Guía del usuario de

AWS PCS

AWS PCS: Guía del usuario de

Table of Contents

¿Qué es el Servicio de Computación AWS Paralela?

AWS El servicio de computación paralela (AWS PCS) es un servicio gestionado que facilita la ejecución y el escalado de las cargas de trabajo de computación de alto rendimiento (HPC) y la creación de modelos científicos y de ingeniería AWS con Slurm. Utilice AWS PCS para crear clústeres de procesamiento que integren los mejores recursos de AWS computación, almacenamiento, redes y visualización de su clase. Ejecute simulaciones o cree modelos científicos y de ingeniería. Optimice y simplifique las operaciones de sus clústeres mediante las funciones integradas de administración y observabilidad. Permita que sus usuarios se centren en la investigación y la innovación al permitirles ejecutar sus aplicaciones y trabajos en un entorno familiar.

Temas

Conceptos en AWS PCS

En AWS PCS, un clúster tiene una o más colas asociadas a al menos un grupo de nodos de cómputo. Los trabajos se envían a colas y se ejecutan en EC2 instancias definidas por grupos de nodos de procesamiento. Puede utilizar estas bases para implementar arquitecturas de HPC sofisticadas.

Clúster

Un clúster es un recurso para gestionar los recursos y ejecutar las cargas de trabajo. Un clúster es un recurso de AWS PCS que define un conjunto de configuraciones de procesamiento, redes, almacenamiento, identidad y programador de tareas. Para crear un clúster, especifique qué programador de tareas desea usar (actualmente Slurm), qué configuración de programador desea, qué controlador de servicios desea administrar el clúster y en qué VPC desea que se lancen los recursos del clúster. El programador acepta y programa los trabajos, y también lanza los nodos de cómputo (EC2 instancias) que procesan esos trabajos.

Grupo de nodos de cómputo

Un grupo de nodos de procesamiento es un conjunto de nodos de procesamiento que AWS PCS utiliza para ejecutar trabajos o proporcionar acceso interactivo a un clúster. Al definir un grupo de nodos de cómputo, se especifican características comunes, como los tipos de EC2 instancias de Amazon, el número mínimo y máximo de instancias, las subredes de VPC de destino, la imagen de

máquina de Amazon (AMI), la opción de compra y la configuración de lanzamiento personalizada. AWS PCS usa esta configuración para lanzar, administrar y terminar de manera eficiente los nodos de cómputo de un grupo de nodos de cómputo.

Cola

Cuando quiere ejecutar un trabajo en un clúster específico, lo envía a una cola determinada (también denominada partición). El trabajo permanece en la cola hasta que AWS PCS lo programe para que se ejecute en un grupo de nodos de procesamiento. Asocia uno o más grupos de nodos de cómputo a cada cola. Se necesita una cola para programar y ejecutar los trabajos en los recursos del grupo de nodos de cómputo subyacentes mediante diversas políticas de programación ofrecidas por el programador de trabajos. Los usuarios no envían los trabajos directamente a un nodo de cómputo o a un grupo de nodos de cómputo.

Administrador de sistemas

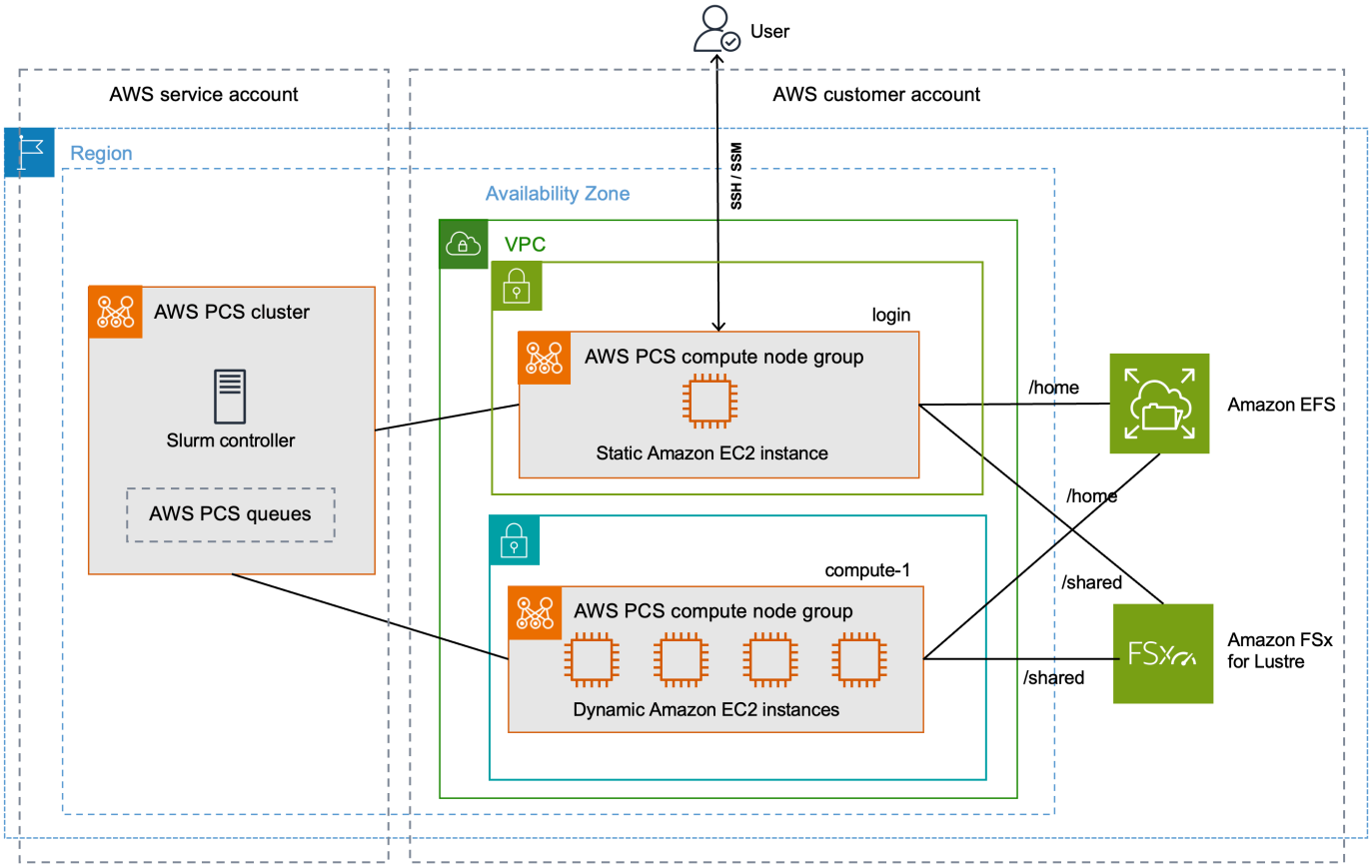
Un administrador del sistema implementa, mantiene y opera un clúster. Pueden acceder a AWS PCS a través de la Consola de administración de AWS API de AWS PCS y el AWS SDK. Tienen acceso a clústeres específicos a través de SSH o AWS Systems Manager, donde pueden ejecutar tareas administrativas, ejecutar trabajos, administrar datos y realizar otras actividades basadas en el shell. Para obtener más información, consulte la Documentación de AWS Systems Manager.

Usuario final

El usuario final no tiene day-to-day la responsabilidad de implementar u operar un clúster. Utilizan una interfaz de terminal (como SSH) para acceder a los recursos del clúster, ejecutar tareas, administrar datos y realizar otras actividades basadas en el shell.

Comience con AWS Parallel Computing Service

Este es un tutorial para crear un clúster sencillo que puede utilizar para probar AWS PCS. En la siguiente figura se muestra el diseño del clúster.



El tutorial sobre diseño de clústeres incluye los siguientes componentes clave:

- Una VPC y subredes que cumplan con los requisitos de red de AWS PCS.
- Un sistema de archivos Amazon EFS, que se utilizará como directorio principal compartido.
- Un sistema de archivos Amazon FSx for Lustre, que proporciona un directorio compartido de alto rendimiento.
- Un clúster AWS PCS, que proporciona un controlador Slurm.
- 2 grupos de nodos de cómputo AWS PCS.
  - El grupo de `login` nodos, que proporciona acceso interactivo al sistema basado en una consola.

  - El grupo de `compute-1` nodos proporciona instancias que se escalan elásticamente para ejecutar trabajos.
- 1 cola que envía los trabajos a las EC2 instancias del grupo de nodos. `compute-1`

El clúster requiere AWS recursos adicionales, como grupos de seguridad, funciones de IAM y plantillas de EC2 lanzamiento, que no se muestran en el diagrama.

> **Note**
>
> Le recomendamos que complete los pasos de la línea de comandos de este tema en un shell de Bash. Si no está utilizando un intérprete de comandos Bash, algunos comandos de script, como los caracteres de continuación de línea y la forma en que se establecen y utilizan las variables, requieren ajustes para su intérprete de comandos. Además, las reglas de entrecomillado y escape de su intérprete de comandos pueden ser diferentes. Para obtener más información, consulte Comillas y literales con cadenas AWS CLI en la Guía del AWS Command Line Interface usuario de la versión 2.

Temas

- Elimine sus AWS recursos para AWS PCS

Requisitos previos para empezar a utilizar el PCS AWS

Consulte los siguientes temas para preparar su entorno de desarrollo local Cuenta de AWS y el suyo para el AWS PCS.

Temas

- Inscríbase AWS y cree un usuario administrativo
- Instale el AWS CLI para AWS PCS
- Permisos de IAM necesarios para AWS PCS

Inscríbase AWS y cree un usuario administrativo

Complete las siguientes tareas para configurar el Servicio de computación AWS paralela (AWS PCS).

Temas

- Inscríbase en un Cuenta de AWS
- Creación de un usuario con acceso administrativo

Inscríbase en un Cuenta de AWS

Si no tiene uno Cuenta de AWS, complete los siguientes pasos para crearlo.

Para suscribirse a una Cuenta de AWS

1. Abrir https://portal.aws.amazon.com/billing/registro.
2. Siga las instrucciones que se le indiquen.

   Parte del procedimiento de registro consiste en recibir una llamada telefónica o mensaje de texto e indicar un código de verificación en el teclado del teléfono.

   Cuando te registras en un Cuenta de AWS, Usuario raíz de la cuenta de AWSse crea un. El usuario raíz tendrá acceso a todos los Servicios de AWS y recursos de esa cuenta. Como práctica recomendada de seguridad, asigne acceso administrativo a un usuario y utilice únicamente el usuario raíz para realizar tareas que requieren acceso de usuario raíz.

AWS te envía un correo electrónico de confirmación una vez finalizado el proceso de registro. En cualquier momento, puede ver la actividad de su cuenta actual y administrarla accediendo a https://aws.amazon.com/y seleccionando Mi cuenta.

Creación de un usuario con acceso administrativo

Después de crear un usuario administrativo Cuenta de AWS, asegúrelo Usuario raíz de la cuenta de AWS AWS IAM Identity Center, habilite y cree un usuario administrativo para no usar el usuario root en las tareas diarias.

Proteja su Usuario raíz de la cuenta de AWS

1. Inicie sesión Consola de administración de AWScomo propietario de la cuenta seleccionando el usuario root e introduciendo su dirección de Cuenta de AWS correo electrónico. En la siguiente página, escriba su contraseña.

   Para obtener ayuda para iniciar sesión con el usuario raíz, consulte Iniciar sesión como usuario raíz en la Guía del usuario de AWS Sign-In .
2. Active la autenticación multifactor (MFA) para el usuario raíz.

   Para obtener instrucciones, consulte Habilitar un dispositivo MFA virtual para el usuario Cuenta de AWS raíz (consola) en la Guía del usuario de IAM.

Creación de un usuario con acceso administrativo

1. Activar IAM Identity Center.

   Consulte las instrucciones en Activar AWS IAM Identity Center en la Guía del usuario de AWS IAM Identity Center .
2. En IAM Identity Center, conceda acceso administrativo a un usuario.

   Para ver un tutorial sobre su uso Directorio de IAM Identity Center como fuente de identidad, consulte Configurar el acceso de los usuarios con la configuración predeterminada Directorio de IAM Identity Center en la Guía del AWS IAM Identity Center usuario.

Inicio de sesión como usuario con acceso de administrador

- Para iniciar sesión con el usuario de IAM Identity Center, use la URL de inicio de sesión que se envió a la dirección de correo electrónico cuando creó el usuario de IAM Identity Center.

Para obtener ayuda para iniciar sesión con un usuario del Centro de identidades de IAM, consulte Iniciar sesión en el portal de AWS acceso en la Guía del AWS Sign-In usuario.

Concesión de acceso a usuarios adicionales

1. En IAM Identity Center, cree un conjunto de permisos que siga la práctica recomendada de aplicar permisos de privilegios mínimos.

   Para conocer las instrucciones, consulte Create a permission set en la Guía del usuario de AWS IAM Identity Center .
2. Asigne usuarios a un grupo y, a continuación, asigne el acceso de inicio de sesión único al grupo.

   Para conocer las instrucciones, consulte Add groups en la Guía del usuario de AWS IAM Identity Center .

Instale el AWS CLI para AWS PCS

Debe utilizar la última versión de AWS CLI. Para obtener más información, consulte Instalar o actualizar a la última versión de AWS CLI en la Guía del AWS Command Line Interface usuario de la versión 2.

Debe configurar el AWS CLI. Para obtener más información, consulte Configurar el AWS CLI en la Guía del AWS Command Line Interface usuario de la versión 2.

Introduzca el siguiente comando en una línea de comandos para comprobar su AWS CLI estado; debería mostrar información de ayuda.

```
aws pcs help
```

Permisos de IAM necesarios para AWS PCS

El principal de seguridad de IAM que utilice debe tener permisos para trabajar con las funciones de IAM de AWS PCS, las funciones vinculadas a servicios AWS CloudFormation, una VPC y los recursos relacionados. Para obtener más informaciónServicio de Gestión de Identidad y Acceso para Computación AWS Paralela, consulte Crear un rol vinculado a un servicio en la Guía del

usuario.AWS Identity and Access Management Debe completar todos los pasos de esta guía como el mismo usuario. Ejecute el siguiente comando para comprobar el usuario actual:

```
aws sts get-caller-identity
```

Tutorial de uso AWS CloudFormation con el AWS PCS

El tutorial de AWS PCS consta de muchos pasos y su objetivo es ayudarle a comprender las partes de un clúster de AWS PCS y los procedimientos necesarios para crearlo. Le recomendamos que siga los pasos del tutorial al menos una vez. Una vez que comprenda bien lo que implica, puede utilizarlo AWS CloudFormation para crear el clúster de muestras rápidamente y de forma automática.

CloudFormation es un AWS servicio que le permite crear y aprovisionar despliegues de AWS infraestructura de forma predecible y repetitiva. Puede usar una CloudFormation plantilla para aprovisionar automáticamente los AWS recursos del clúster de muestra como una sola unidad, denominada pila. Puede eliminar la pila cuando haya terminado con ella.

Para obtener más información, consulte Comience con un CloudFormation AWS PCS.

Crear una VPC y subredes para PCS AWS

Puede crear una VPC y subredes con una plantilla. CloudFormation Utilice la siguiente URL para descargar la CloudFormation plantilla y, a continuación, cárguela en la CloudFormation consola para crear una pila nueva CloudFormation. Para obtener más información, consulte Uso de la CloudFormation consola en la Guía del AWS CloudFormation usuario.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/net/hpc_large_scale/assets/
main.yaml
```

Con la plantilla abierta en la CloudFormation consola, introduzca las siguientes opciones. Puede utilizar los valores predeterminados que se proporcionan en la plantilla.

- En Proporcione un nombre de pila:
  - En Nombre de pila, ingresa:

    ```
    hpc-networking
    ```
- En Parámetros:

- En VPC:
  - En CidrBlock, introduzca:

```
10.3.0.0/16
```

- En las subredes A:
  - En CidrPublicSubnetA, introduzca:

```
10.3.0.0/20
```

  - En CidrPrivateSubnetA, introduzca:

```
10.3.128.0/20
```

- En las subredes B:
  - En CidrPublicSubnetB, introduzca:

```
10.3.16.0/20
```

  - En CidrPrivateSubnetB, introduzca:

```
10.3.144.0/20
```

- En Subredes C:
  - Para ProvisionSubnetsC, seleccione True
  - En CidrPublicSubnetC, escriba:

```
10.3.32.0/20
```

  - En CidrPrivateSubnetC, introduzca:

```
10.3.160.0/20
```

- En Capacidades:
  - Marque la casilla Reconozco que AWS CloudFormation podría crear recursos de IAM.

Supervisa el estado de la CloudFormation pila. Cuando llegue`CREATE_COMPLETE`, busque el ID del grupo de seguridad predeterminado en la nueva VPC. Utilizará el ID más adelante en el tutorial.

Busque el grupo de seguridad predeterminado para la VPC del clúster

Para buscar el ID del grupo de seguridad predeterminado en la nueva VPC, siga este procedimiento:

- Vaya a la consola de Amazon VPC.
- En el panel de control de VPC, selecciona Filtrar por VPC.
  - Elija la VPC por la que empieza el nombre. `hpc-networking`
  - En Seguridad, selecciona Grupos de seguridad.
- Busque el ID del grupo de seguridad del grupo denominado`default`. Tiene la descripción`default VPC security group`. El ID se utilizará más adelante para configurar las plantillas de lanzamiento de EC2.

Crear grupos de seguridad para AWS PCS

AWS PCS se basa en los grupos de seguridad para administrar el tráfico de red que entra y sale de un clúster y sus grupos de nodos de cómputo. Para obtener información detallada sobre este tema, consulteRequisitos y consideraciones sobre los grupos de seguridad.

En este paso, utilizará una CloudFormation plantilla para crear dos grupos de seguridad.

- Un grupo de seguridad de clúster, que permite las comunicaciones entre el controlador AWS PCS, los nodos de cómputo y los nodos de inicio de sesión.
- Un grupo de seguridad SSH entrante, que puede añadir de forma opcional a sus nodos de inicio de sesión para admitir el acceso SSH

Cree los grupos de seguridad para PCS AWS

Puede utilizar una CloudFormation plantilla para crear los grupos de seguridad. Utilice la siguiente URL para descargar la CloudFormation plantilla y, a continuación, cárguela en la CloudFormation consola para crear una CloudFormation pila nueva. Para obtener más información, consulte Uso de la CloudFormation consola en la Guía del AWS CloudFormation usuario.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/getting_started/assets/pcs-cluster-sg.yaml
```

Con la plantilla abierta en la AWS CloudFormation consola, introduzca las siguientes opciones. Tenga en cuenta que algunas opciones se rellenarán automáticamente en la plantilla; simplemente puede dejarlas como valores predeterminados.

- En Indique un nombre para la pila
  - En Nombre de pila, ingresa:

    ```
    getstarted-sg
    ```
- En Parámetros
  - En VpcId, elija la VPC por la que empieza el nombre. `hpc-networking`
  - (Opcional) En ClientIpCidr, introduzca un rango de IP más restrictivo para el grupo de seguridad SSH entrante. Le recomendamos que lo restrinja con su propia IP/subred (x.x.x.x/32 para su propia IP o x.x.x.x/24 para el rango). Sustituya x.x.x.x por su propia IP PÚBLICA. Puede obtener su IP pública mediante herramientas como https://ifconfig.co/)

Supervisa el estado de la CloudFormation pila. Cuando llegue al grupo `CREATE_COMPLETE` de seguridad, los recursos estarán listos.

