

Service Name

AWS Parallel Computing Service

Service Overview

AWS Parallel Computing Service (AWS PCS) is a managed service that simplifies running and scaling high performance computing (HPC) workloads and building scientific and engineering models on AWS using Slurm. The service integrates AWS compute, storage, networking, and visualization resources while streamlining cluster operations through built-in management and observability capabilities. AWS PCS enables users to focus on research and innovation by providing a familiar environment for running applications and jobs without the operational overhead of managing HPC infrastructure. The service utilizes Slurm for workload management and automatically handles cluster scaling, maintenance, and updates.

Key Use Cases

- Computational chemistry and chemical formulation discovery for pharmaceutical research and materials science
- Scientific simulations including computational fluid dynamics (CFD), structural analysis, and engineering modeling
- Genomic sequencing, medical imaging analysis, and clinical research simulations for healthcare organizations
- Cryogenic electron microscopy (Cryo-EM) workloads for drug discovery and molecular structure determination
- Aerospace simulation workloads including mission planning, space operations, and vehicle innovation
- Multi-node MPI jobs requiring elastic scaling across distributed compute resources

Features

- **Managed Slurm Scheduler:** Fully managed Slurm workload manager with automated scaling and maintenance
- **Elastic Compute Node Groups:** Automatically scaling EC2 instances with customizable instance types and purchase options
- **Queue Management:** Job scheduling and execution across compute node groups with customizable policies
- **Built-in Observability:** Job completion logs, monitoring, and cluster management through AWS console
- **Slurm Customization:** Over 60 configurable Slurm parameters for scheduling, resource allocation, and access control

- **Managed Accounting:** Slurm accounting feature for usage tracking, resource limits, and chargeback reporting
- **Capacity Blocks Integration:** Support for EC2 Capacity Blocks to reserve GPU instances up to 8 weeks in advance
- **REST API Support:** HTTP interface for programmatic interaction with Slurm workload manager

Value Proposition

AWS PCS offers significant advantages over traditional on-premises HPC solutions and self-managed alternatives like AWS ParallelCluster. The service eliminates operational overhead by providing fully managed infrastructure with automatic scaling, patching, and updates:

- Organizations benefit from faster time-to-science with instant access to virtually unlimited computational resources, avoiding long queue times typical of traditional HPC environments
- AWS PCS provides enterprise-grade reliability with full AWS support and service-level agreements, making it ideal for production workloads requiring uptime guarantees
- The service offers cost optimization through elastic scaling, spot instance support, and pay-as-you-go pricing, eliminating the need for large capital investments in depreciating hardware
- Integration with the broader AWS ecosystem enables seamless access to storage, networking, and visualization tools

Service Integration

AWS PCS integrates deeply with core AWS services to provide comprehensive HPC solutions. Amazon EC2 provides the underlying compute infrastructure with support for various instance types including GPU-accelerated instances:

- Amazon EFS and Amazon FSx for Lustre deliver shared storage solutions for home directories and high-performance workloads
- Amazon VPC enables secure networking with customizable subnets and security groups
- AWS IAM provides identity and access management for cluster resources and user permissions
- Amazon CloudWatch delivers monitoring and logging capabilities including job completion logs
- AWS Systems Manager enables administrative access and management tasks
- AWS Secrets Manager handles cluster secret rotation for secure communications
- The service also integrates with AWS Backup for data protection and AWS PrivateLink for secure private connectivity

Onboarding and Offboarding

Customers begin by completing prerequisites including AWS account setup, AWS CLI installation, and required IAM permissions. The getting started process involves creating a VPC with appropriate subnets, configuring security groups, and setting up shared storage using Amazon EFS and FSx for Lustre:

- Users then create an AWS PCS cluster specifying the Slurm scheduler configuration, followed by defining compute node groups with desired instance types and scaling parameters
- A queue must be created to manage job scheduling across compute resources
- AWS provides comprehensive tutorials and CloudFormation templates to automate cluster deployment
- The service offers both console-based and CLI-driven setup options
- Once configured, users can connect via SSH or AWS Systems Manager to submit and manage jobs using standard Slurm commands

Getting Started Guide: <https://docs.aws.amazon.com/pcs/latest/userguide/getting-started.html>

Data Removal

When offboarding from AWS PCS, customers can delete cluster resources through the AWS console or CLI, which automatically terminates all associated EC2 instances and removes cluster configurations. Data stored in customer-owned resources like EFS and FSx file systems must be manually deleted by the customer. Job completion logs and accounting data stored in AWS-managed databases are automatically removed when clusters are deleted. Customers should backup any important data before initiating the deletion process as it is irreversible.

