
- AI model doesn't need context
- Existing LLMs operate at passible levels
- No data science required
- Modest hardware footprint required



## Infrastructure investment

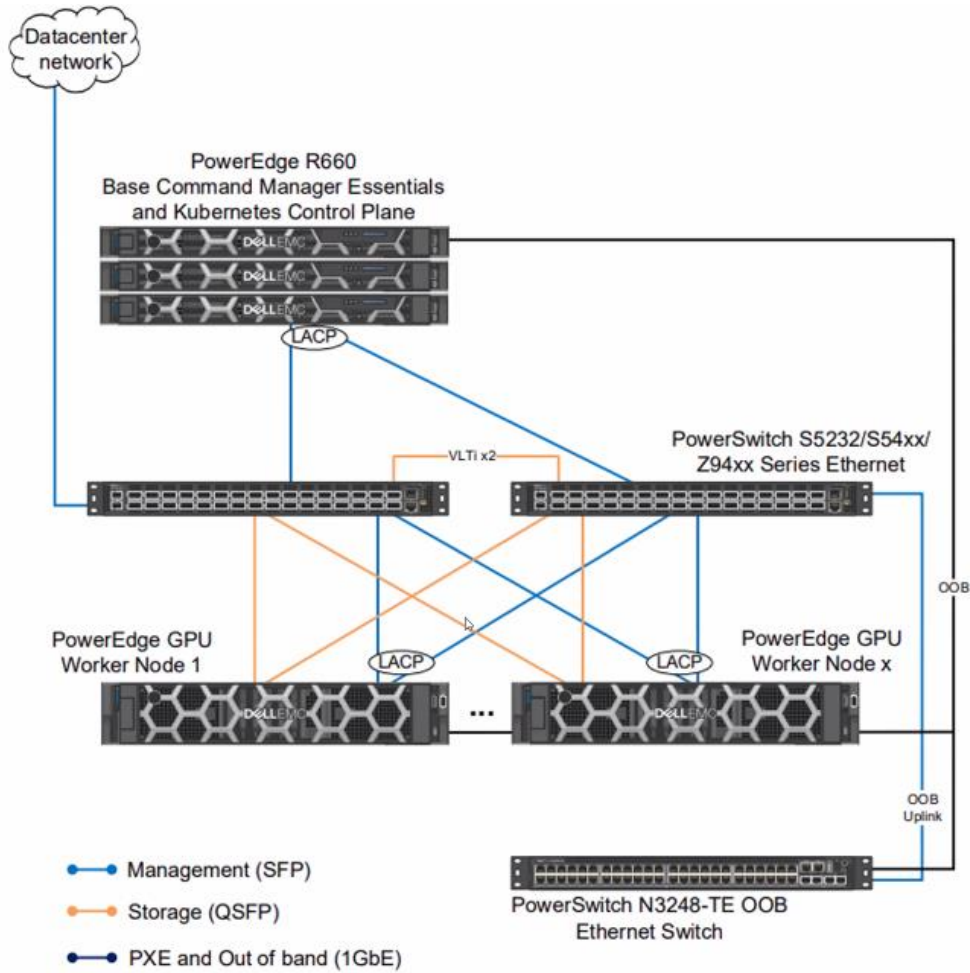
The image displays three categories of Dell hardware: **Laptops** (XPS 15), **Workstations** (Precision Series), and **Servers** (PowerEdge R-Series). The categories are separated by 'OR' indicators, suggesting a choice between them. Below the hardware, there is a 'Want help?' section with three options: **Deployment** (gear icon), **Advisory** (person icon), and **Data Science** (flask icon).

# Design Guide for Generative AI – Inference Use Case





- AI Models
  - Opensource and commercial models
  - Selection of outcomes
- AI Operations
  - Model Lifecycle and Management (MLOps)
  - AI Frameworks and Libraries
- Infrastructure
  - Software
    - Kubernetes Deployment and management
    - Accelerator Operators
    - Operating Systems
  - Hardware
    - AI Optimized Dell PowerEdge Servers
    - Dell PowerSwitch Networking
    - NVIDIA Accelerators

Validated



# Use case example | Design and data creation

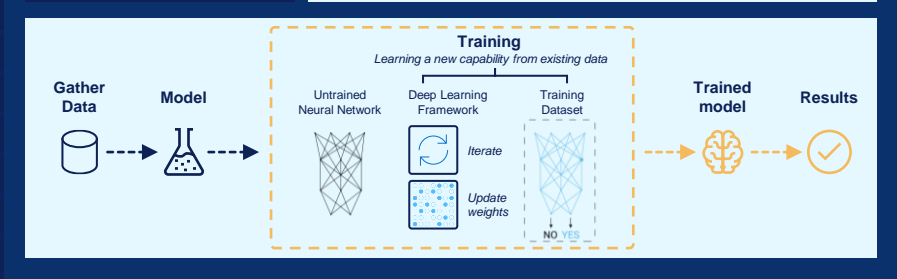
Use case	GenAI purpose	Users	Benefits
 <p>Design and data creation</p>	<ul style="list-style-type: none"> <li>Solving high-specialized, bleeding-edge problems</li> <li>Example: Drug discovery and design                             <ul style="list-style-type: none"> <li>Accelerate the process of predicting which parts of the genome may impact the growth of cancerous cells and how to treat them in a more targeted and localized manner.</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Data scientists</li> <li>Researchers</li> <li>Doctors</li> <li># of users</li> <li>Etc.</li> </ul>	<ul style="list-style-type: none"> <li>Offers transformational opportunities and demonstrates innovations</li> <li>Introduces potential resale opportunities and/or differentiated offers</li> <li>Eliminates the 'black-box'</li> </ul>



Model training

### Why create a new model?

- For use cases that off-the-shelf models are failing to answer accurately (even w/ fine-tuning)
- Best, most accurate results
- Most differentiated value



### Infrastructure investment

Dell Validated Designs

Networking    Compute    Storage

Workstations

Want help?

