aws

API Reference

AWS PCS

API Version 2023-02-10

AWS PCS: API Reference

Table of Contents

Welcome

AWS Parallel Computing Service (AWS PCS) is a managed service that makes it easier for you to run and scale your high performance computing (HPC) workloads, and build scientific and engineering models on AWS using Slurm. For more information, see the AWS Parallel Computing Service User Guide.

This reference describes the actions and data types of the service management API. You can use the AWS SDKs to call the API actions in software, or use the AWS Command Line Interface (AWS CLI) to call the API actions manually. These API actions manage the service through an AWS account.

The API actions operate on AWS PCS resources. A *resource* is an entity in AWS that you can work with. AWS services create resources when you use the features of the service. Examples of AWS PCS resources include clusters, compute node groups, and queues. For more information about resources in AWS, see Resource in the *AWS Resource Explorer User Guide*.

An AWS PCS *compute node* is an Amazon EC2 instance. You don't launch compute nodes directly. AWS PCS uses configuration information that you provide to launch compute nodes in your AWS account. You receive billing charges for your running compute nodes. AWS PCS automatically terminates your compute nodes when you delete the AWS PCS resources related to those compute nodes.

This document was last published on April 10, 2026.

Actions

The following actions are supported:

- CreateCluster
- CreateComputeNodeGroup
- CreateQueue
- DeleteCluster
- DeleteComputeNodeGroup
- DeleteQueue
- GetCluster
- GetComputeNodeGroup
- GetQueue
- ListClusters
- ListComputeNodeGroups
- ListQueues
- ListTagsForResource
- RegisterComputeNodeGroupInstance
- TagResource
- UntagResource
- UpdateCluster
- UpdateComputeNodeGroup
- UpdateQueue

CreateCluster

Creates a cluster in your account. AWS PCS creates the cluster controller in a service-owned account. The cluster controller communicates with the cluster resources in your account. The subnets and security groups for the cluster must already exist before you use this API action.

> **Note**
>
> It takes time for AWS PCS to create the cluster. The cluster is in a `Creating` state until it is ready to use. There can only be 1 cluster in a `Creating` state per AWS Region per AWS account. `CreateCluster` fails with a `ServiceQuotaExceededException` if there is already a cluster in a `Creating` state.

Request Syntax

```
{
   "clientToken": "string",
   "clusterName": "string",
   "networking": {
      "networkType": "string",
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   },
   "scheduler": {
      "type": "string",
      "version": "string"
   },
   "size": "string",
   "slurmConfiguration": {
      "accounting": {
         "defaultPurgeTimeInDays": number,
         "mode": "string"
      },
      "cgroupCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "scaleDownIdleTimeInSeconds": number,
```

```
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "slurmdbdCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "slurmRest": {
         "mode": "string"
      }
   },
   "tags": {
      "string" : "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterName

A name to identify the cluster. Example: `MyCluster`

Type: String

Length Constraints: Minimum length of 3. Maximum length of 40.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

networking

The networking configuration used to set up the cluster's control plane.

Type: NetworkingRequest object

Required: Yes

scheduler

The cluster management and job scheduling software associated with the cluster.

Type: SchedulerRequest object

Required: Yes

size

A value that determines the maximum number of compute nodes in the cluster and the maximum number of jobs (active and queued).

- `SMALL`: 32 compute nodes and 256 jobs
- `MEDIUM`: 512 compute nodes and 8192 jobs
- `LARGE`: 2048 compute nodes and 16,384 jobs

Type: String

Valid Values: `SMALL | MEDIUM | LARGE`

Required: Yes

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: ClusterSlurmConfigurationRequest object

Required: No

tags

1 or more tags added to the resource. Each tag consists of a tag key and tag value. The tag value is optional and can be an empty string.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Response Syntax

```
{
   "cluster": {
      "arn": "string",
      "createdAt": "string",
      "endpoints": [
         {
            "ipv6Address": "string",
            "port": "string",
            "privateIpAddress": "string",
            "publicIpAddress": "string",
            "type": "string"
         }
      ],
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "id": "string",
      "modifiedAt": "string",
      "name": "string",
      "networking": {
         "networkType": "string",
         "securityGroupIds": [ "string" ],
         "subnetIds": [ "string" ]
```

```
      },
      "scheduler": {
         "type": "string",
         "version": "string"
      },
      "size": "string",
      "slurmConfiguration": {
         "accounting": {
            "defaultPurgeTimeInDays": number,
            "mode": "string"
         },
         "authKey": {
            "secretArn": "string",
            "secretVersion": "string"
         },
         "cgroupCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "jwtAuth": {
            "jwtKey": {
               "secretArn": "string",
               "secretVersion": "string"
            }
         },
         "scaleDownIdleTimeInSeconds": number,
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmdbdCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmRest": {
            "mode": "string"
         }
      },
```

```
        "status": "string"
    }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

cluster

The cluster resource.

Type: Cluster object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.

- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateComputeNodeGroup

Creates a managed set of compute nodes. You associate a compute node group with a cluster through 1 or more AWS PCS queues or as part of the login fleet. A compute node group includes the definition of the compute properties and lifecycle management. AWS PCS uses the information you provide to this API action to launch compute nodes in your account. You can only specify subnets in the same Amazon VPC as your cluster. You receive billing charges for the compute nodes that AWS PCS launches in your account. You must already have a launch template before you call this API. For more information, see Launch an instance from a launch template in the *Amazon Elastic Compute Cloud User Guide for Linux Instances*.

Request Syntax

```
{
   "amiId": "string",
   "clientToken": "string",
   "clusterIdentifier": "string",
   "computeNodeGroupName": "string",
   "customLaunchTemplate": {
      "id": "string",
      "version": "string"
   },
   "iamInstanceProfileArn": "string",
   "instanceConfigs": [
      {
         "instanceType": "string"
      }
   ],
   "purchaseOption": "string",
   "scalingConfiguration": {
      "maxInstanceCount": number,
      "minInstanceCount": number
   },
   "slurmConfiguration": {
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ]
   },
```

```
    "spotOptions": {
       "allocationStrategy": "string"
    },
    "subnetIds": [ "string" ],
    "tags": {
       "string" : "string"
    }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

amiId

The ID of the Amazon Machine Image (AMI) that AWS PCS uses to launch compute nodes (Amazon EC2 instances). If you don't provide this value, AWS PCS uses the AMI ID specified in the custom launch template.

Type: String

Pattern: `ami-[a-z0-9]+`

Required: No

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster to create a compute node group in.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupName

A name to identify the cluster. Example: `MyCluster`

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

customLaunchTemplate

An Amazon EC2 launch template AWS PCS uses to launch compute nodes.

Type: CustomLaunchTemplate object

Required: Yes

iamInstanceProfileArn

The Amazon Resource Name (ARN) of the IAM instance profile used to pass an IAM role when launching EC2 instances. The role contained in your instance profile must have the `pcs:RegisterComputeNodeGroupInstance` permission and the role name must start with `AWSPCS` or must have the path `/aws-pcs/`. For more information, see IAM instance profiles for AWS PCS in the *AWS PCS User Guide*.

Type: String

Pattern: `arn:aws([a-zA-Z-]{0,10})?:iam::[0-9]{12}:instance-profile/([!-~]{1,510}/)?([\w+=,.@-]{1,128})`

Required: Yes

instanceConfigs

A list of EC2 instance configurations that AWS PCS can provision in the compute node group.

Type: Array of InstanceConfig objects

Required: Yes

purchaseOption

Specifies how EC2 instances are purchased on your behalf. AWS PCS supports On-Demand Instances, Spot Instances, and Amazon EC2 Capacity Blocks for ML. For more information, see Amazon EC2 billing and purchasing options in the *Amazon Elastic Compute Cloud User Guide*. For more information about AWS PCS support for Capacity Blocks, see Using Amazon EC2 Capacity Blocks for ML with AWS PCS in the *AWS PCS User Guide*. If you don't provide this option, it defaults to On-Demand.

Type: String

Valid Values: `ONDEMAND | SPOT | CAPACITY_BLOCK`

Required: No

scalingConfiguration

Specifies the boundaries of the compute node group auto scaling.

Type: ScalingConfigurationRequest object

Required: Yes

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: ComputeNodeGroupSlurmConfigurationRequest object

Required: No

spotOptions

Additional configuration when you specify `SPOT` as the `purchaseOption` for the `CreateComputeNodeGroup` API action.

Type: SpotOptions object

Required: No

subnetIds

The list of subnet IDs where the compute node group launches instances. Subnets must be in the same VPC as the cluster.

Type: Array of strings

Required: Yes

tags

1 or more tags added to the resource. Each tag consists of a tag key and tag value. The tag value is optional and can be an empty string.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Response Syntax

