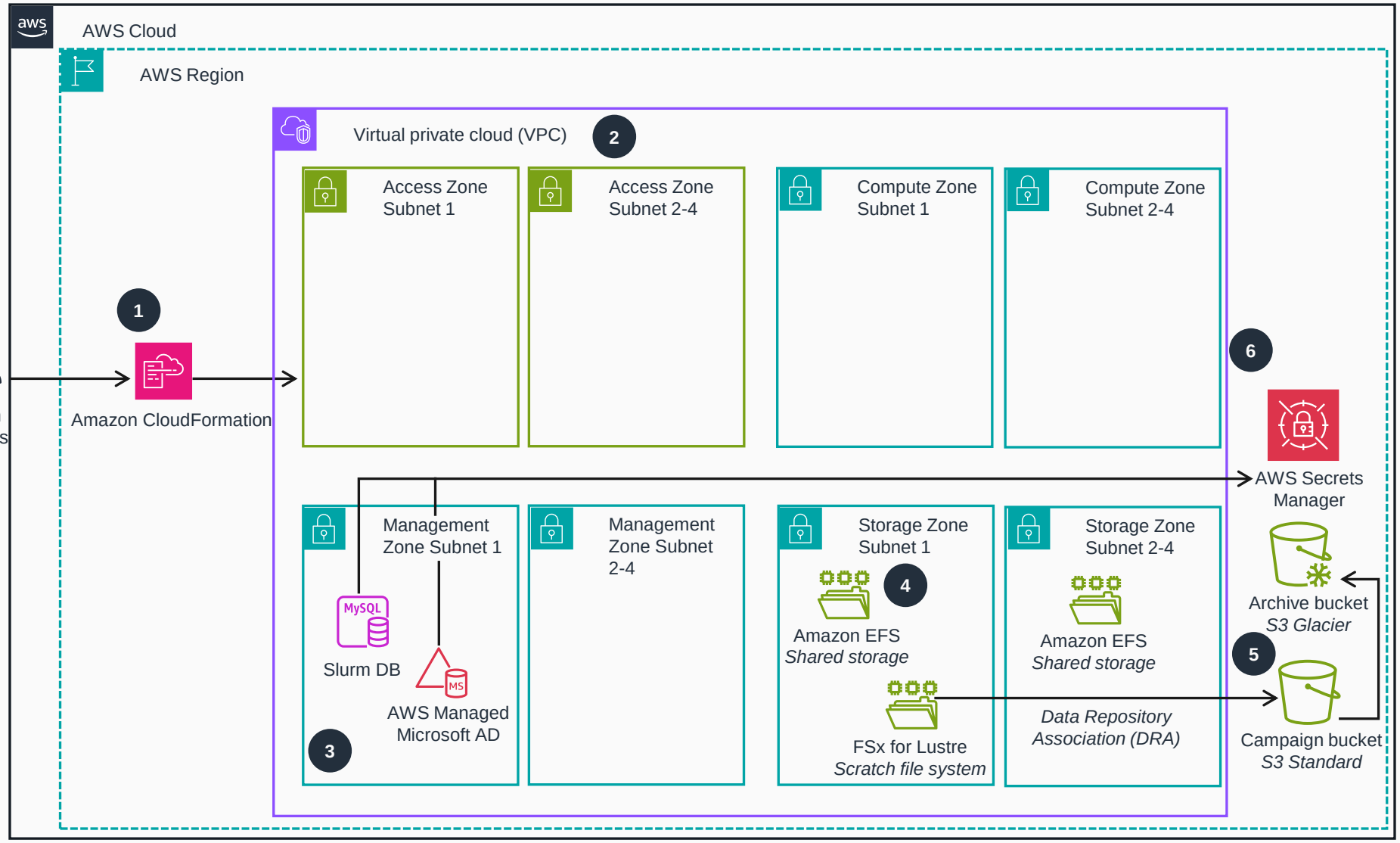
Guidance for Deploying High Performance Computing Clusters on AWS

Network, security, and infrastructure deployment

This architecture diagram shows how to deploy this Guidance using AWS CloudFormation templates that provision networking resources, security, and storage components. The next slide shows how HPC resources are deployed using the AWS ParallelCluster CloudFormation stack.



1. Admins/DevOps users can deploy this architecture using a series of **AWS CloudFormation** templates. These templates provision networking resources, including **Amazon Virtual Private Cloud** (Amazon VPC) and subnets. The templates also provision resources for security and storage, such as **Amazon Simple Storage Service** (Amazon S3), **Amazon Elastic File System** (Amazon EFS), and **Amazon FSx for Lustre**. There are optional templates included to deploy a Slurm accounting database (DB) and a Microsoft Active Directory user directory.

2. Four logical subnets (zones) are created, each in multiple Availability Zones (AZs), based on the target AWS Region. All required networking, networking access control list (ACLs), routes, and security resources are deployed. The four zones are: 1) Access Zone (public subnet), 2) Compute Zone, 3) Management Zone, and 4) Storage Zone (all private subnets).

3. An **Amazon RDS for MySQL** instance is created that will be used as the Slurm Accounting Database. This is set up in a single zone, or can be modified to be multi-AZ if preferred. One **AWS Directory Service** user directory is created across two AZs.

4. An **Amazon EFS** file system is created for shared cluster storage that is mounted in all of the deployed subnets for the Storage Zone. An **FSx for Lustre** file system is created that is used as a highly performant scratch file system in the preferred AZ.

5. Two **Amazon S3** buckets are created: one for campaign storage using **Amazon S3** Intelligent-Tiering, and one for archival storage using **Amazon S3 Glacier**.

