CISCO AI-OPTIMIZED FABRIC TEST DRIVE

Service Overview:

Netnology specializes in the design, deployment and enablement of Cisco AI-Optimized Fabrics. The Test Drive helps customers build a baseline understanding of the Cisco AI-Optimized Fabric and its value proposition for AI/ML infrastructure.

This 2–4-hour session demonstrates how a Cisco Nexus based AI fabric is provisioned, how it delivers lossless, high-performance transport for GPU-to-GPU (RoCEv2) traffic, and how it is monitored and operated using Cisco Nexus Dashboard. It also highlights the ease of configuration and manageability of this integrated solution.

This Test Drive is built and delivered by Netnology subject matter experts (SMEs) with extensive experience in Cisco data center and AI fabric deployments.

Solution Overview:

AI training and inference clusters place demands on the data center network that general-purpose fabrics were not designed for. GPU-to-GPU communication relies on RDMA over Converged Ethernet (RoCEv2), which generates synchronized, high-bandwidth flows that are highly sensitive to packet loss, latency and congestion. A single congested link or dropped packet can stall collective operations across the entire GPU cluster, extending job completion time and reducing the return on the GPU investment.

The Cisco AI-Optimized Fabric is built on Cisco Nexus 9000 Series switches at 400G and 800G, deployed in a non-blocking spine-leaf design with separate back-end (GPU-to-GPU), front-end and storage networks. It delivers lossless Ethernet transport using Priority Flow Control (PFC) and Explicit Congestion Notification (ECN), load balancing that spreads RDMA traffic evenly across all available paths, and granular telemetry for congestion visibility. Cisco Nexus Dashboard provides automated fabric provisioning with built-in AI/ML QoS and queuing policies, together with day-2 visibility and analytics, while Cisco UCS GPU servers managed through Cisco Intersight connect the compute nodes into the fabric.

- Lossless Ethernet: PFC and ECN provide drop-free transport for RoCEv2 GPU-to-GPU traffic.
- High-Performance Switching: Cisco Nexus 9000 Series 400G/800G non-blocking spine-leaf fabrics with low latency.
- Intelligent Load Balancing: Spreads large RDMA flows evenly across fabric paths to avoid hotspots.
- Automation with Cisco Nexus Dashboard: Template-driven fabric provisioning with built-in AI/ML QoS policies.
- End-to-End Visibility: Telemetry for congestion, PFC/ECN events and flow-level analytics.
- Integrated Compute: Cisco UCS GPU servers managed through Cisco Intersight.

Service Benefits:

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

- AI Fabric Architecture: Back-end, front-end and storage network design options sized for your GPU cluster.
- Seamless Integration: Integrating the AI fabric with your existing data center network and

operations.
- Optimized Performance: Tuning lossless QoS, congestion management and load balancing for AI traffic.

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 Cisco Nexus Dashboard
    - AI fabric creation and switch onboarding
    - Fabric underlay and overlay
    - AI/ML QoS policy deployment
  - Lossless Ethernet for RoCEv2
    - PFC and ECN in action
    - Load balancing across fabric paths
  - GPU Server Fabric Integration
    - Cisco Intersight policies and server profiles
    - Server NIC RoCEv2 settings
  - AI Traffic Validation and Visibility
    - RDMA and NCCL collective test traffic across the fabric
    - Congestion and telemetry visibility in Cisco Nexus Dashboard
    - Link failure and recovery
- Lessons learned and enhancements

Target Audience:

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

Service Deliverables:

No.	Deliverable	Service Details
1.	Concepts	• Solution Overview • Environment Overview • Use-Case Overview
2.	Value-Prop	• Ease of configuration • Ease of manageability • Performance and visibility for AI traffic
3.	Execution	• Fabric provisioning with Cisco Nexus Dashboard • Lossless Ethernet (RoCEv2) • AI traffic validation and visibility
4.	Wrap-up	• Lessons Learned • Q&A and next steps