Se han creado dos grupos de seguridad, con los nombres:

- `cluster-getstarted-sg`— este es el grupo de seguridad del clúster
- `inbound-ssh-getstarted-sg`— se trata de un grupo de seguridad que permite el acceso SSH entrante

Crear un clúster en AWS PCS

En AWS PCS, un clúster es un recurso persistente para administrar recursos y ejecutar cargas de trabajo. Se crea un clúster para un programador específico (AWS PCS actualmente admite Slurm) en una subred de una VPC nueva o existente. El clúster acepta y programa los trabajos, y también lanza los nodos de cómputo (EC2 instancias) que procesan esos trabajos.

Cómo crear el clúster

1. Abra la consola AWS PCS y seleccione Crear clúster.
2. En la sección de detalles del clúster, introduzca los siguientes campos:

- Nombre del clúster: introduzca `get-started`
- Planificador: seleccione la versión 25.05 de Slurm
- Tamaño del controlador: seleccione Pequeño

3. En la sección Redes, seleccione valores para los siguientes campos:
   - VPC: elija la VPC denominada `hpc-networking:Large-Scale-HPC`
   - Subred: seleccione la subred donde comienza el nombre `hpc-networking:PrivateSubnetA`
   - Grupos de seguridad: seleccione el nombre del grupo de seguridad del clúster `cluster-getstarted-sg`
4. Elija Create cluster.

> **Note**
>
> El campo Estado muestra la opción Crear mientras se aprovisiona el clúster. La creación del clúster puede tardar varios minutos.

Cree almacenamiento compartido para AWS PCS en Amazon Elastic File System

Amazon Elastic File System (Amazon EFS) es un AWS servicio que proporciona almacenamiento de archivos totalmente elástico y sin servidor para que pueda compartir datos de archivos sin aprovisionar ni administrar la capacidad de almacenamiento y el rendimiento. Para obtener más información, consulte ¿Qué es Amazon Elastic File System? en la Guía del usuario de Amazon Elastic File System.

El clúster de demostración de AWS PCS utiliza un sistema de archivos EFS para proporcionar un directorio principal compartido entre los nodos del clúster. Cree un sistema de archivos EFS en la misma VPC que el clúster.

Cree su sistema de archivos de Amazon EFS

1. Vaya a la consola de Amazon EFS.

2. Asegúrese de que está configurada de la misma manera en la Región de AWS que probará AWS PCS.
3. Seleccione Crear sistema de archivos.
4. En la página Crear sistema de archivos, defina los siguientes parámetros:
   - En Nombre, introduzca `getstarted-efs`.
   - En Virtual Private Cloud (VPC), elige la VPC denominada `hpc-networking:Large-Scale-HPC`
   - Seleccione Crear. De este modo, volverá a la página de sistemas de archivos.
5. Anote el ID del sistema de archivos del sistema de `getstarted-efs` archivos. Usará esta información más tarde.

Cree almacenamiento compartido para AWS PCS en Amazon FSx for Lustre

Amazon FSx for Lustre hace que sea fácil y rentable lanzar y ejecutar el popular sistema de archivos Lustre de alto rendimiento. Utiliza Lustre para cargas de trabajo en las que la velocidad es importante, como el machine learning, la computación de alto rendimiento (HPC), el procesamiento de vídeo y el modelado financiero. Para obtener más información, consulta ¿Qué es Amazon FSx for Lustre? en la Guía del usuario FSx de Amazon for Lustre.

El clúster de demostración de AWS PCS puede utilizar un sistema de archivos FSx for Lustre para proporcionar un directorio compartido de alto rendimiento entre los nodos del clúster. Cree un sistema de archivos FSx para Lustre en la misma VPC que su clúster.

Para crear su sistema de archivos FSx para Lustre

1. Ve a la FSx consola de Amazon.
2. Asegúrese de que la consola esté configurada para usar lo Región de AWS mismo que su clúster.
3. Seleccione Crear sistema de archivos.
   - En Seleccione el tipo de sistema de archivos, elija Amazon FSx for Lustre y, a continuación, elija Siguiente.
4. En la página Especificar los detalles del sistema de archivos, defina los siguientes parámetros:

- En Detalles del sistema de archivos
  - En Nombre, introduzca `getstarted-fsx`.
  - Para el tipo de implementación y almacenamiento, selecciona Persistente, SSD
  - Para obtener un rendimiento por unidad de almacenamiento, elija 125 MB/s/TiB
  - Para Capacidad de almacenamiento, introduzca 1,2 TiB
  - Para la configuración de metadatos, elija Automática
  - Para el tipo de compresión de datos, elija LZ4
- En Red y seguridad
  - Para Virtual Private Cloud (VPC), elija la VPC denominada `hpc-networking:Large-Scale-HPC`
  - Para los grupos de seguridad de VPC, deje el grupo de seguridad denominado `default`
  - En Subred, elige la subred en la que comience el nombre `hpc-networking:PrivateSubnetA`
- Deje las demás opciones con sus valores predeterminados.
- Elija Next (Siguiente).

5. En la página Revisar y crear, elija Crear sistema de archivos. De este modo, volverá a la página Sistemas de archivos.
6. Navegue a la página de detalles del sistema de archivos FSx para Lustre que creó.
7. Anote el ID del sistema de archivos y el nombre del montaje. Usará esta información más tarde.

> **Note**
>
> El campo Estado muestra la opción Crear mientras se aprovisiona el sistema de archivos. La creación del sistema de archivos puede tardar varios minutos. Espere a que se complete antes de continuar con el resto del tutorial.

Cree grupos de nodos de cómputo en AWS PCS

Un grupo de nodos de cómputo es un conjunto virtual de nodos de cómputo (instancias EC2) que AWS PCS lanza y administra. Al definir un grupo de nodos de procesamiento, se especifican características comunes, como los tipos de instancias EC2, el número mínimo y máximo de instancias, las subredes de VPC de destino, la opción de compra preferida y la configuración

de lanzamiento personalizada. AWS PCS lanza, administra y termina eficientemente los nodos de cómputo de un grupo de nodos de cómputo, de acuerdo con estos ajustes. El clúster de demostración utiliza un grupo de nodos de cálculo para proporcionar nodos de inicio de sesión para el acceso de los usuarios y un grupo de nodos de cálculo independiente para procesar los trabajos. En los temas siguientes se describen los procedimientos para configurar estos grupos de nodos de procesamiento en el clúster.

Temas

Crear un perfil de instancia para AWS PCS

Los grupos de nodos de cómputo requieren un perfil de instancia cuando se crean. Si utiliza la Consola de administración de AWS para crear un rol para Amazon EC2, la consola crea automáticamente un perfil de instancias y le da el mismo nombre que al rol. Para obtener más información, consulta Cómo usar perfiles de instancia en la Guía del AWS Identity and Access Management usuario.

En el siguiente procedimiento, se utiliza Consola de administración de AWS para crear un rol para Amazon EC2, que también crea el perfil de instancia para los grupos de nodos de cómputo.

Para crear el perfil de rol y de instancia

- Vaya a la consola de IAM.
- En Administración de acceso, seleccione Políticas.
  - Elija Create Policy.
  - En Especificar permisos, en el editor de políticas, selecciona JSON.
  - Sustituya el contenido del editor de texto por lo siguiente:

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
```

```
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

  - Elija Siguiente.
  - En Revisar y crear, escriba el nombre de la políticaAWSPCS-getstarted-policy.
  - Elija Crear política.
- En Access management (Administración de acceso), elija Roles (Roles).
- Elija Crear rol.
- En Seleccionar entidad de confianza:
  - En el tipo de entidad de confianza, seleccione AWS servicio
  - En Caso de uso, seleccione EC2.
    - A continuación, en Elija un caso de uso para el servicio especificado, elija EC2.
  - Elija Siguiente.
- En Añadir permisos:
  - En Políticas de permisos, busque AWSPCS-getstarted-policy.
  - Marque la casilla situada junto a AWSPCS-getstarted-policy para añadirla al rol.
  - En Políticas de permisos, busca Amazon SSMManaged InstanceCore.
  - Marca la casilla situada junto SSMManaged InstanceCore a Amazon para añadirlo al rol.
  - Elija Siguiente.
- En Nombre, revisa y crea:
  - En Detalles del rol:
    - En Role name (Nombre del rol), introduzca AWSPCS-getstarted-role.
  - Elija Create role (Crear rol).

Cree plantillas de lanzamiento para AWS PCS

Al crear un grupo de nodos de cómputo, se proporciona una plantilla de lanzamiento de EC2 que AWS PCS utiliza para configurar las instancias de EC2 que lanza. Esto incluye ajustes como los grupos de seguridad y los scripts que se ejecutan cuando se lanza la instancia.

En este paso, se utilizará una CloudFormation plantilla para crear dos plantillas de lanzamiento de EC2. Una plantilla se usará para crear nodos de inicio de sesión y la otra se usará para crear nodos de cómputo. La diferencia clave entre ellos es que los nodos de inicio de sesión se pueden configurar para permitir el acceso SSH entrante.

Accede a la plantilla CloudFormation

Usa la siguiente URL para descargar la CloudFormation plantilla y, a continuación, cárgala en la CloudFormation consola para crear una CloudFormation pila nueva. Para obtener más información, consulte Uso de la CloudFormation consola en la Guía del AWS CloudFormation usuario.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/getting_started/assets/pcs-lt-efs-fsxl.yaml
```

Utilice la CloudFormation plantilla para crear plantillas de lanzamiento de EC2

Utilice el siguiente procedimiento para completar la CloudFormation plantilla en la consola CloudFormation

- En Proporcione un nombre de pila:
  - En Nombre de pila, ingresa`getstarted-lt`.
- En Parámetros:
  - Bajo seguridad
    - Para VpcSecurityGroupId, seleccione el grupo de seguridad nombrado `default` en la VPC de su clúster.
    - Para ClusterSecurityGroupId, seleccione el grupo denominado `cluster-getstarted-sg`
    - Para SshSecurityGroupId, seleccione el grupo denominado `inbound-ssh-getstarted-sg`
    - Para SshKeyName, selecciona tu key pair de claves SSH preferido.
  - En Sistemas de archivos
    - Para EfsFilesystemId, introduzca el ID del sistema de archivos del sistema de archivos EFS que creó anteriormente en el tutorial.

    - Para FSxLustreFilesystemId, introduzca el ID del sistema de archivos del sistema FSx de archivos de Lustre que creó anteriormente en el tutorial.
    - Para FSxLustreFilesystemMountName, introduzca el nombre de montaje correspondiente al mismo sistema FSx de archivos Lustre.
- Seleccione Siguiente y, a continuación, vuelva a seleccionar Siguiente.
- Seleccione Enviar.

Supervisa el estado de la CloudFormation pila. Cuando llegue a `CREATE_COMPLETE` la plantilla de lanzamiento estará lista para ser utilizada.

> **Note**
>
> Para ver todos los recursos que creó la CloudFormation plantilla, abre la CloudFormation consola. Elija la pila `getstarted-lt` y, a continuación, elija la pestaña Resources (Recursos).

Cree un grupo de nodos de cómputo para los nodos de inicio de sesión en AWS PCS

Un grupo de nodos de cómputo es un conjunto virtual de nodos de cómputo (instancias EC2) que AWS PCS lanza y administra. Al definir un grupo de nodos de procesamiento, se especifican características comunes, como los tipos de instancias EC2, el número mínimo y máximo de instancias, las subredes de VPC de destino, la opción de compra preferida y la configuración de lanzamiento personalizada. AWS PCS lanza, administra y termina eficientemente los nodos de cómputo de un grupo de nodos de cómputo, de acuerdo con estos ajustes.

En este paso, lanzará un grupo de nodos de computación estáticos que proporciona acceso interactivo al clúster. Puede utilizar SSH o Amazon EC2 Systems Manager (SSM) para iniciar sesión en él y, a continuación, ejecutar comandos de shell y gestionar los trabajos de Slurm.

Para crear el grupo de nodos de cómputo

- Abra la consola AWS PCS y vaya a Clústeres.
- Seleccione el clúster denominado `get-started`
- Vaya a Compute nodes groups y elija Crear.

- En la sección de configuración del grupo de nodos de Compute, proporciona lo siguiente:
  - Nombre del grupo de nodos de cómputo: `login` introdúzcalo.
- En Configuración informática, introduzca o seleccione estos valores:
  - Plantilla de lanzamiento de EC2: elija la plantilla de lanzamiento con el nombre `login-getstarted-lt`
  - Perfil de instancia de IAM: elija el nombre del perfil de instancia `AWSPCS-getstarted-role`
  - Subredes: seleccione la subred por la que comienza el nombre. `hpc-networking:PublicSubnetA`
  - Instancias: seleccione. `c6i.xlarge`
  - Configuración de escalado: introduzca `1` el número mínimo de instancias. Para el recuento máximo de instancias, introduzca. `1`
- En Configuración adicional, especifique lo siguiente:
  - ID de AMI: seleccione la AMI que desee usar y que tenga un nombre con el siguiente formato:

    ```
    aws-pcs-sample_ami-amzn2-platform-slurm-version
    ```

    Para obtener más información sobre el ejemplo AMIs, consulteUso de Amazon Machine Images (AMIs) de muestra con AWS PCS.
- Seleccione Crear grupo de nodos de cómputo.

El campo Estado muestra la opción Crear mientras se aprovisiona el grupo de nodos de cómputo. Puede continuar con el siguiente paso del tutorial mientras está en curso.

Cree un grupo de nodos de cómputo para ejecutar trabajos de cómputo en AWS PCS

En este paso, lanzará un grupo de nodos de cómputo que se escale de forma elástica para ejecutar los trabajos enviados al clúster.

Para crear el grupo de nodos de cómputo

- Abra la consola AWS PCS y vaya a Clústeres.
- Seleccione el clúster denominado `get-started`
- Vaya a Compute nodes groups y elija Crear.

- En la sección de configuración del grupo de nodos de Compute, proporciona lo siguiente:
  - Nombre del grupo de nodos de cómputo: `compute-1` introdúzcalo.
- En Configuración informática, introduzca o seleccione estos valores:
  - Plantilla de lanzamiento de EC2: elija la plantilla de lanzamiento con el nombre `compute-getstarted-lt`
  - Perfil de instancia de IAM: elija el nombre del perfil de instancia `AWSPCS-getstarted-role`
  - Subredes: seleccione la subred por la que comienza el nombre. `hpc-networking:PrivateSubnetA`
  - Instancias: seleccione. `c6i.xlarge`
  - Configuración de escalado: para el recuento mínimo de instancias, introduzca**`0`**. Para el recuento máximo de instancias, introduzca. `4`
- En Configuración adicional, especifique lo siguiente:
  - ID de AMI: seleccione la AMI que desee usar y que tenga un nombre con el siguiente formato:

    ```
    aws-pcs-sample_ami-amzn2-platform-slurm-version
    ```

    Para obtener más información sobre el ejemplo AMIs, consulteUso de Amazon Machine Images (AMIs) de muestra con AWS PCS.
- Seleccione Crear grupo de nodos de cómputo.

El campo Estado muestra la opción Crear mientras se aprovisiona el grupo de nodos de cómputo.

> **Important**
>
> Espere a que el campo Estado muestre Activo antes de continuar con el siguiente paso de este tutorial.

Cree una cola para administrar los trabajos en AWS PCS

Debe enviar un trabajo a una cola para ejecutarlo. El trabajo permanece en la cola hasta que AWS PCS lo programe para que se ejecute en un grupo de nodos de cómputo. Cada cola está asociada a uno o más grupos de nodos de cómputo, que proporcionan las EC2 instancias necesarias para realizar el procesamiento.

En este paso, creará una cola que utilice el grupo de nodos de procesamiento para procesar los trabajos.

Para crear una cola

- Abra la consola AWS PCS.
- Seleccione el clúster denominado `get-started`.
- Vaya a Compute node groups y asegúrese de que el estado del `compute-1` grupo sea Activo.

> **Important**
>
> El estado del `compute-1` grupo debe ser Activo antes de continuar con el siguiente paso.

- Navegue hasta Colas y elija Crear cola.
  - En la sección de configuración de colas, proporciona los siguientes valores:
    - Nombre de la cola: introduzca lo siguiente: `demo`
    - Grupos de nodos de cómputo: seleccione el nombre `compute-1` del grupo de nodos de cómputo.
- Elige Crear cola.

El campo Estado muestra la opción Crear mientras se crea la cola.

> **Important**
>
> Espere a que el campo Estado muestre Activo antes de continuar con el siguiente paso de este tutorial.

Conéctese a su clúster AWS PCS

Cuando el estado del grupo de nodos de `login` cómputo pase a ser Activo, podrás conectarte a la EC2 instancia que creó.

Para conectarse al nodo de inicio de sesión

- Abra la consola AWS PCS y vaya a Clusters.
- Seleccione el clúster denominado `get-started`.

- Elija grupos de nodos de cómputo.
- Navegue hasta el grupo de nodos de cómputo denominado`login`.
- Busque el ID del grupo de nodos de cómputo.
- En otra ventana o pestaña del navegador, abre la EC2 consola de Amazon.
  - Elija Instances.
  - Busca EC2 instancias con la siguiente etiqueta. *`node-group-id`*Sustitúyalo por el valor del ID del grupo de nodos de Compute del paso anterior. Debe haber 1 instancia.

    ```
    aws:pcs:compute-node-group-id=node-group-id
    ```

  - Conéctese a la EC2 instancia. Puedes usar el administrador de sesiones o SSH.

    Session Manager

    - Seleccione la instancia.
    - Elija Conectar.
    - En Conectar a la instancia, selecciona Administrador de sesiones.
    - Elija Conectar.
    - Elija Conectar. Se abrirá un terminal interactivo en el navegador.

    SSH

    - Seleccione la instancia.
    - Elija Conectar.
    - En Conectar a la instancia, selecciona Cliente SSH.
    - Sigue las instrucciones que te proporciona la consola.

      > **Note**
      >
      > El nombre de usuario de la instancia **`ec2-user`**no lo es`root`.

Explore el entorno de clústeres en AWS PCS

Una vez que haya iniciado sesión en el clúster, podrá ejecutar comandos de shell. Por ejemplo, puede cambiar de usuario, trabajar con datos en sistemas de archivos compartidos e interactuar con Slurm.

Cambiar de usuario

Si ha iniciado sesión en el clúster mediante el Administrador de sesiones, es posible que esté conectado como`ssm-user`. Se trata de un usuario especial que se creó para el Administrador de sesiones. Cambie al usuario predeterminado en Amazon Linux 2 mediante el siguiente comando. No necesitará hacer esto si se conectó mediante SSH.