6. Random passwords are generated for both the Slurm accounting database and the **Directory Service** that are stored securely in **AWS Secrets Manager**.

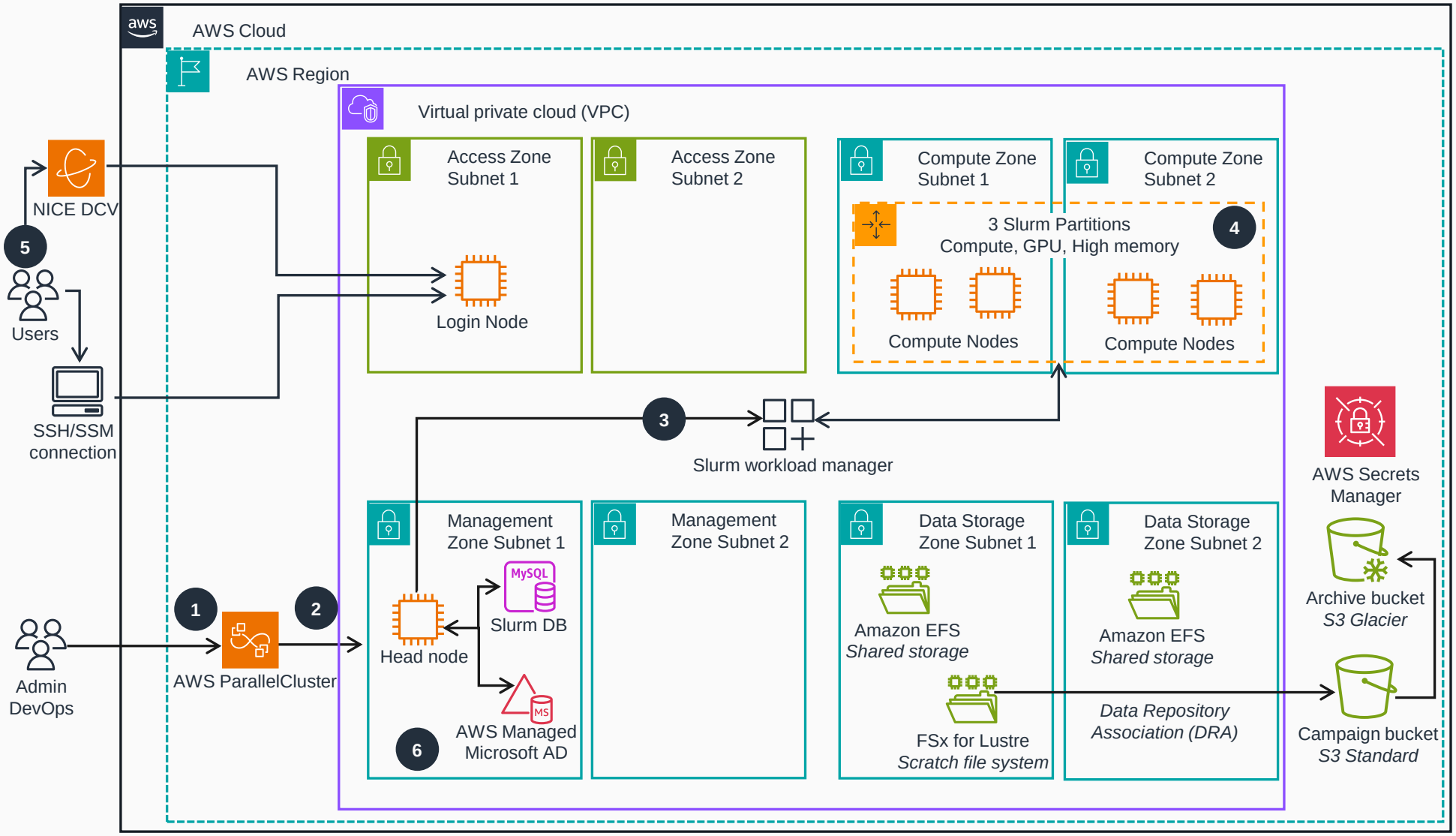
Reviewed for technical accuracy December 18, 2024

AWS Reference Architecture

Guidance for Deploying High Performance Computing Clusters on AWS

HPC cluster deployment

This architecture diagram shows how HPC resources are deployed using the AWS ParallelCluster CloudFormation stack. It references the network, storage, security, database, and user directory components from the previous slide.



1. Admins/DevOps users use the **AWS ParallelCluster AWS CloudFormation** stack to deploy HPC resources. Resources can reference the network, storage, security, database, and user directory from the previously launched **CloudFormation** stacks.
2. The **AWS ParallelCluster CloudFormation** template provisions a sample cluster configuration, which includes a head node deployed in a single Availability Zone within the Management zone. It also provisions a login node deployed in a single Availability Zone within the Access zone.
3. The Slurm workload manager is deployed on the head node and used for managing the HPC workflow processes.
4. The sample cluster configuration included creates two Slurm queues that provision compute nodes within the Compute zone. One queue uses compute-optimized **Amazon Elastic Compute Cloud** (Amazon EC2) instances, while the other queue utilizes GPU-accelerated **Amazon EC2** instances.
5. Users access this Guidance by establishing a connection to the deployed login node within the Access zone, utilizing either a NICE DCV, SSH, or an **AWS Systems Manager** Session Manager.
6. Users authenticate to the log in node using a username and password stored in the AWS Managed Microsoft Active Directory.

Reviewed for technical accuracy December 18, 2024

AWS Reference Architecture