AI Network Topology Analysis

Scaling Considerations for Training & Inference

The Infrastructure Challenge

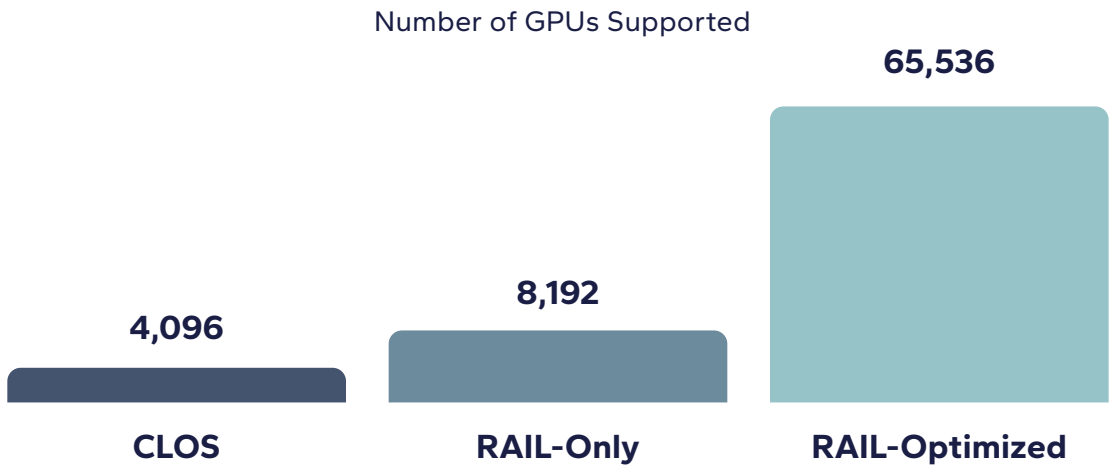
Modern AI workloads demand unprecedented network bandwidth.
Traditional CLOS architectures are being pushed beyond their breaking point, requiring purpose-built topologies for AI training and inference.

Three Network Topologies Compared

FLEXIBLE	COST EFFECTIVE	SCALABLE
Leaf-Spine	**RAIL-Only**	**RAIL-Optimized**
Connectivity: Full Any-to-Any	Connectivity: Optimized RAIL	Connectivity: Full with RAIL
Maximum Scale: 4,096 GPUs	Maximum Scale: 8,192 GPUs	Maximum Scale: 65,536 GPUs
Total Switches: 128 Switches	Total Switches: 8 Switches	Total Switches: 192 Switches
High Cost Low Scale	Low Cost High Scale	Medium Cost Very High Scale

GPU Scalability Comparison

Number of GPUs Supported



16x RAIL-Only delivers 2x more servers than CLOS with only 6% of the switches (8 vs 128).

Mixture-of-Experts (MoE) Impact

Massive Performance Gains

Llama 4 MoE models deliver 2-3x better tokens/sec than dense models of similar size, enabling faster inference and training.

Cost Optimization

With MoE models, the 5-10% performance trade-off of RAIL-Only vs RAIL-Optimized becomes economically justified by massive cost savings.

Better Resource Utilization

Sparse activation patterns mean fewer GPUs active per token, reducing power consumption and thermal load.

Dell Open Networking: Cost Advantages

70% Potential reduction in network infrastructure costs moving from CLOS to RAIL-Only architecture.

Fewer Switches Required

RAIL-Only uses 8 switches vs 128 for CLOS at comparable scale, dramatically reducing CapEx on switching hardware.

Lower Optics Costs

Optical transceivers account for most network costs. RAIL-Only requires far fewer interconnects between switches.

Reduced Power & Cooling

Fewer active network components translate to lower OpEx for power consumption and data center cooling requirements.

Simplified Managemen

Less complex topology means easier deployment, monitoring, and maintenance, reducing operational overhead.

Decision Framework

Choose RAIL-Only

when building dedicated AI infrastructure primarily for inference, MoE workloads, or when TCO optimization is critical for competitive advantage.

Consider RAIL-Optimized

when requiring maximum flexibility for experimental architectures, diverse legacy models, or extreme scalability for large-scale AI training.

Consider CLOS Networks

when building heterogeneous infrastructure with limited AI workloads or when simplified routing and management are preferred.

Ready to Optimize Your AI Infrastructure?

Purpose-built architectures that precisely align with modern ML workloads define the next generation of AI infrastructure.

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