```
sudo su - ec2-user
```

Trabaje con sistemas de archivos compartidos

Puede confirmar que el sistema de archivos EFS y los sistemas FSx de archivos Lustre están disponibles con el comando. `df -h` El resultado del clúster debe ser similar al siguiente:

```
[ec2-user@ip-10-3-6-103 ~]$ df -h
Filesystem                Size  Used Avail Use% Mounted on
devtmpfs                  3.8G     0  3.8G   0% /dev
tmpfs                     3.9G     0  3.9G   0% /dev/shm
tmpfs                     3.9G  556K  3.9G   1% /run
tmpfs                     3.9G     0  3.9G   0% /sys/fs/cgroup
/dev/nvme0n1p1             24G   18G  6.6G  73% /
127.0.0.1:/               8.0E     0  8.0E   0% /home
10.3.132.79@tcp:/zlshxbev 1.2T  7.5M  1.2T   1% /shared
tmpfs                     780M     0  780M   0% /run/user/0
tmpfs                     780M     0  780M   0% /run/user/1000
```

El `/home` sistema de archivos monta 127.0.0.1 y tiene una capacidad muy grande. Este es el sistema de archivos EFS que creó anteriormente en el tutorial. Todos los archivos que se escriban aquí estarán disponibles `/home` en todos los nodos del clúster.

El `/shared` sistema de archivos monta una IP privada y tiene una capacidad de 1,2 TB. Este es el sistema de archivos FSx para Lustre que creó anteriormente en el tutorial. Todos los archivos que se escriban aquí estarán disponibles `/shared` en todos los nodos del clúster.

Interactúa con Slurm

Temas

- Enumere las colas y los nodos

- Mostrar trabajos

Enumere las colas y los nodos

Puede enumerar las colas y los nodos a los que están asociadas. `sinfo` El resultado del clúster debe ser similar al siguiente:

```
[ec2-user@ip-10-3-6-103 ~]$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
demo         up   infinite      4  idle~ compute-1-[1-4]
[ec2-user@ip-10-3-6-103 ~]$
```

Anote el nombre de la particióndemo. Su estado es up y tiene un máximo de 4 nodos. Está asociado a los nodos del grupo de `compute-1` nodos. Si edita el grupo de nodos de cómputo y aumenta el número máximo de instancias a 8, el número de nodos se leerá 8 y la lista de nodos se leerá`compute-1-[1-8]`. Si creara un segundo grupo de nodos de cómputo `test` con un nombre de 4 nodos y lo añadiera a la demo cola, esos nodos también aparecerían en la lista de nodos.

Mostrar trabajos

Puede enumerar todos los trabajos, en cualquier estado, del sistema con`squeue`. El resultado del clúster debe ser similar al siguiente:

```
[ec2-user@ip-10-3-6-103 ~]$ squeue
JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON)
```

Intente `squeue` volver a ejecutarlo más tarde, cuando tenga un trabajo de Slurm pendiente o en ejecución.

Ejecute un trabajo de un solo nodo en AWS PCS

Para ejecutar un trabajo con Slurm, debe preparar un script de envío que especifique los requisitos del trabajo y enviarlo a una cola con el comando. `sbatch` Por lo general, esto se hace desde un directorio compartido, de modo que los nodos de inicio de sesión y de procesamiento tengan un espacio común para acceder a los archivos.

Conéctese al nodo de inicio de sesión de su clúster y ejecute los siguientes comandos en su intérprete de comandos.

- Conviértete en el usuario predeterminado. Cambie al directorio compartido.

```
sudo su - ec2-user
cd /shared
```

- Utilice los siguientes comandos para crear un ejemplo de script de trabajo:

```
cat << EOF > job.sh
#!/bin/bash
#SBATCH -J single
#SBATCH -o single.%j.out
#SBATCH -e single.%j.err

echo "This is job \${SLURM_JOB_NAME} [\${SLURM_JOB_ID}] running on \
${SLURMD_NODENAME}, submitted from \${SLURM_SUBMIT_HOST}" && sleep 60 && echo "Job
 complete"
EOF
```

- Envíe el script de trabajo al programador de Slurm:

```
sbatch -p demo job.sh
```

- Cuando se envíe el trabajo, devolverá una ID de trabajo en forma de número. Usa ese identificador para comprobar el estado del trabajo. Sustituya *job-id* el siguiente comando por el número devuelto desde`sbatch`.

```
squeue --job job-id
```

  Example

```
squeue --job 1
```

  El `squeue` comando devuelve un resultado similar al siguiente:

```
JOBID PARTITION NAME USER     ST TIME NODES NODELIST(REASON)
1     demo      test ec2-user CF 0:47 1     compute-1
```

- Continúe comprobando el estado del trabajo hasta que alcance el estado `R` (en ejecución). El trabajo está hecho cuando `squeue` no devuelve nada.
- Inspeccione el contenido del `/shared` directorio.

```
ls -alth /shared
```

El resultado del comando es similar al siguiente:

```
-rw-rw-r- 1 ec2-user ec2-user 107 Mar 19 18:33 single.1.out
-rw-rw-r- 1 ec2-user ec2-user 0 Mar 19 18:32 single.1.err
-rw-rw-r- 1 ec2-user ec2-user 381 Mar 19 18:29 job.sh
```

Uno de los nodos de cómputo del clúster `single.1.err` asignó un nombre `single.1.out` a los archivos y los escribió. Como el trabajo se ejecutó en un directorio compartido (`/shared`), también están disponibles en su nodo de inicio de sesión. Por eso configuró un sistema de archivos FSx para Lustre para este clúster.

- Inspeccione el contenido del `single.1.out` archivo.

```
cat /shared/single.1.out
```

El resultado es similar al siguiente:

```
This is job test [1] running on compute-1, submitted from ip-10-3-13-181
Job complete
```

Ejecute un trabajo de MPI de varios nodos con Slurm en PCS AWS

Estas instrucciones muestran el uso de Slurm para ejecutar un trabajo de interfaz de paso de mensajes (MPI) en PCS. AWS

Ejecute los siguientes comandos en el intérprete de comandos de su nodo de inicio de sesión.

- Conviértete en el usuario predeterminado. Cambie a su directorio principal.

```
sudo su - ec2-user
cd ~/
```

- Cree el código fuente en el lenguaje de programación C.

```
cat > hello.c << EOF
// * mpi-hello-world - https://www.mpitutorial.com
```

```
// Released under MIT License
//
// Copyright (c) 2014 MPI Tutorial.
//
// Permission is hereby granted, free of charge, to any person obtaining a copy
// of this software and associated documentation files (the "Software"), to
// deal in the Software without restriction, including without limitation the
// rights to use, copy, modify, merge, publish, distribute, sublicense, and/or
// sell copies of the Software, and to permit persons to whom the Software is
// furnished to do so, subject to the following conditions:
// The above copyright notice and this permission notice shall be included in
// all copies or substantial portions of the Software.
//
// THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
// IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
// FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE
// AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
// LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING
// FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER
// DEALINGS IN THE SOFTWARE.

#include <mpi.h>
#include <stdio.h>
#include <stddef.h>

int main(int argc, char** argv) {
  // Initialize the MPI environment. The two arguments to MPI Init are not
  // currently used by MPI implementations, but are there in case future
  // implementations might need the arguments.
  MPI_Init(NULL, NULL);

  // Get the number of processes
  int world_size;
  MPI_Comm_size(MPI_COMM_WORLD, &world_size);

  // Get the rank of the process
  int world_rank;
  MPI_Comm_rank(MPI_COMM_WORLD, &world_rank);

  // Get the name of the processor
  char processor_name[MPI_MAX_PROCESSOR_NAME];
  int name_len;
  MPI_Get_processor_name(processor_name, &name_len);
```

```
  // Print off a hello world message
  printf("Hello world from processor %s, rank %d out of %d processors\n",
         processor_name, world_rank, world_size);

  // Finalize the MPI environment. No more MPI calls can be made after this
  MPI_Finalize();
}
EOF
```

- Cargue el módulo OpenMPI.

```
module load openmpi
```

- Compila el programa C.

```
mpicc -o hello hello.c
```

- Escribe un guion de envío de trabajos de Slurm.

```
cat > hello.sh << EOF
#!/bin/bash
#SBATCH -J multi
#SBATCH -o multi.out
#SBATCH -e multi.err
#SBATCH --exclusive
#SBATCH --nodes=4
#SBATCH --ntasks-per-node=1

srun $HOME/hello
EOF
```

- Cambie al directorio compartido.

```
cd /shared
```

- Envíe el script de trabajo.

```
sbatch -p demo ~/hello.sh
```

- `squeue`Úselo para supervisar el trabajo hasta que esté terminado.
- Compruebe el contenido de`multi.out`:

```
cat multi.out
```

El resultado es similar al siguiente. Tenga en cuenta que cada rango tiene su propia dirección IP porque se ejecutó en un nodo diferente.

```
Hello world from processor ip-10-3-133-204, rank 0 out of 4 processors
Hello world from processor ip-10-3-128-219, rank 2 out of 4 processors
Hello world from processor ip-10-3-141-26, rank 3 out of 4 processors
Hello world from processor ip-10-3-143-52, rank 1 out of 4 processor
```

Elimine sus AWS recursos para AWS PCS

Una vez que haya terminado con los grupos de clústeres y nodos que creó para este tutorial, debe eliminar los recursos que creó.

> **Important**
>
> Recibirás cargos de facturación por todos los recursos que estén en funcionamiento en tu Cuenta de AWS

Para eliminar los recursos de AWS PCS que creó para este tutorial

- Abra la consola AWS PCS.
- Navegue hasta el clúster denominado get-started.
- Navegue hasta la sección Colas.
- Seleccione la cola llamada demo.
- Elija Eliminar.

> **Important**
>
> Espere a que se elimine la cola antes de continuar.

- Navegue hasta la sección Grupos de nodos de cómputo.
- Seleccione el grupo de nodos de cómputo denominado compute-1.

- Elija Eliminar.
- Seleccione el grupo de nodos de cómputo denominado login.
- Elija Eliminar.

> **Important**
>
> Espere a que se eliminen ambos grupos de nodos de procesamiento antes de continuar.

- En la página de detalles del clúster para empezar, selecciona Eliminar.

> **Important**
>
> Espere a que se elimine el clúster antes de continuar con los pasos siguientes.

Para eliminar otros AWS recursos que haya creado para este tutorial

- Abra la consola de IAM.
  - Elija Roles.
  - Seleccione el rol denominado AWSPCS-getstarted-role y, a continuación, elija Eliminar.
  - Una vez que se haya eliminado el rol, elija Políticas.
  - Seleccione la política denominada AWSPCS-getstarted-policy y, a continuación, elija Eliminar.
- Abra la consola de CloudFormation.
  - Seleccione la pila denominada getstarted-lt.
  - Elija Eliminar.

> **Important**
>
> Espere a que se elimine la pila antes de continuar.

- Abra la consola de Amazon EFS.
  - Elija Sistemas de archivos.
  - Seleccione el sistema de archivos denominado getstarted-efs.
  - Elija Eliminar.

> **Important**
>
> Espere a que el sistema de archivos elimine antes de continuar.

- Abre la FSx consola de Amazon.
  - Elija Sistemas de archivos.
  - Selecciona el sistema de archivos llamado getstarted-fsx.
  - Elija Eliminar.

> **Important**
>
> Espere a que el sistema de archivos elimine antes de continuar.

- Abra la consola de CloudFormation.
  - Seleccione la pila denominada getstarted-sg.
  - Elija Eliminar.
- Abra la consola de CloudFormation.
  - Seleccione la pila denominada hpc-networking.
  - Elija Eliminar.

Comience con un CloudFormation AWS PCS

Se puede utilizar AWS CloudFormation para crear un clúster de AWS PCS. CloudFormation le permite crear y aprovisionar despliegues de AWS infraestructura de forma predecible y repetitiva. Puede utilizarlos CloudFormation para aprovisionar automáticamente los recursos de muchos AWS servicios para crear aplicaciones altamente confiables, escalables y rentables Nube de AWS sin necesidad de crear ni configurar la infraestructura subyacente AWS . CloudFormation le permite utilizar un archivo de plantilla para crear y eliminar un conjunto de recursos juntos como una sola unidad, denominada pila. Para obtener más información al respecto CloudFormation, consulte ¿Qué es CloudFormation? en la Guía AWS CloudFormation del usuario. Para obtener más información sobre los tipos de recursos de AWS PCS CloudFormation, consulte la referencia de tipos de recursos de AWS PCS en la Guía del AWS CloudFormation usuario.

Temas

Se utiliza CloudFormation para crear un clúster de AWS PCS de muestra

El siguiente procedimiento utiliza una CloudFormation plantilla en el Consola de administración de AWS para crear un clúster de AWS PCS de muestra. Para obtener más información al respecto CloudFormation, consulte ¿Qué es CloudFormation? en la Guía AWS CloudFormation del usuario. Para obtener más información sobre los tipos de recursos de AWS PCS CloudFormation, consulte la referencia de tipos de recursos de AWS PCS en la Guía del AWS CloudFormation usuario.

Para crear el clúster de muestra

1. Elija la opción en la Región de AWS que desea crear el clúster (el enlace abre la CloudFormation consola con la plantilla):

    - EE.UU. Este (Virginia) (us-east-1)

- EE.UU. Este (Ohio) (us-east-2)
- EE.UU. Oeste (Oregón) (us-west-2)
- Asia Pacífico (Singapur) (ap-southeast-1)
- Asia Pacífico (Sídney) (ap-southeast-2)
- Asia Pacífico (Tokio) (ap-northeast-1)
- Europa (Fráncfort) (eu-central-1)
- Europa (Irlanda) (eu-west-1)
- Europa (Londres) (eu-west-2)
- Europa (Estocolmo) (eu-north-1)
- AWS GovCloud (EEUU-Este) (-1) us-gov-east
- AWS GovCloud (EEUU-Oeste) (us-gov-west-1)

2. En Proporcione un nombre de pila, introduzca un nombre descriptivo. Este es el nombre de la CloudFormation pila. La plantilla usa este valor como nombre para el clúster de AWS PCS.
3. En Parámetros:

   a. En SlurmVersion, elige la versión de Slurm que quieres que use tu clúster.

   b. En NodeArchitecture, elija x86 para implementar un clúster que use instancias compatibles con x86_64, o elija Graviton para usar instancias Arm64.

   c. Para KeyName, elija un key pair de claves SSH para acceder a los nodos de inicio de sesión del clúster. Asegúrese de tener el archivo PEM para el key pair que elija.

   d. Para ello ClientIpCidr, introduzca un rango de IP en formato CIDR para controlar el acceso a los nodos de inicio de sesión.

   > **Warning**
   >
   > El valor predeterminado de `0.0.0.0/0` permite el acceso desde todas las direcciones IP.

   e. Deje los valores de HpcRecipesS3Bucket y HpcRecipesBranchsus valores predeterminados.

4. En Capacidades y transformaciones:

   a. Seleccione la casilla de verificación para confirmar que CloudFormation se crearán recursos de IAM.

b. Seleccione la casilla de verificación para confirmar que CloudFormation se crearán recursos de IAM con nombres personalizados.

c. Seleccione la casilla de verificación `CAPABILITY_AUTO_EXPAND` para confirmar la nueva pila. Para obtener más información, consulta CreateStack en la AWS CloudFormation Referencia de la API de .

5. Seleccione Creación de pila.

6. Supervisa el estado de tu pila. Puede conectarse al clúster una vez que se encuentre el estado de la pila`CREATE_COMPLETE`.

Conéctese a un clúster AWS PCS creado con CloudFormation

Tras crear un clúster de AWS PCS a partir de una CloudFormation plantilla, puede utilizar la consola de AWS PCS (en la Consola de administración de AWS) para administrar el clúster. También puede conectarse a uno de los nodos de inicio de sesión del clúster para administrar el clúster, ejecutar trabajos y administrar los datos. La CloudFormation pila proporciona enlaces que puede utilizar para conectarse al clúster.

Para conectarse a su clúster

1. Abra la consola de CloudFormation.

2. Elige la pila que has creado.

3. Seleccione la pestaña Salidas de la pila.

   La pila proporciona los siguientes enlaces:

   - PcsConsoleUrl— Elija este enlace para abrir la consola AWS PCS con el clúster seleccionado. Puede usarlo para explorar las configuraciones del clúster, el grupo de nodos y las colas.
   - Ec2 ConsoleUrl: elija este enlace para abrir la consola Amazon EC2, filtrada para mostrar las instancias que administra el grupo de nodos de inicio de sesión del clúster.

     Desde esta vista, puede seleccionar una instancia y elegir Connect. La instancia del clúster de muestra admite el SSH entrante y AWS Systems Manager las conexiones en un navegador web. Para obtener más información, consulte Conéctese a su clúster AWS PCS.

     Después de conectarte a una instancia de inicio de sesión, puedes seguir el tutorial que se encuentra en. Explore el entorno de clústeres en AWS PCS

Limpiar un clúster de AWS PCS en CloudFormation

Si CloudFormation solía crear su clúster de AWS PCS, puede abrir la CloudFormation consola y eliminar la pila para eliminar el clúster y todos sus recursos asociados.

> **Important**
>
> En el caso del clúster de muestra, si ha creado colas o grupos de nodos de procesamiento adicionales en el clúster (además de `compute-1` los grupos `login` y que creó la CloudFormation plantilla de ejemplo), debe utilizar la consola AWS PCS o AWS CLI eliminar esos recursos antes de eliminar la CloudFormation pila. Para obtener más información, consulte Eliminar un clúster en AWS PCS.

Partes de una CloudFormation plantilla para AWS PCS

Una CloudFormation plantilla tiene una o más secciones, cada una de las cuales tiene un propósito específico. CloudFormation define el formato, la sintaxis y el idioma estándar de una plantilla. Para obtener más información, consulte Trabajar con CloudFormation plantillas en la Guía del AWS CloudFormation usuario.

CloudFormation las plantillas son altamente personalizables y, por lo tanto, sus formatos pueden variar. Para comprender las partes necesarias de una CloudFormation plantilla para crear un clúster de AWS PCS, le recomendamos que examine la plantilla de ejemplo que le proporcionamos para crear un clúster de muestra. En este tema se explican brevemente las secciones de esa plantilla de ejemplo.

> **Important**
>
> Los ejemplos de código de este tema no están completos. La presencia de puntos suspensivos (`[...]`) indica que hay código adicional que no se muestra. Para descargar la plantilla completa con formato YAML CloudFormation , consulte. CloudFormation plantillas para crear un clúster de AWS PCS de muestra

Contenido

- Parameters
- Mapeos
- Recursos
- Outputs

Encabezado

```
AWSTemplateFormatVersion: '2010-09-09'
Transform: AWS::Serverless-2016-10-31
Description: AWS Parallel Computing Service "getting started" cluster
```

`AWSTemplateFormatVersion`identifica la versión del formato de plantilla a la que se ajusta la plantilla. Para obtener más información, consulte la sintaxis de la versión del formato de CloudFormation plantilla en la Guía del AWS CloudFormation usuario.

`Transform`especifica una macro que se CloudFormation utiliza para procesar la plantilla. Para obtener más información, consulte la sección Transformación de CloudFormation plantillas en la Guía del AWS CloudFormation usuario. La `AWS::Serverless-2016-10-31` transformación permite CloudFormation procesar una plantilla escrita en la sintaxis AWS Serverless Application Model (AWS SAM). Para obtener más información, consulte la `AWS::Serverless`transformación en la Guía AWS CloudFormation del usuario.

Metadatos

```
### Stack metadata
Metadata:
  AWS::CloudFormation::Interface:
    ParameterGroups:
      - Label:
          default: PCS Cluster configuration
        Parameters:
          - SlurmVersion
          - ManagedAccounting
          - AccountingPolicyEnforcement
      - Label:
          default: PCS ComputeNodeGroups configuration
        Parameters:
          - NodeArchitecture
          - KeyName
```

```
          - ClientIpCidr
      - Label:
          default: HPC Recipes configuration
        Parameters:
          - HpcRecipesS3Bucket
          - HpcRecipesBranch
```

La `metadata` sección de una CloudFormation plantilla proporciona información sobre la propia plantilla. La plantilla de ejemplo crea un clúster de computación de alto rendimiento (HPC) completo que utiliza AWS PCS. La sección de metadatos de la plantilla de ejemplo declara los parámetros que controlan la forma en que CloudFormation se lanza (aprovisiona) la pila correspondiente. Hay parámetros que controlan la elección de la arquitectura (`NodeArchitecture`), la versión de Slurm (`SlurmVersion`) y los controles de acceso (`KeyNameyClientIpCidr`).

Parameters

La `Parameters` sección define los parámetros personalizados de la plantilla. CloudFormation utiliza estas definiciones de parámetros para construir y validar el formulario con el que se interactúa al lanzar una pila desde esta plantilla.

