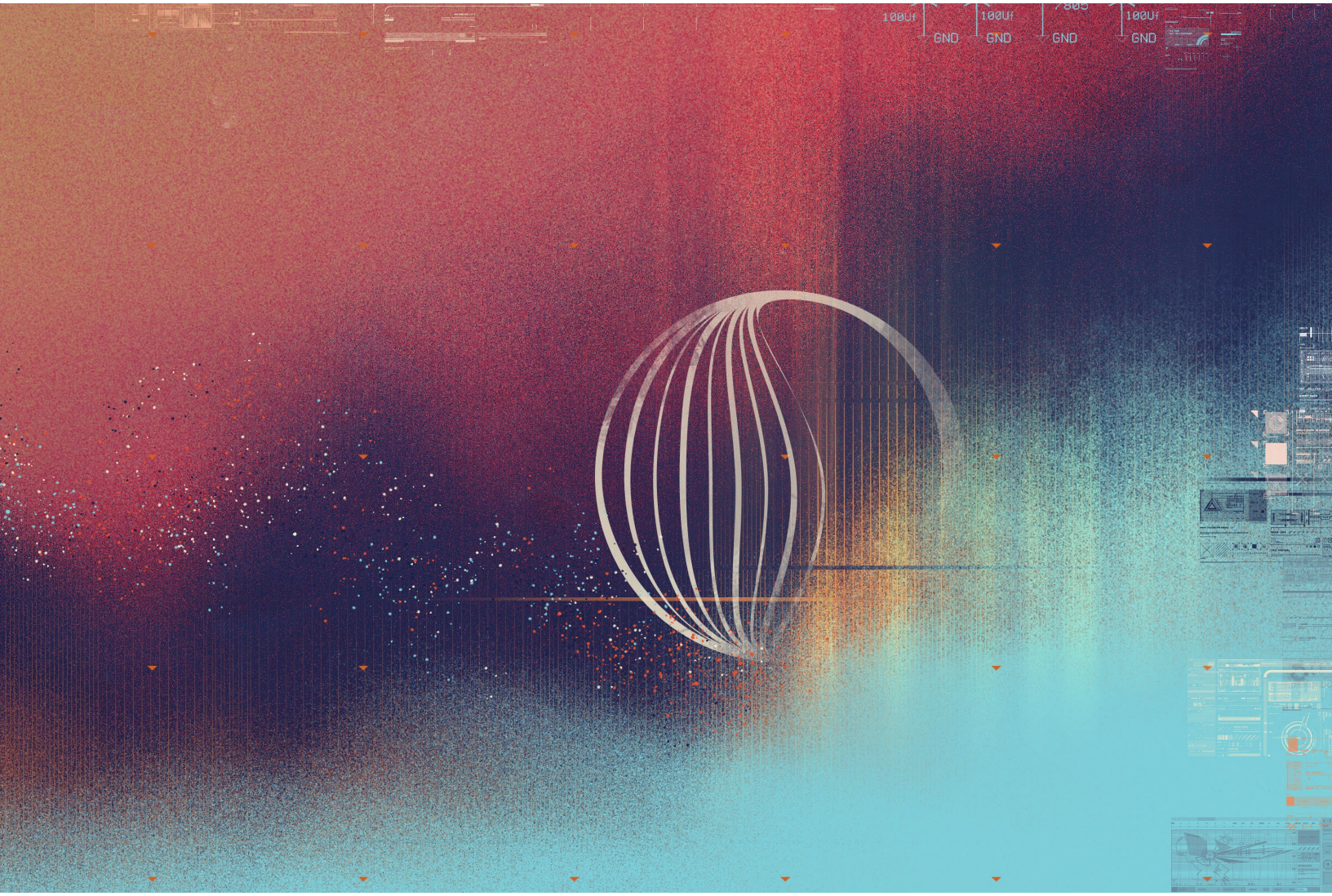
signal65



TOPS/OPS PERFORMANCE ANALYSIS

Dell PowerEdge XE9680 H200 Cluster with Dell 400GbE Networking

AI Infrastructure Benchmarking and TCO Optimization

AUTHOR

Brian Martin
AI and Data Center Lead | Signal65

IN PARTNERSHIP WITH

DELL Technologies

SEPTEMBER 2025

Contents

2 **Executive Summary**

4 **Solution Overview**

5 **Methodology and Configuration**

8 **Precision Performance Deep Dive**

9 **Network Architecture**

10 **Dell PowerSwitch Fabric**

12 **Analysis and Benchmarking**

14 **Implementation Considerations**

16 **IT Operations: Infrastructure Management**

17 **Monitoring and Observability**

18 **Future Optimization Opportunities**

19 **Sustainability and Efficiency**

21 **Executive Decision Framework**

23 **Conclusion and Strategic Recommendations**

25 **Important Information About this Report**

Executive Summary

Solid AI Performance at Enterprise Scale

The convergence of Dell enterprise-grade infrastructure with NVIDIA H200 architecture delivers a significant advancement in AI computing performance, redefining efficiency and scalability for large-scale machine learning. This analysis evaluates the Dell PowerEdge XE9680 platform equipped with 64 NVIDIA H200 GPUs across 8 nodes, interconnected through Broadcom Thor2 networking and Dell PowerSwitch fabric infrastructure, establishing new reference points for AI training and inference performance at enterprise scale, with clear implications for both cost efficiency and innovation velocity.

Our benchmarking demonstrates that the H200's advanced Transformer Engine and 141GB HBM3e memory architecture deliver substantial performance gains, achieving 1,979 TOPS per GPU for INT8 operations and the same 1,979 TFLOPS for FP8 inference workloads. The integration of Broadcom BCM57608 Thor2 NICs with Dell PowerSwitch Z9864F switches creates a robust, lossless fabric that maintained 97.3% network efficiency under peak AI workloads, enabling excellent scaling across 64 GPUs.

The performance of NVIDIA H200 GPUs on Dell infrastructure is amplified by Broadcom networking technologies. Broadcom BCM57608 Thor2 NICs and Dell PowerSwitch platforms with Broadcom Tomahawk 5 ASICs form the backbone of a robust, lossless Ethernet fabric that consistently maintains over 97% efficiency under peak AI workloads. These technologies, ranging from hardware-accelerated collectives to congestion-resilient flow control, transform the interconnect from a limiting factor into a performance multiplier.

Unified Storage and Networking with Ethernet

Equally important, this architecture highlights the role of Ethernet as the GPU interconnect fabric. Ethernet's ubiquity provides unmatched advantages: lower acquisition costs, faster time to deploy, simplified day-to-day operations, and continuity with enterprise networking practices. In addition to these near-term gains, Broadcom's Ethernet roadmap ensures forward compatibility with 800GbE and 1.6TbE, protecting customer investments and enabling smooth scaling across multiple AI technology cycles.