```
{
   "computeNodeGroup": {
      "amiId": "string",
      "arn": "string",
      "clusterId": "string",
      "createdAt": "string",
      "customLaunchTemplate": {
         "id": "string",
         "version": "string"
      },
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "iamInstanceProfileArn": "string",
      "id": "string",
      "instanceConfigs": [
         {
            "instanceType": "string"
         }
```

```
      ],
      "modifiedAt": "string",
      "name": "string",
      "purchaseOption": "string",
      "scalingConfiguration": {
         "maxInstanceCount": number,
         "minInstanceCount": number
      },
      "slurmConfiguration": {
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ]
      },
      "spotOptions": {
         "allocationStrategy": "string"
      },
      "status": "string",
      "subnetIds": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

computeNodeGroup

A compute node group associated with a cluster.

Type: ComputeNodeGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

CreateQueue

Creates a job queue. You must associate 1 or more compute node groups with the queue. You can associate 1 compute node group with multiple queues.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string",
   "computeNodeGroupConfigurations": [
      {
         "computeNodeGroupId": "string"
      }
   ],
   "queueName": "string",
   "slurmConfiguration": {
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ]
   },
   "tags": {
      "string" : "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster for which to create a queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupConfigurations

The list of compute node group configurations to associate with the queue. Queues assign jobs to associated compute node groups.

Type: Array of ComputeNodeGroupConfiguration objects

Required: No

queueName

A name to identify the queue.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: QueueSlurmConfigurationRequest object

Required: No

tags

1 or more tags added to the resource. Each tag consists of a tag key and tag value. The tag value is optional and can be an empty string.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: No

Response Syntax

```
{
   "queue": {
      "arn": "string",
      "clusterId": "string",
      "computeNodeGroupConfigurations": [
         {
            "computeNodeGroupId": "string"
         }
      ],
      "createdAt": "string",
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "id": "string",
      "modifiedAt": "string",
      "name": "string",
      "slurmConfiguration": {
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ]
```

```
      },
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

queue

A queue resource.

Type: Queue object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.

- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteCluster

Deletes a cluster and all its linked resources. You must delete all queues and compute node groups associated with the cluster before you can delete the cluster.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster to delete.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteComputeNodeGroup

Deletes a compute node group. You must delete all queues associated with the compute node group first.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string",
   "computeNodeGroupIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster of the compute node group.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupIdentifier

The name or ID of the compute node group to delete.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.

- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteQueue

Deletes a job queue. If the compute node group associated with this queue isn't associated with any other queues, AWS PCS terminates all the compute nodes for this queue.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string",
   "queueIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster of the queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

queueIdentifier

The name or ID of the queue to delete.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.

- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetCluster

Returns detailed information about a running cluster in your account. This API action provides networking information, endpoint information for communication with the scheduler, and provisioning status.

Request Syntax

```
{
   "clusterIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clusterIdentifier

The name or ID of the cluster.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

Response Syntax

```
{
   "cluster": {
      "arn": "string",
      "createdAt": "string",
      "endpoints": [
         {
            "ipv6Address": "string",
            "port": "string",
            "privateIpAddress": "string",
            "publicIpAddress": "string",
            "type": "string"
```

```
      }
   ],
   "errorInfo": [
      {
         "code": "string",
         "message": "string"
      }
   ],
   "id": "string",
   "modifiedAt": "string",
   "name": "string",
   "networking": {
      "networkType": "string",
      "securityGroupIds": [ "string" ],
      "subnetIds": [ "string" ]
   },
   "scheduler": {
      "type": "string",
      "version": "string"
   },
   "size": "string",
   "slurmConfiguration": {
      "accounting": {
         "defaultPurgeTimeInDays": number,
         "mode": "string"
      },
      "authKey": {
         "secretArn": "string",
         "secretVersion": "string"
      },
      "cgroupCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "jwtAuth": {
         "jwtKey": {
            "secretArn": "string",
            "secretVersion": "string"
         }
      },
      "scaleDownIdleTimeInSeconds": number,
      "slurmCustomSettings": [
```

```
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmdbdCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmRest": {
            "mode": "string"
         }
      },
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

cluster

The cluster resource.

Type: Cluster object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.

- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2

- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetComputeNodeGroup

Returns detailed information about a compute node group. This API action provides networking information, EC2 instance type, compute node group status, and scheduler (such as Slurm) configuration.

Request Syntax

```
{
   "clusterIdentifier": "string",
   "computeNodeGroupIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clusterIdentifier

The name or ID of the cluster.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupIdentifier

The name or ID of the compute node group.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

Response Syntax

```
{
   "computeNodeGroup": {
```

```
      "amiId": "string",
      "arn": "string",
      "clusterId": "string",
      "createdAt": "string",
      "customLaunchTemplate": {
         "id": "string",
         "version": "string"
      },
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "iamInstanceProfileArn": "string",
      "id": "string",
      "instanceConfigs": [
         {
            "instanceType": "string"
         }
      ],
      "modifiedAt": "string",
      "name": "string",
      "purchaseOption": "string",
      "scalingConfiguration": {
         "maxInstanceCount": number,
         "minInstanceCount": number
      },
      "slurmConfiguration": {
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ]
      },
      "spotOptions": {
         "allocationStrategy": "string"
      },
      "status": "string",
      "subnetIds": [ "string" ]
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

computeNodeGroup

A compute node group associated with a cluster.

Type: ComputeNodeGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetQueue

Returns detailed information about a queue. The information includes the compute node groups that the queue uses to schedule jobs.

Request Syntax