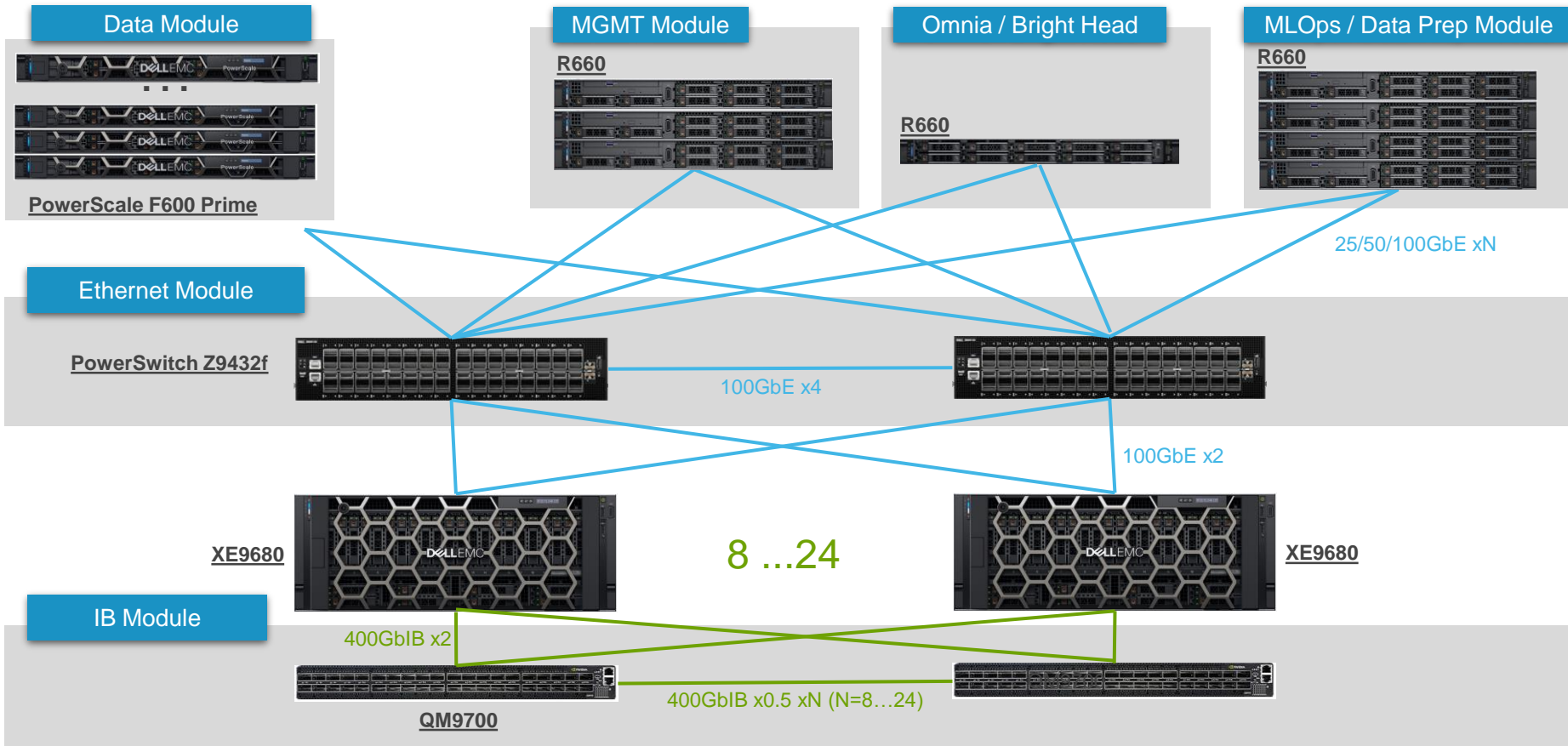
- Deployment
- Advisory
- Data Science

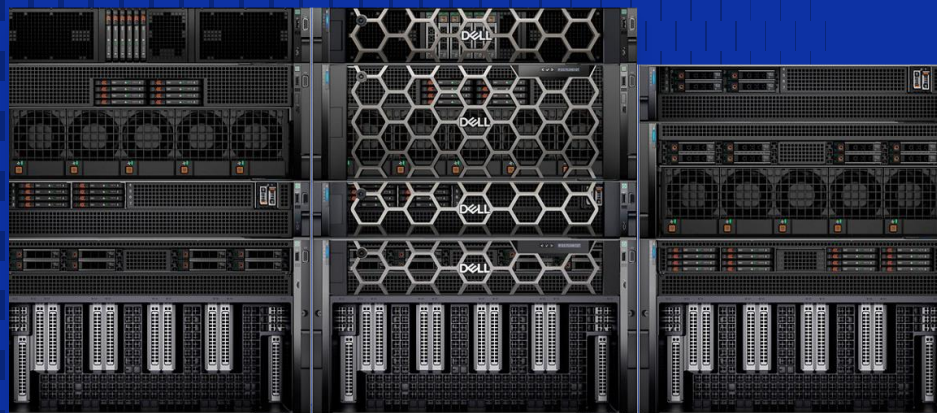
# Training with H100 GPU

- Requirement:  $0.4 * 10^{24}$  flops
- H100/SXM (fp16): 989.4 Tflop/s ~ 40-52% efficiency => ~0.4 Pflop/s ( $0.4 * 10^{15}$ )
- For one NVIDIA H100/SXM GPU:  $0.4 * 10^{24} / 0.4 * 10^{15} = 10^9$  seconds
  - Or 277777 hours, or 11574 days!
- Given a dense interconnection between GPUs (think InfiniBand), almost linear scaling: 80 8-way GPU nodes would train that model in ~20 days

Server model	kW / server	Number of nodes	Total kW	Energy cost	TeqCO <sup>2</sup>
PowerEdge XE9680	11.5	64	736	52k€	15.7
PowerEdge XE8640	5.8	128	742	53k€	15.9
PowerEdge XE9640	4.5	128	576	34k€	10.3

# Use Case : Large Model Training





**#1** worldwide  
in AI server and infrastructure<sup>1</sup>

**Highest performance**  
training and inferencing for AI  
operations<sup>2</sup>

## Unleash your GenAI advantage with PowerEdge



Simplify & streamline AI operations with acceleration-optimized compute



Deploy a tailored, scalable GenAI infrastructure for all your needs



Develop trusted AI with cyber-resilient and secure platforms



# PowerEdge.Next GPU Acceleration Server Portfolio

PCIe Optimized



4-way SXM



4-way Dense



8-way SXM



## R760XA

- 2U monolithic
- 2-socket Sapphire Rapids CPU
- Up to 4 x double-wide GPUs
- Up to 12 x single-wide GPUs
- Full PCIe GPU portfolio supported
- Air cooled with optional liquid cooling for CPU

## XE8640

- 4U monolithic
- 2-socket Sapphire Rapids CPU
- 4 x Nvidia H100 SXM NVLink GPUs;
- Air cooled

## XE9640

- 2U monolithic
- 2-socket Sapphire Rapids CPU
- 4 x Nvidia H100 SXM NVLink GPUs
- or-
- 4 x Intel Data Center Max 1550 OAM XeLink GPUs
- Direct liquid cooled CPUs and GPUs

## XE9680

- 6U monolithic
- 2-socket Sapphire Rapids CPU
- 8 x Nvidia H100 SXM NVLink GPUs
- or-
- 8 x Nvidia A100 SXM NVLink GPUs
- Air cooled
- + AMD Instinct MI300X 8-way

## USE CASES

- AI/ML Inferencing
- AI/ML Training
- Rendering/Perf. Gfx
- VDI

- AI/ML Training
- HPC Modeling & Simulation

- HPC Modeling & Simulation
- AI/ML Training

- Gen AI Training
- Large Language Model Training
- Recommendation engines
- Neural Networks

## POWER CONSUMPTION

~up to 3kW, config dependent

~5.2kW

~4.5kW

~10.2kW

## MAX MODEL SIZE BEFORE SCALE OUT

320GB

320GB

320GB (H100 80GB) or 376GB (H100 94GB)  
512GB (Intel Max 1550)

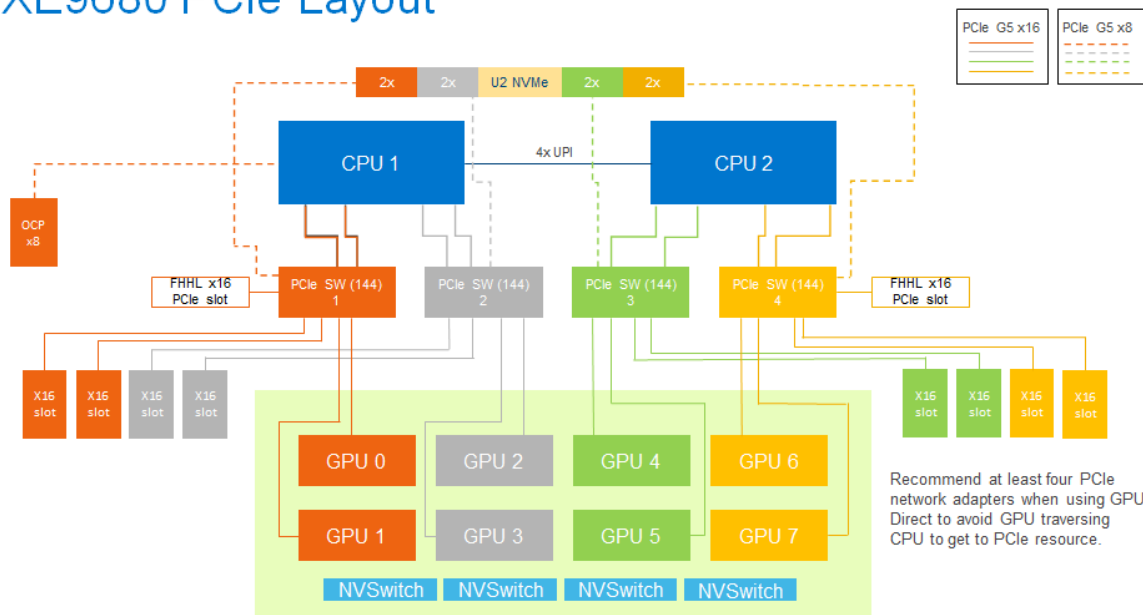
640GB/1.536GB

# 10.c. The actual components in an XE9680



Important topics: GPUs, NVLink Switch, PCIe, memory, RDMA over something, GPU Direct Storage, Latency

