Dell Technologies

Solution Brief

Dell AI Factory with NVIDIA

Dell Technologies AI Factory with NVIDIA

Accelerate AI outcomes with modular, automated and proven enterprise AI.
NVIDIA 2-8-9-400 configuration, ERA endorsed for the Dell AI Factory with NVIDIA

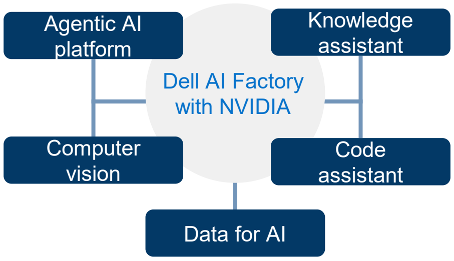
The Dell AI Factory with NVIDIA is how organizations turn AI potential into real-world progress. This partnership combines Dell Technologies industry leading enterprise infrastructure with NVIDIA accelerated computing to create powerful, full-stack solutions. The PowerEdge XE9680 based Dell AI Factory with NVIDIA is endorsed by NVIDIA Enterprise Reference Architectures (Enterprise RAs) aligned with 2-8-9-400 configuration. It provides validated, repeatable platforms for building and scaling high-performance AI infrastructure.

- **Up to 1,225% ROI over four years.[1]**
- **Up to 75% more cost-effective than public cloud IaaS[1]**

The Dell AI Factory with NVIDIA Difference

Multi-workload versatility, built-in automation, and scalable infrastructure provides modern enterprises the solution they need to support diverse AI workloads, from model training to inference. The Dell AI Factory with NVIDIA is engineered to handle this complexity with a modular architecture where compute, storage, and networking scale independently. This flexibility is managed by a powerful automation platform that streamlines deployment, reduces manual effort, and minimizes configuration drift.

One solution, multiple workloads, many outcomes

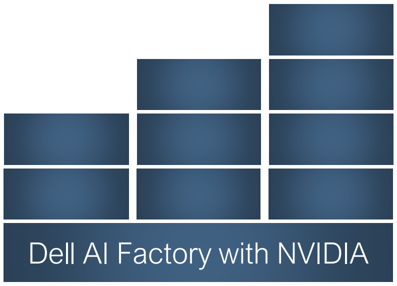


Common infrastructure starting point for enterprise AI production

Deploy essential compute, storage, networking, software, and automation required to launch quickly.

Right sized to workload requirements, from training, fine tuning, and long thinking inference