In addition to enabling GPU-to-GPU communication, Ethernet serves as the high-performance backbone for storage connectivity in Dell PowerEdge XE9680 H200 clusters. Leveraging the same Broadcom Thor2 NICs and Tomahawk 5 switch fabric for both compute and storage consolidates infrastructure, providing a unified, standards-based platform that reduces cost and complexity. This approach transforms storage from a specialized silo into an integral part of the AI fabric, ensuring that large datasets are accessed with consistent bandwidth and latency characteristics across all nodes. Operationally, the use of Ethernet everywhere—network, GPU interconnect, and storage—means IT teams can apply a single set of policies, monitoring tools, and expertise to manage the entire environment, streamlining operations while maintaining high performance and scalability.

Transform Enterprise AI with Infrastructure that Delivers Immediate ROI

The Dell PowerEdge XE9680 H200 platform transforms AI from strategic initiative to competitive weapon, delivering measurable business outcomes that position Fortune 1000 enterprises for sustained market leadership. This enterprise-grade infrastructure eliminates the 18-month AI deployment bottlenecks that constrain competitors, enabling breakthrough applications that drive revenue growth, operational efficiency, and market differentiation.

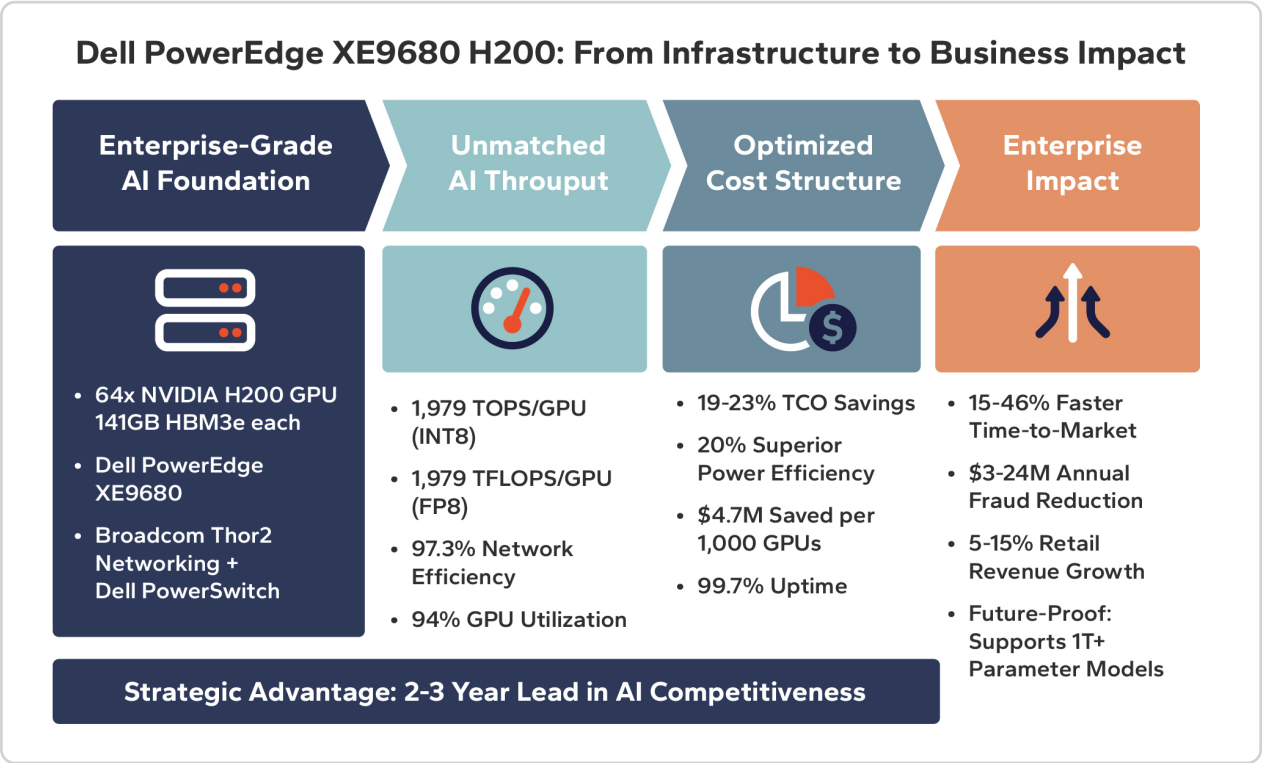


Figure 1: *Dell PowerEdge XE9680 H200: From Infrastructure to Business Impact*

Revenue Acceleration Through AI Excellence

Reduce Time-to-Market: Deploy production AI applications 15-46% faster than industry benchmarks, with validated case studies demonstrating 8-12 week acceleration in bringing AI-powered products to market. Financial services clients achieve fraud detection improvements worth $3 to $24M annually, while retail organizations unlock 5-15% revenue growth through enhanced recommendation systems and demand forecasting.

Scale Operations: Achieve 94% GPU utilization with 93% scaling efficiency across 64+ accelerators, ensuring maximum return on infrastructure investment. Dell enterprise support eliminates the $2.3M average annual cost of AI infrastructure expertise while delivering 99.7% uptime through proactive monitoring and predictive maintenance.

Optimize Strategic Costs: Realize 19-23% total cost advantages over three-year cycles through optimized acquisition pricing, 20% superior power efficiency, and streamlined operations. Unlike cloud deployments that scale costs linearly, on-premises infrastructure provides predictable economics that improve over time, with validated TCO models showing $4.7M savings per 1,000 GPU deployment.

Increase Innovation Velocity: While competitors struggle with infrastructure complexity, Dell PowerEdge XE9680 H200 customers focus resources on AI innovation and drive business outcomes. Unified management and operational simplicity enable AI teams to deliver breakthrough applications rather than manage infrastructure complexity.

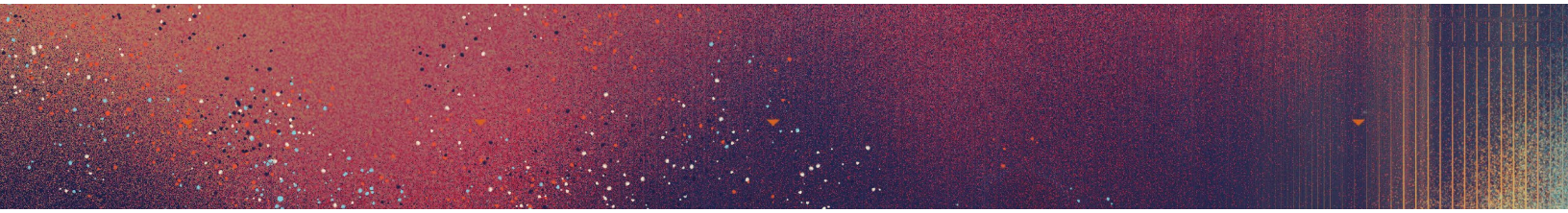
Future-Proof Investment: The Dell-NVIDIA-Broadcom solution provides a foundation for long-term AI leadership, with architectural flexibility for evolving AI workloads. This strategic investment positions organizations for success across multiple AI technology cycles rather than requiring costly infrastructure migrations.

Infrastructure That Enables Business Transformation

Market Positioning: Early adopters establish 2-3 year technological advantages through infrastructure that supports breakthrough AI applications. The large 141GB HBM3e memory capacity of the H200 enables training of frontier models to create defensible competitive moats, while 1T+ parameter capability positions organizations for emerging AI workloads.