## XE9680 PCIe Layout

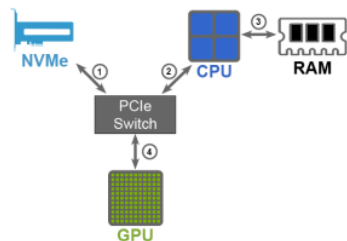


# GPU Direct Storage

## Data Transfer with GPUDirect Storage (GDS)

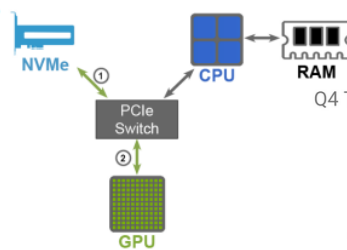
### Traditional Data Transfer

1. `fd = open("file.txt", O_RDONLY, ...);`
2. `h_buf = malloc(size);` No need for a "bounce buffer"
3. `pread(fd, h_buf, size, 0);`
4. `cudaMalloc(d_buf, size);`
5. `cudaMemcpy(d_buf, h_buf, size, cudaMemcpyHostToDevice);`

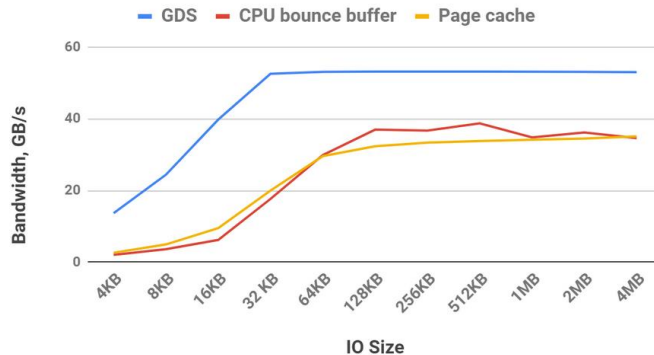


### NVIDIA GPU Direct Storage: IO directly from DMA/RDMA capable storage to/from user allocated GPU memory on NVIDIA GPUs

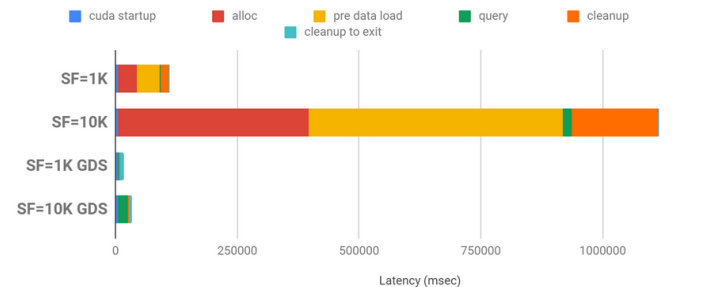
1. `fd = open("file.txt", O_RDONLY | O_DIRECT, ...);`
2. `cudaMalloc(d_buf, size);`
3. `cuFileRead(fd, d_buf, size, 0, 0);`



## Comparison of Transfer Methods



### Q4 TPC-H Benchmark Work Breakdown





The world's **most secure**  
NAS storage array<sup>1</sup>



Over **2 TB/s** of  
**read throughput** to a  
**massive GPU farm**<sup>2</sup>

**186PB**  
in a **252-node cluster**<sup>3</sup>

Unlock the full potential of  
your data with PowerScale



Exceptional storage performance for the most demanding AI workloads



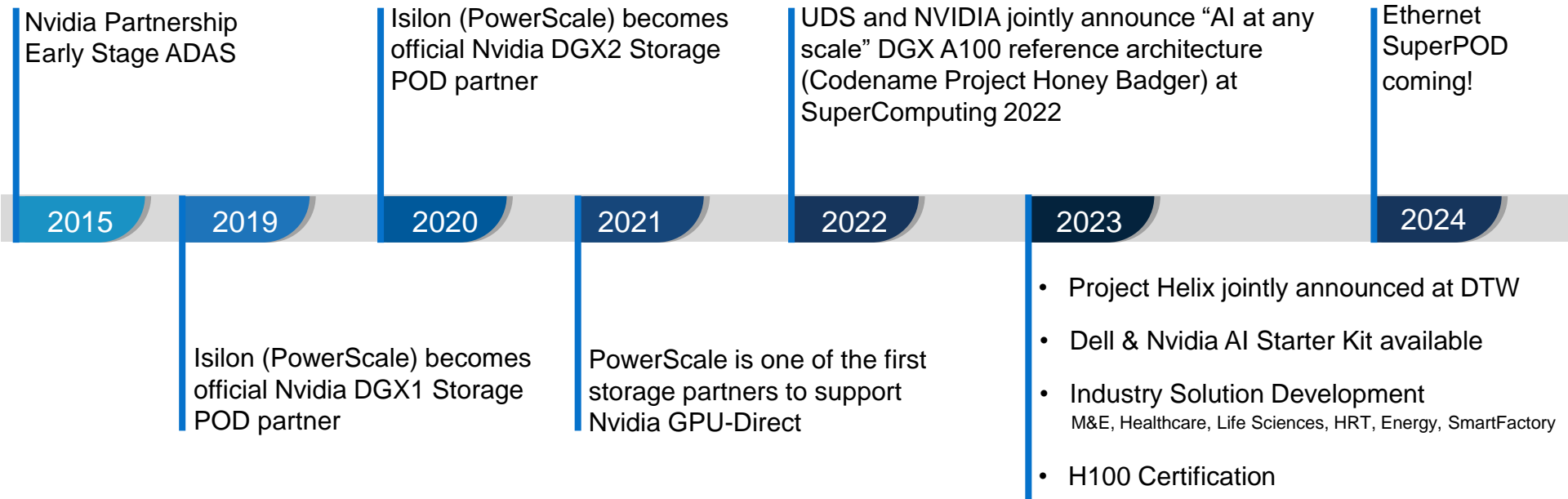
Optimize your AI agility with a data platform that scales with your data



