NVIDIA INFINIBAND AI FABRIC TEST DRIVE

Service Overview:

Netnology specializes in the design, deployment and enablement of NVIDIA InfiniBand fabrics for AI. The Test Drive helps customers build a baseline understanding of NVIDIA Quantum InfiniBand at 400 Gb/s (NDR) and 800 Gb/s (XDR) and its value proposition for large-scale AI training and inference.

This 2–4-hour session demonstrates how an InfiniBand fabric is provisioned and managed with NVIDIA Unified Fabric Manager (UFM), how adaptive routing, congestion control and SHARP in-network computing accelerate GPU-to-GPU collective operations, and how the fabric is monitored and operated day to day.

This Test Drive is built and delivered by Netnology subject matter experts (SMEs) with hands-on experience in NVIDIA InfiniBand and AI cluster deployments.

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.

- Natively Lossless: Credit-based flow control delivers drop-free transport without PFC/ECN tuning.
- 400G and 800G: Quantum-2 NDR (64 × 400 Gb/s) and Quantum-X800 XDR (up to 144 × 800 Gb/s) switches.
- Adaptive Routing and Congestion Control: Spreads traffic across all fat-tree paths and throttles congesting senders at the source.
- In-Network Computing: NVIDIA SHARP performs reductions in the switches, cutting data movement and collective latency.
- Centralized Management with UFM: Subnet management, topology validation, telemetry and health monitoring from one console.
- Host Integration: ConnectX-7 and ConnectX-8 adapters with DOCA-Host drivers, NCCL and HPC-X.

Service Benefits:

Netnology has a team of world-class engineers who specialize in NVIDIA InfiniBand fabrics for AI and are passionate about customer success. Netnology will partner with you to provide guidance on:

- InfiniBand Fabric Architecture: Fat-tree, rail-optimized and scalable-unit design options sized for

your GPU cluster.
- NDR or XDR: Choosing between 400 Gb/s and 800 Gb/s generations, optics and cabling.
- Seamless Integration: Connecting the InfiniBand compute fabric to your Ethernet front-end, storage networks and operations tooling.
- Optimized Performance: Tuning adaptive routing, congestion control and SHARP for AI collectives.

Scope of Services:

As part of the 2–4-hour Test Drive engagement, Netnology will provide the following services:

- Solution Overview
- Solution Value Proposition
- Solution Use-case Demo
  - Fabric Provisioning with NVIDIA UFM
    - Fabric discovery and topology validation
    - Subnet manager, routing engine and partition (PKey) configuration
  - Performance Features
    - Adaptive routing and congestion control in action
    - NVIDIA SHARP in-network all-reduce
  - GPU Server Fabric Integration
    - ConnectX adapter firmware and InfiniBand link configuration
    - Host-to-fabric connectivity verification
  - AI Traffic Validation and Visibility
    - RDMA and NCCL collective test traffic, with and without SHARP
    - Telemetry, link health and congestion visibility in UFM
    - Link failure and recovery
- Lessons learned and enhancements

Target Audience:

This service is designed for IT leaders, architects and network engineers evaluating NVIDIA InfiniBand for AI infrastructure. No prior experience with InfiniBand is required.

2026 Netnology and Affiliates | Confidential

Service Deliverables:

No.	Deliverable	Service Details
1.	Concepts	• Solution Overview • Environment Overview • Use-Case Overview
2.	Value-Prop	• Ease of provisioning and management with UFM • Performance for AI collectives (adaptive routing, SHARP) • Visibility and fabric health
3.	Execution	• Fabric provisioning with NVIDIA UFM • Adaptive routing, congestion control and SHARP • AI traffic validation and visibility
4.	Wrap-up	• Lessons Learned • Q&A and next steps

2026 Netnology and Affiliates | Confidential