# CISCO AI-OPTIMIZED FABRIC MENTORED INSTALL

## Service Overview:

At Netnology, we specialize in Cisco High-Performance AI-Optimized Fabric enablement and implementation services to expedite the adoption of AI solutions. As part of this Mentored Install service offer, our subject matter experts (SME) partner with your team to ensure smooth deployment and provide knowledge transfer to equip your staff with the necessary skills to configure and manage the Cisco High-Performance AI-Optimized Fabric effectively.

## Solution Overview:

In today's data centers, the implementation of AI/ML solutions is becoming increasingly common. Machine learning, a fundamental aspect of AI, enables systems to derive insights and make predictions based on data analysis. The advent of GPU-accelerated servers has greatly facilitated the creation and training of sophisticated deep learning models. Languages like Python and C/C++, along with frameworks such as PyTorch, TensorFlow, and JAX, natively support GPUs, streamlining the development of these applications. Neural networks often leverage extensive GPU clusters, typically configured with multiple GPUs per server, and are interconnected using dual 100Gbps network interfaces to ensure robust connectivity and meet stringent networking demands.

The Cisco Nexus 9000 series switches provide the necessary low latency, advanced congestion management, and telemetry capabilities to support the demanding requirements of AI/ML workloads. These switches, in conjunction with Cisco UCS X-Series Modular Systems and the C240 and C220 rack servers, which feature AI capable CPU and GPU, offer versatile options for deploying inference solutions both in centralized data centers and at the network edge. This integrated solution is perfectly suited for establishing a high-performance AI/ML network infrastructure.

## Service Benefits:

Netnology has a team of world class engineers who specialize in AI Solution on High-Speed Cisco Infrastructure and are passionate about customer success. This Mentored Install engagement will provide information on how to deploy, configure and integrate AI solution on Cisco Infrastructure.

## Service Scope:

As part of the 10-day (up to 80 hours) engagement, Netnology will provide the following services:

- AI-Optimized Cisco High-Performance Fabric
  - Configure Network Fabric Infrastructure
    - Design a lossless fabric
    - Configure Underlay and Overlay
    - Set Up VLANS/VTEPs
    - Configure RoCEv2
  - Configure Compute infrastructure (UCS-X)
    - Initialize FI in IMM mode
    - Discover FI/Chassis/Blades

    - IP Pool Policy
    - Basic Domain Profile Setup
    - Port Policy, VLANS/vSAN, Network Connectivity policy
    - Server Profile Templates / Server Profiles
        - BIOS / Boot / LAN
        - vNIC / vHBA
- Operating System Installation on Nodes
    - Determine the compatible OS for AI workload (Ubuntu, Red Hat).
    - Create an Image repository.
    - Perform OS installation, including setting up disk partitions and configuring basic network settings.
    - Install and configure required drivers (e.g., GPU drivers).
    - Configure firewall rules for security.
- AI Model Sizing
    - Assess and document infrastructure requirements for deploying AI workloads (e.g., LLaMA 2/3).
    - Define KV_Cache needs based on model size, sequence length, and concurrency.
    - Determine CPU and RAM requirements for preprocessing, tokenization, and model execution.
    - Evaluate GPU requirements (type, count, and memory) for both training and inference.
    - Recommend cluster configurations and scaling strategies to meet throughput, latency, and efficiency targets.
- AI Workload Deployment (LLAMA 2/LLAMA 3)
    - Obtain the LLAMA 2/3 model files and required dependencies from the official repository or source.
    - Create a virtual environment using venv or conda.
    - Install the necessary Python packages, such as transformers, torch, or other libraries.
    - Create a code to load the LLAMA 2/3 model into memory.
- Configure Data Pipeline and Inference
    - Set up the data preprocessing pipeline, including tokenization, normalization, and batching.
    - Develop or configure an inference script.
- Test and Validation
    - Solution validation by running sample inputs.
- Knowledge Transfer

## Target Audience:

This service is designed for Network Architects, Network Engineers and Administrators configuring, deploying, and managing the AI Infrastructure.

## Prerequisites:

- Basic knowledge of Cisco Nexus, Cisco UCS, Large Language Models (LLMs).
- Customers need to ensure that all equipment and devices are racked and stacked, cabled, and powered up prior to the kick-off.
- Customers also need to acquire the necessary software licenses for the deployment of the infrastructure.

## Service Deliverables:

| No | Deliverable | Service Details |
|---|---|---|
| 1. | Project Kickoff | • Project Overview<br>• Solution Overview<br>• Gather customer requirements |
| 2. | Pre-Requisite Validation | • Review/Confirm Hardware Readiness (Server and Network)<br>• License Validation<br>• Network Readiness (Bandwidth, latency and redundancy)<br>• Intersight Readiness |
| 3. | High Level Design | • Develop High Level Design (HLD) Document |
| 4. | Configure Network Fabric Infrastructure | • Design a lossless fabric<br>• Configure Underlay and Overlay<br>• Set Up VLANS/VTEPs<br>• Configure RoCEv2<br>• Configure Network and Security Policies |
| 5. | Configure Compute Infrastructure | • Initializing FI in IMM mode<br>  ○ Discover FI/Chassis/Blades<br>• IP Pool Policy<br>• Basic Domain Profile Setup<br>• Port Policy, VLANS/vSAN, Network Connectivity policy<br>• Server Profile Templates / Server Profiles<br>  ○ BIOS / Boot / LAN<br>  ○ vNIC / vHBA |
| 6. | Operating System Installation on Nodes | • Determine the compatible OS for AI workload (Ubuntu, Red hat)<br>• Create an Image repository.<br>• Perform OS installation, including setting up disk partitions and configuring basic network settings.<br>• Install and configure required drivers (e.g., GPU drivers).<br>• Configure firewall rules for security. |
| 7. | AI Model Sizing | • Assess and document infrastructure requirements for deploying AI workloads (e.g., LLaMA 2/3).<br>• Define KV_Cache needs based on model size, sequence length, and concurrency.<br>• Determine CPU and RAM requirements for preprocessing, tokenization, and model execution.<br>• Evaluate GPU requirements (type, count, and memory) for both training and inference.<br>• Recommend cluster configurations and scaling strategies to meet throughput, latency, and efficiency targets. |
| 8. | AI Workload Deployment (LLAMA 2/LLAMA 3) | • Obtain the LLAMA 2/3 model files and required dependencies from the official repository or source.<br>• Create a virtual environment using venv or conda.<br>• Install the necessary Python packages, such as transformers, torch, or other libraries.<br>• Create a code to load the LLAMA 2/3 model into memory. |

| | | |
|---|---|---|
| 9. | Configure Data Pipeline and Inference | • Set up the data preprocessing pipeline, including tokenization, normalization, and batching.<br>• Develop or configure an inference script. |
| 10. | Test and Validation | • Solution validation by running sample inputs. |
| 11. | Knowledge Transfer | • Explain how to configure and manage the deployed solution in the respective customers' environment. |