Industry-leading security for the data fueling your GenAI models

# The History of PowerScale and AI

1,500+ customers running GPU workloads with UDS



# Data Platform for all your AI needs

Complete storage solution for AI and Gen-AI workloads

## Simple

Deploy, Operate, and Maintain at Scale

## Multi-Protocol Access

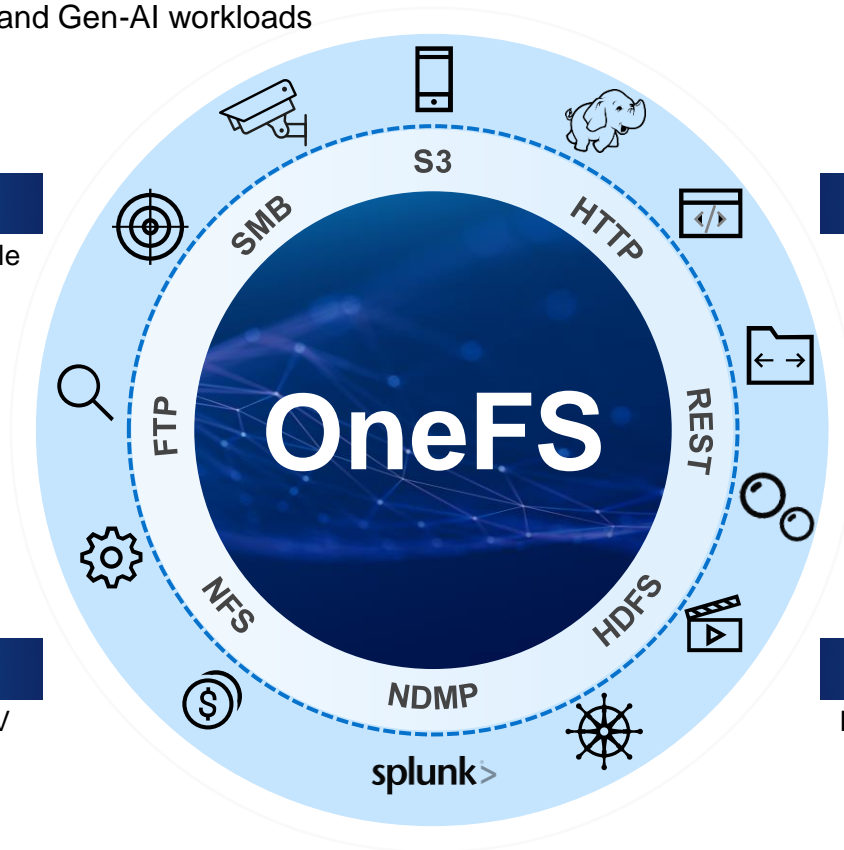
Seamless access to datasets

## Deployment Flexibility

On-Prem, Public Cloud, APEX

## Store and Protect Data

Snapshots, Replication, Backup, DR, AV



## Security

Zero Trust, Federal certs, integrated ransomware protection

## Multi-tenancy

Secure segregation of resources for customers or LOBs

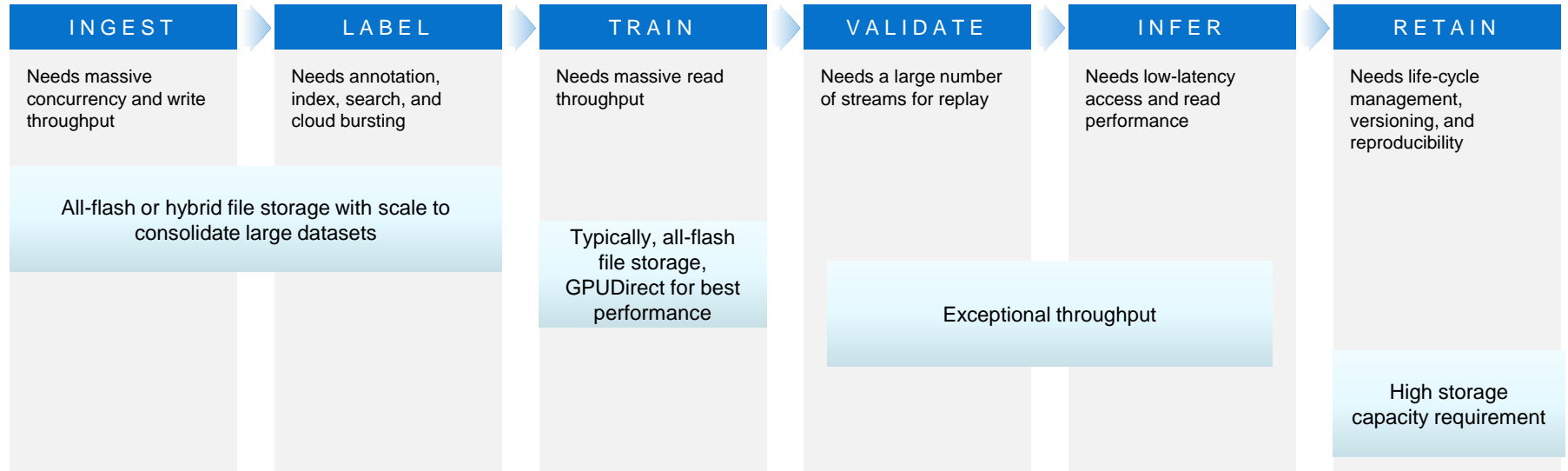
## Performance

Faster time to outcomes for better ROI

## Scalability and Flexibility

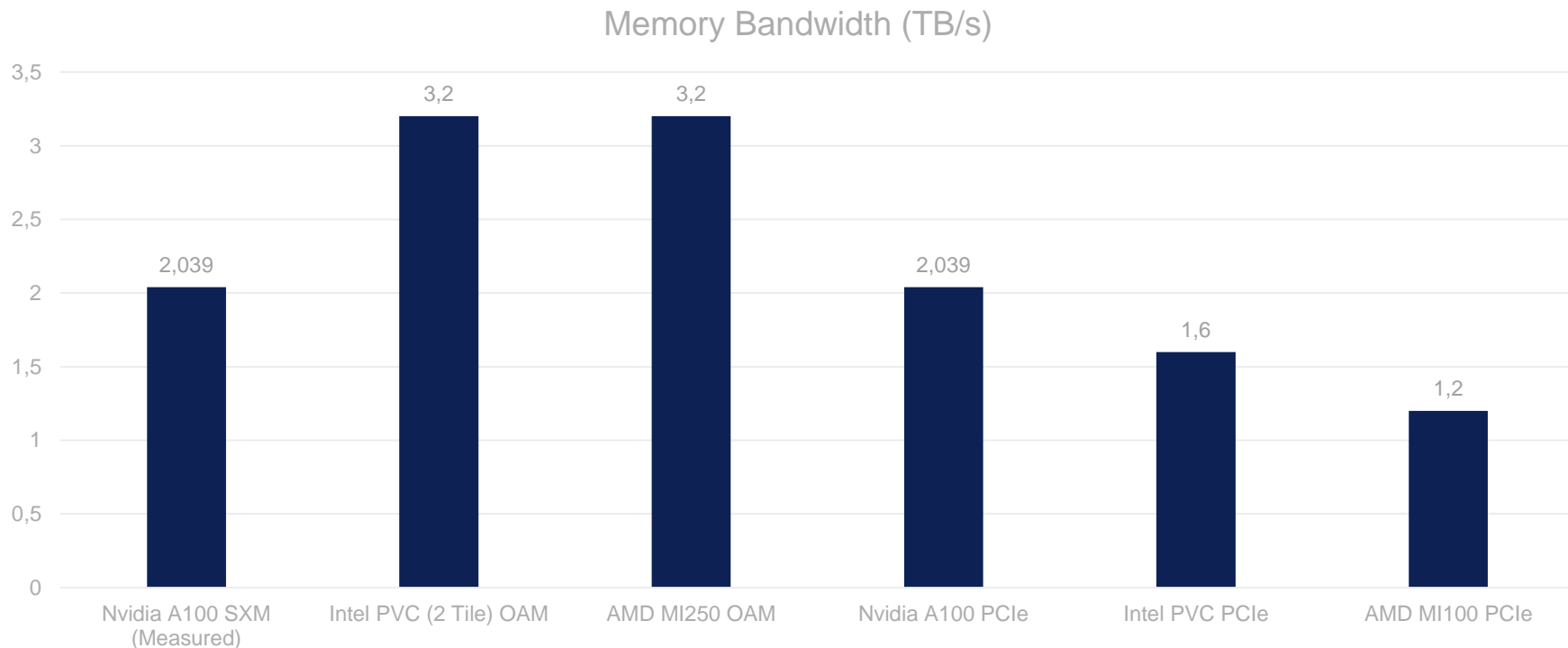
Linearly grow with the workflow needs

# Storage requirements from ingest to retention in the AI development



Dell PowerScale delivers on all the storage demands for AI development

# GPU Memory Bandwidth Performance





# PowerScale Exceeds the Most Demanding Requirements

- Meets the Nvidia SuperPOD Published Metrics with significant headroom to scale **w/o IB**
- Achieve *Better, Best* ML Commons Benchmarks with low node counts
- GPUDirect storage for high-speed, efficient access to data
- NFSoRDMA for high speed read/write performance for data collection, pre-processing & AI training
- Unrivalled Scalability with PowerScale
  - 252 Nodes