Expand and scale predictably

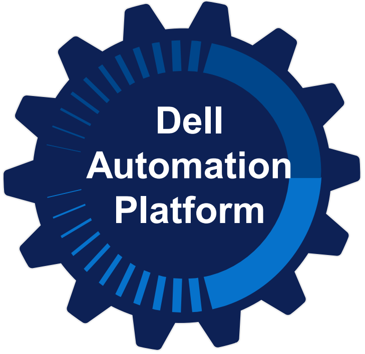


Scale quickly with validated, pre-configured expansion modules

Future-ready solutions, leveraging current and future NVIDIA GPUs

Manage data growth with validated and curated blocks of storage

Automated to accelerate and optimize



Deploy AI outcomes quickly and easily

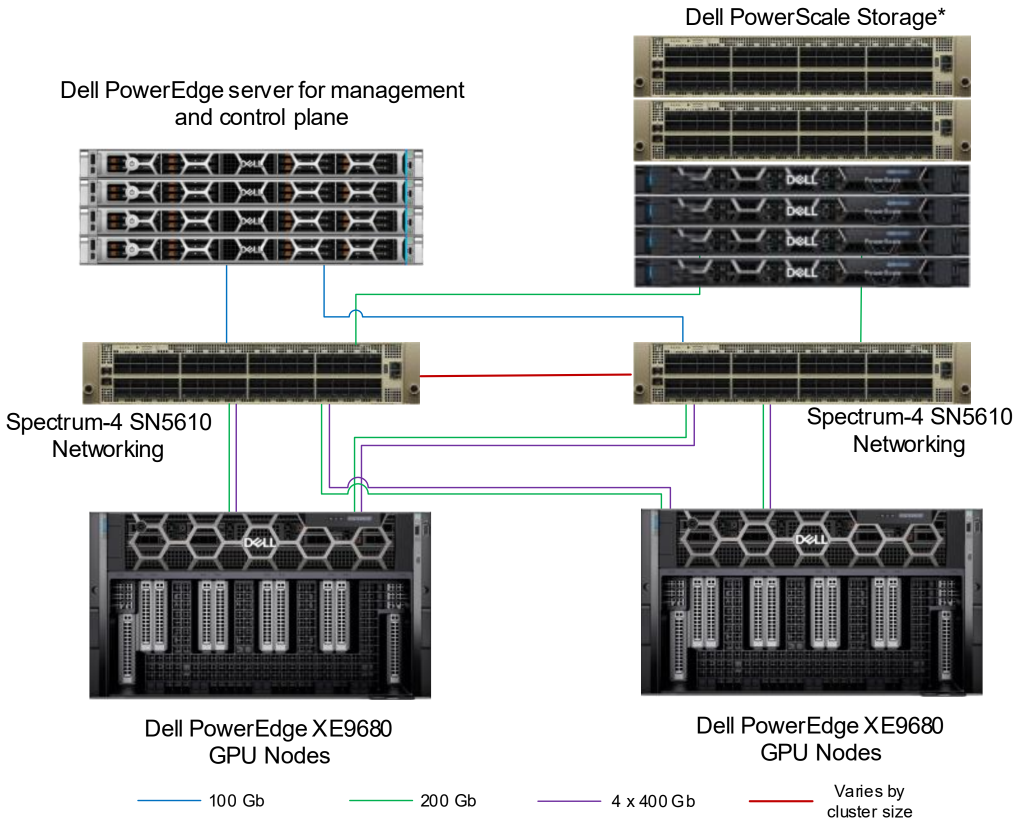
De-risk outcome deployment with standardized and validated deployment blueprints

Speed ROI and reduce skills gap from the infrastructure to outcome

Aligned with NVIDIA 2-8-9-400 Enterprise Reference Architecture and Dell PowerEdge XE9680

Dell AI Factory with NVIDIA AI Enterprise and Red Hat OpenShift AI

This architecture leverages NVIDIA-Certified Dell PowerEdge XE9680 servers, aligning with the 2–8–9–400 Enterprise RA design pattern to ensure optimal performance and validated scalability for enterprise AI workloads. Endorsed as part of NVIDIA's Enterprise Reference Architectures, it is purpose-built to accelerate demanding AI workloads, such as machine learning, deep learning, and large language model training, with unmatched efficiency and scalability.It also enables seamless integration with NVIDIA and open source for AI workload deployment and operational efficiency.


Dell PowerScale Storage*
Dell PowerEdge server for management and control plane
Spectrum-4 SN5610 Networking
Spectrum-4 SN5610 Networking
Dell PowerEdge XE9680 GPU Nodes
Dell PowerEdge XE9680 GPU Nodes
100 Gb
200 Gb
4 x 400 Gb
Varies by cluster size

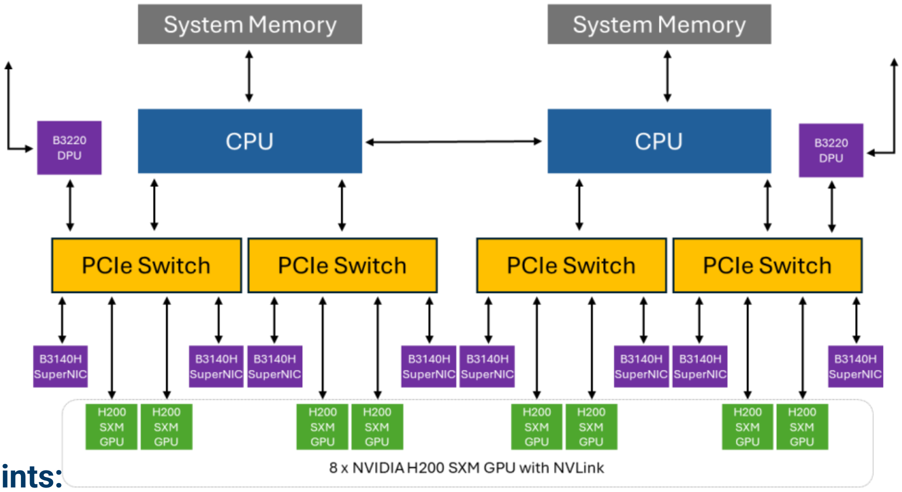
* Dell PowerScale storage included as per ERA recommendations

