Large-scale GPU cluster reference architecture with RoCE v2 networking

The convergence of NVIDIA's GB200/GB300 architectures with RoCE v2 Ethernet networking represents a fundamental shift in how hyperscale AI infrastructure is designed and deployed. [fb] Based on extensive analysis of production deployments including xAI's 100,000 GPU Colossus cluster [SemiAnalysis +2] and Meta's Research SuperCluster, [meta +2] this report provides comprehensive technical specifications and configuration parameters for building 10,000 to 100,000 GPU deployments using production-proven designs.

NVIDIA GB200 and GB300 architectural foundations

The **GB200 NVL72 system** forms the cornerstone of modern AI factories, packaging 36 Grace CPUs and 72 Blackwell B200 GPUs into a liquid-cooled rack consuming 120kW. [nvidia] Each rack delivers 1,440 PFLOPS FP4 performance with 130 TB/s total NVLink bandwidth, enabling 30x faster inference for trillion-parameter models compared to H100 systems. [nvidia +2] The architecture employs 18 compute trays housing dual GB200 Superchips, each combining 72-core Grace Arm processors with two Blackwell GPUs via **900 GB/s NVLink-C2C** interconnects. [Wccftech +3] Memory architecture provides 13.4 TB HBM3e plus 17 TB LPDDR5X per rack, with hardware decompression engines achieving 800 GB/s throughput to effectively triple memory bandwidth. [nvidia] [nvidia]

The **GB300 enhancement** arriving in late 2025 increases performance by 50% through architectural improvements including 288 GB HBM3e per GPU (versus 192 GB in GB200), 12-layer chip stacks, and enhanced tensor cores delivering 2x attention-layer acceleration. [nvidia +3] Power consumption rises to 142kW per rack, but the system achieves 10x tokens per second per user while improving performance per watt by 5x. [StockTitan +2] These systems scale through DGX SuperPOD architectures organizing 64 DGX systems into Scalable Units (SUs) of 512 GPUs, supporting deployments up to 64 SUs totaling 32,768 GPUs in a single administrative domain. [NVIDIA +3]

For the **target 1,024 node pod configuration**, the architecture comprises 16 SUs delivering 8,192 GPUs across 256 compute racks. This configuration requires 30.7-36.6 MW total power with 26-32 MW cooling capacity, implementing 128 Spectrum-4 SN5600 leaf switches and 64 Quantum-3 spine switches in a non-blocking topology. The NVLink domain configuration utilizes 16 NVL72 domains with 144 NVSwitch chips providing 2.08 PB/s aggregate NVLink bandwidth within the pod. [NVIDIA +2]

RoCE v2 networking architecture specifications

The RoCE v2 implementation for AI/ML clusters requires careful attention to lossless Ethernet configuration and congestion management. [Wikipedia +6] The **Clos network topology** for 1,024 nodes employs a two-tier design with 32-64 leaf switches and 16-32 spine switches, achieving the critical **≤1.2:1 oversubscription ratio** for RDMA traffic. [The Elegant Network] [Dell Technologies Info Hub] With 48-

port 400G leaf switches, the architecture requires 256 leaf switches handling 32 server-facing ports and 16 spine uplinks each, connecting to 64 spine switches for full fabric connectivity.

QoS mapping assigns RoCE v2 traffic to DSCP 26 (Traffic Class 3) as a lossless priority, with Congestion Notification Packets (CNP) using DSCP 48 (Traffic Class 7) at highest priority. World Wide Technology +2 Configuration implements dedicated queue assignments with 50% bandwidth allocation for RoCE traffic and strict priority queuing for CNP traffic. **Priority Flow Control (PFC)** operates on Traffic Class 3 with watchdog timers set to 200ms detection and 400ms recovery, preventing head-of-line blocking through storm protection mechanisms. juniper +3

Explicit Congestion Notification (ECN) marking begins at 10-25% buffer utilization with WRED curves configured for minimum threshold at 2,000 segments and maximum at 10,000 segments. Juniper Networks +2 The **DCQCN algorithm parameters** include Rp=50 Mbps for rate increase, Rai=5 Mbps for normal increase, and Gd=1/256 for multiplicative decrease, with CNP timers at 1-10ms intervals and rate recovery timers at 55ms. FS Community +3 **EVPN-MH** at the ToR level implements all-active redundancy mode with 10-byte Ethernet Segment Identifiers per ToR pair, using flow-based ECMP across ESI links for optimal load distribution. Nokia

Server NIC configurations and cabling architecture

The **NVIDIA ConnectX-7** provides dual 400GbE connectivity through models like MCX75310AAS-NEAT (single-port OSFP) and MCX75310AAS-NDAT (dual-port QSFP112), requiring PCIe Gen5 x16 interfaces. Juniper Networks nvidia These adapters deliver sub-microsecond latency with advanced RoCE support, hardware-based NVMe-oF acceleration, and GPUDirect Storage capabilities. NVIDIA Power consumption averages 15W with support for in-line encryption at full 400Gb/s line rate. Fibermall

The **ConnectX-8 SuperNIC** doubles bandwidth to 800Gb/s using PCIe Gen6 x16 primary interfaces plus optional auxiliary connections, totaling 48 PCIe lanes with integrated switching. Wccftech +2 Model C8180 provides single-port 800Gb/s while maintaining backward compatibility with 400Gb/s operation. NADDOD The integrated RISC-V data path accelerator and Spectrum-X switch capabilities position these NICs for next-generation deployments, ServeTheHome though at significantly higher power consumption requiring robust cooling infrastructure.