Backup/Restore and Disaster Recovery

AWS PCS leverages integrated AWS services for backup and disaster recovery capabilities. Customer data stored in Amazon EFS and FSx for Lustre can be backed up using AWS Backup for automated, policy-driven protection. EC2 instances can be protected through EBS snapshots and AMI creation. Job completion logs are automatically stored in CloudWatch Logs or S3 for durability. The service supports cross-region deployment for disaster recovery scenarios, though cluster configurations must be recreated in alternate regions.

Backup Capabilities

AWS PCS does not provide native backup capabilities as a service but integrates with AWS Backup for comprehensive data protection. Customer workload data stored on shared file systems (EFS, FSx) can be automatically backed up using AWS Backup policies:

- Compute node data can be protected through EBS snapshots and custom AMI creation
- Job metadata and logs are durably stored in AWS-managed services with built-in redundancy

Disaster Recovery

AWS PCS supports disaster recovery through multi-region deployment strategies and integration with AWS disaster recovery services. While clusters cannot be directly replicated across regions, customers can deploy identical cluster configurations in multiple regions using CloudFormation templates:

- Data replication can be achieved through EFS and FSx cross-region backup and restore capabilities
- AWS Elastic Disaster Recovery can protect compute workloads and data

Recovery Objectives

AWS PCS helps customers meet RPO and RTO objectives through automated scaling and rapid cluster deployment capabilities. New clusters can be provisioned within minutes using pre-configured templates, supporting aggressive RTO requirements. Data RPO objectives are met through integration with AWS Backup for frequent, automated snapshots of shared storage. The service's elastic nature allows rapid scaling to handle recovery workloads and catch-up processing after incidents.

Service Constraints

Service Quotas: <https://docs.aws.amazon.com/pcs/latest/userguide/service-endpoints-quotas.html>

FAQ: <https://aws.amazon.com/pcs/faqs/>

Technical Requirements

Service Documentation: <https://docs.aws.amazon.com/pcs/latest/userguide/what-is-service.html>

User Guide: <https://docs.aws.amazon.com/pcs/latest/userguide/getting-started.html>

API Reference: <https://docs.aws.amazon.com/pcs/latest/APIReference/Welcome.html>

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: <https://aws.amazon.com/pcs/pricing/>

Cost Optimization

AWS PCS offers several unique cost optimization features including elastic auto-scaling that automatically adds and removes compute nodes based on job queue demand, eliminating idle resource costs. The service supports EC2 Spot instances for significant cost savings on fault-tolerant workloads, and mixed instance types within compute node groups for optimal price-performance ratios:

- Capacity Blocks integration enables reserving GPU instances up to 8 weeks in advance with discounted rates
- The managed accounting feature provides detailed usage tracking for chargeback and resource optimization
- Pay-as-you-go pricing eliminates capital expenditure requirements, and the service automatically terminates unused resources to prevent unnecessary charges

Savings Considerations

Primary cost savings come from eliminating on-premises HPC infrastructure capital costs and operational overhead. Organizations save on hardware depreciation cycles typically lasting 5-8 years and reduce staffing costs for cluster management and maintenance:

- The elastic scaling model ensures customers only pay for compute resources during active job execution, avoiding the high utilization costs of traditional HPC systems
- Spot instance support can reduce compute costs by up to 90% for appropriate workloads
- The managed service eliminates costs associated with Slurm expertise, system administration, and infrastructure troubleshooting
- Multi-tenancy capabilities allow sharing resources across projects and teams, improving overall utilization efficiency and reducing per-user costs