```
{
   "clusterIdentifier": "string",
   "queueIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clusterIdentifier

The name or ID of the cluster of the queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

queueIdentifier

The name or ID of the queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

Response Syntax

```
{
   "queue": {
      "arn": "string",
```

```
      "clusterId": "string",
      "computeNodeGroupConfigurations": [
         {
            "computeNodeGroupId": "string"
         }
      ],
      "createdAt": "string",
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "id": "string",
      "modifiedAt": "string",
      "name": "string",
      "slurmConfiguration": {
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ]
      },
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

queue

A queue resource.

Type: Queue object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListClusters

Returns a list of running clusters in your account.

Request Syntax

```
{
   "maxResults": number,
   "nextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

maxResults

The maximum number of results that are returned per call. You can use `nextToken` to obtain further pages of results. The default is 10 results, and the maximum allowed page size is 100 results. A value of 0 uses the default.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

Required: No

Response Syntax

```
{
   "clusters": [
      {
         "arn": "string",
         "createdAt": "string",
         "id": "string",
         "modifiedAt": "string",
         "name": "string",
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

clusters

The list of clusters.

Type: Array of ClusterSummary objects

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListComputeNodeGroups

Returns a list of all compute node groups associated with a cluster.

Request Syntax

```
{
   "clusterIdentifier": "string",
   "maxResults": number,
   "nextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clusterIdentifier

The name or ID of the cluster to list compute node groups for.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

maxResults

The maximum number of results that are returned per call. You can use `nextToken` to obtain further pages of results. The default is 10 results, and the maximum allowed page size is 100 results. A value of 0 uses the default.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the

returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

Required: No

Response Syntax

```
{
   "computeNodeGroups": [
      {
         "arn": "string",
         "clusterId": "string",
         "createdAt": "string",
         "id": "string",
         "modifiedAt": "string",
         "name": "string",
         "status": "string"
      }
   ],
   "nextToken": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

computeNodeGroups

The list of compute node groups for the cluster.

Type: Array of ComputeNodeGroupSummary objects

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination

token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListQueues

Returns a list of all queues associated with a cluster.

Request Syntax

```
{
   "clusterIdentifier": "string",
   "maxResults": number,
   "nextToken": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clusterIdentifier

The name or ID of the cluster to list queues for.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

maxResults

The maximum number of results that are returned per call. You can use `nextToken` to obtain further pages of results. The default is 10 results, and the maximum allowed page size is 100 results. A value of 0 uses the default.

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 100.

Required: No

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the

returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

Required: No

Response Syntax

```
{
   "nextToken": "string",
   "queues": [
      {
         "arn": "string",
         "clusterId": "string",
         "createdAt": "string",
         "id": "string",
         "modifiedAt": "string",
         "name": "string",
         "status": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

nextToken

The value of `nextToken` is a unique pagination token for each page of results returned. If `nextToken` is returned, there are more results available. Make the call again using the returned token to retrieve the next page. Keep all other arguments unchanged. Each pagination token expires after 24 hours. Using an expired pagination token returns an `HTTP 400 InvalidToken` error.

Type: String

queues

The list of queues associated with the cluster.

Type: Array of QueueSummary objects

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

ListTagsForResource

Returns a list of all tags on an AWS PCS resource.

Request Syntax

```
{
   "resourceArn": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceArn

The Amazon Resource Name (ARN) of the resource for which to list tags.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `arn:aws.*:pcs:.*:[0-9]{12}:.*/[a-z0-9_\-]+`

Required: Yes

Response Syntax

```
{
   "tags": {
      "string" : "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

tags

1 or more tags added to the resource. Each tag consists of a tag key and tag value. The tag value is optional and can be an empty string.

Type: String to string map

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

RegisterComputeNodeGroupInstance

> **Important**
>
> This API action isn't intended for you to use.

AWS PCS uses this API action to register the compute nodes it launches in your account.

Request Syntax

```
{
   "bootstrapId": "string",
   "clusterIdentifier": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

bootstrapId

The client-generated token to allow for retries.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1000.

Pattern: `[\S]+`

Required: Yes

clusterIdentifier

The name or ID of the cluster to register the compute node group instance in.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

Response Syntax

```
{
   "endpoints": [
      {
         "ipv6Address": "string",
         "port": "string",
         "privateIpAddress": "string",
         "publicIpAddress": "string",
         "type": "string"
      }
   ],
   "nodeID": "string",
   "sharedSecret": "string"
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

endpoints

The list of endpoints available for interaction with the scheduler.

Type: Array of Endpoint objects

nodeID

The scheduler node ID for this instance.

Type: String

sharedSecret

For the Slurm scheduler, this is the shared Munge key the scheduler uses to authenticate compute node group instances.

Type: String

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

TagResource

Adds or edits tags on an AWS PCS resource. Each tag consists of a tag key and a tag value. The tag key and tag value are case-sensitive strings. The tag value can be an empty (null) string. To add a tag, specify a new tag key and a tag value. To edit a tag, specify an existing tag key and a new tag value.

Request Syntax

```
{
   "resourceArn": "string",
   "tags": {
      "string" : "string"
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceArn

The Amazon Resource Name (ARN) of the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `arn:aws.*:pcs:.*:[0-9]{12}:.*/[a-z0-9_\-]+`

Required: Yes

tags

1 or more tags added to the resource. Each tag consists of a tag key and tag value. The tag value is optional and can be an empty string.

Type: String to string map

Map Entries: Maximum number of 200 items.

Key Length Constraints: Minimum length of 1. Maximum length of 128.

Value Length Constraints: Minimum length of 0. Maximum length of 256.

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UntagResource

Deletes tags from an AWS PCS resource. To delete a tag, specify the tag key and the Amazon Resource Name (ARN) of the AWS PCS resource.

Request Syntax

```
{
   "resourceArn": "string",
   "tagKeys": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

resourceArn

The Amazon Resource Name (ARN) of the resource.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1011.

Pattern: `arn:aws.*:pcs:.*:[0-9]{12}:.*/[a-z0-9_\-]+`

Required: Yes

tagKeys

1 or more tag keys to remove from the resource. Specify only tag keys and not tag values.

Type: Array of strings

Array Members: Minimum number of 1 item. Maximum number of 200 items.

Length Constraints: Minimum length of 1. Maximum length of 128.

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Error Types.

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python

- AWS SDK for Ruby V3

UpdateCluster

Updates a cluster configuration. You can modify Slurm scheduler settings, accounting configuration, and security groups for an existing cluster.

> **Note**
>
> You can only update clusters that are in `ACTIVE`, `UPDATE_FAILED`, or `SUSPENDED` state. All associated resources (queues and compute node groups) must be in `ACTIVE` state before you can update the cluster.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string",
   "slurmConfiguration": {
      "accounting": {
         "defaultPurgeTimeInDays": number,
         "mode": "string"
      },
      "cgroupCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "scaleDownIdleTimeInSeconds": number,
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
      "slurmdbdCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ],
```

```
      "slurmRest": {
         "mode": "string"
      }
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster to update.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: UpdateClusterSlurmConfigurationRequest object

Required: No

Response Syntax

