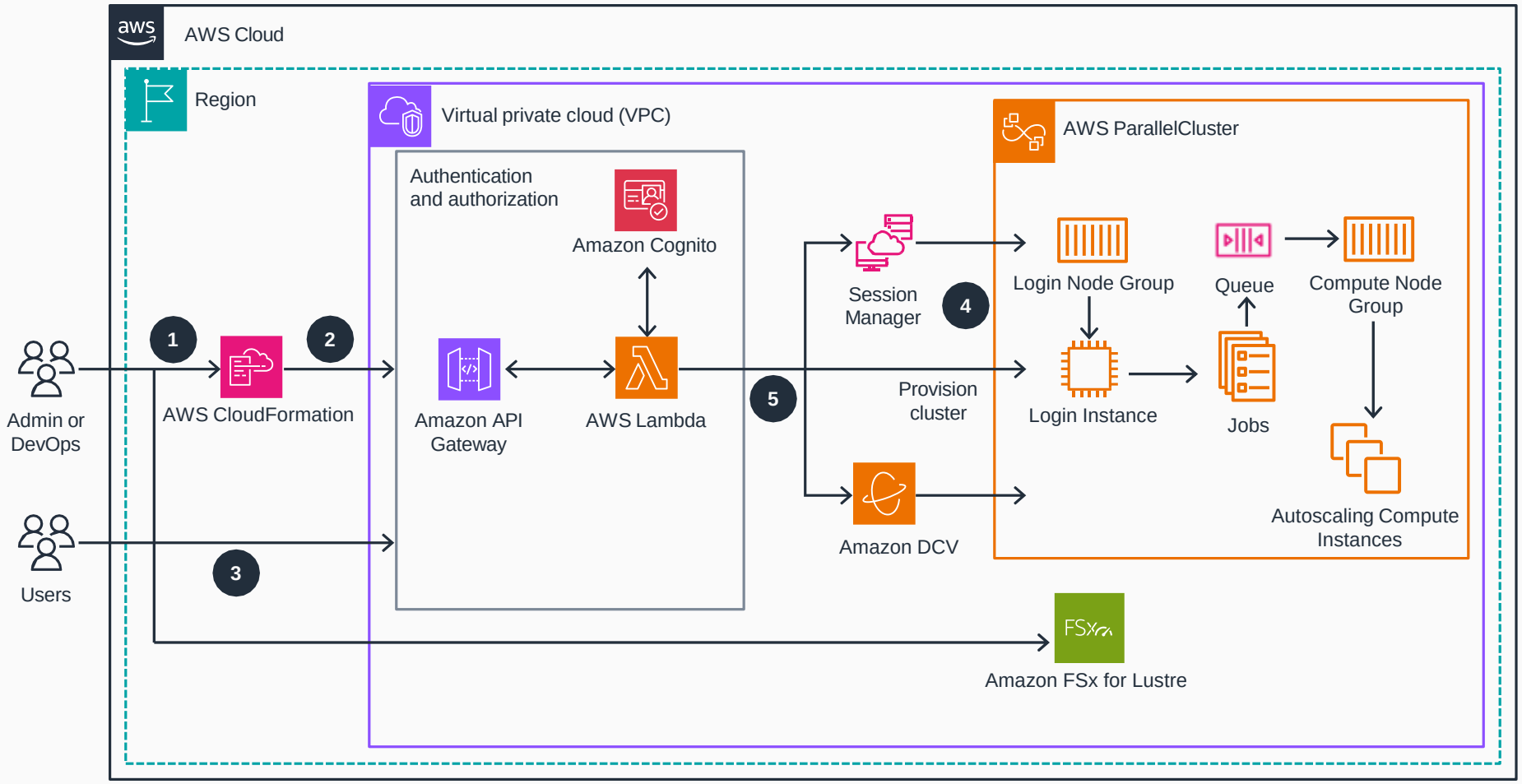
Guidance for Building a High-Performance Numerical Weather Prediction System on AWS – HPC Cluster Deployment

This architecture diagram shows how to provision the AWS ParallelCluster user interface (UI) and configure an HPC cluster with compute and storage capabilities. This slide shows how to set up the HPC cluster deployment.



