

# NVIDIA AI Enterprise Deployment on BareMetal Kubernetes

Instructor-led Training

## Why should I take this course?

---

The NVIDIA AI Enterprise Deployment on BareMetal Kubernetes training provides a comprehensive, hands-on experience designed to equip BareMetal IT professionals with the skills needed to deploy, manage, and validate AI workloads in production environments using the NVIDIA AI Enterprise solution.

## What will I learn?

---

In this course, you'll learn to deploy and manage NVIDIA AI Enterprise on BareMetal Kubernetes environments. The training follows a scenario-based approach, starting with Kubernetes deployment on BareMetal infrastructure, followed by NVIDIA AI Enterprise implementation, comprehensive GPU validation to ensure readiness for accelerated AI workloads, and concluding with a deployment of a standard Retrieval Augmented Generation (RAG) use case.

## Target Audience

---

This course is designed for:

- IT professionals
- DevOps/MLOps engineers
- Anyone responsible for deploying and managing AI infrastructure on BareMetal.

## Training Duration

---

3 sessions of 4 hours each

## Training Delivery Method

---

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

## Course Objectives:

---

Upon completion, you should be able to:

- Recall the key features and benefits of deploying NVIDIA AI Enterprise on BareMetal.
- Identify the prerequisites for deploying NVIDIA AI Enterprise on BareMetal.
- Implement the steps to deploy NVIDIA AI Enterprise on BareMetal with K8s.
- Construct a basic K8s cluster on BareMetal and configure it for use with NVIDIA AI Enterprise.
- Apply NVIDIA AI Enterprise deployment techniques on K8s using Helm charts.
- Utilize the NVIDIA GPU Operator for GPU configuration deployment.
- Evaluate GPU deployment effectiveness by implementing and testing an example training job.
- Analyze an enterprise RAG use case through practical workflow deployment.
- Monitor and manage NVIDIA AI Enterprise deployments on K8s.
- Design scaling strategies and update procedures for NVIDIA AI Enterprise deployments on K8s.
- Create solutions for troubleshooting common issues and effectively accessing support for NVIDIA AI Enterprise on BareMetal K8s deployments.

## Course Prerequisites

---

To gain the most value from this course, the target audience should have knowledge of the following domains:

- Working knowledge of data center infrastructure:
  - Servers
  - Storage
  - Networking
  - GPUs
  - Operating systems
- Familiarity with containerization and Kubernetes:

- Docker
- Kubernetes - Cloud Native Computing Foundation (CNCF)

## Training Topics

---

- Overview of the learning objectives, course outline, prerequisite knowledge, and hands-on lab details.
- **NVIDIA AI Enterprise on BareMetal Overview:** This unit introduces NVIDIA AI Enterprise and its implementation on bare-metal infrastructure. You'll learn about NVIDIA AI Blueprints, NVIDIA Inference Microservice (NIM) for LLMs, and access essential documentation resources.
- **NVIDIA AI Enterprise BareMetal Deployment Overview:** This unit covers deployment methods, and hardware and software prerequisites for bare-metal implementations. You'll complete a hands-on lab preparing a bare-metal platform for NVIDIA AI Enterprise deployment.
- **NVIDIA AI Enterprise Deployment:** This unit explores NVIDIA AI Enterprise platform prerequisites and NGC resources including Catalog, CLI, and API. You'll perform a hands-on lab deploying NVIDIA AI Enterprise on Kubernetes in a BareMetal environment.
- **Management & Monitoring:** This unit addresses tools and techniques for monitoring, scaling, and maintaining NVIDIA AI Enterprise deployments. You'll learn about workload management options including Kubernetes and Base Command Manager Essentials.
- **Reference Use Case:** This unit demonstrates a practical enterprise RAG workflow deployment. You'll complete a hands-on lab implementing a RAG workflow on Kubernetes and access related documentation resources.
- **Troubleshooting and Support:** This unit covers common issues, resolution techniques, and support resources for bare-metal deployments. You'll learn troubleshooting methodologies specifically tailored for bare-metal NVIDIA AI Enterprise implementations.