```
{
   "cluster": {
      "arn": "string",
      "createdAt": "string",
      "endpoints": [
         {
            "ipv6Address": "string",
            "port": "string",
            "privateIpAddress": "string",
            "publicIpAddress": "string",
            "type": "string"
         }
      ],
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "id": "string",
      "modifiedAt": "string",
      "name": "string",
      "networking": {
         "networkType": "string",
         "securityGroupIds": [ "string" ],
         "subnetIds": [ "string" ]
      },
      "scheduler": {
         "type": "string",
         "version": "string"
      },
      "size": "string",
      "slurmConfiguration": {
         "accounting": {
            "defaultPurgeTimeInDays": number,
            "mode": "string"
         },
         "authKey": {
            "secretArn": "string",
            "secretVersion": "string"
         },
```

```
         "cgroupCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "jwtAuth": {
            "jwtKey": {
               "secretArn": "string",
               "secretVersion": "string"
            }
         },
         "scaleDownIdleTimeInSeconds": number,
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmdbdCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ],
         "slurmRest": {
            "mode": "string"
         }
      },
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

cluster

The cluster resource and configuration.

Type: Cluster object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateComputeNodeGroup

Updates a compute node group. You can update many of the fields related to your compute node group including the configurations for networking, compute nodes, and settings specific to your scheduler (such as Slurm).

Request Syntax

```
{
   "amiId": "string",
   "clientToken": "string",
   "clusterIdentifier": "string",
   "computeNodeGroupIdentifier": "string",
   "customLaunchTemplate": {
      "id": "string",
      "version": "string"
   },
   "iamInstanceProfileArn": "string",
   "purchaseOption": "string",
   "scalingConfiguration": {
      "maxInstanceCount": number,
      "minInstanceCount": number
   },
   "slurmConfiguration": {
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ]
   },
   "spotOptions": {
      "allocationStrategy": "string"
   },
   "subnetIds": [ "string" ]
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

amiId

The ID of the Amazon Machine Image (AMI) that AWS PCS uses to launch instances. If not provided, AWS PCS uses the AMI ID specified in the custom launch template.

Type: String

Pattern: `ami-[a-z0-9]+`

Required: No

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster of the compute node group.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupIdentifier

The name or ID of the compute node group.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

customLaunchTemplate

An Amazon EC2 launch template AWS PCS uses to launch compute nodes.

Type: CustomLaunchTemplate object

Required: No

iamInstanceProfileArn

The Amazon Resource Name (ARN) of the IAM instance profile used to pass an IAM role when launching EC2 instances. The role contained in your instance profile must have the `pcs:RegisterComputeNodeGroupInstance` permission and the role name must start with `AWSPCS` or must have the path `/aws-pcs/`. For more information, see IAM instance profiles for AWS PCS in the *AWS PCS User Guide*.

Type: String

Pattern: `arn:aws([a-zA-Z-]{0,10})?:iam::[0-9]{12}:instance-profile/([!-~]{1,510}/)?([\w+=,.@-]{1,128})`

Required: No

purchaseOption

Specifies how EC2 instances are purchased on your behalf. AWS PCS supports On-Demand Instances, Spot Instances, and Amazon EC2 Capacity Blocks for ML. For more information, see Amazon EC2 billing and purchasing options in the *Amazon Elastic Compute Cloud User Guide*. For more information about AWS PCS support for Capacity Blocks, see Using Amazon EC2 Capacity Blocks for ML with AWS PCS in the *AWS PCS User Guide*. If you don't provide this option, it defaults to On-Demand.

Type: String

Valid Values: `ONDEMAND | SPOT | CAPACITY_BLOCK`

Required: No

scalingConfiguration

Specifies the boundaries of the compute node group auto scaling.

Type: ScalingConfigurationRequest object

Required: No

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: UpdateComputeNodeGroupSlurmConfigurationRequest object

Required: No

spotOptions

Additional configuration when you specify `SPOT` as the `purchaseOption` for the `CreateComputeNodeGroup` API action.

Type: SpotOptions object

Required: No

subnetIds

The list of subnet IDs where the compute node group provisions instances. The subnets must be in the same VPC as the cluster.

Type: Array of strings

Required: No

Response Syntax

```
{
   "computeNodeGroup": {
      "amiId": "string",
      "arn": "string",
      "clusterId": "string",
      "createdAt": "string",
      "customLaunchTemplate": {
         "id": "string",
         "version": "string"
      },
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
```

```
            }
         ],
         "iamInstanceProfileArn": "string",
         "id": "string",
         "instanceConfigs": [
            {
               "instanceType": "string"
            }
         ],
         "modifiedAt": "string",
         "name": "string",
         "purchaseOption": "string",
         "scalingConfiguration": {
            "maxInstanceCount": number,
            "minInstanceCount": number
         },
         "slurmConfiguration": {
            "slurmCustomSettings": [
               {
                  "parameterName": "string",
                  "parameterValue": "string"
               }
            ]
         },
         "spotOptions": {
            "allocationStrategy": "string"
         },
         "status": "string",
         "subnetIds": [ "string" ]
      }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

computeNodeGroup

A compute node group associated with a cluster.

Type: ComputeNodeGroup object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++

- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateQueue

Updates the compute node group configuration of a queue. Use this API to change the compute node groups that the queue can send jobs to.

Request Syntax

```
{
   "clientToken": "string",
   "clusterIdentifier": "string",
   "computeNodeGroupConfigurations": [
      {
         "computeNodeGroupId": "string"
      }
   ],
   "queueIdentifier": "string",
   "slurmConfiguration": {
      "slurmCustomSettings": [
         {
            "parameterName": "string",
            "parameterValue": "string"
         }
      ]
   }
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters.

The request accepts the following data in JSON format.

clientToken

A unique, case-sensitive identifier that you provide to ensure the idempotency of the request. Idempotency ensures that an API request completes only once. With an idempotent request, if the original request completes successfully, the subsequent retries with the same client token return the result from the original successful request and they have no additional effect. If you don't specify a client token, the AWS CLI and SDK automatically generate 1 for you.

Type: String

Length Constraints: Minimum length of 8. Maximum length of 100.

Required: No

clusterIdentifier

The name or ID of the cluster of the queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,40})`

Required: Yes

computeNodeGroupConfigurations

The list of compute node group configurations to associate with the queue. Queues assign jobs to associated compute node groups.

Type: Array of ComputeNodeGroupConfiguration objects

Required: No

queueIdentifier

The name or ID of the queue.

Type: String

Pattern: `(pcs_[a-zA-Z0-9]+|[A-Za-z][A-Za-z0-9-]{2,25})`

Required: Yes

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: UpdateQueueSlurmConfigurationRequest object

Required: No

Response Syntax

```
{
   "queue": {
      "arn": "string",
```

```
      "clusterId": "string",
      "computeNodeGroupConfigurations": [
         {
            "computeNodeGroupId": "string"
         }
      ],
      "createdAt": "string",
      "errorInfo": [
         {
            "code": "string",
            "message": "string"
         }
      ],
      "id": "string",
      "modifiedAt": "string",
      "name": "string",
      "slurmConfiguration": {
         "slurmCustomSettings": [
            {
               "parameterName": "string",
               "parameterValue": "string"
            }
         ]
      },
      "status": "string"
   }
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

queue

A queue resource.

