Turnkey Deployment of NVIDIA AI Data Centers

- Hands-on

Course Overview

This hands-on course teaches students how to stand up a complete, production-style NVIDIA AI data center from bare metal. Students work inside isolated tenant environments and use NVIDIA Base Command Manager (BCM) to provision servers over Redfish/PXE, bootstrap a Kubernetes cluster, deploy the NVIDIA GPU Operator, install the ClearML MLOps platform, and configure advanced GPU orchestration (hardware MIG slicing and software fractional GPUs).

By the end of the course, students will have built a fully functional, multi-tenant AI platform that includes persistent storage, secure public ingress, model serving, and a timed disaster-recovery drill — the same operational patterns used in real NVIDIA DGX and enterprise AI clusters.

Who Should Attend

- Platform engineers and SREs responsible for standing up or operating NVIDIA GPU clusters
- DevOps and infrastructure teams adopting Base Command Manager or migrating from traditional HPC schedulers to Kubernetes
- MLOps engineers who need to understand the full stack beneath ClearML (from bare metal to dynamic GPU partitioning)
- Systems administrators and architects building multi-tenant AI platforms for internal users or customers
- Technical staff preparing for NVIDIA DGX SuperPOD, Base Command, or similar enterprise AI infrastructure roles

What You'll Learn

- Provision bare-metal nodes using BCM and Redfish/PXE automation
- Bootstrap a production Kubernetes cluster with BCM's native tooling
- Deploy and configure the NVIDIA GPU Operator for containerized GPU access
- Install and operate the full ClearML Server suite (web, API, file server) with external persistent storage
- Configure both hardware Multi-Instance GPU (MIG) slicing via ClearML Dynamic MIG Operator (CDMO) and software fractional GPU time-slicing
- Implement multi-tenant isolation using Kubernetes namespaces and NetworkPolicies
- Deploy model inference endpoints with ClearML Serving
- Execute a complete disaster-recovery workflow under time pressure

Outline

1. Base Command Manager (BCM) & Out-of-Band Setup

- Map Redfish Endpoints in BCM
- Build OS Profiles & PXE Boot Images

2. Automated Bare-Metal Provisioning & K8s Onboarding
    - Execute Automated Redfish PXE Boot
    - Bootstrap Kubernetes via BCM

3. Ingress Routing, TLS & NVIDIA GPU Operator
    - Configure Public Ingress Routing & TLS
    - Deploy NVIDIA GPU Operator via Helm

4. ClearML Server & Infrastructure Control Plane
    - Deploy Storage Dependencies & Persistent Volumes
    - Install ClearML Server Suite via Helm
    - Configure Ingress & Initial Web UI Authentication

5. GPU Orchestration (Dynamic MIG vs. Fractional GPUs)
    - Attach Student Agent to Real Physical GPU Host
    - Deploy ClearML Dynamic MIG Operator (CDMO)
    - Configure Software Fractional GPU Time-Slicing

6. Enterprise Operations, Multi-Tenancy & Disaster Recovery
    - Multi-Tenant Namespace & Network Policy Isolation
    - Deploy Inference Endpoint via ClearML Serving
    - Timed Disaster Recovery Drill (30-Minute Cap)

Prerequisites

- Solid Linux administration skills
- Basic Kubernetes familiarity (pods, deployments, kubectl)
- Comfort working at the command line
- Prior experience with Ansible, Helm, or GPU systems is helpful but not required

Next Courses

- ClearML Infrastructure Engineer