Cabling requirements differ substantially between speeds and distances. Direct Attach Copper (DAC) cables support 400GbE to 5m and 800GbE to 3m with lowest latency and cost, while Active Optical Cables (AOC) extend 400GbE reach to 100m typical (300m extended) and 800GbE from 0.5m to 2km depending on variant. Juniper Networks +2 OSFP transceivers consume 12-15W with integrated heat sinks supporting 36 ports per 1U (14.4Tbps), while QSFP-DD uses 7-12W in smaller form factors enabling higher port density but limiting thermal capacity. FiberMall +2

Leaf port calculations for the 1,024 node pod with dual 400GbE NICs require 2,048 400GbE ports total. Using NVIDIA Spectrum-4 SN5400 switches with 64 QSFP-DD ports delivering 25.6 Tbps, approximately 32 leaf switches accommodate server connections while reserving 25% of ports for spine uplinks, NVIDIA achieving 800 Tbps bisectional bandwidth at 2:1 to 3:1 oversubscription typical for AI workloads.

Network failure handling and high availability

Fast failover mechanisms leverage **Bidirectional Forwarding Detection (BFD)** with 300ms minimum intervals for centralized sessions and 100ms for distributed BFD, Juniper Networks juniper achieving sub-second detection when combined with hardware acceleration. Arista **LACP configuration** implements fast mode with 1-second timeouts and minimum link thresholds at 50% for redundancy, while **MLAG/VPC** architectures provide dual control planes eliminating single points of failure. Juniper Networks IBM Cloud's implementation demonstrates this with dual-port ConnectX-7 NICs achieving 3.2 Tbps aggregate throughput at 97% line rate through Virtual Rail architecture. ibm

Convergence targets achieve Layer 2 failover immediately through MLAG for dual-homed devices, while Layer 3 BGP convergence with optimized 3-second keepalive and 9-second hold timers plus BFD delivers sub-second detection. Arista +3 **RDMA-aware convergence** requires special consideration as NCCL operations exhibit extreme sensitivity to packet loss, demanding BGP Prefix Independent Convergence (PIC) for deterministic behavior in large clusters. nvidia The architecture achieves sub-100ms hardware-accelerated BFD detection Juniper Networks juniper with pre-computed backup paths enabling sub-50ms convergence.

Maintenance strategies implement In-Service Software Upgrade (ISSU) with zero data plane downtime and 50-90 second control plane interruption, reduced to 3 seconds with Enhanced ISSU. Cisco Blogs Retail News & More Graceful shutdown procedures utilize BGP community 65535:0 for automated traffic drainage while preserving forwarding state during control plane restarts. Juniper Networks Arista **Dual A/B fabric designs** provide complete redundancy with independent control planes, separate power and cooling systems, and isolated management networks. IBM's Virtual Rail architecture demonstrates active-active operation achieving full bandwidth utilization with dynamic flow redistribution on congestion detection. ibm

L3/ECMP routing configurations support up to 64-128 parallel paths with BGP multipath and add-path for diversity. FRRouting Cisco However, traditional 5-tuple hashing creates imbalance with RDMA traffic patterns, necessitating **flowlet switching** that breaks flows into sub-flows based on 16μs packet gaps, achieving near-perfect distribution. fb Each flowlet independently selects ECMP paths based on real-time utilization while maintaining packet ordering within flowlets. World Wide Technology +2

Hyperscale deployment best practices

Analysis of xAI's Colossus (100,000 H100s), SemiAnalysis +2 Meta's RSC (16,000 A100s), meta +2 and other hyperscale deployments reveals why **~1,024 nodes emerges as the optimal pod size.** fb SDxCentral This sweet spot balances multiple constraints: InfiniBand Quantum-2 switches require 4-

tier architecture beyond 1k nodes with 1.33x more transceivers; SemiAnalysis Substack single cooling zones typically support 500-1,000 high-density racks; and power distribution limits practical pods to 30-50MW matching datacenter building capacity. Network fabric costs increase substantially beyond 1k nodes as maintaining full bisection bandwidth becomes prohibitive, Substack while switch radix limitations create natural breakpoints favoring 3-tier over 4-tier designs.

Statistical failure analysis from Meta's 11-month study of 150M A100 GPU hours shows 6.50 failures per thousand node-days, with GPU failures accounting for 30.1% of training disruptions and HBM3 memory failures contributing 17.2%. arXiv Mean Time Between Errors improved from 0.88 hours for A100s to 2.2 hours for H100s, though node availability remains at 99.3-99.4%, arXiv **requiring 5% GPU overprovisioning** to achieve 99.9% job-level availability.

