DCLLM - IMPLEMENTING AND OPERATING LLM INFERENCING SYSTEMS WITH CISCO AND NVIDIA DATA CENTER TECHNOLOGIES

Course Code: 860052

This comprehensive training equips participants with the knowledge and skills required to design, deploy, and optimize Large Language Models (LLMs) using NVIDIA GPUs and Cisco infrastructure. Through in-depth modules, hands-on labs, and real-world case studies, participants will learn how to manage data preparation, build scalable pipelines, optimize performance, ensure security, and migrate from cloud to on-premises deployments. The course provides a holistic approach to mastering the technical complexities of LLM systems while leveraging cutting-edge NVIDIA and Cisco technologies for scalability, efficiency, and security.

What You'll Learn

By the end of this course, participants will:

- Master the Foundations of LLMs: Gain an in-depth understanding of LLM architecture, scaling principles, and design trade-offs.
- Prepare and Manage Large Datasets: Learn techniques for sourcing, preprocessing, and managing large-scale, high-quality datasets for LLM training.
- Deploy LLMs for Production: Use NVIDIA TensorRT and Cisco Nexus Dashboard to build efficient, low-latency inferencing pipelines.
- Optimize LLM Performance: Apply advanced optimization techniques like quantization, pruning, and dynamic batching to improve throughput and reduce latency.
- Design Scalable Pipelines: Build fault-tolerant, high-performance pipelines for real-time and batch inferencing.
- Monitor and Maintain Systems: Use NVIDIA and Cisco tools to monitor GPU and network performance, ensuring reliability and uptime.
- Ensure Security and Privacy: Implement robust security measures using Cisco Nexus Dashboard, Cisco XDR, and NVIDIA encryption tools.
- Build On-Premises Data Centers: Design and implement LLM inferencing systems using NVIDIA GPUs and Cisco UCS for maximum scalability and efficiency.
- Migrate Cloud Models to On-Premise: Transition cloud-trained LLMs to

on-premise infrastructure while optimizing performance and costs.

Who Needs to Attend

This course is tailored for professionals involved in designing and managing AI and data infrastructure, including:

- **Systems Architects:** To understand the integration of LLM systems into broader IT environments.
- **Network Architects:** To optimize network configurations for high-speed LLM training and inferencing.
- **Storage Architects:** To manage the storage and retrieval of large-scale datasets used in LLM systems.
- **AI Infrastructure Architects:** To build robust and scalable AI platforms optimized for LLM workloads.
- **Data Scientists:** To prepare high-quality datasets and fine-tune LLMs for specific use cases.
- **Machine Learning Engineers:** To deploy and optimize LLMs for real-world applications with low latency and high throughput.

Prerequisites

Participants should possess basic knowledge of LLM models, server infrastructure, Cloud knowledge, networking concepts and virtualization fundamentals

DCLLM - IMPLEMENTING AND OPERATING LLM INFERENCING SYSTEMS WITH CISCO AND NVIDIA DATA CENTER TECHNOLOGIES

Course Code: 860052

CLASSROOM LIVE	$3,995 USD	5 Day

Classroom Live Outline

Module 1: Large Language Model (LLM) Foundations

Objectives:

- Understand the architecture and mathematical principles of LLMs.
- Learn design trade-offs for scalability and performance.
- Explore emerging innovations in LLM development.

Topics:

- Transformer architecture, self-attention mechanism, and positional encoding.
- Types of LLMs: Encoder-only, decoder-only, and encoder-decoder.
- Training objectives: Masked language modeling (MLM), causal language modeling (CLM), and sequence-to-sequence modeling.
- Scaling laws and challenges: Parameter size, dataset size, and compute.
- Emerging architectures: Reformer, Longformer, and multi-modal LLMs.

Module 2: Data Collection and Preparation for LLM Training

Objectives:

- Understand data requirements for LLMs and their impact on performance.
- Learn techniques for sourcing, cleaning, and managing large-scale datasets.
- Explore NVIDIA and Cisco tools for efficient data handling.

Topics:

- Data sourcing: Open-source, proprietary, and domain-specific datasets.
- Preprocessing: Cleaning, deduplication, tokenization, and filtering.
- Data management: Sharding, scalable storage, and high-speed data transfer.
- Ethical considerations: Bias detection, privacy compliance, and fairness.

Module 3: Deployment of LLMs for Inferencing

Objectives:

- Deploy LLMs for production inferencing with high performance and scalability.
- Use NVIDIA TensorRT and Cisco Nexus Dashboard for optimized deployment.

Topics:

- Deployment architectures: On-premises, cloud, and hybrid.
- Optimizing inferencing with NVIDIA TensorRT: Precision calibration, layer fusion, and batching.
- Traffic management and load balancing with Cisco Nexus Dashboard.
- Exposing LLM APIs: RESTful and gRPC endpoints with security mechanisms.

Module 4: Optimizing LLM Models for Inferencing

Objectives:

- Optimize LLM inferencing pipelines for low latency and high throughput.
- Learn techniques like quantization, pruning, and model compression.