Type: Queue object

Errors

For information about the errors that are common to all actions, see Common Error Types.

AccessDeniedException

You don't have permission to perform the action.

Examples

- The launch template instance profile doesn't pass `iam:PassRole` verification.
- There is a mismatch between the account ID and cluster ID.
- The cluster ID doesn't exist.
- The EC2 instance isn't present.

HTTP Status Code: 400

ConflictException

Your request has conflicting operations. This can occur if you're trying to perform more than 1 operation on the same resource at the same time.

Examples

- A cluster with the same name already exists.
- A cluster isn't in `ACTIVE` status.
- A cluster to delete is in an unstable state. For example, because it still has `ACTIVE` node groups or queues.
- A queue already exists in a cluster.

resourceId

The unique identifier of the resource that caused the conflict exception.

resourceType

The type or category of the resource that caused the conflict exception."

HTTP Status Code: 400

InternalServerException

AWS PCS can't process your request right now. Try again later.

HTTP Status Code: 500

ResourceNotFoundException

The requested resource can't be found. The cluster, node group, or queue you're attempting to get, update, list, or delete doesn't exist.

Examples

resourceId

The unique identifier of the resource that was not found.

resourceType

The type or category of the resource that was not found.

HTTP Status Code: 400

ServiceQuotaExceededException

You exceeded your service quota. Service quotas, also referred to as limits, are the maximum number of service resources or operations for your AWS account. To learn how to increase your service quota, see Requesting a quota increase in the *Service Quotas User Guide*

Examples

- The max number of clusters or queues has been reached for the account.
- The max number of compute node groups has been reached for the associated cluster.
- The total of `maxInstances` across all compute node groups has been reached for associated cluster.

quotaCode

The **quota code** of the service quota that was exceeded.

resourceId

The unique identifier of the resource that caused the quota to be exceeded.

resourceType

The type or category of the resource that caused the quota to be exceeded.

serviceCode

The service code associated with the quota that was exceeded.

HTTP Status Code: 400

ThrottlingException

Your request exceeded a request rate quota. Check the resource's request rate quota and try again.

retryAfterSeconds

The number of seconds to wait before retrying the request.

HTTP Status Code: 400

ValidationException

The request isn't valid.

Examples

- Your request contains malformed JSON or unsupported characters.
- The scheduler version isn't supported.
- There are networking related errors, such as network validation failure.
- AMI type is `CUSTOM` and the launch template doesn't define the AMI ID, or the AMI type is AL2 and the launch template defines the AMI.

fieldList

A list of fields or properties that failed validation.

reason

The specific reason or cause of the validation error.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface V2
- AWS SDK for .NET V4
- AWS SDK for C++
- AWS SDK for Go v2
- AWS SDK for Java V2
- AWS SDK for JavaScript V3
- AWS SDK for Kotlin

- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Data Types

The AWS Parallel Computing Service API contains several data types that various actions use. This section describes each data type in detail.

> **Note**
>
> The order of each element in a data type structure is not guaranteed. Applications should not assume a particular order.

The following data types are supported:

- Accounting
- AccountingRequest
- CgroupCustomSetting
- Cluster
- ClusterSlurmConfiguration
- ClusterSlurmConfigurationRequest
- ClusterSummary
- ComputeNodeGroup
- ComputeNodeGroupConfiguration
- ComputeNodeGroupSlurmConfiguration
- ComputeNodeGroupSlurmConfigurationRequest
- ComputeNodeGroupSummary
- CustomLaunchTemplate
- Endpoint
- ErrorInfo
- InstanceConfig
- JwtAuth
- JwtKey
- Networking
- NetworkingRequest

- Queue
- QueueSlurmConfiguration
- QueueSlurmConfigurationRequest
- QueueSummary
- ScalingConfiguration
- ScalingConfigurationRequest
- Scheduler
- SchedulerRequest
- SlurmAuthKey
- SlurmCustomSetting
- SlurmdbdCustomSetting
- SlurmRest
- SlurmRestRequest
- SpotOptions
- UpdateAccountingRequest
- UpdateClusterSlurmConfigurationRequest
- UpdateComputeNodeGroupSlurmConfigurationRequest
- UpdateQueueSlurmConfigurationRequest
- UpdateSlurmRestRequest
- ValidationExceptionField

Accounting

The accounting configuration includes configurable settings for Slurm accounting. It's a property of the **ClusterSlurmConfiguration** object.

Contents

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means Slurm accounting is enabled.

Type: String

Valid Values: `STANDARD | NONE`

Required: Yes

defaultPurgeTimeInDays

The default value for all purge settings for `slurmdbd.conf`. For more information, see the slurmdbd.conf documentation at SchedMD.

The default value for `defaultPurgeTimeInDays` is `-1`.

A value of `-1` means there is no purge time and records persist as long as the cluster exists.

> **Important**
>
> `0` isn't a valid value.

Type: Integer

Valid Range: Minimum value of -1. Maximum value of 10000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

AccountingRequest

The accounting configuration includes configurable settings for Slurm accounting. It's a property of the **ClusterSlurmConfiguration** object.

Contents

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means Slurm accounting is enabled.

Type: String

Valid Values: `STANDARD | NONE`

Required: Yes

defaultPurgeTimeInDays

The default value for all purge settings for `slurmdbd.conf`. For more information, see the slurmdbd.conf documentation at SchedMD.

The default value for `defaultPurgeTimeInDays` is `-1`.

A value of `-1` means there is no purge time and records persist as long as the cluster exists.

> **Important**
>
> `0` isn't a valid value.

Type: Integer

Valid Range: Minimum value of -1. Maximum value of 10000.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CgroupCustomSetting

Additional settings that directly map to Cgroup settings.

> **Important**
>
> AWS PCS supports a subset of Cgroup settings. For more information, see Configuring custom Cgroup settings in AWS PCS in the *AWS PCS User Guide*.

Contents

parameterName

AWS PCS supports custom Cgroup settings for clusters. For more information, see Configuring custom Cgroup settings in AWS PCS in the *AWS PCS User Guide*.

Type: String

Required: Yes

parameterValue

The values for the configured Cgroup settings.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Cluster

The cluster resource and configuration.

Contents

arn

The unique Amazon Resource Name (ARN) of the cluster.

Type: String

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

id

The generated unique ID of the cluster.

Type: String

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the cluster.

Type: String

Required: Yes

networking

The networking configuration for the cluster's control plane.

Type: Networking object

Required: Yes

scheduler

The cluster management and job scheduling software associated with the cluster.

Type: Scheduler object

Required: Yes

size

The size of the cluster.

- `SMALL`: 32 compute nodes and 256 jobs
- `MEDIUM`: 512 compute nodes and 8192 jobs
- `LARGE`: 2048 compute nodes and 16,384 jobs

Type: String

Valid Values: `SMALL | MEDIUM | LARGE`

Required: Yes

status

The provisioning status of the cluster.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the cluster.