PowerScale exceeds SuperPOD published performance metrics

NVIDIA Category	NVIDIA Published R/W GBs			PowerScale F600 Nodes		
	Good	Better	Best	Good	Better	Best
Single Node- read/write	4/2	8/4	40/20	✓	✓	✓
Single SU Aggregate- read/write	15/7	40/20	125/62	✓	✓	✓
4SU Aggregate- read/write	60/30	160/80	500/250	✓	✓	✓

PowerScale exceeds MLCommons benchmark

Training Model Type	Resnet-50		BERT-Large		GPT3		DLRM-DCNv2		3D U-Net		MaskRCNN	
	Image Classification	Transformer Based Language Model	Large Language Model	Recommendation Model	3D Image Segmentation for Medical	Object Detection	Better	Best	Better	Best	Better	Best
SuperPOD Storage Perf Level	Better	Best	Better	Best	Better	Best	Better	Best	Better	Best	Better	Best
Est. Perf	44K Img/sec 5 GB/sec 8x H100	4.5M Img/sec 500 GB/sec 808x H100	8K Seq/sec 0.015 GB/sec 8x H100	1M Seq/sec 1.9 GB/sec 1024x H100	4.6K Seq/sec 0.04 GB/sec 8x H100	595K Seq/sec 4.5 GB/sec 1024x H100	9.6M Reqs/sec 8.3 GB/sec 8x H100	576M Reqs/sec 500 GB/sec 481x H100	441 Img/sec 40 GB/sec 8x H100	5.5K Img/sec 500 GB/sec 97x H100	1.2K Img/sec 0.2 GB/sec 8x H100	147K Img/Sec 22.5 GB/sec 1024x H100
# F600 Nodes	✓ 3	✓ 52	✓ 3	✓ 3	✓ 3	✓ 3	✓ 3	✓ 3	✓ 5	✓ 52	✓ 3	✓ 3

Assumes up to 100% of PowerScale read bandwidth is used for AI activity  
PowerScale has a 3-node minimum cluster size

Copyright © Dell Inc. All Rights Reserved

**DELL**Technologies

**DELL**Technologies

# Connecting in the AI Era

Accelerator proliferation requires a new networking architecture

## Dell's Ethernet fabric for GenAI



### Lossless

Preserves data integrity and improves reliability



### High performance

Delivers high speed data transfer without congestion



### Scalable

Accommodates various AI environment sizes



# Ultra Ethernet

Consortium

MEMBERSHIP

Co-developing modern networking to meet the demands of AI

PowerSwitch Z9664F-ON

Powered by:

Dell Technologies

PowerSwitch

Handle massive data volumes and complexity using leading Ethernet technology

Dell Technologies / Open-Source

Enterprise SONiC

Efficiently scale, automate and operate network fabrics with an end-to-end, unified open-source-based OS (NOS)

BE BEYONDEDGE ANSIBLE

BeyondEdge + Ansible

Orchestrate the network fabric with automation and graphical representations

augtera networks

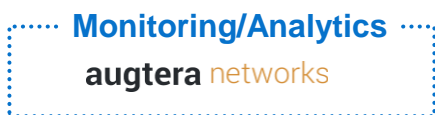
Augtera Networks

Enhance network visibility, anomaly detection and traffic management

Dell Technologies



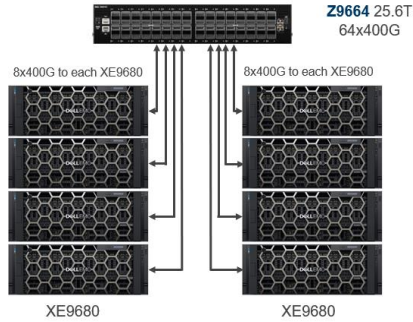
# Offering Summary



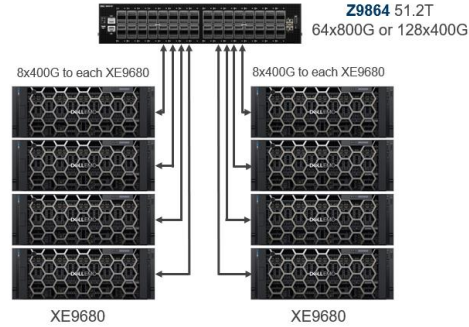
## Phase 1 Solutions – Based on Z9664F (Shipping)

## Phase 2 Solutions – Based on Z9864F (Roadmap)

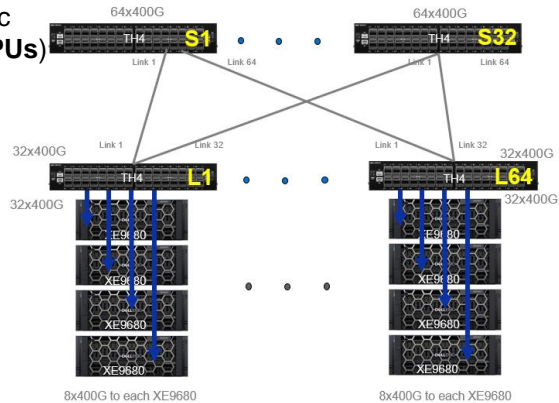
TH4 Single System  
(up to 64 GPUs)



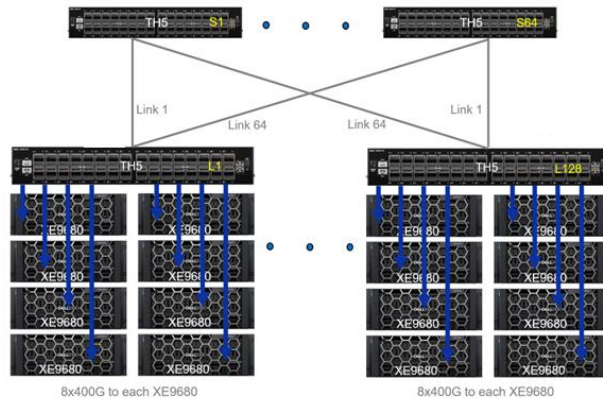
TH5 Single System  
(up to 128 GPUs)



TH4 Fabric  
(up to 2K GPUs)



TH5 Fabric  
(up to 8K GPU support)



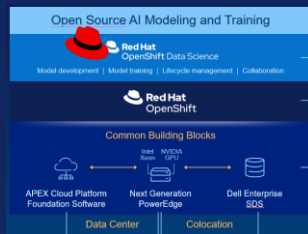
# Expanding AI capabilities

Enabling customer success with flexible starting foundations and partners



Meta  
-  
Hugging Face

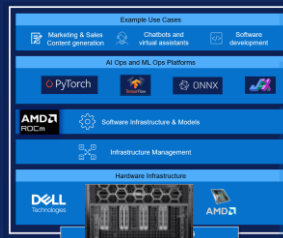
- Open-source, cost-effective models
- Enable easier and faster to deploy GenAI projects on premises
- Tested and validated in Dell Validated Designs



