# Orchestration and Job Scheduling

Unit 14 - Summary

- **Orchestration and Scheduling**: This unit gives an overview of orchestration and scheduling. IT compares the concepts and gives examples of common tools for each. It gives an overview of Machine Learning operations (MLOps).

- **Kubernetes:** Kubernetes is an open-source container orchestration system for automating software deployment, scaling, and management.

- **SLURM:** SLURM is an open-source cluster management and job scheduling system for Linux clusters.

---

### What is the difference between orchestration and scheduling?

Orchestration is container-based, designed for micro-services, and adapted for AI. It scales up/down for inferencing, manages entire workflows and processes, and load balances to distribute traffic across containers. Scheduling is bare-metal based, supports containers, and is designed for HPC. It has advanced scheduling features built-in, assigns tasks and jobs to available resources, and determines hosts with available resources to run containers.

### What is Kubernetes?

Kubernetes (K8s) is an open-source container orchestration system for automating software deployment, scaling, and management. It is often used in container based environments that need to scale to meet user needs and is useful for inference in AI clusters.

### What is SLURM?

SLURM is an open-source cluster management and job scheduling system for Linux clusters. It is highly scalable, fault-tolerant, and requires no kernel modifications. It schedules jobs to run on a subset of cluster resources and is excellent for AI training.

### What is MLOps?

MLOps is short for Machine Learning Operations. MLOps tools help to improve user productivity and speed up workflow, maximize utilization of resources, and allow projects to scale. MLOps tools help to bring consistency and repeatability to AI and ML workloads.