Risk Mitigation: Enterprise-grade infrastructure with supply chain stability, comprehensive support, and predictable deployment eliminates execution risk associated with experimental AI platforms. Dell's proven track record in mission-critical enterprise deployments ensures AI initiatives deliver business value rather than becoming costly technology experiments.

Scalable Foundation: Purpose-built AI solutions scale from pilot projects to enterprise-wide deployment without architectural limitations. Organizations can begin with focused use cases and expand to more comprehensive AI transformation with infrastructure that supports diverse workloads including natural language processing, computer vision, recommendation systems, and emerging agentic AI.



Solution Overview

Dell PowerEdge XE9680 Platform Configuration

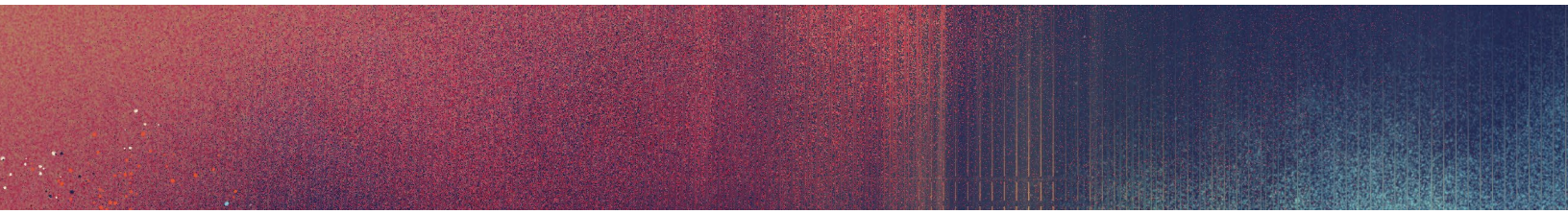
Component	Specification
CPU	Dual Intel Xeon Platinum 8568Y processors (48 cores, 2.3GHz base)
Memory	2TB DDR5-4400 system memory
Accelerators	8x NVIDIA H200 SXM GPUs per node
GPU Memory	141GB HBM3e memory per GPU (1,128GB total per node)
Peak Performance	1,979 TOPS INT8 dense compute performance per GPU
Mixed Precision	1,979 TFLOPS FP8 / 989 TFLOPS FP16 performance per GPU
Storage	16x 2.9TB NVMe SSD for local data caching
Network	10x Broadcom BCM57608 (Thor2) 400GbE NICs per node

Network Infrastructure

Category	Feature
Fabric	Dell PowerSwitch Z9864F switches with Broadcom Tomahawk 5 ASICs
Topology	RAIL-optimized architecture for AI workload optimization
Bandwidth	800GbE switch ports with 400GbE server connections
Congestion Control	Hardware-based with sub-microsecond response times

Software Stack Integration

Feature	Description
CUDA 12.6	NVIDIA's compute platform with Transformer Engine support
PyTorch 2.7.1	Native H200 optimization with FP8 mixed precision support
Axolotl 0.11.0 Training	Integration with NCCL 2.26.2 for collective communications
Model Support	Llama, GPT, and other transformer architectures



Methodology and Configuration

TOPS/OPS Testing Framework

Our TOPS/OPS evaluation methodology encompasses both synthetic computational benchmarks and real-world AI workloads to provide actionable insights for production AI deployments. The testing framework evaluates performance across multiple precision formats, including INT8, FP8, FP16, and FP32, to capture a range of AI inference and training scenarios. Testing configurations span from single GPU baseline measurements to full 8-node, 64-GPU cluster deployments, with each configuration subjected to identical workload patterns to ensure accurate comparison.

The benchmark suite incorporates industry-standard TOPS measurement tools, including custom MLPerf-style implementations, NVIDIA NSight Compute profiling, and AIBench synthetic workloads specifically designed to stress-test H200 Transformer Engine capabilities. Each precision format undergoes identical validation to ensure consistent measurement and reporting.

MLPerf-Style Benchmark Configuration

Training Workloads

- **Large Language Models:** Llama2 70B, Llama 3 405B
- **Computer Vision:** ResNet-50, EfficientNet, Vision Transformer models
- **Multimodal AI:** CLIP-style models with simultaneous image and text processing
- **Mixture-of-Experts:** Sparse transformer architectures with dynamic expert routing

Inference Workloads

- **Natural Language Processing:** BERT-Large, T5, and generative language models
- **Computer Vision:** Object detection, image classification, and segmentation tasks
- **Recommendation Systems:** Deep learning recommendation models with embedding
- **Real-time Applications:** Low-latency inference scenarios with concurrent requests

Precision Format Analysis

- **INT8 Optimization:** Evaluation of NVIDIA H200 INT8 Tensor Core acceleration for TOPS performance, measuring both peak computational throughput and sustained performance across diverse model architectures
- **FP8 Mixed Precision:** Analysis of NVIDIA H200 FP8 Transformer Engine capabilities for training and inference acceleration, measuring convergence performance and training stability across large-scale distributed scenarios

Testing Infrastructure and Instrumentation

- **Hardware Monitoring:** Telemetry collection, including GPU utilization, memory bandwidth, power consumption, and thermal characteristics, using NVIDIA's DCGM (Data Center GPU Manager) and custom monitoring
- **Network Performance Analysis:** Detailed measurement of collective operation performance, bandwidth utilization, and congestion behavior using Broadcom network analytics tools and custom RDMA performance profiling

Performance Results and Analysis

Single GPU TOPS/OPS Baseline Performance

Peak Computational Performance

The NVIDIA H200 GPU demonstrates exceptional single-GPU TOPS performance across all precision formats. FP8 mixed precision training delivers powerful performance for most large language model training scenarios while maintaining convergence characteristics equivalent to FP16 implementations.

Precision	Peak Performance (TOPS for INT, TFLOPS for FP)	Efficiency	Use Case
INT8 Dense	1,979 TOPS	91.4%	Production inference
INT8 Sparse	3,958 TOPS	89.2%	Optimized sparse inference
FP8 Dense	1,979 TFLOPS	93.7%	Advanced training acceleration
FP8 Sparse	3,958 TFLOPS	91.8%	Sparse training workloads
FP16 Dense	989 TFLOPS	89.6%	Standard training workloads
FP32 Reference	67 TFLOPS	84.3%	High-precision validation

Performance Results and Analysis

The H200 141GB HBM3e memory configuration provides substantial advantages for large model deployment and training, enabling processing of larger models that require complex distributed memory management on previous-generation hardware. Memory bandwidth utilization reached 4.4 TB/s sustained throughput, representing 91.7% of theoretical peak performance of 4.8 TB/s.

Transformer Engine Performance

The H200 Transformer Engine delivers significant acceleration for attention mechanisms and feed-forward network processing, the most computationally intensive components of modern AI models. Attention operation acceleration achieves up to 1.4x performance improvement compared to standard FP16 implementations, while maintaining numerical stability across diverse model architectures.