Dell Validated Design for Red Hat OpenShift Data Science on APEX Cloud Platform

Dell Technologies | A P E X

- Simplify AI deployment and operations
- Deliver full MLOps across private cloud
- Accelerate AI Value with Dell Services



PowerEdge XE9680 with AMD Instinct™ MI300x GPU & Dell Validated Design for Generative AI with AMD


- Leading bandwidth and HBM capacity for LLMs
- Power workflows at scale with open-source SW stack and ecosystem
- Open portability & minimal rework to existing application code




Use cases & Strategies:  
Chatbots  
Retrieval Augmented Generation (RAG)

- Chatbot demo w/ APEX ACP, a simplified approach to AI-assisted customer support
- RAG enables company data into models for more accurate results

# Use case example | Content creation

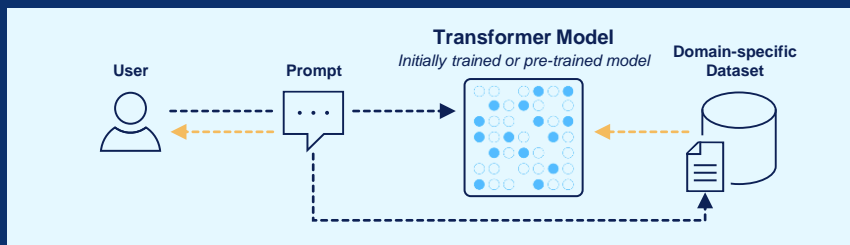
Use case	GenAI purpose	Users	Benefits
 <p>Content creation</p>	<ul style="list-style-type: none"> <li>Support IT staff and Service Reliability Engineers (SRE) responding to help desk tickets</li> <li>Utilize info from historical ticketing databases to understand past context and provide recommendations</li> </ul>	<ul style="list-style-type: none"> <li>IT</li> <li>Internal users</li> <li># of users</li> <li>Etc.</li> </ul>	<ul style="list-style-type: none"> <li>Identify, learn from and extend best practices</li> <li>Improves the speed, accuracy and/or completeness of recommended actions</li> <li>Captures your organization's voice and context to deliver brand-aligned responses</li> </ul>



Model Augmentation with RAG

## Why RAG?

- Provides contextually aware answers
- Unique patterns have been established based on institutional trends
- No data science needed but keeps humans-in-the-loop (HITL)



## Infrastructure investment



XPS 15

**Laptops**

OR



Precision Series

**Workstations**

OR



PowerEdge R-Series

**Servers**

Want help?



Deployment



Advisory



Data Science

## Virtualizing GPUs for AI with VMware and NVIDIA

Based on Dell Infrastructure

March 2022  
H18904.2

## Implementing a Digital Assistant with Red Hat OpenShift AI on Dell APEX Cloud Platform for Red Hat OpenShift

With a Language Model (LLM) and the Retrieval Augmented Generation (RAG) framework

November 2023  
H19833

### Design Guide

#### Abstract

This design guide describes the Validated Design for Virtualizing and Tanzu and NVIDIA AI. It describes the reference architecture and performance characteristics.

Dell Technologies Solutions

Dell Technologies  
Validated Design

## Generative AI in the Enterprise – Model Customization

A Scalable and Modular Production Infrastructure with NVIDIA for AI Large Language Model Customization

November 2023  
H19825.1

### Design Guide

#### Abstract

This design guide describes the architecture and design of the Dell Validated Design for Generative AI Model Customization with NVIDIA, a collaboration between Dell Technologies and NVIDIA to enable high performance, scalable, and modular full-stack generative AI model customization solutions for large language models in the enterprise.

Dell Generative AI Solutions

Dell  
Validated Design

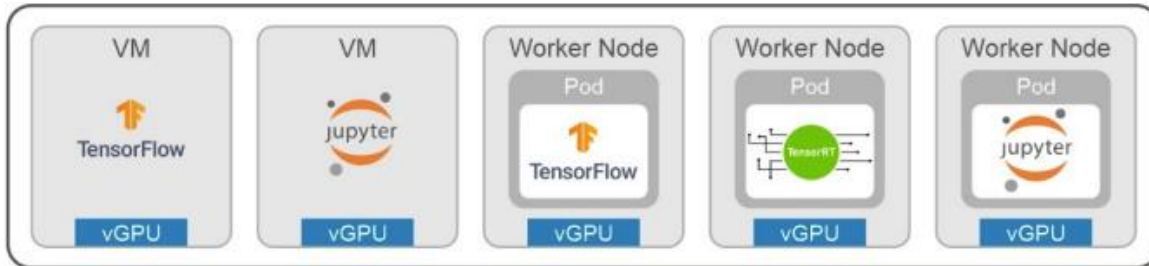
### Design Guide

#### Abstract

This design guide describes the architecture and design of the Dell Technologies Validated Design for deploying a digital assistant on Dell APEX Cloud Platform for Red Hat OpenShift using Red Hat OpenShift AI. This solution leverages a LLM and the RAG technique in combination with a set of vectorized documents.