> **Important**
>
> The resource enters the `SUSPENDING` and `SUSPENDED` states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated.

The resources still count toward your service quotas. You can delete a resource if its status is SUSPENDED. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

endpoints

The list of endpoints available for interaction with the scheduler.

Type: Array of Endpoint objects

Required: No

errorInfo

The list of errors that occurred during cluster provisioning.

Type: Array of ErrorInfo objects

Required: No

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: ClusterSlurmConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ClusterSlurmConfiguration

Additional options related to the Slurm scheduler.

Contents

accounting

The accounting configuration includes configurable settings for Slurm accounting.

Type: Accounting object

Required: No

authKey

The shared Slurm key for authentication, also known as the **cluster secret**.

Type: SlurmAuthKey object

Required: No

cgroupCustomSettings

Additional Cgroup-specific configuration that directly maps to Cgroup settings.

Type: Array of CgroupCustomSetting objects

Required: No

jwtAuth

The JWT authentication configuration for Slurm REST API access.

Type: JwtAuth object

Required: No

scaleDownIdleTimeInSeconds

The time (in seconds) before an idle node is scaled down.

Default: `600`

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 10000000.

Required: No

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

slurmdbdCustomSettings

Additional SlurmDBD-specific configuration that directly maps to SlurmDBD settings.

Type: Array of SlurmdbdCustomSetting objects

Required: No

slurmRest

The Slurm REST API configuration for the cluster.

Type: SlurmRest object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

accounting

The accounting configuration includes configurable settings for Slurm accounting.

Type: AccountingRequest object

Required: No

cgroupCustomSettings

Additional Cgroup-specific configuration that directly maps to Cgroup settings.

Type: Array of CgroupCustomSetting objects

Required: No

scaleDownIdleTimeInSeconds

The time (in seconds) before an idle node is scaled down.

Default: 600

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 10000000.

Required: No

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

slurmdbdCustomSettings

Additional SlurmDBD-specific configuration that directly maps to SlurmDBD settings.

Type: Array of SlurmdbdCustomSetting objects

Required: No

slurmRest

The Slurm REST API configuration for the cluster.

Type: SlurmRestRequest object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ClusterSummary

The object returned by the `ListClusters` API action.

Contents

arn

The unique Amazon Resource Name (ARN) of the cluster.

Type: String

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

id

The generated unique ID of the cluster.

Type: String

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the cluster.

Type: String

Required: Yes

status

The provisioning status of the cluster.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the cluster.

> **Important**
>
> The resource enters the SUSPENDING and SUSPENDED states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated. The resources still count toward your service quotas. You can delete a resource if its status is SUSPENDED. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeNodeGroup

A compute node group associated with a cluster.

Contents

arn

The unique Amazon Resource Name (ARN) of the compute node group.

Type: String

Required: Yes

clusterId

The ID of the cluster of the compute node group.

Type: String

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

customLaunchTemplate

An Amazon EC2 launch template AWS PCS uses to launch compute nodes.

Type: CustomLaunchTemplate object

Required: Yes

iamInstanceProfileArn

The Amazon Resource Name (ARN) of the IAM instance profile used to pass an IAM role when launching EC2 instances. The role contained in your instance profile must have the `pcs:RegisterComputeNodeGroupInstance` permission and the role name must start with `AWSPCS` or must have the path `/aws-pcs/`. For more information, see IAM instance profiles for AWS PCS in the *AWS PCS User Guide*.

Type: String

Pattern: `arn:aws([a-zA-Z-]{0,10})?:iam::[0-9]{12}:instance-profile/([!-~]{1,510}/)?([\w+=,.@-]{1,128})`

Required: Yes

id

The generated unique ID of the compute node group.

Type: String

Required: Yes

instanceConfigs

A list of EC2 instance configurations that AWS PCS can provision in the compute node group.

Type: Array of InstanceConfig objects

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the compute node group.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

scalingConfiguration

Specifies the boundaries of the compute node group auto scaling.

Type: ScalingConfiguration object

Required: Yes

status

The provisioning status of the compute node group.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the compute node group.

> **Important**
>
> The resource enters the SUSPENDING and SUSPENDED states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated. The resources still count toward your service quotas. You can delete a resource if its status is SUSPENDED. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | DELETED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

subnetIds

The list of subnet IDs where instances are provisioned by the compute node group. The subnets must be in the same VPC as the cluster.

Type: Array of strings

Array Members: Minimum number of 1 item.

Pattern: `subnet-\w{8,17}`

Required: Yes

amiId

The ID of the Amazon Machine Image (AMI) that AWS PCS uses to launch instances. If not provided, AWS PCS uses the AMI ID specified in the custom launch template.

Type: String

Pattern: `ami-[a-z0-9]+`

Required: No

errorInfo

The list of errors that occurred during compute node group provisioning.

Type: Array of ErrorInfo objects

Required: No

purchaseOption

Specifies how EC2 instances are purchased on your behalf. AWS PCS supports On-Demand Instances, Spot Instances, and Amazon EC2 Capacity Blocks for ML. For more information, see Amazon EC2 billing and purchasing options in the *Amazon Elastic Compute Cloud User Guide*. For more information about AWS PCS support for Capacity Blocks, see Using Amazon EC2 Capacity Blocks for ML with AWS PCS in the *AWS PCS User Guide*. If you don't provide this option, it defaults to On-Demand.

Type: String

Valid Values: `ONDEMAND | SPOT | CAPACITY_BLOCK`

Required: No

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: ComputeNodeGroupSlurmConfiguration object

Required: No

spotOptions

Additional configuration when you specify `SPOT` as the `purchaseOption` for the `CreateComputeNodeGroup` API action.

Type: SpotOptions object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeNodeGroupConfiguration

The compute node group configuration for a queue.

Contents

computeNodeGroupId

The compute node group ID for the compute node group configuration.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeNodeGroupSlurmConfiguration

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeNodeGroupSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ComputeNodeGroupSummary

The object returned by the `ListComputeNodeGroups` API action.

Contents

arn

The unique Amazon Resource Name (ARN) of the compute node group.

Type: String

Required: Yes

clusterId

The ID of the cluster of the compute node group.

Type: String

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

id

The generated unique ID of the compute node group.

Type: String

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the compute node group.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

status

The provisioning status of the compute node group.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the compute node group.

> **Important**
>
> The resource enters the SUSPENDING and SUSPENDED states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated. The resources still count toward your service quotas. You can delete a resource if its status is SUSPENDED. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | DELETED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomLaunchTemplate

An Amazon EC2 launch template AWS PCS uses to launch compute nodes.

Contents

id

The ID of the EC2 launch template to use to provision instances.

Example: `lt-xxxx`

Type: String

Required: Yes

version

The version of the EC2 launch template to use to provision instances.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Endpoint

An endpoint available for interaction with the scheduler.

Contents

port

The endpoint's connection port number.

Example: `1234`

Type: String

Required: Yes

privateIpAddress

For clusters that use IPv4, this is the endpoint's private IP address.

Example: `10.1.2.3`

For clusters configured to use IPv6, this is an empty string.