Topics:

- Quantization: FP16, INT8, and mixed precision.
- Pruning and knowledge distillation for lightweight models.
- TensorRT optimization: Dynamic batching and asynchronous execution.
- Benchmarking tools: NVIDIA Triton Inference Server, TensorRT Profiler.

Module 5: Scalable Pipeline Design for LLM Inferencing

Objectives:

- Build robust, scalable, and fault-tolerant pipelines for inferencing.
- Use batching, caching, and dynamic scaling for efficient pipelines.

Topics:

- Pipeline components: Batching, caching, and queuing.
- Load balancing with Cisco Nexus Dashboard for traffic optimization.
- Fault tolerance: Automatic failover and disaster recovery plans.
- Monitoring pipeline performance with NVIDIA DCGM and Cisco Nexus Dashboard.

Module 6: Monitoring, Logging, and Maintenance for LLM Systems

Objectives:

- Monitor and maintain LLM deployments using NVIDIA and Cisco tools.

Topics:

- Key metrics: Latency, throughput, GPU utilization, and memory usage.
- Monitoring tools: NVIDIA DCGM and Cisco Nexus Dashboard Insights.
- Maintenance workflows for hardware and software reliability.

Module 7: Security and Privacy Considerations in LLM Training and Inferencing

Objectives:

- Secure LLM pipelines using Cisco Nexus Dashboard, Cisco XDR, and NVIDIA tools.

Topics:

- NVIDIA runtime encryption and secure boot.
- Cisco Robust Intelligence for adversarial defense and vulnerability detection.
- Cisco XDR for unified threat detection and automated response.
- Traffic segmentation and endpoint authentication.

Module 8: Migrating from Cloud-Based Training to On-Premises Inferencing

Objectives:

- Transition LLM models from cloud training to on-premises Cisco infrastructure.

Topics:

- Migration strategies for exporting and deploying models.
- Data transfer optimization using Cisco Nexus Dashboard.
- Integrating models with on-premises inferencing pipelines.

Module 9: On-Premises Data Center Design for LLM Inferencing Systems

Objectives:

- Design an on-premises data center with Cisco and NVIDIA technologies.

Topics:

- Cisco UCS and NVIDIA GPUs for high-performance compute.
- Network design and automation with Cisco Nexus Dashboard.
- Storage solutions for large-scale data management.

Module 10: On-Premises Data Center Implementation for LLM Inferencing Systems

Objectives:

- Implement and configure an LLM inferencing data center using NVIDIA and Cisco technologies.

Topics:

- Physical setup: NVIDIA GPUs on Cisco UCS and Nexus networking configuration.
- Performance testing and validation of inferencing pipelines

Classroom Live Labs

Labs are designed to assure learners a whole practical experience, through the following practical activities:

- Exploring Transformer model architectures
- Compute attention scores manually for a small sequence.
- Preprocessing lab: Clean, deduplicate, and tokenize a dataset using NVIDIA RAPIDS.
- Tokenization exercise: Implement and analyze subword tokenization methods.
- Deploy an LLM as a REST API using NVIDIA TensorRT.
- Configure traffic policies in Cisco Nexus Dashboard for inferencing workloads.

- Apply quantization and pruning to optimize a pre-trained LLM.
- Benchmark latency, memory usage, and accuracy of optimized models.
- Design a scalable pipeline with batching and caching strategies.
- Configure routing and scaling policies for GPU nodes using Nexus Dashboard.
- Configure dashboards for real-time monitoring of GPU and network performance.
- Simulate hardware failures and evaluate maintenance workflows.
- Analyze and secure an LLM using Cisco Robust Intelligence.
- Configure Cisco XDR to monitor and respond to threats across pipelines.
- Export a cloud-trained model and deploy it on Cisco UCS for inferencing.
- Optimize data transfer pipelines for low-latency inferencing.
- Design a complete data center architecture for LLM inferencing.

Skillsoft global knowledge™

DCLLM - IMPLEMENTING AND OPERATING LLM INFERENCING SYSTEMS WITH CISCO AND NVIDIA DATA CENTER TECHNOLOGIES

Course Code: 860052

VIRTUAL CLASSROOM LIVE	$3,995 USD	5 Day

Virtual Classroom Live Outline

Module 1: Large Language Model (LLM) Foundations

Objectives:

- Understand the architecture and mathematical principles of LLMs.
- Learn design trade-offs for scalability and performance.
- Explore emerging innovations in LLM development.

Topics:

- Transformer architecture, self-attention mechanism, and positional encoding.
- Types of LLMs: Encoder-only, decoder-only, and encoder-decoder.
- Training objectives: Masked language modeling (MLM), causal language modeling (CLM), and sequence-to-sequence modeling.
- Scaling laws and challenges: Parameter size, dataset size, and compute.
- Emerging architectures: Reformer, Longformer, and multi-modal LLMs.

Module 2: Data Collection and Preparation for LLM Training

Objectives:

- Understand data requirements for LLMs and their impact on performance.
- Learn techniques for sourcing, cleaning, and managing large-scale datasets.
- Explore NVIDIA and Cisco tools for efficient data handling.