```
Parameters:

  NodeArchitecture:
    Type: String
    Default: x86
    AllowedValues:
      - x86
      - Graviton
    Description: Processor architecture for the login and compute node instances

  SlurmVersion:
    Type: String
    Default: 25.05
    Description: Version of Slurm to use
    AllowedValues:
         - 24.11
         - 25.05

  ManagedAccounting:
    Type: String
    Default: 'disabled'
    AllowedValues:
```

```
      - 'enabled'
      - 'disabled'
    Description: Monitor cluster usage, manage access control, and enforce resource
limits with Slurm accounting. Requires Slurm 24.11 or newer.

  AccountingPolicyEnforcement:
    Description: Specify which Slurm accounting policies to enforce
    Type: String
    Default: none
    AllowedValues:
      - none
      - 'associations,limits,safe'

  KeyName:
    Description: SSH keypair to log in to the head node
    Type: AWS::EC2::KeyPair::KeyName
    AllowedPattern: ".+"  # Required

  ClientIpCidr:
    Description: IP(s) allowed to access the login node over SSH. We recommend that
you restrict it with your own IP/subnet (x.x.x.x/32 for your own ip or x.x.x.x/24 for
range. Replace x.x.x.x with your own PUBLIC IP. You can get your public IP using tools
such as https://ifconfig.co/)
    Default: 127.0.0.1/32
    Type: String
    AllowedPattern: (\d{1,3})\.(\d{1,3})\.(\d{1,3})\.(\d{1,3})/(\d{1,2})
    ConstraintDescription: Value must be a valid IP or network range of the form
x.x.x.x/x.

  HpcRecipesS3Bucket:
    Type: String
    Default: aws-hpc-recipes
    Description: HPC Recipes for AWS S3 bucket
    AllowedValues:
         - aws-hpc-recipes
         - aws-hpc-recipes-dev
  HpcRecipesBranch:
    Type: String
    Default: main
    Description: HPC Recipes for AWS release branch
    AllowedPattern: '^(?!.*/\.git$)(?!.*/\.)(?!.*\\.\.)[a-zA-Z0-9-_\.]+$'
```

Mapeos

La `Mappings` sección define pares clave-valor que especifican valores en función de determinadas condiciones o dependencias.

```
Mappings:

  Architecture:
    AmiArchParameter:
      Graviton: arm64
      x86: x86_64
    LoginNodeInstances:
      Graviton: c7g.xlarge
      x86: c6i.xlarge
    ComputeNodeInstances:
      Graviton: c7g.xlarge
      x86: c6i.xlarge
```

Recursos

La `Resources` sección declara los AWS recursos que se van a aprovisionar y configurar como parte de la pila.

```
Resources:

  [...]
```

La plantilla aprovisiona la infraestructura de clústeres de muestra en capas. Comienza con la `Networking` configuración de VPC. El almacenamiento lo proporcionan dos sistemas: `EfsStorage` para almacenamiento compartido y `FSxLStorage` almacenamiento de alto rendimiento. El clúster principal se establece mediante`PCSCluster`.

```

  Networking:
    Type: AWS::CloudFormation::Stack
    Properties:
      Parameters:
        ProvisionSubnetsC: "False"
      TemplateURL: !Sub 'https://${HpcRecipesS3Bucket}.s3.amazonaws.com/
${HpcRecipesBranch}/recipes/net/hpc_large_scale/assets/main.yaml'
```

```
  EfsStorage:
    Type: AWS::CloudFormation::Stack
    Properties:
      Parameters:
        SubnetIds: !GetAtt [ Networking, Outputs.DefaultPrivateSubnet ]
        SubnetCount: 1
        VpcId: !GetAtt [ Networking, Outputs.VPC ]
      TemplateURL: !Sub 'https://${HpcRecipesS3Bucket}.s3.amazonaws.com/
${HpcRecipesBranch}/recipes/storage/efs_simple/assets/main.yaml'

  FSxLStorage:
    Type: AWS::CloudFormation::Stack
    Properties:
      Parameters:
        PerUnitStorageThroughput: 125
        SubnetId: !GetAtt [ Networking, Outputs.DefaultPrivateSubnet ]
        VpcId: !GetAtt [ Networking, Outputs.VPC ]
      TemplateURL: !Sub 'https://${HpcRecipesS3Bucket}.s3.amazonaws.com/
${HpcRecipesBranch}/recipes/storage/fsx_lustre/assets/persistent.yaml'

  [...]

  # Cluster
  PCSCluster:
    Type: AWS::PCS::Cluster
    Properties:
      Name: !Sub '${AWS::StackName}'
      Size: SMALL
      Scheduler:
        Type: SLURM
        Version: !Ref SlurmVersion
      Networking:
        SubnetIds:
          - !GetAtt [ Networking, Outputs.DefaultPrivateSubnet ]
        SecurityGroupIds:
          - !GetAtt [ PCSSecurityGroup, Outputs.ClusterSecurityGroupId ]
```

En el caso de los recursos informáticos, la plantilla crea dos grupos de nodos: `PCSNodeGroupLogin` para un único nodo de inicio de sesión y `PCSNodeGroupCompute` para un máximo de cuatro nodos informáticos. Estos grupos de nodos son compatibles con los permisos y, `PCSInstanceProfile` por `PCSLaunchTemplate` ejemplo, con las configuraciones.

```
  # Compute Node groups
  PCSInstanceProfile:
    Type: AWS::CloudFormation::Stack
    Properties:
      Parameters:
        # We have to regionalize this in case CX use the template in more than one
 region. Otherwise,
        # the create action will fail since instance-role-${AWS::StackName} already
 exists!
        RoleName: !Sub '${AWS::StackName}-${AWS::Region}'
      TemplateURL: !Sub 'https://${HpcRecipesS3Bucket}.s3.amazonaws.com/
${HpcRecipesBranch}/recipes/pcs/getting_started/assets/pcs-iip-minimal.yaml'

  PCSLaunchTemplate:
    Type: AWS::CloudFormation::Stack
    Properties:
      Parameters:
        VpcDefaultSecurityGroupId: !GetAtt [ Networking, Outputs.SecurityGroup ]
        ClusterSecurityGroupId: !GetAtt [ PCSSecurityGroup,
 Outputs.ClusterSecurityGroupId ]
        SshSecurityGroupId: !GetAtt [ PCSSecurityGroup,
 Outputs.InboundSshSecurityGroupId ]
        EfsFilesystemSecurityGroupId: !GetAtt [ EfsStorage, Outputs.SecurityGroupId ]
        FSxLustreFilesystemSecurityGroupId: !GetAtt [ FSxLStorage,
 Outputs.FSxLustreSecurityGroupId ]
        SshKeyName: !Ref KeyName
        EfsFilesystemId: !GetAtt [ EfsStorage, Outputs.EFSFilesystemId ]
        FSxLustreFilesystemId: !GetAtt [ FSxLStorage, Outputs.FSxLustreFilesystemId ]
        FSxLustreFilesystemMountName: !GetAtt [ FSxLStorage,
 Outputs.FSxLustreMountName ]
      TemplateURL: !Sub 'https://${HpcRecipesS3Bucket}.s3.amazonaws.com/
${HpcRecipesBranch}/recipes/pcs/getting_started/assets/cfn-pcs-lt-efs-fsxl.yaml'

  # Compute Node groups - Login Nodes
  PCSNodeGroupLogin:
    Type: AWS::PCS::ComputeNodeGroup
    Properties:
      ClusterId: !GetAtt [PCSCluster, Id]
      Name: login
      ScalingConfiguration:
        MinInstanceCount: 1
        MaxInstanceCount: 1
```

```
      IamInstanceProfileArn: !GetAtt [ PCSInstanceProfile, Outputs.InstanceProfileArn ]
      CustomLaunchTemplate:
        TemplateId: !GetAtt [ PCSLaunchTemplate, Outputs.LoginLaunchTemplateId ]
        Version: 1
      SubnetIds:
        - !GetAtt [ Networking, Outputs.DefaultPublicSubnet ]
      AmiId: !GetAtt [PcsSampleAmi, AmiId]
      InstanceConfigs:
        - InstanceType: !FindInMap [ Architecture, LoginNodeInstances, !Ref
NodeArchitecture ]

 # Compute Node groups - Compute Nodes
 PCSNodeGroupCompute:
   Type: AWS::PCS::ComputeNodeGroup
   Properties:
     ClusterId: !GetAtt [PCSCluster, Id]
     Name: compute-1
     ScalingConfiguration:
       MinInstanceCount: 0
       MaxInstanceCount: 4
     IamInstanceProfileArn: !GetAtt [ PCSInstanceProfile, Outputs.InstanceProfileArn ]
     CustomLaunchTemplate:
       TemplateId: !GetAtt [ PCSLaunchTemplate, Outputs.ComputeLaunchTemplateId ]
       Version: 1
     SubnetIds:
       - !GetAtt [ Networking, Outputs.DefaultPrivateSubnet ]
     AmiId: !GetAtt [PcsSampleAmi, AmiId]
     InstanceConfigs:
       - InstanceType: !FindInMap [ Architecture, ComputeNodeInstances, !Ref
NodeArchitecture ]
```

La programación de trabajos se gestiona mediante`PCSQueueCompute`.

```
 PCSQueueCompute:
   Type: AWS::PCS::Queue
   Properties:
     ClusterId: !GetAtt [PCSCluster, Id]
     Name: demo
     ComputeNodeGroupConfigurations:
       - ComputeNodeGroupId: !GetAtt [PCSNodeGroupCompute, Id]
```

La selección de la AMI se realiza automáticamente a través de la función Pcs AMILookup Fn Lambda y los recursos relacionados.

```
  PcsAMILookupRole:
    Type: AWS::IAM::Role
    [...]

  PcsAMILookupFn:
    Type: AWS::Lambda::Function
    Properties:
      Runtime: python3.12
      Handler: index.handler
      Role: !GetAtt PcsAMILookupRole.Arn
      Code:
        [...]
      Timeout: 30
      MemorySize: 128

  # Example of using the custom resource to look up an AMI
  PcsSampleAmi:
    Type: Custom::AMILookup
    Properties:
      ServiceToken: !GetAtt PcsAMILookupFn.Arn
      OperatingSystem: 'amzn2'
      Architecture: !FindInMap [ Architecture, AmiArchParameter, !Ref
 NodeArchitecture ]
      SlurmVersion: !Ref SlurmVersion
```

Outputs

La plantilla genera la identificación y administración de los clústeres URLs mediante `ClusterIdPcsConsoleUrl`, y. `Ec2ConsoleUrl`

```
Outputs:
  ClusterId:
    Description: The Id of the PCS cluster
    Value: !GetAtt [ PCSCluster, Id ]

  PcsConsoleUrl:
    Description: URL to access the cluster in the PCS console
    Value: !Sub
```

```
      - https://${ConsoleDomain}/pcs/home?region=${AWS::Region}#/clusters/${ClusterId}
      - { ConsoleDomain: !If [ GovCloud, 'console.amazonaws-us-gov.com', !If [ China,
 'console.amazonaws.cn', !Sub '${AWS::Region}.console.aws.amazon.com']],
          ClusterId: !GetAtt [ PCSCluster, Id ]
        }
    Export:
      Name: !Sub ${AWS::StackName}-PcsConsoleUrl

  Ec2ConsoleUrl:
    Description: URL to access instance(s) in the login node group via Session Manager
    Value: !Sub
      - https://${ConsoleDomain}/ec2/home?region=
${AWS::Region}#Instances:instanceState=running;tag:aws:pcs:compute-node-group-id=
${NodeGroupLoginId}
      - { ConsoleDomain: !If [ GovCloud, 'console.amazonaws-us-gov.com', !If [ China,
 'console.amazonaws.cn', !Sub '${AWS::Region}.console.aws.amazon.com']],
          NodeGroupLoginId: !GetAtt [ PCSNodeGroupLogin, Id ]
        }
    Export:
      Name: !Sub ${AWS::StackName}-Ec2ConsoleUrl
```

CloudFormation plantillas para crear un clúster de AWS PCS de muestra

Región de AWS nombre	Región de AWS	Ver fuente	Pila de lanzamiento
Este de EE. UU. (Norte de Virginia)	us-east-1	Descarga YAML	
Este de EE. UU. (Ohio)	us-east-2	Descarga YAML	
Oeste de EE. UU. (Oregón)	us-west-2	Descarga YAML	
Asia-Pacífico (Singapur)	ap-southeast-1	Descarga YAML	

Región de AWS nombre	Región de AWS	Ver fuente	Pila de lanzamiento
Asia-Pacífico (Sídney)	ap-southeast-2	Descarga YAML	
Asia-Pacífico (Tokio)	ap-northeast-1	Descarga YAML	
Europa (Fráncfort)	eu-central-1	Descarga YAML	
Europa (Irlanda)	eu-west-1	Descarga YAML	
Europa (Londres)	eu-west-2	Descarga YAML	
Europa (Estocolmo)	eu-north-1	Descarga YAML	
AWS GovCloud (Este de EE. UU.)	us-gov-east-1	Descarga YAML	
AWS GovCloud (US-Oeste)	us-gov-west-1	Descarga YAML	

AWS Clústeres PCS

Un clúster AWS PCS consta de los siguientes componentes:

- Instancias gestionadas del software programador del sistema HPC, como el daemon de control Slurm (). `slurmctld`
- Componentes que se integran con el programador del sistema HPC para aprovisionar y gestionar las instancias de Amazon EC2.
- Componentes que se integran con el programador del sistema HPC para transmitir registros y métricas a Amazon. CloudWatch

Estos componentes se ejecutan en una cuenta gestionada por. AWS Trabajan juntos para gestionar las EC2 instancias de Amazon en tu cuenta de cliente. AWS PCS proporciona interfaces de red elásticas en su subred de Amazon VPC para proporcionar conectividad desde el software del programador a las EC2 instancias de Amazon (por ejemplo, para permitir la programación de trabajos por lotes en ellas y permitir a los usuarios ejecutar comandos del programador para enumerar y administrar esos trabajos).

Temas

Creación de un clúster en AWS PCS

En este tema se proporciona una descripción general de las opciones disponibles y se describe lo que se debe tener en cuenta al crear un clúster en AWS Parallel Computing Service (AWS PCS). Si es la primera vez que crea un clúster de AWS PCS, le recomendamos que haga lo siguienteComience con AWS Parallel Computing Service. El tutorial puede ayudarle a crear un sistema HPC que funcione sin tener que ampliar todas las opciones y arquitecturas de sistema disponibles.

> **Note**
>
> Tras crear un clúster, puede modificar muchos ajustes de configuración sin tener que reconstruir la infraestructura. Para obtener más información, consulte Actualización de un clúster en AWS PCS.

> **Note**
>
> Puede configurar los ajustes de Slurm personalizados para implementar políticas de programación avanzadas y administración de recursos. Para obtener más información, consulte Configuración de ajustes de Slurm personalizados en PCS AWS.

Requisitos previos

- Una VPC y una subred existentes que cumplen los requisitos. AWS Redes PCS Antes de implementar un clúster para su uso en producción, le recomendamos que conozca a fondo los requisitos de VPC y subred. Para crear una VPC y una subred, consulte. Creación de una VPC para su AWS clúster de PCS
- Un director de IAM con permisos para crear y administrar AWS los recursos de PCS. Para obtener más información, consulte Servicio de Gestión de Identidad y Acceso para Computación AWS Paralela.

Cree un clúster de AWS PCS

Puede usar Consola de administración de AWS o AWS CLI para crear un clúster.

Consola de administración de AWS

Para crear un clúster

1. Abra la consola AWS PCS en https://console.aws.amazon.com/pcs/home#/clusters y seleccione Crear clúster.
2. En la sección Configuración del clúster, introduzca los siguientes campos:

- Nombre del clúster: un nombre para el clúster. El nombre solo puede contener caracteres alfanuméricos (con distinción de mayúsculas y minúsculas) y guiones. Debe empezar por un carácter alfabético y no puede tener más de 40 caracteres. El nombre debe ser único dentro del grupo Región de AWS y en el Cuenta de AWS que se va a crear el clúster.
- Planificador: elige un programador y una versión. Para obtener más información, consulte Versiones de Slurm en PCS AWS.
- Tamaño del mando: elige un tamaño para el mando. Esto determina cuántos trabajos y nodos de cómputo simultáneos puede administrar el clúster de AWS PCS. Solo puede establecer el tamaño de la controladora cuando se crea el clúster. Para obtener más información sobre el tamaño, consulteTamaño del clúster en AWS PCS.

3. En la sección Redes, seleccione valores para los siguientes campos:

- Tipo de red: elija el tipo de dirección IP del clúster. El clúster puede usar una IPv4 o ambas IPv6, pero no ambas. La VPC y las subredes deben usar el mismo tipo de dirección de red. El bloque de direcciones IP que utilice para cada subred debe tener al menos 1 dirección disponible. AWS reserva algunas de las direcciones de cada subred. Para obtener más información, consulte bloques CIDR de subred en la Guía del usuario de Amazon VPC.
- VPC: elija una VPC existente que cumpla con los requisitos de PCS. AWS Para obtener más información, consulte AWS Requisitos y consideraciones sobre la VPC y la subred. Después de crear el clúster, no puede cambiar su VPC. VPCs Si no aparece ninguna, primero debe crear una.
- Subred: se muestran todas las subredes disponibles en la VPC seleccionada. Elija una subred que cumpla con los requisitos de subred del PCS AWS . Para obtener más información, consulte AWS Requisitos y consideraciones sobre la VPC y la subred. Le recomendamos que seleccione una subred privada para evitar exponer los puntos finales de su programador a la Internet pública.
- Grupos de seguridad: especifique los grupos de seguridad que desea que AWS PCS asocie a las interfaces de red que crea para su clúster. Debe seleccionar al menos un grupo de seguridad que permita la comunicación entre el clúster y sus nodos de procesamiento. Puede seleccionar Crear rápidamente un grupo de seguridad para que AWS PCS cree uno con la configuración necesaria en la VPC seleccionada o seleccionar un grupo de seguridad existente. Para obtener más información, consulte Requisitos y consideraciones sobre los grupos de seguridad.

4. (Opcional) En la sección de configuración de la contabilidad de Slurm, puede activar la contabilidad de Slurm y establecer los parámetros de la contabilidad. Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.
5. (Opcional) En la sección de configuración de Slurm, puede añadir pares de nombre y valor de los parámetros para configurar otros ajustes de Slurm. Para obtener una lista completa de los parámetros admitidos, consulte. Configuración de Slurm personalizada para AWS clústeres de PCS
6. (Opcional) En Etiquetas, añada cualquier etiqueta al clúster de AWS PCS.
7. Elija Create cluster. El campo Estado se muestra `Creating` mientras el AWS PCS crea el clúster. Este proceso puede tardar varios minutos.

> **Important**
>
> Solo puede haber 1 clúster en un `Creating` estado Región de AWS por cada uno Cuenta de AWS. AWS Cuando intenta crear un clúster, PCS devuelve un error si ya existe un clúster en ese `Creating` estado.

AWS CLI

Para crear un clúster

1. Creación del clúster con el siguiente comando. Antes de ejecutar el comando, realice los siguientes reemplazos:
   - Sustitúyalo por el ID *`region`* con el Región de AWS que desea crear el clúster, por ejemplo`us-east-1`.
   - Reemplace *`my-cluster`* por el nombre del clúster. El nombre solo puede contener caracteres alfanuméricos (con distinción de mayúsculas y minúsculas) y guiones. Debe empezar por un carácter alfabético y no puede tener más de 40 caracteres. El nombre debe ser único dentro del clúster Región de AWS y en el Cuenta de AWS lugar en el que vaya a crearlo.
   - *`25.05`*Sustitúyalo por cualquier versión compatible de Slurm.

> **Note**
>
> AWS Actualmente, PCS es compatible con Slurm 25.05 y 24.11.

- Sustitúyalo por cualquier *SMALL* tamaño de clúster compatible. Esto determina cuántos trabajos y nodos de cómputo simultáneos puede administrar el clúster de AWS PCS. Solo se puede configurar cuando se crea el clúster. Para obtener más información sobre el tamaño, consulteTamaño del clúster en AWS PCS.
- Sustituya el valor de por `subnetIds` el suyo propio. Le recomendamos que seleccione una subred privada para evitar exponer los puntos finales de su programador a la Internet pública.
- Especifique lo `securityGroupIds` que desea que el AWS PCS asocie a las interfaces de red que crea para el clúster. Los grupos de seguridad deben estar en la misma VPC que el clúster. Debe seleccionar al menos un grupo de seguridad que permita la comunicación entre el clúster y sus nodos de procesamiento. Para obtener más información, consulte Requisitos y consideraciones sobre los grupos de seguridad.

