CISCO AI-OPTIMIZED FABRIC MENTORED POV PLUS

Service Overview:

Netnology specializes in the design, deployment and enablement of Cisco AI-Optimized Fabrics, the high-performance, lossless Ethernet networks that connect GPU clusters. As part of this Mentored POV Plus service, our subject matter experts (SMEs) partner with your team to stand up a Cisco AI-Optimized Fabric in a representative environment and prove its value against agreed success criteria (lossless performance, congestion handling, visibility and ease of operation) before you commit to a production rollout.

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.

Service Benefits:

Netnology has a team of world-class engineers who specialize in high-performance Cisco data center 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 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 Cisco Nexus Dashboard (up to 5 switches)
  - Onboard Cisco Nexus 9000 switches and create the AI fabric

    - Configure underlay routing and MTU (jumbo frames)

- Lossless Ethernet for RoCEv2
    - Configure QoS classification and queuing for RDMA traffic
    - Configure Priority Flow Control (PFC) and ECN
    - Configure load balancing across fabric links

- GPU Server Fabric Integration (up to 3 nodes)
    - Onboard GPU servers to Cisco Intersight and apply server policies and profiles
    - Configure server NICs for RoCEv2 (DSCP/PFC priority, ECN, MTU)

- Telemetry and Visibility
    - Enable fabric telemetry and congestion visibility in Cisco Nexus Dashboard
    - Deploy the Splunk OpenTelemetry Collector on one GPU server and send host and GPU (NVIDIA DCGM) metrics to Splunk Observability Cloud, where licensed, with one dashboard and one GPU detector with email alerting, viewable alongside Nexus Dashboard congestion views.

- Results Readout and Knowledge Transfer

Target Audience:

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

Prerequisites:

- Working knowledge of Cisco Nexus (NX-OS) and data center networking (BGP, VXLAN EVPN, QoS). Familiarity with Cisco UCS and Cisco Intersight 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 software licenses (Cisco Nexus, Cisco Nexus Dashboard, Cisco Intersight) are procured and available.
- Cisco Nexus Dashboard is deployed and reachable, or compute resources are available for its deployment.
- GPU servers are installed with a supported operating system, GPU and NIC drivers/firmware, and RDMA/NCCL test tools to support fabric validation.
- Remote or on-site access to the environment is provided to Netnology engineers.
- Where the Splunk Observability Cloud module applies, the customer provides a Splunk Observability Cloud organization and ingest access token. The GPU servers have outbound HTTPS access, a supported OS, NVIDIA drivers and the DCGM exporter installed.

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.
- Storage system configuration beyond its connectivity to the fabric.

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) • NX-OS software and license validation • Cisco Nexus Dashboard and Cisco Intersight readiness
3.	Fabric Bring-Up	• Onboard Nexus 9000 switches (up to 5 switches) • Configure underlay routing and MTU
4.	Lossless Ethernet (RoCEv2) Configuration	• QoS classification and queuing • PFC and ECN • Load balancing across fabric links
5.	GPU Server Fabric Integration (up to 3 nodes)	• Cisco Intersight onboarding, server policies and profiles • Server NIC RoCEv2 settings
6.	Telemetry and Visibility	• Fabric telemetry and congestion visibility in Cisco Nexus Dashboard • GPU server host and GPU metrics in Splunk Observability Cloud, where licensed (dashboard and detector)
7.	Knowledge Transfer	• Explain how to configure and manage the AI fabric in the customer's environment