DGX Onboarding for AI Practitioners

OUTLINE

Training Overview

The **DGX Onboarding for AI Practitioners** introduces users to NVIDIA DGX systems and NVIDIA AI Enterprise software. Participants will learn how these two technologies can be combined to accelerate the development and deployment of AI applications.

Throughout the course, learners will gain essential knowledge in key software tools to accelerate AI workloads. This training explores diverse topics including model training and optimization, deployment of GenAI models, and other critical aspects of building data-centric software solutions with accelerated computing. Our objective is to empower users to maximize their AI cluster's capabilities by familiarizing them with the expansive software ecosystem within the NVIDIA platform.

Training Delivery Method

Two instructor-led remote training sessions, via NVIDIA Academy Teams platform.

Hands-on lab exercises focused on deploying NVIDIA AI Enterprise containers using NVIDIA LaunchPad and DGX systems

Target Audience

This course is aimed at AI practitioners to be able to successfully run NVIDIA AI Enterprise training and inference workloads.

Training Duration

2 sessions of 4 hours each

Training Prerequisites

- A working knowledge of:
  - Docker and containerization
  - Linux administration
- Familiarity with GPU-accelerated workloads

Training Outline

Session 1

Introduction to DGX Systems

- Overview of DGX system architecture and specifications
- DGX operating system and software stack

NVIDIA AI Enterprise Software Overview

- Introduction to the NVIDIA AI Enterprise software suite
- AI use cases, frameworks, and pretrained models
- AI development workflows

Accessing and Installing AI Software

- Navigating the NGC & NVIDIA AI Enterprise Catalogs
- Downloading and sharing AI software
- Managing access through the NGC Enterprise Catalog
- Using the NGC private registry for collaboration

AI Workflows

- AI workflow overview
- AI workflow benefits
- NVIDIA AI workflows

AI Inference

- Triton inference server overview
- Triton inference server deployment
- LLM NIMs

Session 2

NeMo Framework

- Overview of NeMo suite for training, customizing, and GenAI models
- Deploying multi-GPU, multi-node LLM fine-tuning with NeMo

- Demo: LLM fine-tuning with NeMo

NVIDIA NIMs (Inference Microservices)

- Overview of NIM microservices, their key components, and role in AI deployment
- Deploying NIM microservices
- Domain specific NIMs and best practices
- Demo: NIM deployment workflow

Additional NVIDIA AI Tools (as time permits)