```
aws pcs create-cluster --region region \
    --cluster-name my-cluster \
    --scheduler type=SLURM,version=25.05 \
    --size SMALL \
    --networking subnetIds=subnet-ExampleId1,securityGroupIds=sg-ExampleId1
```

- para usarlo IPv6, `networkType=IPV6` agréguelo a la `--networking` configuración.

```
  --networking networkType=IPV6,subnetIds=subnet-
ExampleId1,securityGroupIds=sg-ExampleId1
```

- Si lo desea, puede añadir la `--slurm-configration` opción de personalizar el comportamiento de Slurm y especificar las opciones de configuración de Slurm. El siguiente ejemplo establece el tiempo de inactividad reducido en 60 minutos (3600 segundos), habilita la contabilidad de Slurm y especifica la configuración como valor para. `slurm.conf` `slurmCustomSettings` Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.

> **Note**
>
> Se admite la contabilidad en Slurm 24.11 o versiones posteriores.

```
aws pcs create-cluster --region region \
    --cluster-name my-cluster \
    --scheduler type=SLURM,version=25.05 \
    --size SMALL \
    --networking subnetIds=subnet-ExampleId1,securityGroupIds=sg-ExampleId1
    --slurm-configuration
 scaleDownIdleTimeInSeconds=3600,accounting='{mode=STANDARD}',slurmCustomSettings='[{p
```

2. El aprovisionamiento del clúster puede tardar varios minutos. Puede consultar el estado del clúster con el siguiente comando. No proceda a crear colas ni a calcular grupos de nodos hasta que aparezca el campo de estado del clúster. `ACTIVE`

```
aws pcs get-cluster --region region --cluster-identifier my-cluster
```

> **Important**
>
> Solo puede haber 1 clúster en un `Creating` estado Región de AWS por cada uno. Cuenta de AWS AWS Cuando intenta crear un clúster, PCS devuelve un error si ya existe un clúster en ese `Creating` estado.

Próximos pasos recomendados para su clúster

- Agregue grupos de nodos de cómputo.
- Añada colas.
- Habilitar el registro.

Actualización de un clúster en AWS PCS

AWS El PCS permite actualizar las configuraciones de los clústeres tras su creación a través de la UpdateCluster API o la consola. Puede modificar la configuración del clúster sin tener que reconstruir la infraestructura, lo que reduce la sobrecarga operativa y minimiza las interrupciones.

Ventajas de las actualizaciones de clústeres

La actualización de los clústeres de AWS PCS le permite adaptar la infraestructura de HPC a los nuevos requisitos sin interrumpir el servicio. Los cambios de configuración tardan unos minutos en lugar de la hora o más necesaria para reconstruir los clústeres. Esta capacidad es importante para los entornos de producción que requieren un tiempo de inactividad mínimo y para los equipos que necesitan ajustar la configuración del clúster a medida que cambian los patrones de carga de trabajo.

Cambios de configuración compatibles

Puede modificar tres categorías principales de ajustes:

- Configuración de la contabilidad: active o desactive la contabilidad gestionada y configure los ajustes de retención.
- Comportamiento de reducción: ajuste el `scaleDownIdleTime` parámetro, que controla cuánto tiempo permanecen inactivas las instancias dinámicas antes de que AWS PCS las termine automáticamente.
- Configuración personalizada de Slurm: modifique cualquiera de las configuraciones de Slurm compatibles que se apliquen a nivel de clúster, incluidas Prolog, Epilog y. SelectTypeParameters

Limitaciones

No puede modificar determinadas configuraciones después de la creación del clúster. Entre ellos se incluyen:

- Configuraciones de grupos de seguridad
- Selección de subredes de VPC
- Tamaño del clúster
- Versión Slurm
- Cluster name (Nombre del clúster)

Estos ajustes son fundamentales para la arquitectura del clúster y requieren la creación de un clúster nuevo para modificarlos.

Requisitos previos para las actualizaciones del clúster

Antes de actualizar un clúster, asegúrese de que se cumplan las siguientes condiciones:

- El clúster debe estar en `ACTIVE`, `UPDATE_FAILED`, o `SUSPENDED` estado
- Todos los recursos asociados (colas, grupos de nodos de cómputo) deben estar en estado `ACTIVE`
- Debe tener los permisos de IAM adecuados para la operación UpdateCluster
- No se puede realizar ninguna otra operación de actualización

El impacto en el proceso de actualización y en el trabajo

Durante una operación de actualización, los nodos de cómputo siguen ejecutando las tareas existentes incluso cuando no se puede acceder al controlador de clúster por un momento. Sin embargo, el sistema no puede aceptar nuevas solicitudes de trabajo ni tomar decisiones de programación durante este período.

Puede supervisar las actualizaciones del clúster a través de la consola y de la interfaz API. El clúster pasará por los siguientes estados durante una actualización:

- `UPDATING`- Actualización en curso
- `ACTIVE`- La actualización se ha completado correctamente
- `UPDATE_FAILED`- Se ha detectado un error en la actualización

Facturación durante las actualizaciones

Los cargos por hora estándar para su clúster de AWS PCS continúan durante las operaciones de actualización. Al actualizar un clúster para deshabilitar la contabilidad, la facturación de la función de contabilidad se detiene en cuanto el clúster entra en `UPDATING` estado. Al habilitar la contabilidad, la facturación no comienza hasta que el clúster completa correctamente la actualización y vuelve al `ACTIVE` estado.

Temas

- Actualizar un clúster de AWS PCS

- Preguntas frecuentes sobre la actualización de clústeres en AWS PCS
- Solución de problemas de actualizaciones del clúster AWS PCS

Actualizar un clúster de AWS PCS

Siga estos pasos para modificar la configuración del programador, la configuración de la contabilidad y la configuración personalizada de Slurm en su clúster. Para obtener más información, consulte Configuración de Slurm personalizada para AWS clústeres de PCS.

Requisitos previos

- El clúster debe estar en`ACTIVE`, `UPDATE_FAILED` o estado `SUSPENDED`
- Todos los recursos asociados (colas, grupos de nodos de cómputo) deben estar en estado `ACTIVE`
- No se puede realizar ninguna otra operación de actualización

Procedimiento

Consola de administración de AWS

1. Abra la consola AWS PCS en https://console.aws.amazon.com/pcs/.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Seleccione el clúster que desee actualizar.
4. Seleccione Editar.
5. En la página Editar clúster, modifique la configuración deseada:
    - En la configuración del programador, actualice el tiempo de inactividad de Scale-down para controlar cuánto tiempo permanecen inactivas las instancias dinámicas antes de la finalización automática.
    - Modifique la configuración de los parámetros de Prolog, Epilog y Select-type según sea necesario.
    - Habilite, deshabilite o configure el tiempo de retención para la contabilidad gestionada.
    - En Configuración adicional del programador, agrega, edita o elimina la configuración personalizada de Slurm. Para obtener más información sobre los parámetros compatibles, consulte. Configuración de Slurm personalizada para AWS clústeres de PCS

> **Note**
>
> Los campos que no se pueden editar se muestran como de solo lectura y muestran sus valores actuales.

6. Elija Actualizar para enviar los cambios.
7. Supervise el estado del clúster, que aparece como «Actualizando» durante el proceso. El estado cambia cuando la actualización se completa correctamente.

AWS CLI

1. Abre un terminal o una línea de comandos.
2. Compruebe el estado del clúster mediante el siguiente comando:

```
aws pcs get-cluster --cluster-identifier my-cluster
```

3. Envíe una solicitud de actualización mediante uno de los siguientes ejemplos:

   - Para habilitar la contabilidad gestionada:

   ```
   aws pcs update-cluster --cluster-identifier my-cluster \
   --slurm-configuration 'accounting={mode=STANDARD}'
   ```

   - Para actualizar la configuración de Slurm Prolog:

   ```
   aws pcs update-cluster --cluster-identifier my-cluster \
   --slurm-configuration \
   'SlurmCustomSettings=[{parameterName=Prolog,parameterValue="/path/to/
   prolog.sh"}]'
   ```

   - Para actualizar el tiempo de inactividad reducido:

   ```
   aws pcs update-cluster --cluster-identifier my-cluster \
   --slurm-configuration 'scaleDownIdleTimeInSeconds=300'
   ```

4. Supervise el progreso de la actualización comprobando el estado del clúster:

```
aws pcs get-cluster --cluster-identifier my-cluster
```

Tras una solicitud de actualización correcta, el comando devuelve el objeto de clúster con todos los cambios. El estado del clúster cambia de `UPDATING` a `ACTIVE` cuando está completo.

Preguntas frecuentes sobre la actualización de clústeres en AWS PCS

Obtenga respuestas a preguntas frecuentes sobre la actualización de las configuraciones de los clústeres en AWS PCS.

¿Qué ajustes puedo modificar?

Puede modificar la configuración contable (activar/desactivar la contabilidad gestionada), el comportamiento de reducción (parámetro scaleDownIdle Time) y cualquiera de los ajustes personalizados de Slurm compatibles que se apliquen a nivel de clúster. No puede modificar los grupos de seguridad, las subredes de VPC, el tamaño del clúster, la versión de Slurm ni el nombre del clúster.

¿Puedo poner en cola varias actualizaciones?

¿No?. Debe esperar a que el clúster vuelva a su `ACTIVE` estado antes de enviar otra actualización. Todos los recursos asociados (colas, grupos de nodos de cómputo) también deben estar en `ACTIVE` estado.

¿Puedo cancelar una operación de actualización de un clúster?

No, no puede cancelar una operación de actualización de clústeres en curso.

¿Puedo enviar trabajos mientras mi clúster se actualiza?

Le recomendamos que evite enviar trabajos durante las actualizaciones del clúster. Es posible que el controlador Slurm no esté disponible durante el proceso de actualización.

¿Seguirán ejecutándose mis trabajos durante las actualizaciones del clúster?

Sí, los trabajos en ejecución siguen ejecutándose en los nodos de procesamiento incluso cuando el controlador de clúster deja de estar accesible por un momento durante el proceso de actualización. Sin embargo, es posible que el estado del trabajo no se actualice hasta que el controlador vuelva a estar disponible.

¿Cómo se ve afectada la facturación durante las actualizaciones?

Los cargos por hora estándar continúan durante las operaciones de actualización. Al deshabilitar la contabilidad, la facturación se detiene cuando el clúster entra en `UPDATING` estado. Al habilitar la contabilidad, la facturación comienza cuando el clúster vuelve a su `ACTIVE` estado correcto.

Solución de problemas de actualizaciones del clúster AWS PCS

Este tema le ayuda a identificar y resolver los problemas más comunes que se pueden producir al actualizar las configuraciones del clúster.

La actualización falla debido a un error de configuración de la contabilidad

Causa habitual

El clúster entra en `UPDATE_FAILED` estado y el mensaje de error indica un problema de configuración contable. Esto suele ocurrir cuando la configuración contable no es compatible con la versión actual de Slurm o contiene una configuración no válida.

Resolución

Revise la configuración de contabilidad para comprobar si es compatible con la versión de Slurm del clúster y envíe una solicitud de actualización corregida con los parámetros de configuración válidos.

La actualización falla debido a un error de configuración personalizada

Causa habitual

El clúster entra en `UPDATE_FAILED` estado y el mensaje de error indica un problema con la configuración personalizada de Slurm. Esto ocurre cuando se proporcionan valores de parámetros de Slurm no válidos o combinaciones de parámetros no compatibles.

Resolución

Valide su configuración personalizada de Slurm con los parámetros admitidos y envíe una solicitud de actualización corregida con valores y combinaciones de parámetros válidos.

No se puede enviar la solicitud de actualización

Causa habitual

El botón de actualización está deshabilitado en la consola o la API devuelve un error de nivel 400. Esto ocurre cuando el clúster no está en un estado adecuado, los recursos asociados no están activos o hay errores de validación en la configuración.

Resolución

Espere a que el clúster y todos los recursos asociados alcancen el `ACTIVE` estado y, a continuación, revise la configuración para ver si hay errores de validación antes de volver a enviar la solicitud de actualización.

Errores de validación

Causa habitual

El comando vuelve inmediatamente con un error HTTP de nivel 400 y un mensaje descriptivo. Esto se debe a que el estado del clúster, el estado de los recursos o los parámetros de configuración no son válidos.

Resolución

Corrija el error de validación específico mencionado en la respuesta y vuelva a intentar la operación de actualización.

Eliminar un clúster en AWS PCS

En este tema se proporciona información general sobre cómo eliminar un clúster de AWS PCS.

Consideraciones a la hora de eliminar un clúster de AWS PCS

- Se deben eliminar todas las colas asociadas al clúster antes de poder eliminar el clúster. Para obtener más información, consulte Eliminar una cola en PCS AWS.
- Todos los grupos de nodos de procesamiento asociados al clúster deben eliminarse antes de poder eliminar el clúster. Para obtener más información, consulte Eliminar un grupo de nodos de cómputo en AWS PCS.

Elimine el clúster

Puede usar Consola de administración de AWS o AWS CLI para eliminar un clúster.

Consola de administración de AWS

Para eliminar un clúster

1. Abra la consola AWS PCS.

2. Seleccione el clúster que desee eliminar.
3. Elija Eliminar.
4. Aparece el campo Estado del clúster`Deleting`. Puede tardar varios minutos en completarse.

AWS CLI

Para eliminar un clúster

1. Use el siguiente comando para eliminar un clúster, con estas sustituciones:
   - *region-code*Reemplácelo por el Región de AWS que contiene su clúster.
   - *my-cluster*Sustitúyalo por el nombre o el ID de tu clúster.

```
aws pcs delete-cluster --region region-code --cluster-identifier my-cluster
```

2. Eliminar el clúster puede tardar varios minutos. Puede comprobar el estado del clúster con el siguiente comando.

```
aws pcs get-cluster --region region-code --cluster-identifier my-cluster
```

Tamaño del clúster en AWS PCS

AWS El PCS proporciona clústeres seguros y de alta disponibilidad, a la vez que automatiza tareas clave como la aplicación de parches, el aprovisionamiento de nodos y las actualizaciones.

Al crear un clúster, se selecciona su tamaño en función de dos factores:

- La cantidad de nodos de cómputo que administrará
- La cantidad de trabajos activos y en cola que espera ejecutar en el clúster

> **Important**
>
> No puede cambiar el tamaño del clúster después de crearlo. Si necesita cambiar el tamaño, debe crear un clúster nuevo.

Tamaño del clúster de Slurm	Número de instancias administradas	Número de trabajos activos y en cola
Small	Hasta 32	Hasta 256
Medio	Hasta 512	Hasta 8192
Grande	Hasta 2048	Hasta 16384

Ejemplos

- Si su clúster tendrá hasta 24 instancias administradas y ejecutará hasta 100 trabajos, elija Small.
- Si el clúster tendrá hasta 24 instancias administradas y ejecutará hasta 1000 trabajos, elija Medium.
- Si el clúster tendrá hasta 1000 instancias administradas y ejecutará hasta 100 trabajos, elija Grande.
- Si el clúster tendrá hasta 1000 instancias administradas y ejecutará hasta 10 000 trabajos, elija Grande.

Trabajar con secretos de clústeres en AWS PCS

Como parte de la creación de un clúster, AWS PCS crea un secreto de clúster que es necesario para conectarse al programador de tareas del clúster. También se crean grupos de nodos de cómputo de AWS PCS, que definen conjuntos de instancias que se van a lanzar en respuesta a eventos de escalado. AWS PCS configura las instancias lanzadas por esos grupos de nodos de cómputo con el secreto del clúster para que puedan conectarse al programador de tareas. Hay casos en los que es posible que desee configurar los clientes de Slurm manualmente. Los ejemplos incluyen la creación de un nodo de inicio de sesión persistente o la configuración de un administrador de flujo de trabajo con capacidades de administración de trabajos.

AWS PCS almacena el secreto del clúster como un secreto gestionado con el prefijo puestopcs!. AWS Secrets Manager El coste del secreto está incluido en el coste del uso del AWS PCS. Puede filtrar los secretos de los clústeres AWS Secrets Manager para mantener el cumplimiento de las normas de seguridad y corregir posibles problemas de seguridad.

Temas

Úselo AWS Secrets Manager para encontrar el secreto del clúster

Consola de administración de AWS

1. Vaya a la consola del administrador de secretos.
2. Selecciona Secretos y busca el `pcs!` prefijo.

> **Note**
>
> El secreto de un clúster de AWS PCS tiene un nombre en el formato `pcs!slurm-secret-`*`cluster-id`* donde *`cluster-id`* está el ID del clúster de AWS PCS.

AWS CLI

Cada secreto de clúster de AWS PCS también está etiquetado con`aws:pcs:`*`cluster-id`*. Puede obtener el identificador secreto de un clúster con el siguiente comando. Realice estas sustituciones antes de ejecutar el comando:

- *`region`* Región de AWS Sustitúyalo por el para crear el clúster, por ejemplo`us-east-1`.
- *`cluster-id`*Sustitúyalo por el ID del clúster de AWS PCS para encontrar el secreto del clúster.