Accelerating inference workloads at scale requires scalable computational power, and that's exactly what this design delivers. It powers demanding AI tasks like real-time translation, image recognition, and large language model deployment with speed and efficiency. At the core of this infrastructure are up to 12 Dell PowerEdge XE9680 servers, each equipped with two high-performance processors and eight NVIDIA H200 SXM GPUs. This architecture combines dense GPU power with minimal physical footprint, ensuring the speed and efficiency needed for both training and inference at scale.

Effective management is key to maintaining a stable and responsive environment, and we ensure it doesn't compete with resources needed for your critical AI workloads. This is achieved with four PowerEdge R670 servers, which provide dedicated resources for orchestration and system administration. Supporting this setup, Spectrum-4 Ethernet switches create a non-blocking, spine-leaf topology for data transfer. This advanced networking design delivers high-bandwidth, low-latency pathways between GPUs, ensuring efficient data flow.

To turn your data into insights and gain a competitive advantage, you need infrastructure that can perform without delay. Our design includes high-throughput, scalable storage from PowerScale F710, which is crucial for modern AI models. The software stack runs on NVIDIA AI Enterprise, a suite of optimized tools that unlocks the hardware's full potential. We use orchestration through Red Hat OpenShift or Upstream Kubernetes for flexible deployment and management of containerized AI applications.

Aligned with NVIDIA 2-8-9-400 Enterprise Reference Architecture and Dell PowerEdge XE9680



Salient Design points:

- High-performance GPU network for distributed AI workloads: Spectrum-X provides a non-blocking, rail-optimized spine-leaf fabric that ensures deterministic, low-diameter GPU paths. With RDMA/RoCEv2, PFC, ECN, adaptive routing, and congestion control, the fabric delivers low-latency, predictable, and highly scalable performance for LLM training workloads.

- Enterprise network architecture for data-center-scale operations: The platform uses an EVPN-VXLAN fabric that supports scalable north-south traffic, workload mobility, and tenant segmentation through a modern BGP-based control plane. EVPN-Multihoming (EVPN-MH) provides active/active server connectivity without MLAG, improving link utilization and failover behavior.

- Cloud-native infrastructure deployment and configuration: Using Kubernetes operators—including the NVIDIA GPU Operator, NVIDIA Network Operator, NVIDIA NIM Operator and Dell CSI Operator—the environment can be provisioned, configured, and managed using declarative, cloud-native patterns. Cluster expansion, lifecycle management, and routine operations are streamlined, enabling enterprises to maintain consistent configuration and accelerate day-to-day management.

