NVIDIA INFINIBAND AI FABRIC MENTORED POV PLUS

Service Overview:

Netnology specializes in the design, deployment and enablement of NVIDIA Quantum InfiniBand fabrics, the high-performance, natively lossless networks that connect large GPU clusters. As part of this Mentored POV Plus service, our subject matter experts (SMEs) partner with your team to stand up an NDR (400 Gb/s) or XDR (800 Gb/s) InfiniBand fabric in a representative environment and prove its value against agreed success criteria (collective performance, congestion handling, visibility and ease of operation) before you commit to a production rollout.

Solution Overview:

Large-scale AI training and inference clusters place demands on the network that general-purpose fabrics were not designed for. GPU-to-GPU communication generates synchronized, high-bandwidth collective operations (all-reduce, all-gather, all-to-all) that are highly sensitive to latency, jitter and congestion. A single congested link or degraded cable can slow collectives across the entire GPU cluster, extending job completion time and reducing the return on the GPU investment.

NVIDIA Quantum InfiniBand is purpose-built for this traffic. It is a natively lossless, credit-based fabric, centrally managed by a subnet manager, available at 400 Gb/s (Quantum-2 NDR, 64 × 400 Gb/s switches) and 800 Gb/s (Quantum-X800 XDR, Q3200 and Q3400 switches with up to 144 × 800 Gb/s ports). Deployed as a non-blocking, rail-optimized fat tree, it combines adaptive routing and congestion control to keep large flows off congested paths, NVIDIA SHARP in-network computing to offload collective operations into the switches, and ConnectX-7 and ConnectX-8 adapters at the GPU servers. NVIDIA Unified Fabric Manager (UFM) provides fabric provisioning, subnet management, telemetry and day-2 monitoring.

Service Benefits:

Netnology has a team of world-class engineers who specialize in NVIDIA InfiniBand fabrics for AI and are passionate about customer success. This Mentored POV Plus engagement delivers:

- Evidence-based decisions: measurable fabric results against your own success criteria.
- Reduced risk: design, interoperability and performance validated with your GPU servers and ConnectX adapters before production.
- Production blueprint: POV design and findings feed directly into the production fabric design.
- Hands-on experience: your team configures and operates the fabric alongside Netnology SMEs.

Service Scope:

As part of the 5-day (up to 40 hours) engagement, Netnology will provide the following services:

- Fabric Bring-Up with NVIDIA UFM (up to 5 switches)
  - Deploy UFM (or validate the UFM appliance) and discover the fabric
  - Validate switch firmware, cabling and topology with ibdiagnet

- Subnet Management and Routing
  - Configure the subnet manager with high availability and the routing engine for the fat tree
  - Configure partitions (PKeys) and service levels, as required
- Performance Features
  - Enable adaptive routing and congestion control
  - Enable NVIDIA SHARP in-network computing
- GPU Server Fabric Integration (up to 3 nodes)
  - Configure ConnectX adapters (firmware, InfiniBand link type, DOCA-Host drivers)
  - Verify host-to-fabric connectivity and link health
- Telemetry and Visibility
  - Enable UFM telemetry, link health and congestion monitoring
- POV Validation
  - RDMA bandwidth/latency tests and NCCL collective tests (with and without SHARP) against the agreed success criteria
- Results Readout and Knowledge Transfer

Target Audience:

This service is designed for Network Architects, Network Engineers, HPC/AI Infrastructure Engineers and Data Center Administrators responsible for designing, deploying and operating the AI fabric.

Prerequisites:

- Working knowledge of data center networking and Linux. Familiarity with InfiniBand concepts (subnet manager, LIDs, partitions) is helpful.
- All switches, GPU servers, optics and cables are racked, cabled to the agreed cabling plan and powered up prior to the kick-off.
- Required NVIDIA licenses and support entitlements (UFM and switch software support) are procured and available.
- A UFM appliance, or a server that meets UFM requirements, is available with management network access to the fabric.
- GPU servers are installed with a supported operating system, GPU drivers, DOCA-Host (or MLNX_OFED), NCCL, HPC-X/SHARP libraries and RDMA/NCCL test tools to support fabric validation.
- Remote or on-site access to the environment is provided to Netnology engineers.

Out of Scope:

- AI/ML model deployment, LLM tuning, data pipelines and application development.
- Operating system installation and GPU software stack build-out on the servers.
- Ethernet front-end and storage network configuration beyond its connectivity to the GPU servers.
- NVLink and NVLink Switch (scale-up) domain configuration.

Service Deliverables:

No	Deliverable	Service Details
1.	Project Kickoff	• Project overview • Solution overview • Agree POV success criteria • Gather customer requirements
2.	Pre-Requisite Validation	• Review/confirm hardware, optics and cabling readiness (switches and GPU servers) • Switch and adapter firmware and license validation • UFM readiness
3.	Fabric Bring-Up	• UFM deployment and fabric discovery (up to 5 switches) • Topology and cabling validation (ibdiagnet)
4.	Subnet Management and Performance Features	• Subnet manager HA, routing engine and partitions • Adaptive routing and congestion control • NVIDIA SHARP
5.	GPU Server Fabric Integration (up to 3 nodes)	• ConnectX adapter configuration • Host-to-fabric connectivity verification
6.	Telemetry and Visibility	• UFM telemetry, link health and congestion monitoring
7.	POV Validation and Readout	• RDMA and NCCL test results against success criteria • Findings and production recommendations
8.	Knowledge Transfer	• Explain how to configure and manage the InfiniBand fabric in the customer's environment