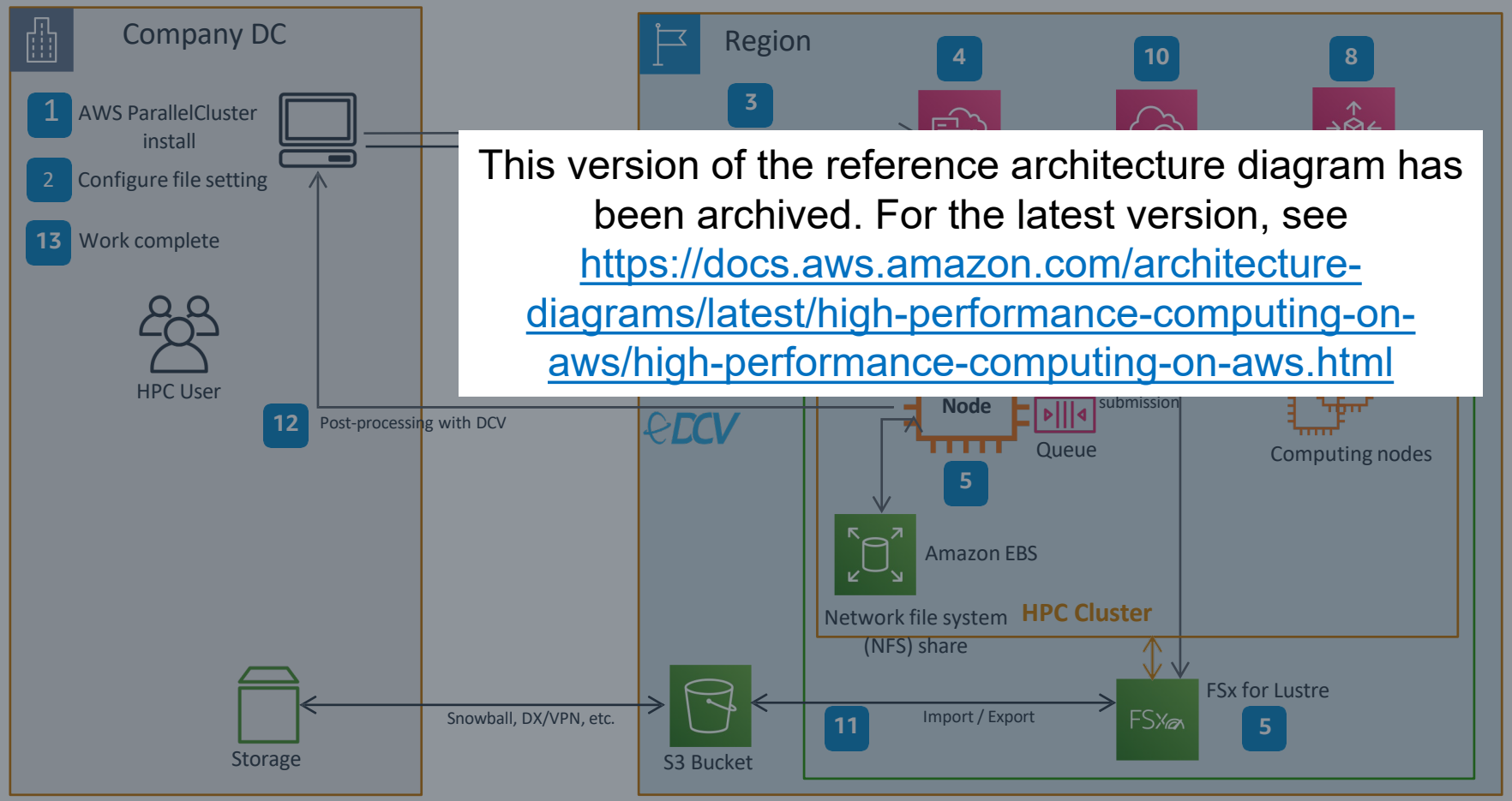
General HPC Architecture on AWS

A series of processes that constitute high-performance computing (HPC) on AWS with lift and shift approach

- This architecture can be applied when the on-premises HPC architecture is migrated to the AWS Cloud using the lift and shift method.
- This method is called a traditional architecture, and its advantage is that users who used HPC systems in an on-premises environment can build and use HPC systems in the AWS Cloud environment without much burden.
- It is almost the same as the on-premises environment, except the resource is defined in the form of a script using **AWS ParallelCluster**.
- Use this architecture to intuitively configure your HPC system in the AWS Cloud environment and use it to perform your simulations.



This version of the reference architecture diagram has been archived. For the latest version, see https://docs.aws.amazon.com/architecture-diagrams/latest/high-performance-computing-on-aws/high-performance-computing-on-aws.html

1. Installs **AWS ParallelCluster**, which is used to provision HPC resources.
2. Use the installed **AWS ParallelCluster** to define the resource you want to provision in the form of a script. This is called the "configure file".
3. Provision the configure file defined in Step 2 with an **AWS ParallelCluster** command.
4. The real provisioning of resources is performed through an infrastructure as code (IaC) service called **AWS CloudFormation** linked with **AWS ParallelCluster**.
5. When provisioning is complete, the defined resources are created. A head node (including defined scheduler) and a file system (**Amazon FSx for Lustre**) are created.
6. To perform the simulation, the user connects to the created head node through a secure shell protocol (SSH) or DCV connection.
7. Create a job script on the head node and submit it to the scheduler already installed on the head node. The job is queued until it is processed.
8. The amount of computing power defined in the job script is allocated to process the job.
9. A compute cluster to process the job in the queue is created, and computing is performed.
10. The created cluster nodes and various HPC resources are monitored through a monitoring service called **Amazon CloudWatch**.
11. The processed results can be stored in **Amazon Simple Storage Service (Amazon S3)**, and sent to the on-premises environment if necessary.
12. If necessary, you can do post-processing with DCV without transmitting the result data into on-premises.
13. When there are no more jobs to process, the cluster is deleted.

AWS Reference Architecture