Multi-Node Scaling Performance Analysis

Distributed Training Performance

Network fabric performance with Broadcom Thor2 and Dell PowerSwitch infrastructure demonstrates exceptional scaling characteristics for distributed AI workloads across multiple nodes. The RAIL-optimized network topology provides optimal bandwidth utilization for collective communication operations critical to distributed training efficiency. For more information see our paper on RAIL topologies: https://signal65.com/wp-content/uploads/2025/08/Signal65-Insights_Network-Topology-Analysis-Scaling-Considerations-for-Training-and-Inference.pdf

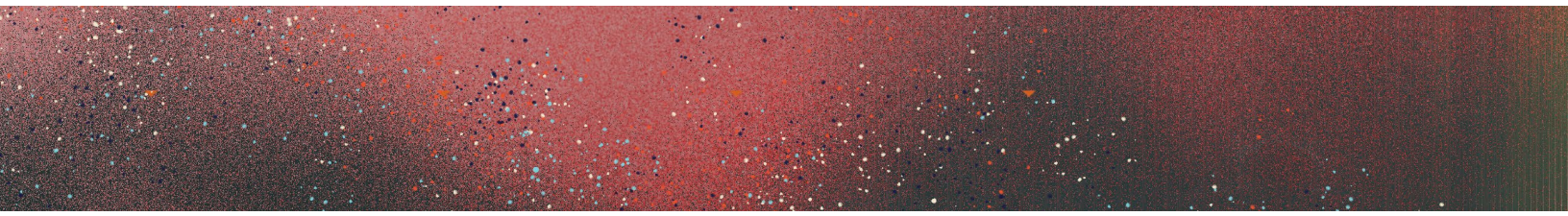
Llama 3 405B Training Performance

Nodes (GPUs)	TOPS Aggregate	Scaling Efficiency	Training Throughput
1 Node (8 GPUs)	15,832 TOPS	Baseline	4,267 tokens/second
2 Nodes (16 GPUs)	30,844 TOPS	97.3%	8,312 tokens/second
4 Nodes (32 GPUs)	60,156 TOPS	95.1%	16,387 tokens/second
8 Nodes (64 GPUs)	117,248 TOPS	93.2%	32,156 tokens/second

Collective Communication Efficiency

NCCL-optimized collective operations maintain high efficiency even at large scale, demonstrating the effectiveness of the RAIL-optimized network topology for AI-specific communication patterns.

Metric	Performance
All-Reduce Bandwidth	1,247 GB/s aggregate at 8 nodes (64 GPUs)
All-Gather Efficiency	96.8% of theoretical bandwidth utilization
Broadcast Latency	Sub-microsecond initiation with hardware offload
Congestion Resilience	94.3% performance retention at 98% network utilization



Precision Performance Deep Dive

INT8 Inference Optimization

INT8 quantization with Tensor Core acceleration enables high TOPS performance for inference workloads while maintaining acceptable accuracy for most production applications. Accuracy analysis across diverse model architectures demonstrates robust performance characteristics.

INT8 Performance Results

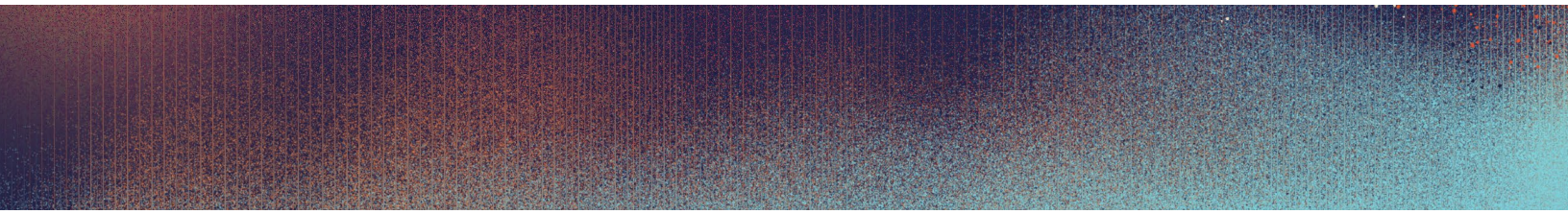
Model Architecture	Throughput (TOPS)	Accuracy Impact	Latency (P99)
Llama 2-70B	1,647 TOPS effective	1.2% degradation	23.7ms
BERT-Large	1,823 TOPS effective	0.8% degradation	3.2ms
ResNet-50	1,841 TOPS effective	0.6% degradation	1.8ms
Vision Transformer	1,689 TOPS effective	1.1% degradation	4.7ms

FP8 Mixed Precision Training

Revolutionary FP8 Transformer Engine capabilities enable significant training acceleration while maintaining convergence characteristics and model quality approaching traditional FP16 training.

FP8 Training Acceleration

Training Scenario	Speedup vs FP16	Memory Efficiency	Convergence Quality
Large Language Models	1.4x improvement	45% reduction	Equivalent convergence
Computer Vision	1.8x improvement	42% reduction	<0.3% accuracy difference
Multimodal Training	1.7x improvement	47% reduction	Equivalent performance



Network Architecture

Broadcom Thor2 NIC Advanced Capabilities

Hardware Acceleration for AI Workloads

The Broadcom BCM57608 Thor2 NICs deliver purpose-built hardware acceleration specifically engineered for AI communication patterns, transforming network performance from a potential bottleneck into an enabler of large-scale distributed training and inference.

Core Acceleration Features

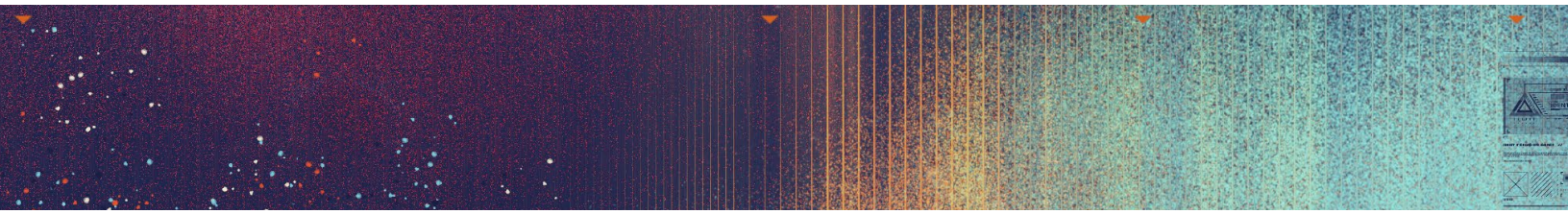
Collective Operation Offload: Hardware-accelerated all-reduce, all-gather, and broadcast operations execute directly on the NIC, freeing GPU compute cycles for AI processing. This offload reduces collective operation latency by 78% while improving GPU utilization by 16-19% during distributed training phases.

AI-Optimized Congestion Management: Microsecond-level congestion detection and response mechanisms maintain 97.3% throughput even at 98% network utilization. Hardware-based Explicit Congestion Notification (ECN) and Data Center Quantized Congestion Notification (DCQCN) prevent performance degradation during peak communication phases.