Blast radius containment implements rail-optimized designs where each GPU rail connects to different first-level switches, preventing cascade failures across the fabric. NVIDIA Job scheduling enforces single-pod-per-node allocation to prevent resource contention, with 60-minute checkpoint intervals and 5-minute restart overhead standard for large training jobs. Software resilience through TorchElastic and custom fault tolerance implementations at every major AI lab enables transparent GPU VM migration using CRIU technology.

Power and thermal constraints drive fundamental design decisions, with **H100 clusters requiring 150MW+ datacenter capacity** for 100k GPUs, Scale Computing consuming 1.59 TWh annually at $123.9M cost. AMAX +2 Liquid cooling becomes mandatory at these scales, Scale Computing with xAI implementing custom Supermicro 1U manifolds between servers and redundant pump systems per rack. Fibermall +3 Industry PUE averages 1.55, though hyperscalers achieve 1.09-1.1 through aggressive optimization, Google with 75% liquid cooling adoption reducing facility power by 18.1% and total datacenter consumption by 10.2%. Vertiv

Storage integration with VAST and NVMe-oF

VAST Data's Disaggregated and Shared Everything (DASE) architecture delivers 140+ GB/s throughput for AI workloads with linear scaling from petabytes to exabytes. The system implements NFS-over-RDMA with multipath support and NVIDIA GPUDirect Storage integration, using dual-homed servers connected to separate storage backend switches via 100GE ports with dynamic VIP allocation for load balancing. Juniper Networks The architecture combines C-Nodes handling storage services through VAST Server Containers with D-Nodes providing storage via DBox chassis, managed through RESTful APIs supporting automation at scale. Juniper Networks

NVMe-oF over RoCE v2 demonstrates substantial performance advantages with 41.22 µs latency for 4K QoS writes (66% lower than TCP) and 49% lower latency for sequential reads. westerndigital Red Hat Configuration requires lossless Ethernet with Data Center Bridging, Priority Flow Control mechanisms, and RDMA-capable NICs with ECN for congestion management. Western Digital Red Hat Queue depth supports up to 64K commands with 64K queues enabling multiple connections per namespace with dynamic provisioning and multi-path support. CodiLime

Tiered storage architecture implements multiple levels: Tier 1 local RAM for sub-millisecond active checkpoints, Tier 2 local NVMe for recent checkpoints under 1ms access, Tier 3 network storage for long-term retention, and Tier 4 object storage for archival. Checkpoint operations for 1 trillion parameter models require 273 GB/s write bandwidth and 1.64 TB/s read bandwidth, with checkpointing every 10-15 minutes during training. nvidia **NVIDIA Magnum IO** with GPUDirect Storage enables direct storage-to-GPU transfers achieving 2-8x throughput improvement by bypassing CPU bottlenecks. VAST Data +2

High availability and multi-tenancy configurations

The dual fabric A/B design implements complete physical separation with independent control and data planes, isolated failure domains, and separate power/cooling infrastructure. Each fabric operates in active-active mode for full bandwidth utilization with automatic failover achieving sub-second convergence through hardware-accelerated detection and pre-computed alternate paths. Cross-fabric connectivity enables controlled communication while maintaining isolation boundaries, supporting both steady-state load distribution and failure scenario capacity.

Network isolation for multi-tenancy leverages EVPN-VXLAN with Pure Type 5 architecture for server isolation, implementing IP-VRF per tenant with routed server links eliminating VLAN requirements. Nvidia +2 The VLAN-aware service model supports per-GPU tenant assignment through both MAC-VRF and IP-VRF per tenant with 8 IRB interfaces per server and distributed anycast gateways. Cisco +3 Security zones enforce intra-tenant east-west communication within VRFs while fusion firewalls control inter-tenant traffic, with border leafs providing external connectivity through VRF-lite peering and dedicated service VRFs enabling shared services access. Cisco

Production implementation roadmap

Deployment follows a phased approach beginning with 4-SU pilot installations (256 nodes, 2,048 GPUs) to validate configurations, scaling to 16-SU production (1,024 nodes, 8,192 GPUs) for initial workloads, then expanding to 100+ SUs for full-scale deployment. NVIDIA nvidia Infrastructure requirements include minimum 50 MW datacenter power for large deployments with 1.3x cooling capacity, 100+ Gbps inter-site backbone connectivity, and 50-100 square feet per rack including support infrastructure.

Operational excellence demands 24/7 monitoring with predictive maintenance capabilities, dedicated liquid cooling expertise with redundant systems, strategic vendor relationships for component supply chains, and NVIDIA Mission Control for unified cluster management. NVIDIA nvidia The total cost of ownership for GB200 NVL72 approaches $3M per rack with GB300 commanding 20-30% premiums, though delivering 2.5x performance per dollar improvement over previous generations. HPCwire Continuumlabs

This architecture enables organizations to build and operate AI factories at unprecedented scale, supporting the training and deployment of next-generation trillion-parameter models while maintaining the reliability, performance, and efficiency demanded by production AI workloads. [fb] The convergence on these design patterns across hyperscale deployments validates these specifications as the foundation for large-scale GPU infrastructure through 2025 and beyond.