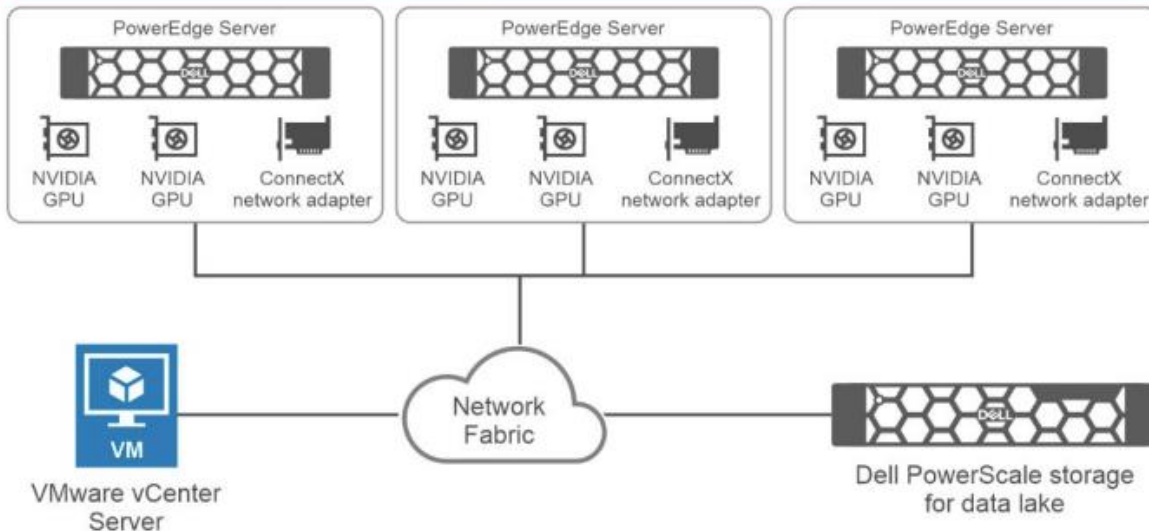
Dell Technologies Solutions

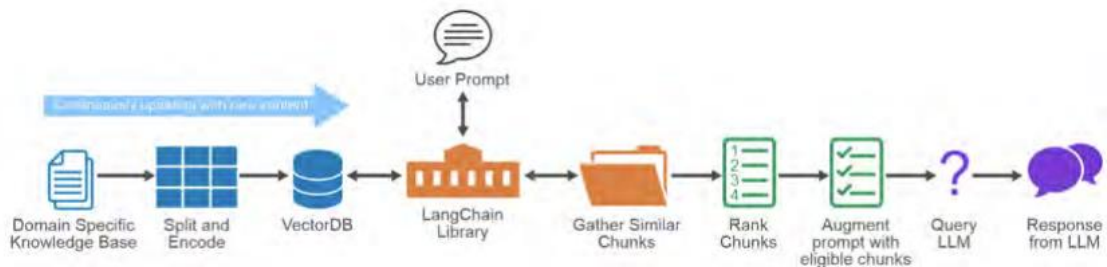
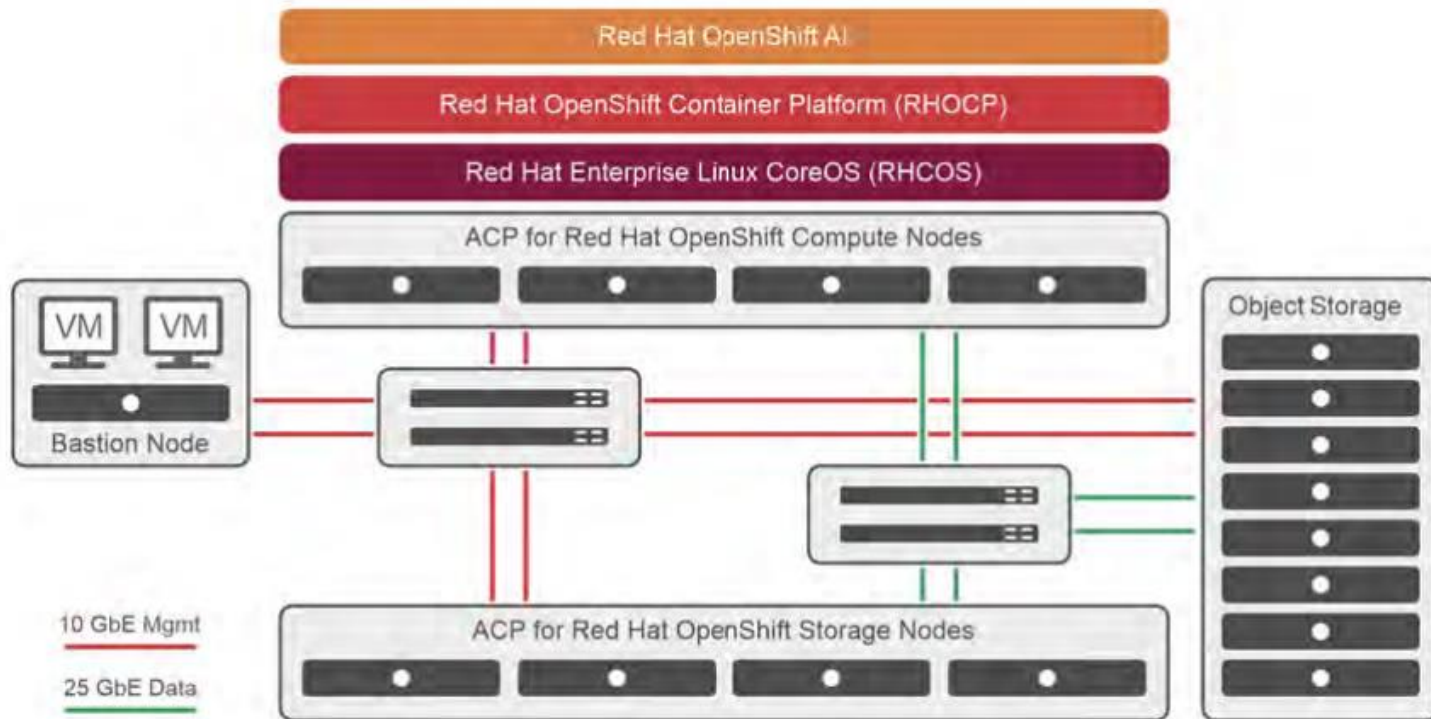
Dell  
Validated Design



**Tanzu Kubernetes Cluster**

**VMware vSphere with Tanzu**

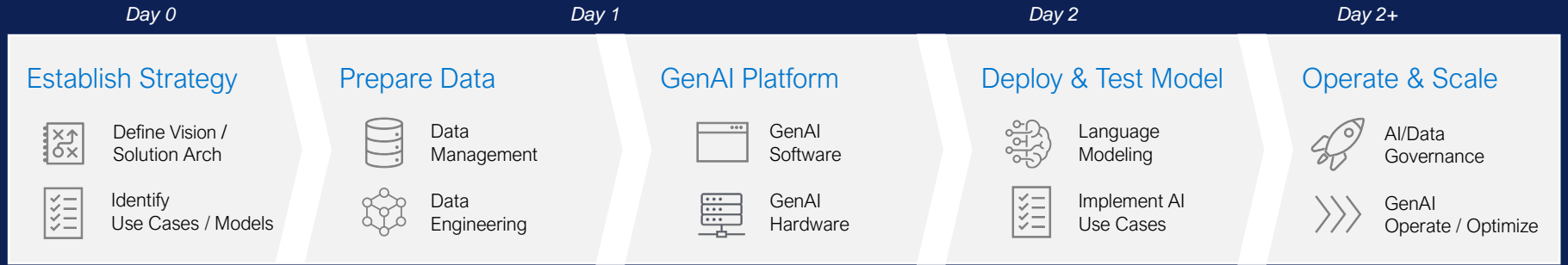






# Training Courses for Generative AI

Learning offers aligned to your GenAI journey



## Training Courses

Artificial Intelligence and Machine Learning

Data Engineering Workshop

LLM Deployment and Customization & NVIDIA HW/SW Admin

Data Governance Security and Privacy for Big Data

Infrastructure Aligned Training

GenAI Bootcamp

Contact your Dell Account Manager to discuss training options, or visit [education.dell.com](https://education.dell.com)

# Project DVD

## Infrastructure foundation



- AI acceleration-optimized PowerEdge servers for training, inferencing, p-tuning
  - XE9680 and R760xa
  - NVIDIA H100, A100, L4 GPUs
- Scalable unstructured data storage, Dell PowerScale and ECS Object Storage
- High performance Dell and NVIDIA Networking

## GenAI framework and management foundation



- NVIDIA AI Enterprise software
- NVIDIA NeMo large language model framework
  - NeMo Guardrails
  - Pretrained models
- Platform Management
  - Dell OpenManage, Dell OneFS
  - Dell CloudIQ
  - NVIDIA Base Command Manager

## Expertise and Advisors



- Dell Technologies Services
  - Platform support
  - Global consulting through deployment
- NVIDIA Advisors
  - Custom Models
  - Fine-tuning

