Deploying Slurm with AWS ParallelCluster on Your AWS Cluster

AWS ParallelCluster is an AWS-supported open source cluster management tool that makes it easy for you to deploy and manage High Performance Computing (HPC) clusters on AWS. ParallelCluster uses a simple text file to model and provision all the resources needed for your HPC applications in an automated and secure manner. It also supports multiple instance types and job submission queues, and job schedulers like Slurm.

ParallelCluster's source code is hosted on the Amazon Web Services repository on GitHub. AWS ParallelCluster is available at no additional charge, and you pay only for the AWS resources needed to run your applications.

SchedMD® and AWS have partnered to optimize Slurm for AWS HPC infrastructure and integrating with AWS ParallelCluster so you can leverage its ease of deployment, autoscaling, and management for HPC clusters as Slurm manages all your large, complex workloads across cloud and on-premise resources. The combined Slurm and ParallelCluster solution enables migration of existing HPC workloads to AWS with little or no modifications and streamlined bursting and hybrid on-prem and cloud workload management for the most cost-effective operations.

Slurm workload manager for AWS is optimally deployed via AWS ParallelCluster using these below easy install and configuration instructions and resources. Familiarity with AWS is recommended when creating a Slurm cluster in AWS. Knowledge of Slurm's installation procedure is not required as Slurm is installed and configured as a step in the Parallel Cluster configuration process. For details on getting support for Slurm, training, or other professional services see [SchedMD's service list](#).

View [Slurm's documentation page](#) to learn more about using Slurm or to help in operating Slurm post-deployment.

Deploying Slurm with AWS ParallelCluster

1. Follow [AWS ParallelCluster User Guide](#) to set-up, install, and configure ParallelCluster best practices.
2. Configure Slurm as the scheduler as outlined in [AWS ParallelCluster User Guide cluster configuration, scheduler section](#). This step will only take a couple of minutes to configure and create the Slurm cluster.

The AWS cost of services are determined by how ParallelCluster is set up and those services chosen to use. For Slurm's operation, the only services that are required are a Slurm controller node (EC2 instance) and the compute nodes (EC2 instance) that execute workload. Using Slurm's autoscaling, nodes will be deprovisioned when not in use.

For an example walkthrough see this [AWS HPC workshop series](#).

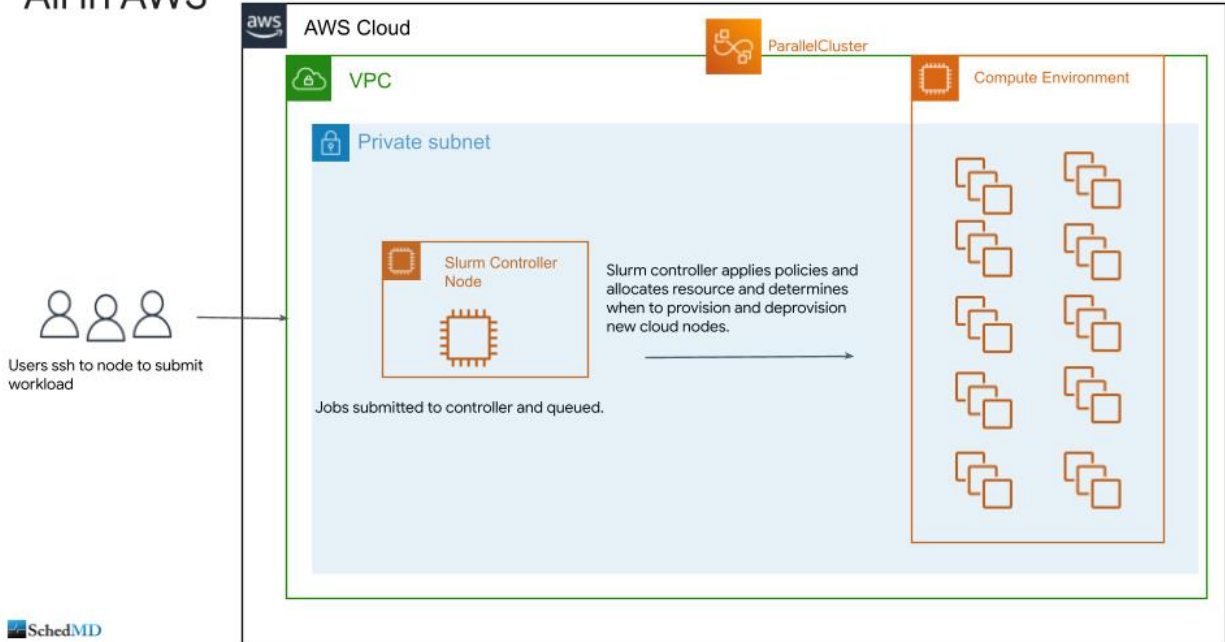**Deploying Slurm in a Hybrid Bursting Model**

With current onprem investments it might make more sense to run Slurm in your onprem environment and burst out to AWS when excess capacity is required or specific resources are needed that are not

available in your onprem environment such as a GPU. See this [blog post](#) and [Git repository](#) for details on setting up a Slurm hybrid cluster.

For more information on the benefits of Slurm, see the [Accelerate HPC and AI Workloads with Slurm on AWS Cloud Solution Brief](#)

Architectural Overview of AWS HPC Cluster with ParallelCluster and Slurm

Bursting to AWS

Onprem Cluster

Login Node

Controller Node

Users ssh to onprem Login
node to submit workload

Jobs submitted to controller
and queued

VPN connection

Slurm controller applies policies and
allocates resource and determines
when to run jobs on prem or burst to
AWS and provision nodes.

aws AWS Cloud

VPC

Private subnet
Compute Environment