

CISCO AI-OPTIMIZED FABRIC TEST DRIVE

Service Overview:

At Netnology, we specialize in AI-Optimized Cisco High-Performance Fabric enablement and implementation services to expedite the adoption of AI solutions. The test-drive offer will help customers develop a baseline understanding of Cisco powered AI Fabric solution and its value proposition.

This half day session will demonstrate the working of Large Language Model (LLM) using Cisco Network and Compute and much more. It will also highlight the ease of configuration and manageability of this integrated solution.

This test-drive is built and delivered by subject matter experts (SMEs) who have extensive experience working with AI Solution.

Solution Overview:

Cisco AI Solution is at the cutting edge of transforming modern data centers with advanced AI/ML technologies. Utilizing GPU-accelerated servers, the solution enables the development and training of sophisticated deep learning models with languages such as Python and C/C++, and frameworks like PyTorch, TensorFlow, and JAX. These models are optimized using extensive GPU clusters, typically featuring multiple GPUs per server, and interconnected with dual 100Gbps network interfaces to ensure robust connectivity and meet high networking demands.

The Cisco Nexus 9000 series switches are integral to supporting the high-performance requirements of AI/ML workloads, providing low latency, advanced congestion management, and comprehensive telemetry. In conjunction with Cisco UCS X-Series Modular Systems and the C240 and C220 rack servers, equipped with AI capable CPU and GPU, the solution offers versatile deployment options for AI inference both in centralized data centers and at the network edge.

- GPU Acceleration: Facilitates the creation and training of deep learning models using GPU clusters.
- High-Performance Networking: Dual 100Gbps interfaces ensure robust connectivity and meet demanding network requirements.
- Cisco Nexus 9000 Series: Provides low latency, advanced congestion management, and telemetry for AI/ML workloads.
- Versatile Deployment: Supports AI inference solutions in both centralized data centers and at the network edge with Cisco UCS X-Series and rack servers.
- Advanced Framework Support: Integrates with popular frameworks such as PyTorch, TensorFlow, and JAX for streamlined model development.

Service Benefits:

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

- Expert AI Strategy: Crafting tailored AI strategies to meet your specific business needs and objectives.
- Seamless Integration: Ensuring smooth integration of AI/ML technologies with your existing infrastructure.
- Optimized Performance: Fine-tuning GPU-accelerated systems and networks for maximum efficiency and performance.

Service Scope:

As part of the half-day (up to 4 hours) Test Drive engagement, Netnology will provide the following services:

- Solution Overview
- Solution Value Proposition
- Solution Use-case Demo
 - Cisco Network and Compute Fabric
 - Fabric Path Validation
 - Fabric Underlay and Overlay
 - RoCEv2
 - IMM Policy and Profiles
 - AI Operating System
 - Ubuntu File System
 - System updates, drivers, and software repositories
 - Package Management
 - LLM Workload
 - Environment Setup
 - AI Model Sizing
 - AI Model Loading
 - Text Generation
 - Conversational AI
 - Interactive Q&A
 - AI-driven Analytics
- Lessons Learned and Next Steps

Target Audience:

This service is designed for candidates with little or no Cisco Network and Compute experience or prior knowledge.

Service Deliverables:

No.	Deliverable	Service Details
1.	Concepts	• Solution Overview • Environment Overview • Use-case Overview
2.	Value-Prop	• Ease of configuration • Ease of manageability
3.	Execution	• Use-case Demos: ○ Cisco Network and Compute Fabric ○ AI Operating System ○ LLM Workload Execution
4.	Wrap-up	• Lessons Learned • Q&A