Type: String

Required: Yes

type

Indicates the type of endpoint running at the specific IP address.

Type: String

Valid Values: `SLURMCTLD | SLURMDBD | SLURMRESTD`

Required: Yes

ipv6Address

The endpoint's IPv6 address.

Example: `2001:db8::1`

Type: String

Required: No

publicIpAddress

The endpoint's public IP address.

Example: `192.0.2.1`

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ErrorInfo

An error that occurred during resource creation.

Contents

code

The short-form error code.

Type: String

Required: No

message

The detailed error information.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

InstanceConfig

An EC2 instance configuration AWS PCS uses to launch compute nodes.

Contents

instanceType

The EC2 instance type that AWS PCS can provision in the compute node group.

Example: `t2.xlarge`

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JwtAuth

The JWT authentication configuration for Slurm REST API access.

Contents

jwtKey

The JWT key for Slurm REST API authentication.

Type: JwtKey object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

JwtKey

The JWT key stored in AWS Secrets Manager for Slurm REST API authentication.

Contents

secretArn

The Amazon Resource Name (ARN) of the AWS Secrets Manager secret containing the JWT key.

Type: String

Required: Yes

secretVersion

The version of the AWS Secrets Manager secret containing the JWT key.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Networking

The networking configuration for the cluster's control plane.

Contents

networkType

The IP address version the cluster uses. The default is `IPV4`.

Type: String

Valid Values: `IPV4 | IPV6`

Required: No

securityGroupIds

The list of security group IDs associated with the Elastic Network Interface (ENI) created in subnets.

The following rules are required:

- Inbound rule 1
  - Protocol: All
  - Ports: All
  - Source: Self
- Outbound rule 1
  - Protocol: All
  - Ports: All
  - Destination: 0.0.0.0/0 (IPv4) or ::/0 (IPv6)
- Outbound rule 2
  - Protocol: All
  - Ports: All
  - Destination: Self

Type: Array of strings

Pattern: `sg-\w{8,17}`

Required: No

subnetIds

The ID of the subnet where AWS PCS creates an Elastic Network Interface (ENI) to enable communication between managed controllers and AWS PCS resources. The subnet must have an available IP address, cannot reside in AWS Outposts, AWS Wavelength, or an AWS Local Zone.

Example: `subnet-abcd1234`

Type: Array of strings

Array Members: Minimum number of 1 item.

Pattern: `subnet-\w{8,17}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

NetworkingRequest

The networking configuration for the cluster's control plane.

Contents

networkType

The IP address version the cluster uses. The default is `IPV4`.

Type: String

Valid Values: `IPV4 | IPV6`

Required: No

securityGroupIds

A list of security group IDs associated with the Elastic Network Interface (ENI) created in subnets.

Type: Array of strings

Pattern: `sg-\w{8,17}`

Required: No

subnetIds

The list of subnet IDs where AWS PCS creates an Elastic Network Interface (ENI) to enable communication between managed controllers and AWS PCS resources. Subnet IDs have the form `subnet-0123456789abcdef0`.

Subnets can't be in AWS Outposts, AWS Wavelength or an AWS Local Zone.

> **Note**
>
> AWS PCS currently supports only 1 subnet in this list.

Type: Array of strings

Array Members: Minimum number of 1 item.

Pattern: `subnet-\w{8,17}`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Queue

A queue resource.

Contents

arn

The unique Amazon Resource Name (ARN) of the queue.

Type: String

Required: Yes

clusterId

The ID of the cluster of the queue.

Type: String

Required: Yes

computeNodeGroupConfigurations

The list of compute node group configurations associated with the queue. Queues assign jobs to associated compute node groups.

Type: Array of ComputeNodeGroupConfiguration objects

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

id

The generated unique ID of the queue.

Type: String

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the queue.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

status

The provisioning status of the queue.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the queue.

> **Important**
>
> The resource enters the `SUSPENDING` and `SUSPENDED` states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated. The resources still count toward your service quotas. You can delete a resource if its status is `SUSPENDED`. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

errorInfo

The list of errors that occurred during queue provisioning.

Type: Array of ErrorInfo objects

Required: No

slurmConfiguration

Additional options related to the Slurm scheduler.

Type: QueueSlurmConfiguration object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueueSlurmConfiguration

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueueSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

QueueSummary

The object returned by the `ListQueues` API action.

Contents

arn

The unique Amazon Resource Name (ARN) of the queue.

Type: String

Required: Yes

clusterId

The ID of the cluster of the queue.

Type: String

Required: Yes

createdAt

The date and time the resource was created.

Type: Timestamp

Required: Yes

id

The generated unique ID of the queue.

Type: String

Required: Yes

modifiedAt

The date and time the resource was modified.

Type: Timestamp

Required: Yes

name

The name that identifies the queue.

Type: String

Length Constraints: Minimum length of 3. Maximum length of 25.

Pattern: `(?!pcs_)^[A-Za-z][A-Za-z0-9-]+`

Required: Yes

status

The provisioning status of the queue.

> **Note**
>
> The provisioning status doesn't indicate the overall health of the queue.

> **Important**
>
> The resource enters the SUSPENDING and SUSPENDED states when the scheduler is beyond end of life and we have suspended the cluster. When in these states, you can't use the cluster. The cluster controller is down and all compute instances are terminated. The resources still count toward your service quotas. You can delete a resource if its status is SUSPENDED. For more information, see Frequently asked questions about Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Type: String

Valid Values: `CREATING | ACTIVE | UPDATING | DELETING | CREATE_FAILED | DELETE_FAILED | UPDATE_FAILED | SUSPENDING | SUSPENDED | RESUMING`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingConfiguration

Specifies the boundaries of the compute node group auto scaling.

Contents

maxInstanceCount

The upper bound of the number of instances allowed in the compute fleet.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

minInstanceCount

The lower bound of the number of instances allowed in the compute fleet.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingConfigurationRequest

Specifies the boundaries of the compute node group auto scaling.

Contents

maxInstanceCount

The upper bound of the number of instances allowed in the compute fleet.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

minInstanceCount

The lower bound of the number of instances allowed in the compute fleet.

Type: Integer

Valid Range: Minimum value of 0.

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Scheduler

The cluster management and job scheduling software associated with the cluster.

Contents

type

The software AWS PCS uses to manage cluster scaling and job scheduling.

Type: String

Valid Values: `SLURM`

Required: Yes

version

The version of the specified scheduling software that AWS PCS uses to manage cluster scaling and job scheduling. For more information, see Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Valid Values: `23.11 | 24.05 | 24.11 | 25.05`

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SchedulerRequest

The cluster management and job scheduling software associated with the cluster.

Contents

type

The software AWS PCS uses to manage cluster scaling and job scheduling.

Type: String

Valid Values: `SLURM`

Required: Yes

version

The version of the specified scheduling software that AWS PCS uses to manage cluster scaling and job scheduling. For more information, see Slurm versions in AWS PCS in the *AWS PCS User Guide*.

Valid Values: `24.11 | 25.05`

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SlurmAuthKey

The shared Slurm key for authentication, also known as the **cluster secret**.

Contents

secretArn

The Amazon Resource Name (ARN) of the shared Slurm key.