# Dell Generative AI solutions: Bring AI to your data

## Flexibility of starting configurations

### Simplified

Go from possible to proven, faster

### Tailored

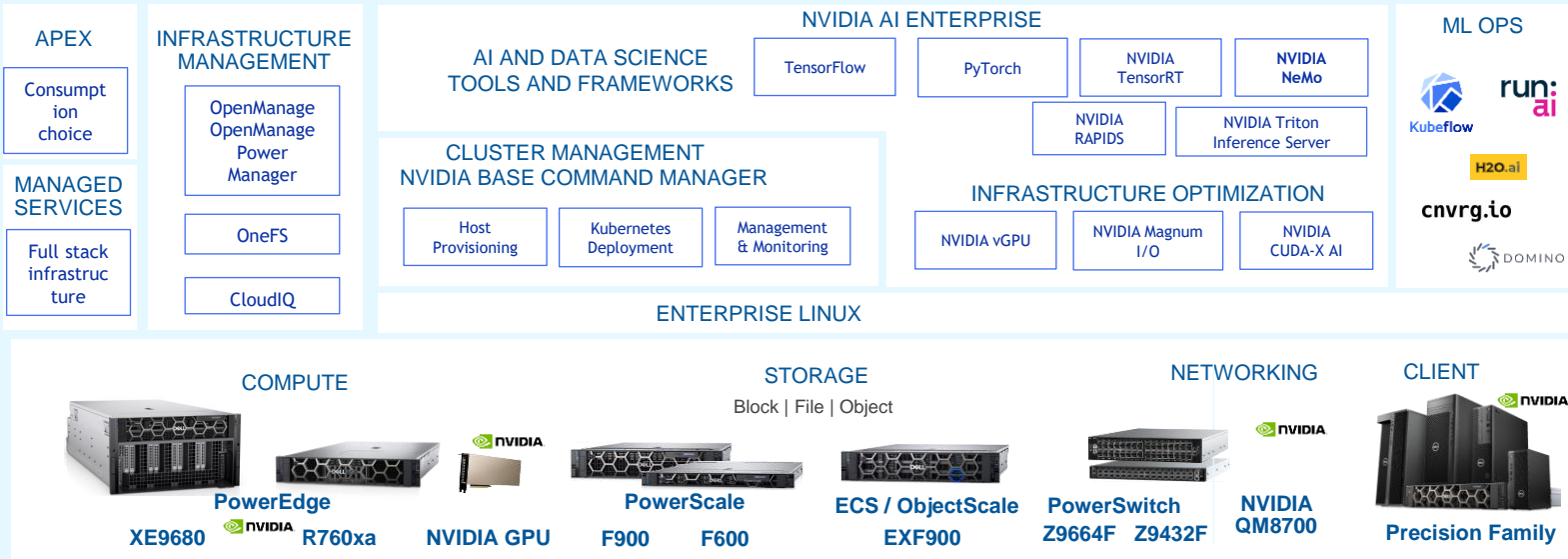
Create better outcomes built specific for you

### Trusted

Protect and sustain your success

SERVICES  
Consulting | Deployment | Support | Residency | Education

SUPPORT  
Customization | Models | Configurations



Data Management & Preparation



Model Augmentation



Training



Model Customization and Tuning



Inferencing

# Dell Technologies has what you need

Dell Technologies can provide you with the power of AI at a price point and commitment level for your project



Expertise and guidance

Validated Designs for GenAI, AI

Customized solutions

# Next steps



## AI Discovery

Dell architects work with you to establish a baseline approach



## Dell Accelerator Workshop

Half-day Services strategy engagement on use cases, requirements, skills and processes (no cost)



## AI Executive Briefing

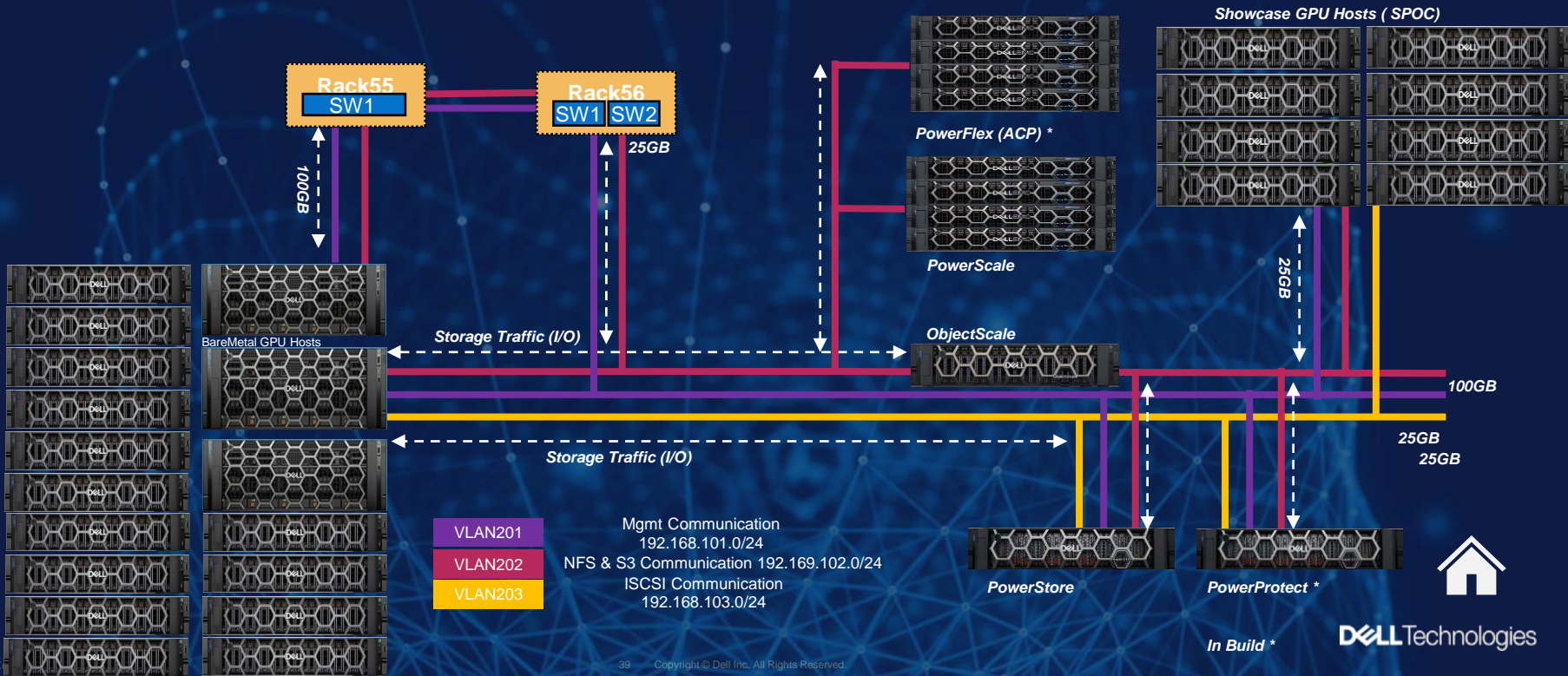
Deep dive into the infrastructure and planning required to make AI real

A comprehensive approach to your **Generative AI** journey

# CSC AI Use Cases – available today

C&C	Digital Cities	Healthcare	Retail	Manufacturing	FSI	Other
Video Management	Urban Planning & Simulation	Digital Pathology	Store Analytics	Lone Worker VR Training	Streaming Analytics	HyperPersonalization using AI
Watchlist Alerting	Integrated Operations Center	Immersive Training	Concealment Detection	Computer Assisted Assembly	Trustworthy AI	GUI Interface for AI Models
Intrusion Detection	Image Enhancement	NLP with Analytics	Loss Prevention		Simulation in Finance	SDP for Datacenter Energy Utilization
Deep Learning Video Analytics	Powerline Monitoring	VR Care Companion	Store Optimization		AI for IT Operations	
Thermal Monitoring		Augmented Reality	Frictionless Experience			
AI Actionable Insights		H&S Computer vision alerts				
		Faster Data Insights				
		Digital Human Clara				

# CSC AI POC High Level Architecture Q1





[www.srce.unizg.hr](http://www.srce.unizg.hr)

Ovo djelo je dano na korištenje pod licencom Creative Commons  
*Imenovanje* 4.0 međunarodna.

[creativecommons.org/licenses/by/4.0/deed](https://creativecommons.org/licenses/by/4.0/deed)



Srce politikom otvorenog pristupa široj javnosti osigurava dostupnost i korištenje svih rezultata rada Srca, a prvenstveno obrazovnih i stručnih informacija i sadržaja nastalih djelovanjem i radom Srca.

[www.srce.unizg.hr/otvoreni-pristup](http://www.srce.unizg.hr/otvoreni-pristup)