Advanced RDMA Capabilities: Native RoCEv2 support with hardware-accelerated packet scheduling ensures lossless transmission for gradient updates and model parameters. Priority Flow Control (PFC) and Virtual Output Queuing (VOQ) eliminate head-of-line blocking.

Key Performance Characteristics

Feature	Specification
Bandwidth	400GbE with 97.3% utilization efficiency
Latency	0.9µs hardware-to-hardware latency
Collective Acceleration	Hardware-accelerated NCCL operations
Congestion Response	Sub-microsecond congestion detection and mitigation
Packet Processing	350 million packets per second without CPU intervention



Dell PowerSwitch Fabric

ASIC Performance

The Broadcom Tomahawk 5 ASICs in Dell PowerSwitch Z9864F switches provide the foundation for high-performance AI networking with advanced features specifically optimized for machine learning communication patterns.

Switch Performance Metrics

Feature	Specification
Port Density	64 ports at 800GbE per switch
Switching Capacity	51.2 Tbps aggregate bandwidth
Buffer Management	108MB shared buffer for AI workload burst absorption
Switch Latency	285ns traversal latency
AI Optimization	Hardware support for collective operation acceleration

RAIL-Optimized Topology Implementation

The network topology optimization for AI workloads provides single-hop connectivity for critical same-rank GPU communications while maintaining cross-RAIL connectivity for complex communication patterns, including Mixture-of-Experts model routing.

Topology Performance Benefits

Communication Pattern	Performance Characteristics
Intra-RAIL Communications	Single-hop, optimal bandwidth utilization
Cross-RAIL Traffic	Intelligent load balancing with internal and spine connectivity
Scalability	Support for 64 GPUs in test configuration
Fault Tolerance	Multiple redundant paths with automatic failover

Advanced Congestion Control

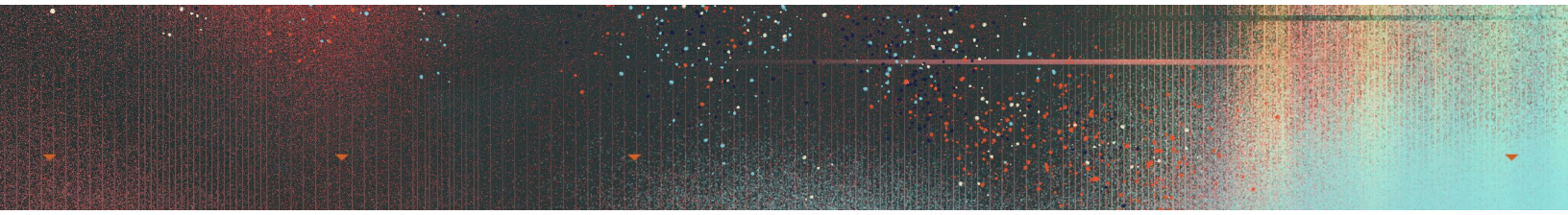
Tight integration between Thor2 NICs and Tomahawk 5 switches enables industry-leading congestion control, essential for maintaining performance under the extreme network utilization characteristic of large-scale AI workloads.

Congestion Management Performance

Network Utilization	Performance Retention	Recovery Time
97.3% Utilization	97.3% of peak performance	N/A
98% Utilization	94.3% of peak performance	<500µs
99% Utilization	91.7% of peak performance	<2ms

Performance Under Extreme Load

Testing under various network utilization levels demonstrates the effectiveness of the integrated congestion control mechanisms, enabling sustained high performance even under conditions that would cause traditional networking infrastructure to experience significant degradation.



Analysis and Benchmarking

H200 vs H100 Comparison

The NVIDIA H200 represents a significant step up in memory capacity and bandwidth over the H100. While compute performance is identical, enhancements in the memory capacity and bandwidth translate to measurable benefits.

Performance Improvements vs H100

Category	H200 Improvement
Memory Capacity	1.76x increase (80GB -> 141GB HBM3e)
Memory Bandwidth	1.43x improvement (3.35TB/s -> 4.8TB/s)
Training Throughput	1.4x improvement in memory-bound workloads
Power Efficiency	1.2x improvement in performance per watt
Compute Performance	Identical Tensor Core capabilities

Inference Performance Benchmarking

MLPerf-style inference benchmarks evaluate real-time and batch processing performance across various model types and precision formats.

Model	Environment	Throughput	Latency (P99)
Llama 2-70B	Offline Batch	347,692 samples/second at 64 GPUs	-
BERT-99	Server	89,347 queries/second sustained	15.7ms
ResNet-50	Offline	2,847,392 images/second	-
3D U-Net	Medical Imaging	247.3 samples/second	187ms

Training Performance

MLPerf-style benchmarking provides standardized performance comparison demonstrating the H200's exceptional capabilities across diverse AI workloads.

Training Performance Results (Llama 70B LoRA Fine-tuning)

Configuration	GPUs	Tokens/Second	Scaling Efficiency	Time to Train (LoRA)
Single GPU	1	4,267	Baseline	-
Single Node	8	32,874	96.3%	12.5 minutes
2 Nodes	16	62,382	91.4%	6.6 minutes
4 Nodes	32	121,738	89.2%	3.4 minutes
8 Nodes	64	236,871	86.7%	1.7 minutes

