Scaling AI with Dell PowerEdge XE9680

Enterprise AI Performance at Scale

Designed for AI at Scale

Dell PowerEdge XE9680 Cluster | 64x NVIDIA H200 GPU 141GB HMB3e each | Broadcom BCM57608 Thor 2 400GbE NICs | Dell PowerSwitch Z9864F

The Dell PowerEdge XE9680 cluster features 64 NVIDIA H200 SXM GPUs deployed across 8 nodes, with each node equipped with 8 GPUs and 141GB of HBM3e memory per GPU delivering 4.8TB/s bandwidth. This enterprise-grade configuration includes dual Intel Xeon Platinum processors, 2TB of DDR5 system memory per node, and 16x 2.9TB NVMe SSDs for high-speed local data caching. The platform is optimized for both large-scale training of frontier models up to 1T+ parameters and high-throughput inference workloads requiring exceptional memory capacity and bandwidth.

Network connectivity is provided through Broadcom BCM57608 Thor 2 400GbE NICs with hardware-accelerated collective operations, integrated into a RAIL-Optimized topology using Dell PowerSwitch Z9864F switches powered by Broadcom Tomahawk 5 ASICs. This architecture organizes GPUs by their local rank into dedicated "rails," enabling single-hop communication for the same-rank traffic patterns that dominate distributed training, while maintaining spine connectivity for cross-rail communications. The result is 97.3% network efficiency under peak AI workloads with sub-microsecond congestion response times.

Network Architecture Excellence

97.3%

Network Efficiency

Sustained under peak AI workloads.

96.8%

All-Gather Efficiency

Theoretical bandwidth utilization.

Training Performance: Llama 3 405B Scaling

Configuration	GPUs	TOPS Aggregate	Scaling Efficiency	Tokens/Second
Single Node	8	15,832 TOPS	Baseline	4,267
2 Nodes	16	30,844 TOPS	97.3%	8,312
4 Nodes	32	60,156 TOPS	95.1%	16,387
8 Nodes	64	117,248 TOPS	93.2%	32,156

Optimized for Enterprise AI Workloads

Large Language Model Training

Train frontier models up to 1T+ parameters with 141GB memory per GPU. Support for Llama, GPT, and transformer architectures with FP8 acceleration.

High-Performance Inference

1,979 TOPS INT8 performance delivers production-grade inference with sub-25ms P99 latency for 70B models at scale.

Computer Vision at Scale

Process 2.8M+ images per second with ResNet-50. Ideal for object detection, classification, and segmentation tasks.

Mixture-of-Experts Models

Sparse transformer architectures with dynamic expert routing benefit from optimized all-to-all communication patterns.

93.2%

Scaling efficiency at 64-GPU cluster scale.

Maintaining exceptional performance across distributed training.

Business Impact & ROI

19-23% TCO Savings

Three-year total cost advantages through optimized acquisition, 20% superior power efficiency, and unified management reducing operational overhead.

15-46% Faster Time-to-Market

Deploy production AI applications faster with validated case studies showing 8-12 week acceleration in bringing AI-powered products to market.

94% GPU Utilization

Maximum return on infrastructure investment with 93% scaling efficiency across 64+ accelerators and unified Dell management tools.

99.7% Uptime

Enterprise support eliminates $2.3M average annual cost of AI infrastructure expertise while delivering reliability through proactive monitoring.

Future-Proof Enterprise AI Infrastructure

Purpose-built for mission-critical AI workloads.

Delivering measurable business outcomes and sustained competitive advantage.

For more, see the full report on the Signal65 website.

In partnership with: Dell Technologies