```
aws secretsmanager list-secrets \
    --region region  \
    --filters Key=tag-key,Values=aws:pcs:cluster-id \
            Key=tag-value,Values=cluster-id
```

Utilice AWS PCS para encontrar el secreto del clúster

Puede utilizar el AWS CLI para encontrar el ARN de un secreto de clúster de AWS PCS. Introduzca el comando siguiente y realice las siguientes sustituciones:

- *region* Región de AWS Sustitúyalo por el para crear el clúster, por ejemplous-east-1.
- *my-cluster*Sustitúyalo por el nombre o identificador del clúster.

```
aws pcs get-cluster --region region --cluster-identifier my-cluster
```

El siguiente ejemplo de salida proviene del `get-cluster` comando. Puedes usar `secretArn` y `secretVersion` juntos para obtener el secreto.

```
{
    "cluster": {
        "name": "get-started",
        "id": "pcs_123456abcd",
        "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_123456abcd",
        "status": "ACTIVE",
        "createdAt": "2024-12-17T21:03:52+00:00",
        "modifiedAt": "2024-12-17T21:03:52+00:00",
        "scheduler": {
            "type": "SLURM",
            "version": "25.05"
        },
        "size": "SMALL",
        "slurmConfiguration": {
            "authKey": {
                "secretArn": "arn:aws:secretsmanager:us-east-1:111122223333:secret:pcs!
slurm-secret-pcs_123456abcd-a12ABC",
                "secretVersion": "ef232370-d3e7-434c-9a87-ec35c1987f75"
            }
        },
        "networking": {
            "subnetIds": [
                "subnet-0123456789abcdef0"
            ],
            "securityGroupIds": [
                "sg-0123456789abcdef0"
            ]
        },
```

```
        "endpoints": [
            {
                "type": "SLURMCTLD",
                "privateIpAddress": "10.3.149.220",
                "port": "6817"
            }
        ]
    }
}
```

Obtén el secreto del cúmulo de Slurm

Puede usar Secrets Manager para obtener la versión actual codificada en base64 de un secreto de clúster de Slurm. El siguiente ejemplo usa el. AWS CLI Realice las siguientes sustituciones antes de ejecutar el comando.

- *`region`* Región de AWS Sustitúyala por la para crear el clúster, por ejemplo`us-east-1`.
- *`secret-arn`*Sustitúyalo `secretArn` por el de un clúster de AWS PCS.

```
aws secretsmanager get-secret-value \
    --region region \
    --secret-id 'secret-arn' \
    --version-stage AWSCURRENT \
    --query 'SecretString' \
    --output text
```

Para obtener información sobre cómo utilizar el secreto del clúster de Slurm, consulte. Uso de instancias independientes como nodos de inicio de sesión de AWS PCS

Permisos

Utiliza un principal de IAM para obtener el secreto del clúster de Slurm. El director de IAM debe tener permiso para leer el secreto. Para obtener más información, consulte los términos y conceptos de las funciones en la Guía del AWS Identity and Access Management usuario.

El siguiente ejemplo de política de IAM permite el acceso a un ejemplo de secreto de clúster.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
```

```
            "Sid": "AllowSecretValueRetrievalAndVersionListing",
            "Effect": "Allow",
            "Action": [
                "secretsmanager:GetSecretValue",
                "secretsmanager:ListSecretVersionIds"
            ],
            "Resource": "arn:aws:secretsmanager:us-east-1:012345678901:secret:pcs!
slurm-secret-s3431v9rx2-FN7tJF"
        }
    ]
}
```

Secretos de clústeres rotativos en AWS PCS

Utilice la rotación AWS Secrets Manager gestionada para rotar los secretos del clúster en AWS PCS. La rotación periódica de los secretos es una buena práctica de seguridad para mantener una sólida postura de seguridad en los entornos de HPC. Esta capacidad le permite cumplir con los estándares de cumplimiento de la industria, incluidos la HIPAA y el FedRAMP, que exigen la rotación regular de credenciales.

El secreto del clúster tiene un doble propósito: autenticar los nodos de cómputo que se unen al clúster y servir de clave JWT para la autenticación de la API REST de Slurm. Cuando se rota, ambos aspectos se ven afectados simultáneamente.

Cómo funciona la rotación secreta de los clústeres

Prepárese manualmente para mantener la estabilidad del clúster durante la rotación secreta:

1. Preparación: escale todos los grupos de nodos de cómputo a una capacidad del 0% y asegúrese de que no se esté ejecutando ningún trabajo
2. Rotación: inicie la rotación a través de la consola o la API de Secrets Manager
3. Supervisión: realice un seguimiento del progreso a lo largo de CloudTrail los eventos
4. Recuperación: escale los grupos de nodos de cómputo hasta alcanzar la capacidad deseada

Durante la rotación, el clúster permanece en ese `ACTIVE` estado y la facturación continúa con normalidad. El proceso suele tardar unos minutos.

Requisitos y limitaciones

Antes de rotar los secretos del clúster, complete estos requisitos:

- El clúster debe estar en `ACTIVE` nuestro `UPDATE_FAILED` estado
- El rol de IAM debe tener permiso `secretsmanager:RotateSecret`
- Todos los grupos de nodos de cómputo deben escalarse a una capacidad igual a 0
- Detenga todos los trabajos antes de la rotación

Limitaciones:

- Se requiere una preparación manual para cada rotación
- Los tokens JWT existentes dejan de ser válidos y deben volver a emitirse
- Los nodos de inicio de sesión BYO requieren una actualización manual del secreto después de la rotación

Temas

Rota el secreto de un clúster en AWS PCS

Cambie el secreto de su clúster para cumplir con los requisitos de seguridad y hacer frente a posibles riesgos. Este proceso requiere poner el clúster en modo de mantenimiento.

Requisitos previos

- Función de IAM con permiso `secretsmanager:RotateSecret`
- Clúster en `ACTIVE` nuestro estado `UPDATE_FAILED`

Procedimiento

1. Notifique a los usuarios del clúster sobre el próximo período de mantenimiento.
2. Coloque el clúster en modo de mantenimiento escalando todos los grupos de nodos de cómputo a una capacidad cero.

    a. Utilice la UpdateComputeNodeGroup API para establecer ambos minInstanceCount valores en 0 maxInstanceCount para todos los grupos de nodos de cómputo.

b. Espere a que se detengan todos los nodos.

c. Opcional: vacíe las colas del programador con los comandos de Slurm antes de cerrar la capacidad para gestionar los trabajos sin problemas.

3. Inicie la rotación a través de Secrets Manager.

- Método de consola:
  - Ve a Secrets Manager, selecciona el secreto de tu clúster y elige Rotar secreto.
- Método de API:
  - Usa la `rotate-secret` API Secrets Manager.

4. Supervise el progreso de la rotación.

a. Realice un seguimiento del progreso a través de CloudTrail los eventos.

b. `lastRotatedDate`Compruébelo en la consola de Secrets Manager o en la `secretsmanager:describeSecret` API.

c. Espere a `RotationSucceeded` nuestro `RotationFailed` CloudTrail evento.

5. Tras una rotación correcta, restaure la capacidad del clúster.

a. Usa la UpdateComputeNodeGroup API para restablecer los grupos de nodos a la min/max capacidad deseada.

b. Para los nodos de inicio de sesión AWS gestionados por PCS: no es necesario realizar ninguna acción adicional.

c. Para los nodos de inicio de sesión BYO:

i. Conéctese a los nodos de inicio de sesión.

ii. Actualiza `/etc/slurm/slurm.key` con el nuevo secreto de Secrets Manager.

iii. Reinicia el daemon Slurm Auth and Cred Kiosk (sackd).

Preguntas frecuentes sobre la rotación secreta de clústeres en AWS PCS

Encuentre respuestas a preguntas frecuentes sobre la rotación de secretos de clústeres en AWS PCS.

¿Qué es un secreto de clúster?

Un secreto de clúster es una credencial segura que permite una comunicación segura entre el controlador Slurm y los nodos de cómputo del AWS PCS. También sirve como clave del token web JSON (JWT) para la autenticación de la API REST de Slurm.

¿Cuál es la diferencia entre el secreto del clúster y la clave JWT?

En AWS PCS, el secreto del clúster y la clave JWT son el mismo recurso y tienen diferentes propósitos. El secreto del clúster autentica las comunicaciones internas de Slurm, mientras que la clave JWT firma los tokens para la autenticación de la API REST. Cuando se gira, ambos aspectos se ven afectados simultáneamente.

¿Cuánto dura la rotación?

El proceso de rotación suele tardar unos minutos. El clúster permanece en estado ACTIVO y la facturación continúa con normalidad durante la rotación.

¿Puedo programar rotaciones automáticas?

Puede activar la rotación programada en Secrets Manager. Sin embargo, la versión inicial requiere una preparación manual (escalando los grupos de nodos a 0) antes de cada rotación.

¿Seguirán funcionando mis fichas JWT actuales después de la rotación?

No, los tokens JWT existentes dejan de ser válidos después de la rotación. Emita nuevos tokens para los clientes de la API REST.

¿Dónde puedo encontrar el secreto de mi clúster?

Puede encontrar el secreto de su clúster en la consola Secrets Manager o en la consola AWS PCS. Para obtener instrucciones detalladas, consulte Úselo AWS Secrets Manager para encontrar el secreto del clúster yUtilice AWS PCS para encontrar el secreto del clúster.

¿Por qué la rotación requiere escalar los grupos de nodos a 0?

La rotación no requiere que se ejecuten instancias para garantizar la estabilidad del clúster durante el proceso de actualización secreta. Esto evita conflictos de autenticación entre los secretos antiguos y los nuevos.

¿Qué requisitos de conformidad admite esta función?

Esta función permite a AWS PCS cumplir con los estándares de cumplimiento de la industria, incluidos la HIPAA y FedRAMP, que exigen la rotación regular de credenciales como parte de sus controles de seguridad.

Solución de problemas de rotación de secretos de clústeres en AWS PCS

La rotación del secreto del clúster falla si el entorno no está preparado adecuadamente. La causa más común son las instancias activas del clúster. Para evitar errores:

1. Establezca todos los grupos de nodos en una capacidad igual a 0.
2. Espere a que los nodos se detengan.
3. Compruebe que el clúster no esté en los siguientes estados: `CREATE_FAILED`,`DELETE_FAILED`,`RESUMING`,`SUSPENDING`, o`SUSPENDED`.

Si la rotación falla:

- Aparece un RotationFailed CloudTrail evento
- El secreto del clúster permanece inalterado
- Consulta el RotationFailed evento CloudTrail para obtener más información
- Complete todos los pasos de preparación para una rotación exitosa

AWS Grupos de nodos de cómputo PCS

Un grupo de nodos de cómputo de AWS PCS es un conjunto lógico de nodos (EC2 instancias de Amazon). Estos nodos se pueden usar para ejecutar tareas informáticas, así como para proporcionar acceso interactivo y basado en shell a un sistema HPC. Un grupo de nodos de cómputo consta de reglas para crear nodos, que incluyen qué tipos de instancias de Amazon EC2 usar, cuántas instancias ejecutar, si usar instancias puntuales o instancias bajo demanda, qué subredes y grupos de seguridad usar y cómo configurar cada instancia cuando se lanza. Cuando se actualizan esas reglas, AWS PCS actualiza los recursos asociados al grupo de nodos de cómputo para que coincidan.

Temas

Creación de un grupo de nodos de cómputo en AWS PCS

En este tema se proporciona una descripción general de las opciones disponibles y se describe lo que se debe tener en cuenta al crear un grupo de nodos de procesamiento en AWS Parallel Computing Service (AWS PCS). Si es la primera vez que crea un grupo de nodos de cómputo en AWS PCS, le recomendamos que siga el tutorial que aparece enComience con AWS Parallel Computing Service. El tutorial puede ayudarle a crear un sistema HPC que funcione sin tener que ampliar todas las opciones y arquitecturas de sistema disponibles.

> **Note**
>
> Puede configurar los ajustes de Slurm personalizados en los grupos de nodos de cómputo para controlar la utilización de los recursos y los comportamientos a nivel de nodo. Para obtener más información, consulte Configuración de ajustes de Slurm personalizados en PCS AWS.

> **Important**
>
> AWS Actualmente, el PCS requiere un núcleo IPv4 compatible con la comunicación entre nodos locales, incluso cuando se utiliza el AWS PCS en una red exclusiva. IPv6 Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Requisitos previos

- Cuotas de servicio suficientes para lanzar el número deseado de instancias de EC2 en su. Región de AWS Puede utilizarlas Consola de administración de AWSpara comprobar y solicitar aumentos en sus cuotas de servicio.
- Una VPC y subredes existentes que cumplen con los requisitos de red de AWS PCS. Le recomendamos que comprenda detenidamente estos requisitos antes de implementar un clúster para su uso en producción. Para obtener más información, consulte AWS Requisitos y consideraciones sobre la VPC y la subred. También puede usar una CloudFormation plantilla para crear una VPC y subredes. AWS proporciona una receta de HPC para la plantilla. CloudFormation Para obtener más información, consulte aws-hpc-recipesen GitHub.
- Un perfil de instancia de IAM con permisos para activar la acción de la `RegisterComputeNodeGroupInstance` API de AWS PCS y acceder a cualquier otro AWS recurso necesario para las instancias de su grupo de nodos. Para obtener más información, consulte Perfiles de instancia de IAM para AWS Parallel Computing Service.
- Una plantilla de lanzamiento para las instancias de tu grupo de nodos. Para obtener más información, consulte Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.
- Para crear un grupo de nodos de cómputo que utilice instancias puntuales de Amazon EC2, debe tener el rol vinculado al servicio AWSServiceRoleForEC2Spot en su. Cuenta de AWS Para obtener más información, consulte Función Amazon EC2 Spot para PCS AWS.

Cree un grupo de nodos de cómputo en PCS AWS

Puede crear un grupo de nodos de cómputo mediante el Consola de administración de AWS o el AWS CLI.

Consola de administración de AWS

Para crear su grupo de nodos de cómputo mediante la consola

1. Abra la consola AWS PCS.
2. Seleccione el clúster en el que desee crear un grupo de nodos de cómputo. Diríjase a los grupos de nodos de cómputo y elija Crear.
3. En la sección de configuración del grupo de nodos de Compute, proporciona un nombre para el grupo de nodos. El nombre solo puede contener caracteres alfanuméricos y guiones que distingan mayúsculas de minúsculas. Debe empezar por un carácter alfabético y no puede tener más de 25 caracteres. El nombre debe ser único en el clúster.
4. En Configuración informática, introduzca o seleccione estos valores:

   a. Plantilla de lanzamiento de EC2: seleccione una plantilla de lanzamiento personalizada para utilizarla en este grupo de nodos. Las plantillas de lanzamiento se pueden utilizar para personalizar la configuración de la red, como la subred y los grupos de seguridad, la configuración de supervisión y el almacenamiento a nivel de instancia. Si no tienes una plantilla de lanzamiento preparada, consulta Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS para aprender a crear una.

   > **Important**
   >
   > AWS PCS crea una plantilla de lanzamiento gestionada para cada grupo de nodos de cómputo. Estos se denominan`pcs-`*`identifier`*`-do-not-delete`. No los seleccione cuando cree o actualice un grupo de nodos de procesamiento, o el grupo de nodos no funcionará correctamente.

   b. Versión de la plantilla de lanzamiento de EC2: debe seleccionar una versión de la plantilla de lanzamiento personalizada. Si cambia la versión más adelante, debe actualizar el grupo de nodos de procesamiento para detectar cambios en la plantilla de lanzamiento. Para obtener más información, consulte Actualización de un grupo de nodos de cómputo AWS PCS.

   c. ID de AMI: si tu plantilla de lanzamiento no incluye un ID de AMI o si quieres anular el valor de la plantilla de lanzamiento, proporciona un ID de AMI aquí. Tenga en cuenta que la AMI utilizada para el grupo de nodos debe ser compatible con el AWS PCS. También puede seleccionar un ejemplo de AMI proporcionado por AWS. Para obtener

más información sobre este tema, consulteAmazon Machine Images (AMIs) para AWS PCS.

d. Perfil de instancia de IAM: elija un perfil de instancia para el grupo de nodos. Un perfil de instancia otorga a la instancia permisos para acceder a los AWS recursos y servicios de forma segura. Si no tiene uno preparado, puede seleccionar Crear un perfil básico para que AWS PCS cree uno automáticamente con la política mínima, o consultePerfiles de instancia de IAM para AWS Parallel Computing Service.

e. Subredes: elija una o más subredes en la VPC en la que se implementa el clúster de AWS PCS. Si seleccionas varias subredes, las comunicaciones EFA no estarán disponibles entre los nodos y es posible que la comunicación entre los nodos de distintas subredes tenga una mayor latencia. Asegúrese de que las subredes que especifique aquí coincidan con las que haya definido en la plantilla de lanzamiento de EC2.

f. Instancias: elija uno o más tipos de instancias para cumplir con las solicitudes de escalado del grupo de nodos. Todos los tipos de instancias deben tener la misma arquitectura de procesador (x86_64 o arm64) y el mismo número de v. CPUs Si las instancias lo tienen GPUs, todos los tipos de instancias deben tener el mismo número de. GPUs

g. Configuración de escalado: especifique la cantidad mínima y máxima de instancias para el grupo de nodos. Puede definir una configuración estática, en la que hay un número fijo de nodos en ejecución, o una configuración dinámica, en la que se pueden ejecutar hasta el número máximo de nodos. Para una configuración estática, defina el mínimo y el máximo en el mismo número, superior a cero. Para una configuración dinámica, establece el número mínimo de instancias en cero y el máximo en un número superior a cero. AWS PCS no admite grupos de nodos de cómputo con una combinación de instancias estáticas y dinámicas.

5. (Opcional) En Configuración adicional, especifique lo siguiente:

a. Opción de compra: seleccione instancias bajo demanda, instancias puntuales o un bloque de capacidad existente. Elija también bajo demanda si planea usar una reserva de capacidad bajo demanda (ODCR). Para obtener más información, consulte Uso ODCRs con AWS PCS. Elija Capacity Block para usar un Bloque de capacidad de Amazon EC2 existente para la reserva de aprendizaje automático. Para obtener más información, consulte Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS.

b. Estrategia de asignación: si ha seleccionado la opción de compra puntual, puede especificar cómo se eligen los grupos de capacidad puntual al lanzar instancias en el grupo de nodos. Para obtener más información, consulte Estrategias de asignación para instancias puntuales en la Guía del usuario de Amazon Elastic Compute Cloud. Esta opción no tiene efecto si ha seleccionado la opción de compra bajo demanda.

6. (Opcional) En la sección de ajustes Slurm personalizados, puede añadir pares de nombre y valor del parámetro para configurar ajustes adicionales de Slurm. Para obtener una lista completa de los parámetros admitidos, consulte. Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS
7. (Opcional) En Etiquetas, añada cualquier etiqueta a su grupo de nodos de cómputo.
8. Selecciona Crear grupo de nodos de cómputo. El campo Estado se muestra `Creating` mientras AWS PCS aprovisiona el grupo de nodos. Esto puede tardar varios minutos.

Siguiente paso recomendado

- Agregue su grupo de nodos a una cola en AWS PCS para que pueda procesar los trabajos.