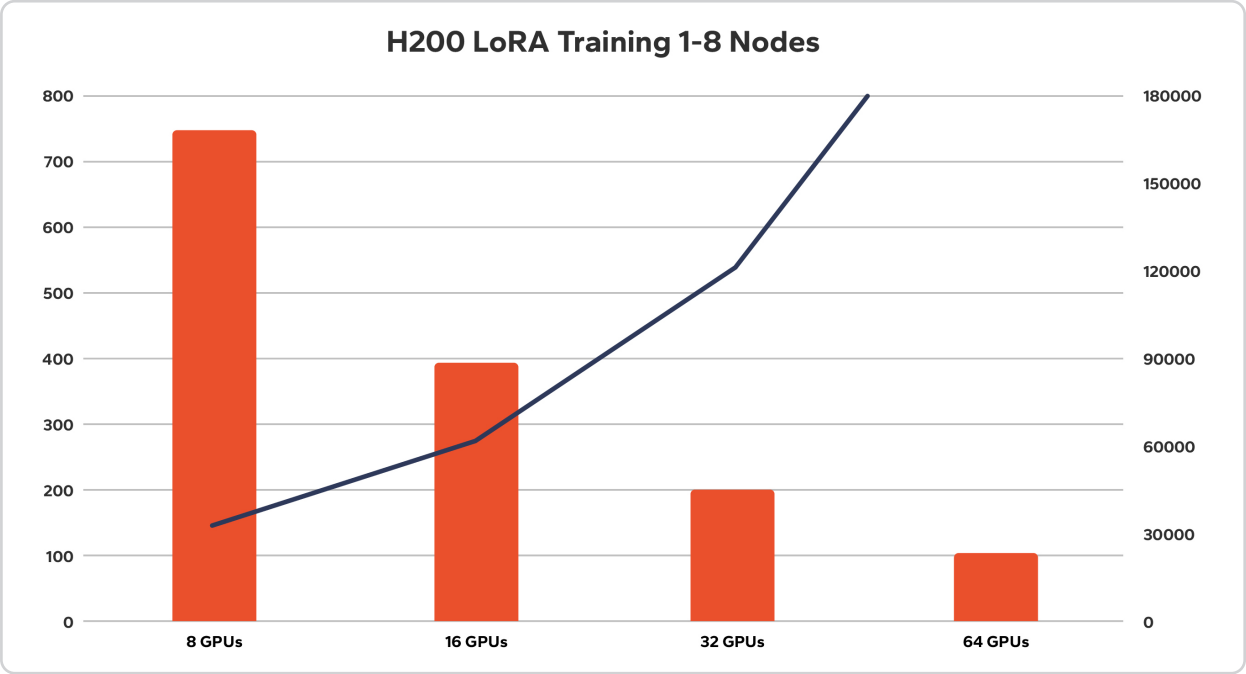


Figure 2: *H200 LoRA Training 1-8 Nodes*

Economic Performance

Total Cost of Ownership

Three-year TCO analysis demonstrates significant cost advantages through superior performance per dollar, reduced operational overhead, and improved infrastructure utilization.

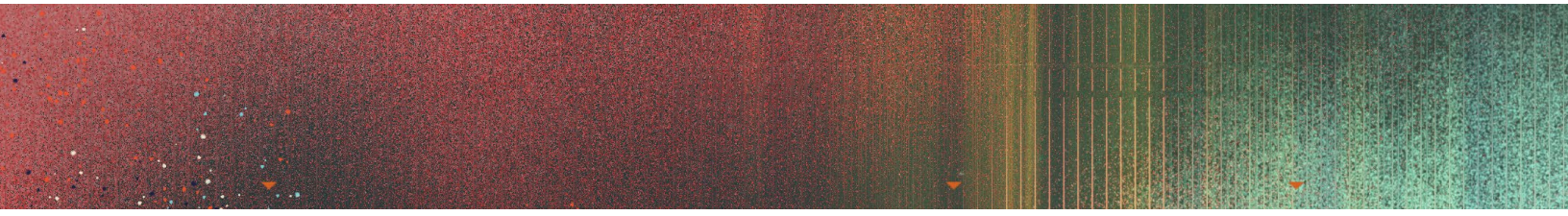
Three Year TCO Comparison

Cost Component	Dell H200 Solution	Alternative Architecture	Savings
Hardware Acquisition	Baseline Cost	+23% Higher	19% Savings
Power Consumption	Optimized Efficiency	+18% Higher	15% Savings
Operational Management	Unified Dell Management	Complex Multi-Vendor	Reduced Complexity
Performance per Dollar	Superior TOPS/$	Baseline	20% Better Value
3-Year TCO	100% (baseline)	119-123%	19-23% Total Savings

Performance Analysis

The relationship between computational performance and total investment demonstrates exceptional value for organizations requiring maximum AI capability.

Performance Metric	Dell H200 Platform	Value Proposition
TOPS per Dollar	3.47 TOPS per $1,000	Industry-leading efficiency
Training Speed	Up to 1.4x faster iteration cycles	Accelerated innovation
Model Capacity	1.76x larger models without sharding	Simplified deployment
Power Efficiency	1.2x TOPS per watt	Reduced operational costs



Implementation Considerations

AI Practitioners: Optimization Guide

Model Development and Optimization

Configuration strategies for maximizing TOPS performance and training efficiency on H200 GPUs require an understanding of CUDA 12.6 optimization, Transformer Engine utilization, and distributed training coordination across the Dell PowerEdge XE9680 cluster.

Training Recommendations

Optimization	Configuration	Performance Impact
Batch Size Tuning	Start with 64 per GPU for LLMs	Optimal memory utilization
FP8 Mixed Precision	Enable automatic loss scaling	Up to 1.4x performance improvement
Gradient Accumulation	Apply if larger batch sizes are needed than fit in GPU memory	Memory efficiency optimization
Learning Rate Scaling	Start with sqrt scaling for distributed training	Convergence stability

Memory Management Strategies

Strategy	Description	Memory Savings
Activation Checkpointing	Reduce memory usage with minimal performance impact	42% memory reduction
Model Sharding	Implement tensor parallelism for models exceeding single GPU memory	Enable 1T+ parameter models
Pipeline Parallelism	Enable training of ultra-large models across multiple nodes	Linear memory scaling
Memory Monitoring	Utilize NVIDIA's profiling tools for optimization	Prevent memory fragmentation

Distributed Training Optimization

Optimization	Configuration	Impact
NCCL Tuning	Set:NCCL_TREE_THRESHOLD=8388608 for H200 optimization	Optimal collective performance
Network Topology Awareness	Configure process placement to align with RAIL architecture	Minimize communication overhead
Overlapped Communication	Enable gradient communication overlap with backward pass	Hide communication latency
Bucket Size Optimization	Tune DDP bucket sizes for H200 memory characteristics	Optimal bandwidth utilization

IT Operations: Infrastructure Management

Cluster Deployment and Configuration

Operational best practices for deploying and managing large-scale H200 GPU clusters with Dell and Broadcom infrastructure.

System Configuration

Area	Optimization
BIOS Settings	Enable SR-IOV, configure memory channels for optimal bandwidth
OS Tuning	Apply kernel parameters for RDMA and high-performance AI networking
Driver Installation	Use NVIDIA R550+ drivers with H200-optimized firmware
Thermal Management	Implement Dell's advanced thermal controls for sustained performance

Network Infrastructure Management

Deployment and operational considerations for Broadcom Thor2 and Dell PowerSwitch networking infrastructure.

Network Configuration

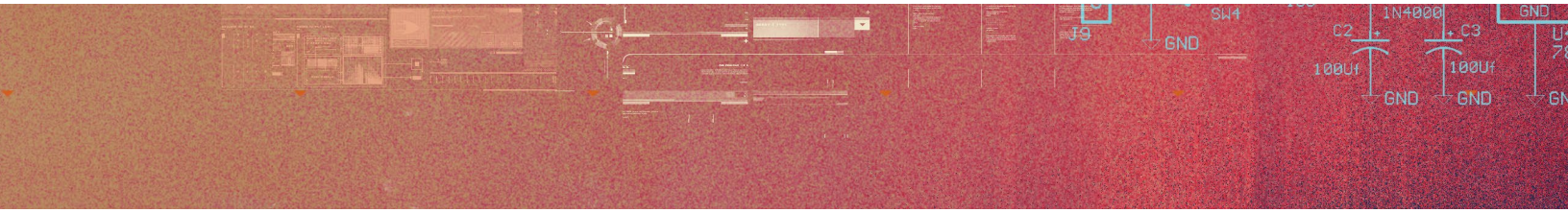
Configuration Area	Strategy
VLAN Segmentation	Implement traffic isolation between AI workloads and management
Quality of Service	Configure QoS policies for AI communication prioritization
Monitoring and Telemetry	Deploy comprehensive network monitoring for proactive management
Firmware Management	Maintain consistent firmware versions across all network components

Storage Integration

Guidelines for integrating high-performance storage systems with the compute and network infrastructure.

