Project Synopsis: HPC Stack Cloud-Based HPC Cluster

Introduction

This project aims to design and implement a scalable, portable, and cost-effective High-Performance Computing (HPC) cluster using Docker containerization of Slurm on a cloud-based infrastructure. By leveraging AWS services and integrating Grafana for monitoring, the project delivers an efficient solution for real-time resource management and analytics.

➢ Key Objectives

1. **Simplified Deployment:** Streamline the deployment of HPC clusters using containerized Slurm components.
2. **Dynamic Scalability:** Facilitate on-demand addition of compute nodes for optimal resource utilization.
3. **Platform Independence:** Support deployment across diverse environments with Docker and AWS integration.
4. **Cost Optimization:** Deliver a cost-effective solution suitable for research and educational purposes.
5. **Enhanced Monitoring:** Provide comprehensive resource and performance analytics with Grafana.

➢ Implementation Highlights

- **Containerized Components:** Essential Slurm components (slurmctld, slurmd, slurmdbd, MySQL) are packaged as Docker containers for simplified deployment and maintenance.
- **AWS Infrastructure:** Leverages AWS EC2 instances for compute resources, VPC for secure network configuration, and EBS for persistent storage.
- **Hybrid HPC-Cloud Workflow:** Combines traditional HPC tools with cloud technologies to support a variety of computational workloads.
- **Real-Time Monitoring:** Integrates Grafana dashboards to visualize metrics such as CPU usage, memory consumption, job queue statistics, and node performance.

➢ Tech Stack

- **Orchestration & Monitoring:** Docker, Docker Compose, Grafana, Prometheus
- **HPC Scheduler:** Slurm Workload Manager
- **Cloud Services:** AWS EC2, VPC, S3, EBS
- **Database:** MySQL for Slurm accounting and resource tracking
- **Networking:** AWS VPC configurations, Security Groups for access control

➢ Milestones

1. **Achieve Slurm Cloud Environment**

Containerize Slurm components and deploy them on AWS infrastructure.
Establish a functional Slurm cluster capable of handling job submissions and resource allocation.

2. **Achieve GPU Support with Docker**

Integrate NVIDIA Docker runtime to enable GPU acceleration for HPC tasks.
Validate GPU-enabled workloads within the containerized environment.

3. **Achieve AI/ML-Related Tasks in DevOps Environment**

Develop pipelines for deploying and managing AI/ML workflows on the HPC cluster.
Utilize DevOps tools for automated job scheduling and resource provisioning.

4. **Achieve Secure VPC Network**

Configure AWS VPC with subnet isolation, routing, and security group policies.
Enforce secure communication between components and implement robust access control measures.

➢ Data Flow Diagram