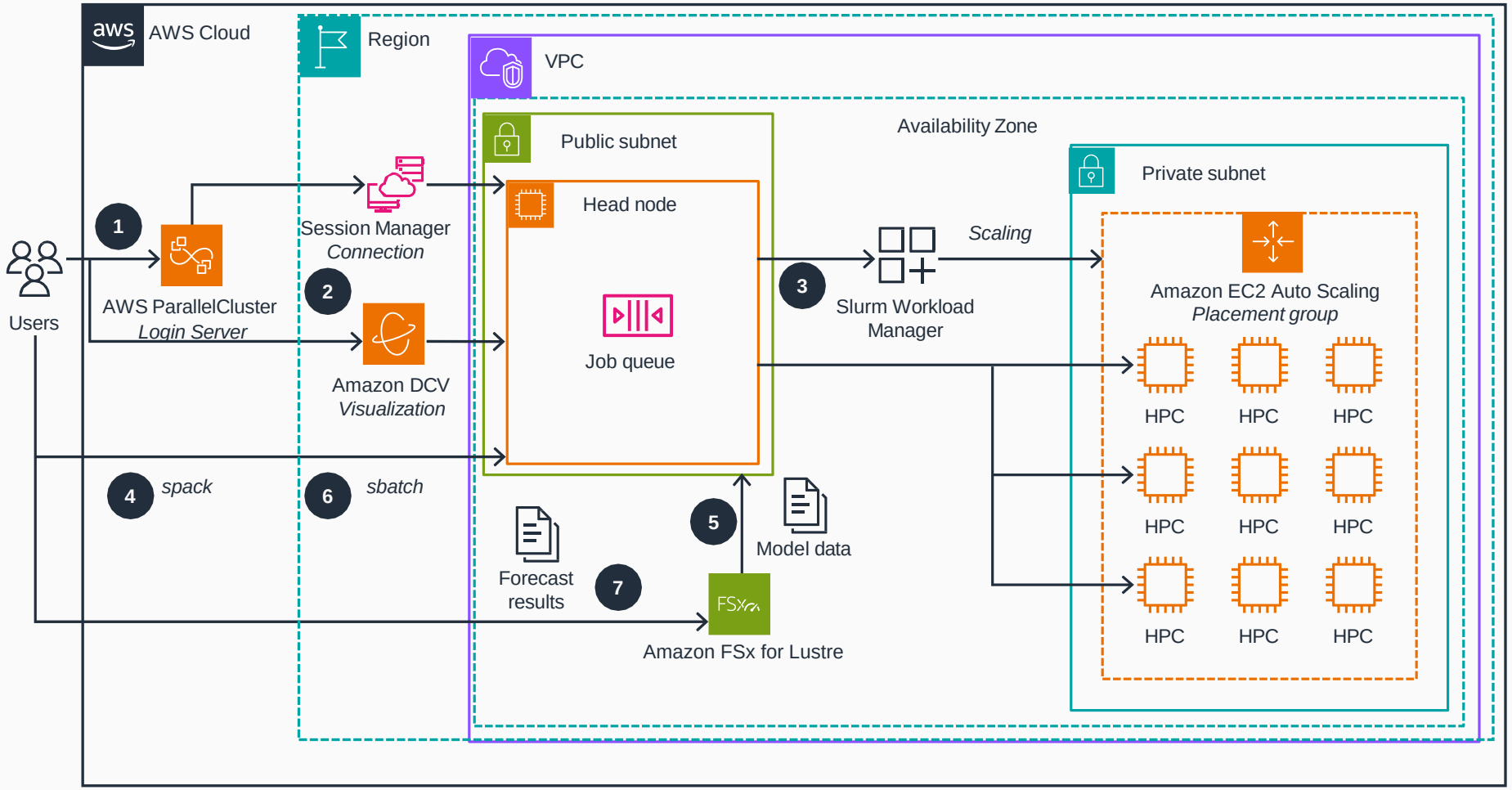
1. Users deploy the Guidance **AWS CloudFormation** stack to provision networking resources (**Amazon Virtual Private Cloud [Amazon VPC]** and subnets), storage (**Amazon FSx for Lustre**), and the **AWS ParallelCluster** UI.
2. The **AWS ParallelCluster** UI endpoint is available for user authentication using **Amazon API Gateway**
3. Users authenticate to the **AWS ParallelCluster** UI endpoint through an **AWS Lambda** function integrated with **Amazon Cognito** to handle log-in details.
4. Authenticated users provision HPC clusters through the **AWS ParallelCluster** UI using cluster specifications available with the Guidance code. Each HPC cluster contains several node groups dynamically provisioned for application workload implementation.
5. Users authenticated through the **AWS ParallelCluster** UI can connect to the HPC clusters either by using Session Manager from **AWS Systems Manager** or by using **Amazon Desktop Cloud Visualization (Amazon DCV)** sessions.

Reviewed for technical accuracy October 14, 2025

AWS Reference Architecture

Guidance for Building a High-Performance Numerical Weather Prediction System on AWS – Prediction Workflow

This architecture diagram shows CONUS weather prediction by deploying the Weather Research & Forecasting (WRF) model on AWS. This demonstrates the numerical weather prediction workflow using the HPC cluster from the previous slide.



1. Users authenticate to the **AWS ParallelCluster** UI (as detailed in the previous HPC Cluster Deployment architecture diagram).
2. Users connect to the HPC cluster using either the **AWS ParallelCluster** UI through the Session Manager or by using a **Amazon DCV** connection
3. Slurm Workload Manager (an HPC resource manager) is used to manage and scale the resources of **AWS ParallelCluster**, such as dynamically provisioned **Amazon Elastic Compute Cloud (Amazon EC2)** instances connected by the **Elastic Fabric Adapter (EFA)** network. Scaling is managed in an **Amazon EC2 Auto Scaling** placement group.
4. Spack (a software package manager for supercomputers) is installed. Spack is used to install necessary compilers, libraries like NCAR Command Language (NCL), and the Weather Research & the Forecasting (WRF) model.
5. **FSx for Lustre** storage was created along with the HPC cluster. The input data used to simulate the WRF test model (12-km CONUS) is copied to a local directory mounted to that storage.
6. Users create an *sbatch* script to run the CONUS 12-km model, submit that job, and monitor its implementation status by using a *squeue* command.
7. Numerical weather prediction results are stored in a locally mounted directory. Users can retrieve and visualize the results using NCL scripts.

Reviewed for technical accuracy October 14, 2025

AWS Reference Architecture