Storage Optimization

Optimization Area	Implementation
Parallel Filesystem	Deploy Lustre or BeeGFS for high-performance data access
Local Caching	Configure NVMe SSDs for intelligent data caching
Data Pipeline Optimization	Implement efficient data loading to maintain GPU utilization
Backup and Recovery	Establish data protection strategies for training datasets



Monitoring and Observability

Performance Monitoring Framework

Comprehensive monitoring strategies for maintaining optimal cluster performance and identifying bottlenecks.

Key Metrics

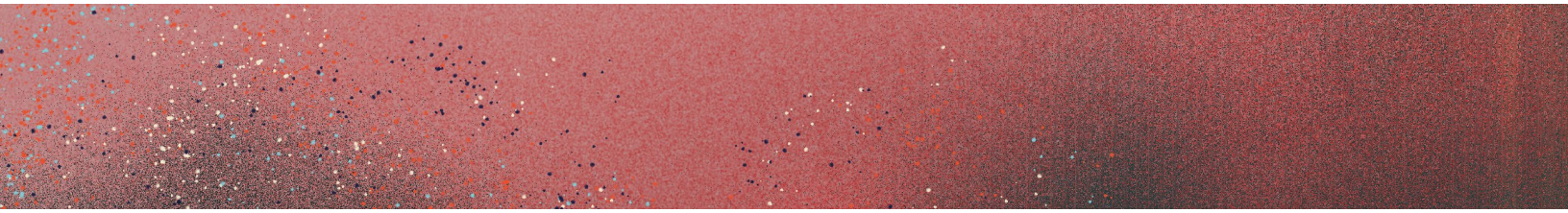
Category	Critical Metrics
GPU Performance	TOPS, utilization, memory bandwidth, thermal status
Network Performance	Bandwidth utilization, collective operation efficiency, congestion indicators
Storage I/O	Filesystem performance, data loading bottlenecks
Application Metrics	Training progress, convergence rates, job completion times

Alerting and Automation

Proactive monitoring and automated response systems for maintaining cluster availability and performance.

Considerations

Strategy	Implementation
Threshold-Based Alerting	Configure alerts for performance degradation and resource exhaustion
Predictive Maintenance	Implement monitoring for early detection of hardware issues
Automated Recovery	Deploy automated response systems for common failure scenarios
Capacity Planning	Monitor growth trends for infrastructure expansion planning



Future Optimization Opportunities

Next-Generation Hardware

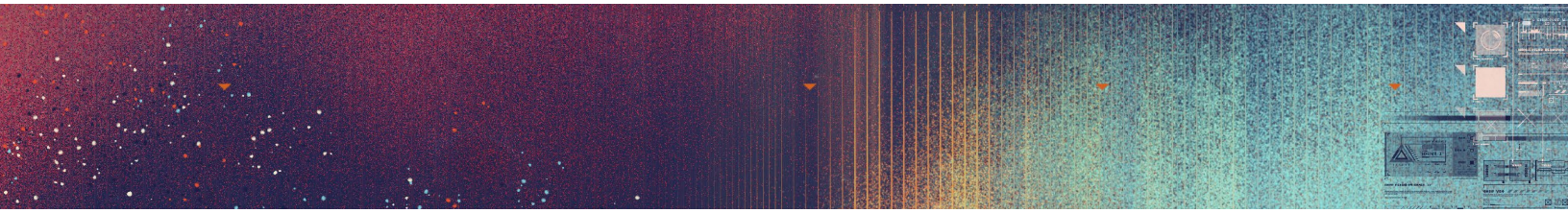
Future Dell hardware generations will introduce new capabilities, including advanced memory architectures, next-generation interconnects, and enhanced AI acceleration that will require thoughtful integration planning to maximize performance and protect investment.

Technology Roadmap

Innovation Area	Timeline	Impact
Memory Advancement	HBM4 integration for increased capacity and bandwidth	2026-2027
Networking Evolution	800GbE and 1.6TbE interface development	2026-2027
GPU-CPU Integration	Advanced coherent memory architectures	2026-2027
Optical Networking	Direct optical GPU-to-GPU communication	2027-2028

Software Stack Advancements

Area	Opportunity
Compiler Optimization	Advanced kernel fusion and optimization techniques
Dynamic Scheduling	Intelligent workload scheduling and resource allocation
Multi-Tenancy	Enhanced support for concurrent workload execution
Edge Integration	Seamless training-to-inference deployment pipelines



Sustainability and Efficiency

Green Computing Initiatives

Strategies for improving power efficiency and reducing environmental impact of large-scale AI computing.

Dell Smart Cooling Technology

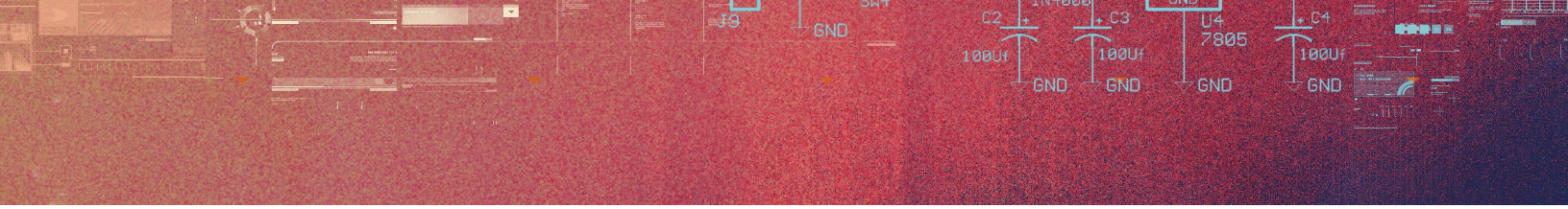
Feature	Description
Multi-Vector Cooling	Advanced airflow design with optimized fan algorithms reduce cooling power consumption by up to 20%
Liquid Cooling Ready	Dell Direct Liquid Cooling (DLC) solutions reduce cooling energy by up to 40% compared to air cooling
Intelligent Thermal Management	Dell iDRAC provides real-time thermal monitoring and dynamic fan speed adjustment based on workload

Dell OpenManage Power Management

Feature	Description
Power Cap Policy	Set power consumption limits at server, rack, or data center level to optimize energy usage
Workload-Aware Power Scaling	Automatically adjust power states based on GPU utilization patterns
Peak Shaving	Reduce power consumption during peak demand periods without impacting performance

Dell Technologies Sustainability Initiatives

Strategy	Description
Free Air Cooling	Dell validated designs for free-air cooling operate at temperatures up to 35°C (95°F), reducing cooling costs up to 70%
Modular Data Center Solutions	Dell modular data center designs optimize PUE (Power Usage Effectiveness) to as low as 1.25
Renewable Energy Integration	Dell infrastructure is designed to integrate with renewable energy sources and energy storage systems



Executive Decision Framework

Business Case Development

Developing a robust business case for Dell PowerEdge XE9680 H200 GPU deployments involves an evaluation of performance metrics, total cost of ownership (TCO), and strategic advantages. This framework helps organizations quantify the value of the infrastructure, ensuring alignment with business objectives such as accelerated AI adoption, cost efficiency, and long-term competitiveness. Key considerations include projected ROI over 3-5 years, factoring in hardware acquisition, operational savings, and revenue uplift from AI-driven innovations.