Type: String

Required: Yes

secretVersion

The version of the shared Slurm key.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SlurmCustomSetting

Additional settings that directly map to Slurm settings.

> **Important**
>
> AWS PCS supports a subset of Slurm settings. For more information, see Configuring custom Slurm settings in AWS PCS in the *AWS PCS User Guide*.

Contents

parameterName

AWS PCS supports custom Slurm settings for clusters, compute node groups, and queues. For more information, see Configuring custom Slurm settings in AWS PCS in the *AWS PCS User Guide*.

Type: String

Required: Yes

parameterValue

The values for the configured Slurm settings.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SlurmdbdCustomSetting

Additional settings that directly map to SlurmDBD settings.

> **Important**
>
> AWS PCS supports a subset of SlurmDBD settings. For more information, see Configuring custom SlurmDBD settings in AWS PCS in the *AWS PCS User Guide*.

Contents

parameterName

AWS PCS supports custom SlurmDBD settings for clusters. For more information, see Configuring custom SlurmDBD settings in AWS PCS in the *AWS PCS User Guide*.

Type: String

Required: Yes

parameterValue

The values for the configured SlurmDBD settings.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SlurmRest

The Slurm REST API configuration includes settings for enabling and configuring the Slurm REST API. It's a property of the **ClusterSlurmConfiguration** object.

Contents

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means the Slurm REST API is enabled.

Type: String

Valid Values: `STANDARD | NONE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SlurmRestRequest

The Slurm REST API configuration includes settings for enabling and configuring the Slurm REST API. It's a property of the **ClusterSlurmConfiguration** object.

Contents

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means the Slurm REST API is enabled.

Type: String

Valid Values: `STANDARD | NONE`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

SpotOptions

Additional configuration when you specify SPOT as the `purchaseOption` for the `CreateComputeNodeGroup` API action.

Contents

allocationStrategy

The Amazon EC2 allocation strategy AWS PCS uses to provision EC2 instances. AWS PCS supports **lowest price**, **capacity optimized**, and **price capacity optimized**. For more information, see Use allocation strategies to determine how EC2 Fleet or Spot Fleet fulfills Spot and On-Demand capacity in the *Amazon Elastic Compute Cloud User Guide*. If you don't provide this option, it defaults to **price capacity optimized**.

Type: String

Valid Values: `lowest-price | capacity-optimized | price-capacity-optimized`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateAccountingRequest

The accounting configuration includes configurable settings for Slurm accounting.

Contents

defaultPurgeTimeInDays

The default value for all purge settings for `slurmdbd.conf`. For more information, see the slurmdbd.conf documentation at SchedMD.

The default value for `defaultPurgeTimeInDays` is `-1`.

A value of `-1` means there is no purge time and records persist as long as the cluster exists.

> **Important**
>
> `0` isn't a valid value.

Type: Integer

Valid Range: Minimum value of -1. Maximum value of 10000.

Required: No

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means Slurm accounting is enabled.

Type: String

Valid Values: `STANDARD | NONE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateClusterSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

accounting

The accounting configuration includes configurable settings for Slurm accounting.

Type: UpdateAccountingRequest object

Required: No

cgroupCustomSettings

Additional Cgroup-specific configuration that directly maps to Cgroup settings.

Type: Array of CgroupCustomSetting objects

Required: No

scaleDownIdleTimeInSeconds

The time (in seconds) before an idle node is scaled down.

Default: `600`

Type: Integer

Valid Range: Minimum value of 1. Maximum value of 10000000.

Required: No

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

slurmdbdCustomSettings

Additional SlurmDBD-specific configuration that directly maps to SlurmDBD settings.

Type: Array of SlurmdbdCustomSetting objects

Required: No

slurmRest

The Slurm REST API configuration for the cluster.

Type: UpdateSlurmRestRequest object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateComputeNodeGroupSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateQueueSlurmConfigurationRequest

Additional options related to the Slurm scheduler.

Contents

slurmCustomSettings

Additional Slurm-specific configuration that directly maps to Slurm settings.

Type: Array of SlurmCustomSetting objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

UpdateSlurmRestRequest

The Slurm REST API configuration includes settings for enabling and configuring the Slurm REST API.

Contents

mode

The default value for `mode` is `NONE`. A value of `STANDARD` means the Slurm REST API is enabled.

Type: String

Valid Values: `STANDARD` | `NONE`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ValidationExceptionField

Stores information about a field in a request that caused an exception.

Contents

message

The message body of the exception.

Type: String

Required: Yes

name

The name of the exception.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see Signing AWS API requests in the *IAM User Guide*.

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: `AWS4-HMAC-SHA256`

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key*/*YYYYMMDD*/*region*/*service*/aws4_request.

For more information, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: `20120325T120000Z`.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see Elements of an AWS API request signature in the *IAM User Guide*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS STS, see AWS services that work with IAM in the *IAM User Guide*.

Condition: If you're using temporary security credentials from AWS STS, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see Create a signed AWS API request in the *IAM User Guide*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Error Types

This section lists common error types that this AWS service may return. Not all services return all error types listed here. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You don't have permission to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 403

ExpiredTokenException

The security token included in the request has expired. Request a new security token and try again.

HTTP Status Code: 403

IncompleteSignature

The request signature doesn't conform to AWS standards. Verify that you're using valid AWS credentials and that your request is properly formatted. If you're using an SDK, ensure it's up to date.

HTTP Status Code: 403

InternalFailure

The request can't be processed right now because of an internal server issue. Try again later. If the problem persists, contact AWS Support.

HTTP Status Code: 500

MalformedHttpRequestException

The request body can't be processed. This typically happens when the request body can't be decompressed using the specified content encoding algorithm. Verify that the content encoding header matches the compression format used.

HTTP Status Code: 400

NotAuthorized

You don't have permissions to perform this action. Verify that your IAM policy includes the required permissions.

HTTP Status Code: 401

OptInRequired

Your AWS account needs a subscription for this service. Verify that you've enabled the service in your account.

HTTP Status Code: 403

RequestAbortedException

The request was aborted before a response could be returned. This typically happens when the client closes the connection.

HTTP Status Code: 400

RequestEntityTooLargeException

The request entity is too large. Reduce the size of the request body and try again.

HTTP Status Code: 413

RequestTimeoutException

The request timed out. The server didn't receive the complete request within the expected time frame. Try again.

HTTP Status Code: 408

ServiceUnavailable

The service is temporarily unavailable. Try again later.

HTTP Status Code: 503

ThrottlingException

Your request rate is too high. The AWS SDKs automatically retry requests that receive this exception. Reduce the frequency of requests.

HTTP Status Code: 400

UnknownOperationException

The action or operation isn't recognized. Verify that the action name is spelled correctly and that it's supported by the API version you're using.

HTTP Status Code: 404

UnrecognizedClientException

The X.509 certificate or AWS access key ID you provided doesn't exist in our records. Verify that you're using valid credentials and that they haven't expired.

HTTP Status Code: 403

ValidationError

The input doesn't meet the required format or constraints. Check that all required parameters are included and that values are valid.

HTTP Status Code: 400