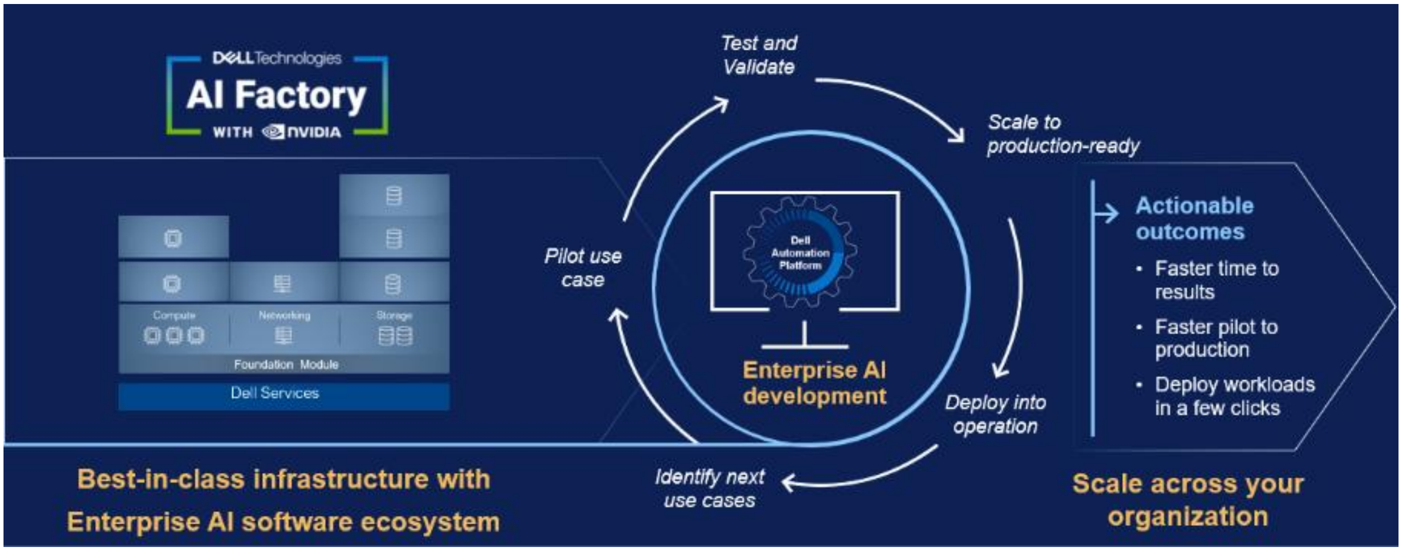
- Validated solution stack: The Dell AI Factory with NVIDIA aligns directly with the NVIDIA AI Enterprise infrastructure support matrix and the Spectrum-X validated solution stack, ensuring that all software components, drivers, and networking layers operate in a fully compatible, jointly validated configuration.

Dell AI Factory with NVIDIA AI Enterprise and Red Hat OpenShift AI

Component	**Upstream Kubernetes**	**Red Hat OpenShift Container Platform**
Management Server	4 x PowerEdge R670	4 x PowerEdge R670
Worker nodes	Up to 12* PowerEdge XE9680 servers	Up to 12 PowerEdge XE9680 servers
Networking	2 x Spectrum-4 SN5610 for converged network 1 x Spectrum SN2201 for OOB	2 x Spectrum-4 SN5610 for converged network 1 x Spectrum SN2201 for OOB
Storage	PowerScale F710	PowerScale F710
Cluster Manager	Base Command Manager	Dell Automation Platform
Container Orchestration	Upstream Kubernetes	OpenShift Container Platform
AI Frameworks and Software	NVIDIA AI Enterprise	NVIDIA AI Enterprise OpenShift AI

*Larger scale clusters supported and require additional network switching

Streamline Enterprise AI innovation with outcome-first modular architectures



Accelerate AI deployment with pre-validated, integrated solutions that reduce setup time.	Scale AI initiatives flexibly, starting with what you need today and expanding as requirements grow.
Streamline operations and free your teams to focus on innovation, not infrastructure	Turn complex data into insights that drive operational efficiency and better decision-making
Minimize risk with a unified platform architected for enterprise security and compliance	Protect intellectual property and sensitive information with robust data governance
Support dynamic growth with a solution built for reliable performance at enterprise scale	

Accelerate AI results

The fastest way to realize AI value is to remove complexity. Dell AI Factory with NVIDIA delivers validated architectures, modular scaling, and integrated automation so teams can confidently move from pilot to production. The platform provides a secure, consistent, and centrally managed AI environment that grows with your needs and ambitions. By focusing on outcomes instead of components, Dell and NVIDIA help you turn AI potential into measurable business progress and impact.

[1] Enterprise Strategy Group, "Analyzing the Economic Benefits of the Dell AI Factory with NVIDIA," August 2025, commissioned by Dell Technologies

Learn more about Dell AI Factory with NVIDIA

Contact a Dell Technologies Expert

View more resources

Join the conversation with #Dell #AIFactory with #NVIDIA