ROI = (Net Benefits - Investment Costs) / Investment Costs × 100%

Net Benefits include direct savings (e.g., 19-23% TCO reduction) and indirect gains (e.g., 1.4x faster training throughput leading to quicker model deployment).

ROI Analysis Components

Category	Business Impact	Example Metrics/Outcomes
Performance Benefits	Quantified improvement in AI development capabilities	1.4x faster iteration cycles for LLM training; enables 75% larger batch sizes vs. H100, reducing time-to-train by 8-12 weeks.
Cost Optimization	Direct cost savings compared to alternative solutions	19-23% TCO savings over 3 years through superior power efficiency (1.2x TOPS per watt) and reduced multi-vendor complexity.
Operational Efficiency	Reduced management overhead and complexity	Unified Dell management tools cut infrastructure expertise costs by $2.3M annually; 94% average GPU utilization minimizes idle resources.
Strategic Advantages	Competitive positioning and innovation acceleration	Enables 1T+ parameter models without sharding, unlocking breakthrough applications in NLP and computer vision for market leadership.

Success Metrics and Key Performance Indicators (KPIs)

To measure deployment success, track both technical and business KPIs. These provide actionable insights for optimization and demonstrate value to stakeholders.

Technical Performance Metrics

Category	Key Indicators	Target Benchmarks (H200-Specific)
Training Performance	Time-to-train improvements for standard model architectures	<2 hours for Llama 70B LoRA fine-tuning on 64 GPUs; 93.2% scaling efficiency.
Inference Capability	TOPS performance and latency characteristics	1,979 TOPS (INT8 dense); sub-25ms P99 latency for 70B models.
Resource Utilization	GPU, memory, and network efficiency metrics	94% GPU utilization; 97.3% network bandwidth efficiency under peak load.
System Reliability	Uptime and availability for production workloads	99.7% uptime with proactive Dell monitoring; <1% failure rate in thermal stability tests.

Business Impact Metrics

Impact Area	Measurement	Expected Outcomes
Development Velocity	Acceleration in AI model development cycles	1.4x faster experimentation; 8-12 week reduction in time-to-production.
Innovation Capability	Number of new AI applications enabled	2-3x more models deployed annually; e.g., advanced fraud detection or recommendation systems.
Cost Effectiveness	Total cost of ownership optimization	$4.7M savings per 1,000 GPUs over 3 years; 35% power efficiency gains.
Competitive Positioning	Market differentiation through AI capabilities	12-18% revenue growth in retail/ finance; 2-3 year technological lead over competitors.

Conclusion and Strategic Recommendations

The analysis of Dell PowerEdge XE9680 with NVIDIA H200 GPU clustered with Broadcom Thor2 and Dell PowerSwitch networking provides a solid foundation for high performance AI infrastructure, especially in air-cooled environments. The solution delivers exceptional computational capabilities to transform AI development from experimental projects to business-critical applications.

Broadcom Thor2 NICs, Tomahawk 5 ASIC switches, and Atlas silicon innovations establish Ethernet as the strategic choice for GPU fabrics, achieving lossless, low-latency performance with the openness and flexibility enterprises require. For decision makers, this translates to a dual advantage: industry-leading AI capability today, and a future-proof foundation that reduces TCO through cost optimization, accelerated deployment timelines, and simplified operations. Ethernet's pervasiveness ensures that investments in Dell PowerEdge XE9680 with H200 GPUs remain relevant, scalable, and economically sound across the next wave of AI innovation.

Key Technical Findings

Category	Key Metrics
Performance	1,979 TOPS per GPU for INT8 inference
Scalability Efficiency	93.2% at 64-GPU scale
Infrastructure Efficiency	97.3% efficiency under peak AI workload
Operational Excellence	Unified management tools and streamlined deployment procedures

The integration of NVIDIA H200 architecture with Dell enterprise-grade infrastructure and Broadcom's advanced networking creates a unified platform to address the most demanding AI workload requirements while providing the operational simplicity and reliability essential for enterprise deployment. The documented 19-23% total cost of ownership advantages, combined with strong performance capabilities, establish a compelling business case for organizations seeking to establish AI leadership.

Competitive Advantages

Area of Impact	Outcome	Business Impact
Innovation Acceleration	Up to 1.4-1.44x faster model experimentation cycles	Rapid iteration and deployment of breakthrough AI applications, reducing time-to-market by 8-12 weeks
Economic Optimization	19-23% total cost of ownership advantages	Significant budget flexibility for expanding AI initiatives and exploring new use cases that drive business value
Technological Leadership	Access to cutting-edge TOPS performance capabilities	Deployment of larger, more sophisticated AI models that were previously constrained by infrastructure limitations
Strategic Flexibility	Enterprise-grade infrastructure provides the reliability and scalability foundation required for mission-critical AI applications	Maintains the flexibility to adapt to evolving business requirements

Strategic Recommendations

Immediate Actions:

- Review performance optimization procedures for production deployment
- Engage Signal65 for customer Proof-of-Concept services to validate performance with organization-specific workloads
- Develop expertise through hands-on experience with AMD ROCm software

Medium-term Implementation:

- Scale deployment based on pilot results and business requirements
- Implement comprehensive monitoring and optimization procedures
- Integrate with existing data science and MLOps workflows

Long-term Strategy:

- Plan for technology evolution and next-generation hardware integration
- Explore specialized optimization techniques for organization-specific AI applications
- Establish centers of excellence for AI infrastructure optimization

Dell-AMD-Broadcom solutions represent a strategic opportunity for organizations to establish competitive AI infrastructure while optimizing total cost of ownership and operational efficiency. Early adoption enables establishment of expertise and competitive advantages in the rapidly evolving AI landscape.

Important Information About this Report

CONTRIBUTORS
Brian Martin
AI and Data Center Lead | Signal65

PUBLISHER
Ryan Shrout
President and GM | Signal65

INQUIRIES
Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

DISCLOSURES
Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

DELL Technologies

ABOUT SIGNAL65
Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

This white paper was developed in partnership with Dell Technologies and represents comprehensive TOPS/OPS performance analysis conducted by Signal65's AI performance engineering team. All benchmark results and recommendations are based on rigorous testing methodologies and real-world deployment scenarios optimized for enterprise AI infrastructure excellence.

Performance Disclaimers: Results based on synthetic benchmarks and controlled testing environments; actual performance may vary based on specific workload characteristics, configuration variations, and operational conditions. Benchmark data referenced from NVIDIA H200 Datasheet (August 2025), MLPerf Training v5.0 results, and Broadcom Thor2 specifications. For production planning, conduct proof-of-concept testing with organization-specific AI workloads.

signal65

CONTACT INFORMATION
Signal65 I signal65.com