AWS CLI

Para crear su grupo de nodos de cómputo mediante AWS CLI

Cree su cola con el siguiente comando. Antes de ejecutar el comando, realice los siguientes reemplazos:

1. *`region`*Sustitúyalo por el ID en el Región de AWS que se va a crear el clúster, por ejemplo. `us-east-1`
2. *`my-cluster`*Sustitúyalo por el nombre o `clusterId` el de tu clúster.
3. *`my-node-group`*Sustitúyalo por el nombre de tu grupo de nodos de procesamiento. El nombre solo puede contener caracteres alfanuméricos (con distinción de mayúsculas y minúsculas) y guiones. Debe empezar por un carácter alfabético y no puede tener más de 25 caracteres. El nombre debe ser único en el clúster.
4. *`subnet-ExampleID1`*Sustitúyala por una o más subredes IDs de la VPC del clúster.
5. *`lt-ExampleID1`*Sustitúyalo por el ID de la plantilla de lanzamiento personalizada. Si no tienes una preparada, consulta Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS para aprender a crearla.

> **Important**
>
> AWS PCS crea una plantilla de lanzamiento gestionado para cada grupo de nodos de cómputo. Estos se denominan`pcs-`*`identifier`*`-do-not-delete`. No los seleccione cuando cree o actualice un grupo de nodos de procesamiento, o el grupo de nodos no funcionará correctamente.

6. *`launch-template-version`*Sustitúyala por una versión de plantilla de lanzamiento específica. AWS PCS asocia su grupo de nodos a esa versión específica de la plantilla de lanzamiento.
7. *`arn:InstanceProfile`*Sustitúyalo por el ARN de tu perfil de instancia de IAM. Si no tiene uno preparado, consulte Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS para obtener orientación.
8. Sustituya *`min-instances`* y *`max-instances`* por valores enteros. Puede definir una configuración estática, en la que hay un número fijo de nodos en ejecución, o una configuración dinámica, en la que se puede ejecutar hasta el número máximo de nodos. Para una configuración estática, defina el mínimo y el máximo en el mismo número, superior a cero. Para una configuración dinámica, establece el número mínimo de instancias en cero y el máximo en un número superior a cero. AWS PCS no admite grupos de nodos de cómputo con una combinación de instancias estáticas y dinámicas.
9. *`t3.large`*Sustitúyala por otro tipo de instancia. Puede añadir más tipos de instancias especificando una lista de `instanceType` ajustes. Por ejemplo, *`--instance-configs instanceType=c6i.16xlarge instanceType=c6a.16xlarge`*. Todos los tipos de instancias deben tener la misma arquitectura de procesador (x86_64 o arm64) y el mismo número de v. CPUs Si las instancias lo tienen GPUs, todos los tipos de instancias deben tener el mismo número de. GPUs

```
aws pcs create-compute-node-group --region region \
    --cluster-identifier my-cluster \
    --compute-node-group-name my-node-group \
    --subnet-ids subnet-ExampleID1 \
    --custom-launch-template id=lt-ExampleID1,version='launch-template-version' \
    --iam-instance-profile-arn=arn:InstanceProfile \
    --scaling-config minInstanceCount=min-instances,maxInstanceCount=max-instance \
    --instance-configs instanceType=t3.large
```

Example— Crear un grupo de nodos de cómputo con una configuración de Slurm personalizada

```
aws pcs create-compute-node-group --region region \
    --cluster-identifier my-cluster \
    --compute-node-group-name my-node-group \
    --subnet-ids subnet-ExampleID1 \
    --custom-launch-template id=lt-ExampleID1,version='launch-template-version' \
    --iam-instance-profile-arn=arn:InstanceProfile \
    --scaling-config minInstanceCount=min-instances,maxInstanceCount=max-instance \
    --instance-configs instanceType=t3.large \
    --slurm-configuration \
    'slurmCustomSettings=[{parameterName=Features,parameterValue="gpu,nvme"}]'
```

Para obtener más información, consulte Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS.

Hay varios ajustes de configuración opcionales que puede añadir al `create-compute-node-group` comando.

- Puede especificar `--amiId` si su plantilla de lanzamiento personalizada no incluye una referencia a una AMI o si desea anular ese valor. Tenga en cuenta que la AMI utilizada para el grupo de nodos debe ser compatible con el AWS PCS. También puede seleccionar un ejemplo de AMI proporcionado por AWS. Para obtener más información sobre este tema, consulteAmazon Machine Images (AMIs) para AWS PCS.
- Úselo `--purchase-option` para elegir la forma en que AWS PCS compra las instancias EC2 para su grupo de nodos de cómputo. La opción predeterminada es On-Demand.
  - `ONDEMAND`— Utilice instancias bajo demanda. Elija también esta opción si planea usar una reserva de capacidad bajo demanda (ODCR). Para obtener más información, consulte Uso ODCRs con AWS PCS.
  - `SPOT`— Utilice instancias puntuales. Si elige instancias puntuales, también puede utilizarlas `--allocation-strategy` para definir cómo AWS PCS elige los grupos de capacidad puntuales cuando lanza instancias en el grupo de nodos. Para obtener más información, consulte Estrategias de asignación para instancias puntuales en la Guía del usuario de Amazon Elastic Compute Cloud.
  - `CAPACITY_BLOCK`— Utilice un bloque de capacidad de Amazon EC2 existente para la reserva de aprendizaje automático. Para obtener más información, consulte Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS.

- Es posible proporcionar opciones de Slurm configuración para los nodos del grupo de nodos utilizando`--slurm-configuration`. Puede establecer el peso (prioridad de programación) y la memoria real. Los nodos con pesos más bajos tienen mayor prioridad y las unidades son arbitrarias. Para obtener más información, consulte Peso en la Slurm documentación. La memoria real establece el tamaño (en GB) de la memoria real en los nodos del grupo de nodos. Se ha diseñado para usarse junto con la `CR_CPU_Memory` opción de clúster en AWS PCS de su Slurm configuración. Para obtener más información, consulte RealMemory en la documentación del Slurm.

> **Important**
>
> La creación del grupo de nodos de cómputo puede tardar varios minutos.

Puede consultar el estado de su grupo de nodos con el siguiente comando. No podrás asociar el grupo de nodos a una cola hasta que se alcance `ACTIVE` su estado.

```
aws pcs get-compute-node-group --region region \
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-node-group
```

Actualización de un grupo de nodos de cómputo AWS PCS

En este tema se proporciona información general sobre las opciones disponibles y se describe lo que se debe tener en cuenta al actualizar un grupo de nodos de cómputo de AWS PCS. Para obtener información sobre la configuración personalizada de Slurm, consulte. Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS

Opciones para actualizar un grupo de nodos de cómputo de AWS PCS

La actualización de un grupo de nodos de procesamiento de AWS PCS le permite cambiar las propiedades de las instancias lanzadas por AWS PCS, así como las reglas de lanzamiento de esas instancias. Por ejemplo, puede reemplazar la AMI de las instancias del grupo de nodos por otra que tenga instalado un software diferente. O bien, puede actualizar los grupos de seguridad para cambiar la conectividad de red entrante o saliente. También puede cambiar la configuración de escalado y la opción de compra preferida.

La siguiente configuración del grupo de nodos no se puede modificar después de su creación:

- Name
- instancias

Consideraciones al actualizar un grupo de nodos de cómputo de AWS PCS

Los grupos de nodos de cómputo definen las instancias EC2 que se utilizan para procesar trabajos, proporcionar acceso interactivo al shell y otras tareas. Suelen estar asociados a una o más colas de AWS PCS. Al actualizar el grupo de nodos de cómputo para cambiar su comportamiento (o el de sus nodos), tenga en cuenta lo siguiente:

- Los cambios en las propiedades del grupo de nodos de cómputo se hacen efectivos cuando el estado del grupo de nodos de cómputo cambia de Actualizado a Activo. Se lanzan nuevas instancias con las propiedades actualizadas.
- Las actualizaciones que no afectan a la configuración de nodos específicos no afectan a los nodos en ejecución. Por ejemplo, añadir una subred y cambiar la estrategia de asignación.
- Si actualiza la plantilla de lanzamiento de un grupo de nodos de cómputo, debe actualizar el grupo de nodos de cómputo para usar la nueva versión.
- Para añadir o eliminar un grupo de seguridad de los nodos de un grupo de nodos de procesamiento, edite su plantilla de lanzamiento y actualice el grupo de nodos de procesamiento. Se lanzan nuevas instancias con el conjunto actualizado de grupos de seguridad.
- Si editas directamente un grupo de seguridad utilizado por un grupo de nodos de procesamiento, esto tendrá efecto inmediato en las instancias en ejecución y en las futuras.
- Si agregas o eliminas permisos del perfil de instancias de IAM utilizado por un grupo de nodos de procesamiento, esto tendrá efecto inmediato en las instancias en ejecución y en las futuras.
- Para cambiar la AMI utilizada por las instancias de un grupo de nodos de cómputo, actualice el grupo de nodos de cómputo (o su plantilla de lanzamiento) para usar la nueva AMI y espere a que AWS PCS sustituya las instancias.
- AWS El PCS reemplaza las instancias existentes en el grupo de nodos tras una operación de actualización del grupo de nodos. Si hay trabajos en ejecución en un nodo, se permite que dichos trabajos se completen antes de que AWS PCS sustituya el nodo. Los procesos de usuario interactivos (por ejemplo, en las instancias de nodos de inicio de sesión) finalizan. El estado del grupo de nodos vuelve `Active` al que el AWS PCS marca las instancias para su reemplazo, pero el reemplazo real se produce cuando las instancias están inactivas.

- Si reduces el número máximo de instancias permitido en un grupo de nodos de cómputo, AWS PCS eliminará los nodos de Slurm para cumplir con el nuevo máximo. AWS El PCS finaliza las instancias en ejecución asociadas a los nodos de Slurm eliminados. Los trabajos en ejecución en los nodos eliminados fallan y vuelven a sus colas.
- AWS PCS crea una plantilla de lanzamiento gestionada para cada grupo de nodos de cómputo. Se nombranpcs-*identifier*-do-not-delete. No los seleccione al crear o actualizar un grupo de nodos de procesamiento, o el grupo de nodos no funcionará correctamente.
- Si actualiza un grupo de nodos de cómputo para usar Spot como opción de compra, debe tener el rol vinculado al servicio de AWSServiceRoleForEC2Spot en su cuenta. Para obtener más información, consulte Función Amazon EC2 Spot para PCS AWS.

Para actualizar un grupo de nodos de cómputo de AWS PCS

Puede actualizar un grupo de nodos mediante la consola de administración de AWS o la CLI de AWS.

Consola de administración de AWS

Para actualizar un grupo de nodos de cómputo

1. Abra la consola PCS de AWS en https://console.aws.amazon.com/pcs/home#/clusters
2. Seleccione el clúster en el que desea actualizar un grupo de nodos de cómputo.
3. Vaya a Grupos de nodos de cómputo, vaya al grupo de nodos que desee actualizar y, a continuación, seleccione Editar.
4. En las secciones Configuración informática, Ajustes adicionales y Ajustes de Slurmpersonalización, actualiza todos los valores excepto:

   - Instancias: no puede cambiar las instancias de un grupo de nodos de cómputo.

   Para obtener más información sobre la configuración personalizada de Slurm, consulte. Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS
5. Elija Actualizar. El campo Estado mostrará la actualización mientras se aplican los cambios.

> **Important**
>
> Las actualizaciones de los grupos de nodos de cómputo pueden tardar varios minutos.

AWS CLI

Para actualizar un grupo de nodos de cómputo

1. Actualice su grupo de nodos de cómputo con el siguiente comando. Antes de ejecutar el comando, realice los siguientes reemplazos:

   a. *region-code*Sustitúyala por la región de AWS en la que deseas crear tu clúster.

   b. *my-node-group*Sustitúyalo por el nombre o `computeNodeGroupId` por el grupo de nodos de cómputo.

   c. *my-cluster*Sustitúyalo por el nombre o `clusterId` el de tu clúster.

```
aws pcs update-compute-node-group --region region-code \
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-node-group
```

   Example— Actualizar un grupo de nodos de cómputo con una configuración de Slurm personalizada

```
aws pcs update-compute-node-group --region region-code \
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-node-group \
    --slurm-configuration \
    'slurmCustomSettings=[{parameterName=Features,parameterValue="gpu,nvme"}]'
```

   Para obtener más información, consulte Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS.

2. Actualice cualquier parámetro del grupo de nodos, excepto: `--instance-configs` Por ejemplo, para configurar un nuevo ID de AMI, pass `--amiId my-custom-ami-id` where *my-custom-ami-id* se sustituye por la AMI que prefiera.

> **Important**
>
> La actualización del grupo de nodos de cómputo puede tardar varios minutos.

Puedes consultar el estado de tu grupo de nodos con el siguiente comando.

```
aws pcs get-compute-node-group --region region-code \
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-node-group
```

Eliminar un grupo de nodos de cómputo en AWS PCS

En este tema se proporciona una descripción general de las opciones disponibles y se describe lo que se debe tener en cuenta al eliminar un grupo de nodos de procesamiento en AWS PCS.

Consideraciones a la hora de eliminar un grupo de nodos de cómputo

Los grupos de nodos de cómputo definen las instancias EC2 que se utilizan para procesar trabajos, proporcionar acceso interactivo al shell y otras tareas. Suelen estar asociados a una o más colas de AWS PCS. Antes de eliminar un grupo de nodos de procesamiento, tenga en cuenta lo siguiente:

- Se cancelará cualquier instancia de EC2 lanzada por el grupo de nodos de procesamiento. Esto cancelará los trabajos que se estén ejecutando en estas instancias y pondrá fin a la ejecución de los procesos interactivos.
- Debe desasociar el grupo de nodos de cómputo de todas las colas antes de poder eliminarlo. Para obtener más información, consulte Actualización de una cola de AWS PCS.

Elimine el grupo de nodos de cómputo

Puede usar Consola de administración de AWS o AWS CLI para eliminar un grupo de nodos de procesamiento.

Consola de administración de AWS

Para eliminar un grupo de nodos de cómputo

1. Abra la consola AWS PCS.

2. Seleccione el clúster del grupo de nodos de cómputo.
3. Vaya a Grupos de nodos de cómputo y seleccione el grupo de nodos de cómputo que desee eliminar.
4. Elija Eliminar.
5. Aparece el campo Estado`Deleting`. Puede tardar varios minutos en completarse.

> **Note**
>
> Puede usar los comandos nativos de su programador para confirmar que se ha eliminado el grupo de nodos de procesamiento. Por ejemplo, usa `sinfo` o `squeue` para Slurm.

AWS CLI

Para eliminar un grupo de nodos de cómputo

- Usa el siguiente comando para eliminar un grupo de nodos de cómputo, con estos reemplazos:
  - *`region-code`*Sustitúyalo por el que se encuentra Región de AWS tu clúster.
  - *`my-node-group`*Sustitúyalo por el nombre o el ID de tu grupo de nodos de procesamiento.
  - *`my-cluster`*Sustitúyalo por el nombre o el ID de tu clúster.

```
aws pcs delete-compute-node-group --region region-code \
      --compute-node-group-identifier my-node-group \
      --cluster-identifier my-cluster
```

Eliminar el grupo de nodos de procesamiento puede tardar varios minutos.

> **Note**
>
> Puede usar los comandos nativos de su programador para confirmar que se ha eliminado el grupo de nodos de procesamiento. Por ejemplo, usa `sinfo` o `squeue` para Slurm.

Obtenga detalles del grupo de nodos de cómputo en AWS PCS

Puede usar Consola de administración de AWS o AWS CLI para obtener detalles sobre un grupo de nodos de cómputo, como su ID de grupo de nodos de cómputo, el nombre de recurso de Amazon (ARN) y el ID de Amazon Machine Image (AMI). Estos detalles suelen ser valores obligatorios para las acciones y configuraciones de la API de AWS PCS.

Consola de administración de AWS

Para obtener los detalles del grupo de nodos de cómputo

1. Abra la consola AWS PCS.
2. Seleccione el clúster.
3. Elija grupos de nodos de cómputo.
4. Elija un grupo de nodos de procesamiento en el panel de lista.

AWS CLI

Para obtener los detalles del grupo de nodos de cómputo

1. Usa la acción ListClustersde la API para buscar el nombre o el ID del clúster.

```
aws pcs list-clusters
```

Ejemplo de salida:

```
{
    "clusters": [
        {
            "name": "get-started-cfn",
            "id": "pcs_abc1234567",
            "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_abc1234567",
            "createdAt": "2025-04-01T20:11:22+00:00",
            "modifiedAt": "2025-04-01T20:11:22+00:00",
            "status": "ACTIVE"
        }
    ]
}
```

2. Usa la acción de la ListComputeNodeGroupsAPI para enumerar los grupos de nodos de procesamiento de un clúster.

```
aws pcs list-compute-node-groups --cluster-identifier cluster-name-or-id
```

Ejemplo de llamada:

```
aws pcs list-compute-node-groups --cluster-identifier get-started-cfn
```

Ejemplo de salida:

```
{
    "computeNodeGroups": [
        {
            "name": "compute-1",
            "id": "pcs_abc123abc1",
            "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_abc1234567/
computenodegroup/pcs_abc123abc1",
            "clusterId": "pcs_abc1234567",
            "createdAt": "2025-04-01T20:19:25+00:00",
            "modifiedAt": "2025-04-01T20:19:25+00:00",
            "status": "ACTIVE"
        },
        {
            "name": "login",
            "id": "pcs_abc456abc7",
            "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_abc1234567/
computenodegroup/pcs_abc456abc7",
            "clusterId": "pcs_abc1234567",
            "createdAt": "2025-04-01T20:19:31+00:00",
            "modifiedAt": "2025-04-01T20:19:31+00:00",
            "status": "ACTIVE"
        }
    ]
}
```

3. Usa la acción de la GetComputeNodeGroupAPI para obtener detalles adicionales de un grupo de nodos de cómputo.

```
aws pcs get-compute-node-group --cluster-identifier cluster-name-or-id --
compute-node-group-identifier compute-node-group-name-or-id
```

Ejemplo de llamada:

```
aws pcs get-compute-node-group --cluster-identifier get-started-cfn --compute-
node-group-identifier compute-1
```

Ejemplo de salida:

```
{
    "computeNodeGroup": {
        "name": "compute-1",
        "id": "pcs_abc123abc1",
        "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_abc1234567/
computenodegroup/pcs_abc123abc1",
        "clusterId": "pcs_abc1234567",
        "createdAt": "2025-04-01T20:19:25+00:00",
        "modifiedAt": "2025-04-01T20:19:25+00:00",
        "status": "ACTIVE",
        "amiId": "ami-0123456789abcdef0",
        "subnetIds": [
            "subnet-abc012345789abc12"
        ],
        "purchaseOption": "ONDEMAND",
        "customLaunchTemplate": {
            "id": "lt-012345abcdef01234",
            "version": "1"
        },
        "iamInstanceProfileArn": "arn:aws:iam::111122223333:instance-profile/
AWSPCS-get-started-cfn-us-east-1",
        "scalingConfiguration": {
            "minInstanceCount": 0,
            "maxInstanceCount": 4
        },
        "instanceConfigs": [
            {
                "instanceType": "c6i.xlarge"
            }
        ]
    }
}
```

Búsqueda de instancias de grupos de nodos de cómputo en AWS PCS

Cada grupo de nodos de cómputo de AWS PCS puede lanzar instancias EC2 con configuraciones compartidas. Puede utilizar las etiquetas EC2 para buscar instancias en un grupo de nodos de procesamiento en Consola de administración de AWS o con. AWS CLI

Consola de administración de AWS

Para encontrar las instancias de tu grupo de nodos de cómputo

1. Abre la consola AWS PCS.
2. Seleccione el clúster.
3. Elija grupos de nodos de cómputo.
4. Busca el ID del grupo de nodos de inicio de sesión que creaste.
5. Navegue hasta la consola EC2 y elija Instances.
6. Busque las instancias con la siguiente etiqueta. *node-group-id*Sustitúyala por el ID (no el nombre) de tu grupo de nodos de cómputo.

