AI Infrastructure – Public Training

OUTLINE

Training Overview

In today's AI-driven era, the ability to deploy AI clusters efficiently and effectively is crucial for organizations.

This course specializes in optimizing efficiency, reliability, and scalability for deploying AI environments. It covers various infrastructure aspects, including compute platforms, networking, storage, and the setup and maintenance of AI infrastructure.

The training focuses on key AI building blocks such as GPUs, CPUs, and BlueField networking platform, providing in-depth knowledge and skills to professionals involved in deploying and managing AI infrastructure.

Training Delivery Method

Instructor-led remote training sessions via NVIDIA Teams platform.
Hands-on lab exercises focused on the data center infrastructure.

Target Audience

The course is designed for administrators, DevOps professionals, and IT-related roles who want to gain the knowledge and skills necessary to deploy and maintain AI data centers.

Training Duration

Remote | 7 sessions of 4 hours

Prerequisites

- Knowledge of networking concepts and principles, including Ethernet and InfiniBand technologies used in data centers and high-performance computing environments.
- Hands-on experience in Linux-like systems administration, such as managing users and permissions, installing software packages, configuring network settings, and troubleshooting common issues in a Linux environment.
- Basic understanding of server hardware components and their roles in a data center environment. This includes knowledge of CPUs, memory, storage devices, and networking interfaces commonly found in servers.
- Knowledge of storage concepts and principles, including different file systems and their characteristics, as well as the functioning and usage of storage protocols in data storage and retrieval.
- Basic understanding of virtualization technologies, including virtual machines (VMs) and containers. You should be familiar with VM creation, management, and the role of hypervisors in virtualized environments.
- Basic understanding of artificial intelligence (AI) concepts and terminology. This may include knowledge of topics such as machine learning, deep learning, neural networks, and common AI applications.
- Before attending the course, we recommend completing the AI Infrastructure and Operation Fundamentals self-paced course. This course will provide the foundations for this training.

Training Outline

AI in the Data Center Overview

- AI Overview
- Data Center Architecture for AI Workloads

Compute Platforms for AI

- AI Compute Platforms Overview
- Scaling AI Compute
- Practice: Installing the NVIDIA GPU driver and using the nvidia-smi tool

Networking for AI

- Networking for AI Data Centers
- Building and Maintaining InfiniBand Infrastructure for AI
- Adapting Ethernet Networks to Run AI Workloads
- AI Data Centers Networks
- Practice: InfiniBand Fabric Management

Storage for AI

- Storage Requirements for AI data Centers
- Storage Architecture
- Practice: Mounting storage and testing performance

BlueField Networking Platform

- BlueField Overview and Uses Cases
- BlueField Bring-up
    - Installing DOCA
    - Firmware Upgrade
    - Management via RShim
    - BlueField Interfaces – Network Interfaces and OVS Bridges
- Practice: BlueField bring-up

AI Data Center Management

- AI Data Center Management Overview
- AI Infrastructure Provisioning and Management with NVIDIA Base Command Manager (BCM)
- Practice: Bringing-up an AI cluster with BCM

Virtualizing GPU Resources

- GPU Temporal Partitioning
- GPU Spatial Partitioning
- Practice: Virtualizing GPU resources with NVIDIA vGPU on VMware vSphere
- Practice: Virtualizing GPU resources using MIG

NVIDIA AI Software

- Using NGC Containers
- NVIDIA AI Enterprise Software Suite
- Practice: Running AI applications in NGC containers