Topics:

- Data sourcing: Open-source, proprietary, and domain-specific datasets.
- Preprocessing: Cleaning, deduplication, tokenization, and filtering.
- Data management: Sharding, scalable storage, and high-speed data transfer.
- Ethical considerations: Bias detection, privacy compliance, and fairness.

Module 3: Deployment of LLMs for Inferencing

Objectives:

- Deploy LLMs for production inferencing with high performance and scalability.
- Use NVIDIA TensorRT and Cisco Nexus Dashboard for optimized deployment.

Topics:

- Deployment architectures: On-premises, cloud, and hybrid.
- Optimizing inferencing with NVIDIA TensorRT: Precision calibration, layer fusion, and batching.
- Traffic management and load balancing with Cisco Nexus Dashboard.
- Exposing LLM APIs: RESTful and gRPC endpoints with security mechanisms.

Module 4: Optimizing LLM Models for Inferencing

Objectives:

- Optimize LLM inferencing pipelines for low latency and high throughput.
- Learn techniques like quantization, pruning, and model compression.

Topics:

- Quantization: FP16, INT8, and mixed precision.
- Pruning and knowledge distillation for lightweight models.
- TensorRT optimization: Dynamic batching and asynchronous execution.
- Benchmarking tools: NVIDIA Triton Inference Server, TensorRT Profiler.

Module 5: Scalable Pipeline Design for LLM Inferencing

Objectives:

- Build robust, scalable, and fault-tolerant pipelines for inferencing.
- Use batching, caching, and dynamic scaling for efficient pipelines.

Topics:

- Pipeline components: Batching, caching, and queuing.
- Load balancing with Cisco Nexus Dashboard for traffic optimization.
- Fault tolerance: Automatic failover and disaster recovery plans.
- Monitoring pipeline performance with NVIDIA DCGM and Cisco Nexus Dashboard.

Module 6: Monitoring, Logging, and Maintenance for LLM Systems

Objectives:

- Monitor and maintain LLM deployments using NVIDIA and Cisco tools.

Topics:

- Key metrics: Latency, throughput, GPU utilization, and memory usage.
- Monitoring tools: NVIDIA DCGM and Cisco Nexus Dashboard Insights.
- Maintenance workflows for hardware and software reliability.

Module 7: Security and Privacy Considerations in LLM Training and Inferencing

Objectives:

- Secure LLM pipelines using Cisco Nexus Dashboard, Cisco XDR, and NVIDIA tools.

Topics:

- NVIDIA runtime encryption and secure boot.
- Cisco Robust Intelligence for adversarial defense and vulnerability detection.
- Cisco XDR for unified threat detection and automated response.
- Traffic segmentation and endpoint authentication.

Module 8: Migrating from Cloud-Based Training to On-Premises Inferencing

Objectives:

- Transition LLM models from cloud training to on-premises Cisco infrastructure.

Topics:

- Migration strategies for exporting and deploying models.
- Data transfer optimization using Cisco Nexus Dashboard.
- Integrating models with on-premises inferencing pipelines.

Module 9: On-Premises Data Center Design for LLM Inferencing Systems

Objectives:

- Design an on-premises data center with Cisco and NVIDIA technologies.

Topics:

- Cisco UCS and NVIDIA GPUs for high-performance compute.
- Network design and automation with Cisco Nexus Dashboard.
- Storage solutions for large-scale data management.

Module 10: On-Premises Data Center Implementation for LLM Inferencing Systems

Objectives:

- Implement and configure an LLM inferencing data center using NVIDIA and Cisco technologies.

Topics:

- Physical setup: NVIDIA GPUs on Cisco UCS and Nexus networking configuration.
- Performance testing and validation of inferencing pipelines

Virtual Classroom Live Labs

Labs are designed to assure learners a whole practical experience, through the following practical activities:

- Exploring Transformer model architectures
- Compute attention scores manually for a small sequence.
- Preprocessing lab: Clean, deduplicate, and tokenize a dataset using NVIDIA RAPIDS.
- Tokenization exercise: Implement and analyze subword tokenization methods.
- Deploy an LLM as a REST API using NVIDIA TensorRT.
- Configure traffic policies in Cisco Nexus Dashboard for inferencing workloads.

- Apply quantization and pruning to optimize a pre-trained LLM.
- Benchmark latency, memory usage, and accuracy of optimized models.
- Design a scalable pipeline with batching and caching strategies.
- Configure routing and scaling policies for GPU nodes using Nexus Dashboard.
- Configure dashboards for real-time monitoring of GPU and network performance.
- Simulate hardware failures and evaluate maintenance workflows.
- Analyze and secure an LLM using Cisco Robust Intelligence.
- Configure Cisco XDR to monitor and respond to threats across pipelines.
- Export a cloud-trained model and deploy it on Cisco UCS for inferencing.
- Optimize data transfer pipelines for low-latency inferencing.
- Design a complete data center architecture for LLM inferencing.

Visit us at www.globalknowledge.com or call us at 1-866-716-6688.

Date created: 5/24/2026 5:18:36 AM