   ```
   aws:pcs:compute-node-group-id=node-group-id
   ```

7. (Opcional) Puede cambiar el valor del estado de la instancia en el campo de búsqueda para buscar las instancias que se estén configurando o que se hayan cancelado recientemente.
8. Busca el ID de instancia y la dirección IP de cada instancia en la lista de instancias etiquetadas.

AWS CLI

Para encontrar las instancias de tu grupo de nodos, usa los siguientes comandos. Antes de ejecutar los comandos, realiza las siguientes sustituciones:

- *region-code*Reemplácelo por el Región de AWS de su clúster. Ejemplo: `us-east-1`
- *node-group-id*Sustitúyalo por el ID (no el nombre) de tu grupo de nodos de procesamiento. Para encontrar el ID de un grupo de nodos de cómputo, consulteObtenga detalles del grupo de nodos de cómputo en AWS PCS.

- `running`Reemplácelo por otros estados de instancia, como `pending` o `terminated` busque instancias de EC2 en otros estados.

```
aws ec2 describe-instances \
     --region region-code --filters \
     "Name=tag:aws:pcs:compute-node-group-id,Values=node-group-id" \
     "Name=instance-state-name,Values=running" \
     --query 'Reservations[*].Instances[*].
{InstanceID:InstanceId,State:State.Name,PublicIP:PublicIpAddress,PrivateIP:PrivateIpAddress}
```

El comando devuelve un resultado similar al siguiente. El valor de `PublicIP` es `null` si la instancia está en una subred privada.

```
[
    [
        {
            "InstanceID": "i-0123456789abcdefa",
            "State": "running",
            "PublicIP": "18.189.32.188",
            "PrivateIP": "10.0.0.1"
        }
    ]
]
```

> **Note**
>
> Si espera `describe-instances` devolver una gran cantidad de instancias, debe usar las opciones para varias páginas. Para obtener más información, consulte la referencia DescribeInstancesde la API de Amazon Elastic Compute Cloud.

Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS

En Amazon EC2, una plantilla de lanzamiento puede almacenar un conjunto de preferencias para que no tenga que especificarlas individualmente al lanzar las instancias. AWS PCS incorpora plantillas de lanzamiento como una forma flexible de configurar grupos de nodos de cómputo. Al crear un grupo de nodos, se proporciona una plantilla de lanzamiento. AWS PCS crea una plantilla de lanzamiento derivada a partir de ella que incluye transformaciones para garantizar que funcione con el servicio.

Entender cuáles son las opciones y las consideraciones a tener en cuenta a la hora de escribir una plantilla de lanzamiento personalizada puede ayudarle a crear una para utilizarla con AWS PCS. Para obtener más información sobre las plantillas de lanzamiento, consulte Lanzamiento de una instancia desde una plantilla de lanzamiento en la Guía del usuario de Amazon EC2.

Temas

Descripción general de las plantillas de lanzamiento en PCS AWS

Hay más de 30 parámetros disponibles que puede incluir en una plantilla de lanzamiento de EC2, que controlan muchos aspectos de la configuración de las instancias. La mayoría son totalmente compatibles con el AWS PCS, pero hay algunas excepciones.

El AWS PCS ignorará los siguientes parámetros de la plantilla de lanzamiento de EC2, ya que el servicio debe administrar directamente estas propiedades:

- Atributos del tipo de type/Specify instancia (`InstanceRequirements`): AWS PCS no admite la selección de instancias basada en atributos.
- Tipo de instancia (`InstanceType`): especifique los tipos de instancia al crear un grupo de nodos.

- Perfil de details/IAM instancia avanzado (`IamInstanceProfile`): lo proporcionas al crear o actualizar el grupo de nodos.
- Terminación avanzada de la details/Disable API (`DisableApiTermination`): AWS PCS debe controlar el ciclo de vida de las instancias del grupo de nodos que lanza.
- Parada de details/Disable API avanzada (`DisableApiStop`): AWS PCS debe controlar el ciclo de vida de las instancias de grupos de nodos que lanza.
- Avanzado details/Stop : comportamiento de hibernación (`HibernationOptions`): AWS PCS no admite la hibernación de instancias.
- details/Elastic GPU avanzada (`ElasticGpuSpecifications`): Amazon Elastic Graphics llegó al final de su vida útil el 8 de enero de 2024.
- details/Elastic Inferencia avanzada (`ElasticInferenceAccelerators`): Amazon Elastic Inference ya no está disponible para clientes nuevos.
- AAdvanced details/Specify CPU options/Threadsper core (`ThreadsPerCore`): AWS PCS establece el número de subprocesos por núcleo en 1.

Estos parámetros tienen requisitos especiales que respaldan la compatibilidad con el AWS PCS:

- Datos de usuario (`UserData`): deben estar codificados en varias partes. Consulte Uso de datos de usuario de Amazon EC2 para PCS AWS.
- Imágenes de la aplicación y del sistema operativo (`ImageId`): puede incluirlas. Sin embargo, si especifica un ID de AMI al crear o actualizar el grupo de nodos, este anulará el valor de la plantilla de lanzamiento. La AMI que proporcione debe ser compatible con AWS PCS. Para obtener más información, consulte "Amazon Machine Images (AMIs) para AWS PCS.
- Red settings/Firewall (grupos de seguridad) (`SecurityGroups`): no se puede configurar una lista de nombres de grupos de seguridad en una plantilla de lanzamiento de AWS PCS. Puede configurar una lista de grupos de seguridad IDs (`SecurityGroupIds`), a menos que defina las interfaces de red en la plantilla de lanzamiento. A continuación, debe especificar el grupo de seguridad IDs para cada interfaz. Para obtener más información, consulte Grupos de seguridad en AWS PCS.
- Configuración de settings/Advanced red (`NetworkInterfaces`): si utiliza instancias EC2 con una sola tarjeta de red y no necesita ninguna configuración de red especializada, AWS PCS puede configurar la red de instancias por usted. Para configurar varias tarjetas de red o habilitar el adaptador Elastic Fabric en sus instancias, utilice`NetworkInterfaces`. Cada interfaz de red debe tener una lista de grupos de seguridad IDs debajo`Groups`. Para obtener más información, consulte Múltiples interfaces de red en AWS PCS.

- Detalles avanzados/reserva de capacidad (`CapacityReservationSpecification`): se puede configurar, pero no puede hacer referencia a un dato específico `CapacityReservationId` cuando se trabaja con AWS PCS. Sin embargo, puede hacer referencia a un grupo de reservas de capacidad, si ese grupo contiene una o más reservas de capacidad. Para obtener más información, consulte Reservas de capacidad en AWS PCS.

Creación de una plantilla de lanzamiento básica

Puede crear una plantilla de lanzamiento mediante el Consola de administración de AWS o el AWS CLI.

Consola de administración de AWS

Para crear una plantilla de lanzamiento

1. Abre la EC2consola de Amazon y selecciona Launch templates.
2. Elija Crear plantilla de inicialización.
3. En Nombre y descripción de la plantilla de lanzamiento, introduce un nombre único y distintivo para el nombre de la plantilla de lanzamiento
4. En Par de claves (inicio de sesión) en Nombre del par de claves, seleccione el par de claves SSH que se usará para iniciar sesión en EC2 las instancias administradas por AWS PCS. Esto es opcional, pero recomendable.
5. En Configuración de red y, a continuación, en Firewall (grupos de seguridad), elija los grupos de seguridad que desee conectar a la interfaz de red. Todos los grupos de seguridad de la plantilla de lanzamiento deben pertenecer a la VPC del clúster de AWS PCS. Como mínimo, elija:
   - Un grupo de seguridad que permite la comunicación con el clúster de AWS PCS
   - Un grupo de seguridad que permite la comunicación entre EC2 instancias lanzadas por AWS PCS
   - (Opcional) Un grupo de seguridad que permite el acceso SSH entrante a instancias interactivas
   - (Opcional) Un grupo de seguridad que permite a los nodos de cómputo establecer conexiones salientes a Internet
   - (Opcional) Grupos de seguridad que permiten el acceso a los recursos de la red, como los sistemas de archivos compartidos o un servidor de bases de datos.

6. Podrás acceder a tu nueva ID de plantilla de lanzamiento en la EC2 consola de Amazon, en Plantillas de lanzamiento. El identificador de la plantilla de lanzamiento incluirá el formulario`lt-0123456789abcdef01`.

Siguiente paso recomendado

- Utilice la nueva plantilla de lanzamiento para crear o actualizar un grupo de nodos de cómputo de AWS PCS.

AWS CLI

Para crear una plantilla de lanzamiento

Cree su plantilla de lanzamiento con el siguiente comando.

- Antes de ejecutar el comando, realice los siguientes reemplazos:
  a. *region-code*Sustitúyala por la Región de AWS que estás trabajando con AWS PCS
  b. *my-launch-template-name*Sustitúyala por un nombre para la plantilla. Debe ser exclusivo del Cuenta de AWS y Región de AWS que está utilizando.
  c. *my-ssh-key-name*Sustitúyala por el nombre de tu clave SSH preferida.
  d. Reemplace *sg-ExampleID1* y *sg-ExampleID2* por un grupo de seguridad IDs que permita la comunicación entre sus EC2 instancias y el programador y la comunicación entre EC2 instancias. Si solo tiene un grupo de seguridad que permite todo este tráfico, puede eliminarlo `sg-ExampleID2` y el carácter de coma que lo precede. También puede añadir más grupos IDs de seguridad. Todos los grupos de seguridad que incluya en la plantilla de lanzamiento deben provenir de la VPC de su clúster de AWS PCS.

```
aws ec2 create-launch-template --region region-code \
    --launch-template-name my-template-name \
    --launch-template-data '{"KeyName":"my-ssh-key-name","SecurityGroupIds":
 ["sg-ExampleID1","sg-ExampleID2"]}'
```

El resultado AWS CLI será un texto parecido al siguiente. El identificador de la plantilla de lanzamiento se encuentra en`LaunchTemplateId`.

```
{
    "LaunchTemplate": {
        "LatestVersionNumber": 1,
        "LaunchTemplateId": "lt-0123456789abcdef01",
        "LaunchTemplateName": "my-launch-template-name",
        "DefaultVersionNumber": 1,
        "CreatedBy": "arn:aws:iam::123456789012:user/Bob",
        "CreateTime": "2019-04-30T18:16:06.000Z"
    }
}
```

Siguiente paso recomendado

- Utilice la nueva plantilla de lanzamiento para crear o actualizar un grupo de nodos de cómputo de AWS PCS.

Uso de datos de usuario de Amazon EC2 para PCS AWS

Puede proporcionar los datos de usuario de EC2 en la plantilla de lanzamiento que `cloud-init` se ejecuta cuando se lanzan las instancias. Los bloques de datos de usuario con ese tipo de contenido se `cloud-config` ejecutan antes de que la instancia se registre en la API de AWS PCS, mientras que los bloques de datos de usuario con ese tipo de contenido se `text/x-shellscript` ejecutan una vez finalizado el registro, pero antes de que se inicie el daemon Slurm. Para obtener más información sobre los tipos de contenido, consulte la documentación de cloud-init.

nuestros datos de usuario pueden realizar escenarios de configuración comunes, incluidos, entre otros, los siguientes:

- Incluidos usuarios o grupos
- Instalación de paquetes
- Creación de particiones y sistemas de archivos
- Montaje de sistemas de archivos de red

Los datos de usuario de las plantillas de lanzamiento deben estar en formato de archivo MIME de varias partes. Esto se debe a que sus datos de usuario se combinan con otros datos de usuario de AWS PCS necesarios para configurar los nodos de su grupo de nodos. Puede combinar varios bloques de datos de usuario en un único archivo multiparte MIME.

Un archivo multiparte MIME consta de los siguientes componentes:

- El tipo de contenido y declaración de límite de partes: `Content-Type: multipart/mixed; boundary="==BOUNDARY=="`
- La declaración de versión de MIME: `MIME-Version: 1.0`
- Uno o más bloques de datos de usuario, que contienen los siguientes componentes:
  - El límite de apertura, que señala el inicio de un bloque de datos de usuario: `--==BOUNDARY==`. Debe dejar en blanco la línea anterior a este límite.
  - La declaración del tipo de contenido del bloque: `Content-Type: text/cloud-config; charset="us-ascii"` o`Content-Type: text/x-shellscript; charset="us-ascii"`. Debe dejar en blanco la línea que sigue a la declaración de tipo de contenido.
  - El contenido de los datos de usuario, por ejemplo, una lista de shell o políticas de `cloud-config`.
- El límite de cierre, que señala el final del archivo multiparte MIME: `--==BOUNDARY==--`. Debe dejar en blanco la línea anterior al límite de cierre.

> **Note**
>
> Si añade datos de usuario a una plantilla de lanzamiento en la consola de Amazon EC2, puede pegarlos como texto sin formato. O bien, puede cargarlos desde un archivo.
> Si usa el SDK AWS CLI o un AWS SDK, primero debe codificar en base64 los datos del usuario y enviar esa cadena como el valor del `UserData` parámetro cuando llame CreateLaunchTemplate, como se muestra en este archivo JSON.

```
{
    "LaunchTemplateName": "base64-user-data",
    "LaunchTemplateData": {
        "UserData":
 "ewogICAgIkxhdW5jaFRlbXBsYXRlTmFtZSI6ICJpbmNyZWFzZS1jb250YWluZXItdm9sdW..."
    }
}
```

Ejemplos

- Ejemplo: instalar software desde un repositorio de paquetes

- Ejemplo: ejecutar scripts desde un bucket de S3
- Ejemplo: establecer variables de entorno globales
- Uso de sistemas de archivos de red con AWS PCS
- Ejemplo: utilizar un sistema de archivos EFS como directorio principal compartido

Ejemplo: instalar software para AWS PCS desde un repositorio de paquetes

Proporcione este script como el valor de `"userData"` en su plantilla de lanzamiento. Para obtener más información, consulte Uso de datos de usuario de Amazon EC2 para PCS AWS.

Este script usa cloud-config para instalar paquetes de software en instancias de grupos de nodos en el momento del lanzamiento. Para obtener más información, consulta los formatos de datos de usuario en la documentación de cloud-init. En este ejemplo, se instala `curl` y. `llvm`

> **Note**
>
> Sus instancias deben poder conectarse a sus repositorios de paquetes configurados.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
- python3-devel
- rust
- golang

--==MYBOUNDARY==--
```

Ejemplo: ejecutar scripts adicionales para AWS PCS desde un bucket de S3

Proporcione este script como el valor de `"userData"` en su plantilla de lanzamiento. Para obtener más información, consulte Uso de datos de usuario de Amazon EC2 para PCS AWS.

El siguiente script de datos de usuario usa cloud-config para importar un script de un bucket de S3 y ejecutarlo en instancias de grupos de nodos en el momento del lanzamiento. Para obtener más información, consulta los formatos de datos de usuario en la documentación de cloud-init.

Sustituya los siguientes valores por sus propios detalles:

- *amzn-s3-demo-bucket*— El nombre de un bucket de S3 desde el que puede leer tu cuenta.
- *object-key*— La clave de objeto S3 del script que se va a importar. Incluye el nombre del script y su ubicación en la estructura de carpetas del depósito. Por ejemplo, `scripts/script.sh`. Para obtener más información, consulte Organizar objetos en la consola de Amazon S3 mediante carpetas en la Guía del usuario de Amazon Simple Storage Service.
- *shell*— El shell de Linux que se utilizará para ejecutar el script, por ejemplo`bash`.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- aws s3 cp s3://amzn-s3-demo-bucket/object-key /tmp/script.sh
- /usr/bin/shell /tmp/script.sh

--==MYBOUNDARY==--
```

El perfil de instancia de IAM del grupo de nodos debe tener acceso al bucket. La siguiente política de IAM es un ejemplo del depósito del script de datos de usuario anterior.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
```

```
            "Resource": [
                "arn:aws:s3:::amzn-s3-demo-bucket",
                "arn:aws:s3:::amzn-s3-demo-bucket/*"
            ]
        }
    ]
}
```

Ejemplo: establecer variables de entorno globales para AWS PCS

Proporcione este script como el valor de `"userData"` en su plantilla de lanzamiento. Para obtener más información, consulte Uso de datos de usuario de Amazon EC2 para PCS AWS.

El siguiente ejemplo se utiliza `/etc/profile.d` para establecer variables globales en instancias de grupos de nodos.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/x-shellscript; charset="us-ascii"

#!/bin/bash
touch /etc/profile.d/awspcs-userdata-vars.sh
echo MY_GLOBAL_VAR1=100 >> /etc/profile.d/awspcs-userdata-vars.sh
echo MY_GLOBAL_VAR2=abc >> /etc/profile.d/awspcs-userdata-vars.sh

--==MYBOUNDARY==--
```

Ejemplo: utilizar un sistema de archivos EFS como directorio principal compartido para AWS PCS

Proporcione este script como valor de `"userData"` en su plantilla de lanzamiento. Para obtener más información, consulte Uso de datos de usuario de Amazon EC2 para PCS AWS.

Este ejemplo amplía el ejemplo de montaje de EFS Uso de sistemas de archivos de red con AWS PCS para implementar un directorio principal compartido. Se hace una copia de seguridad del contenido de /home antes de montar el sistema de archivos EFS. A continuación, el contenido se copia rápidamente en su lugar en el almacenamiento compartido una vez finalizado el montaje.

Sustituya los siguientes valores de este script por sus propios detalles:

- *`/mount-point-directory`*— La ruta de una instancia en la que desea montar el sistema de archivos EFS.
- *`filesystem-id`*— El ID del sistema de archivos del sistema de archivos EFS.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
  - amazon-efs-utils

runcmd:
  - mkdir -p /tmp/home
  - rsync -a /home/ /tmp/home
  - echo "filesystem-id:/ /mount-point-directory efs tls,_netdev" >> /etc/fstab
  - mount -a -t efs defaults
  - rsync -a --ignore-existing /tmp/home/ /home
  - rm -rf /tmp/home/

--==MYBOUNDARY==--
```

Ejemplo: habilitar SSH sin contraseña

Puedes basarte en el ejemplo del directorio principal compartido para implementar conexiones SSH entre instancias del clúster mediante claves SSH. Para cada usuario que utilice el sistema de archivos de inicio compartido, ejecute un script similar al siguiente:

```
#!/bin/bash

mkdir -p $HOME/.ssh && chmod 700 $HOME/.ssh
touch $HOME/.ssh/authorized_keys
chmod 600 $HOME/.ssh/authorized_keys

if [ ! -f "$HOME/.ssh/id_rsa" ]; then
    ssh-keygen -t rsa -b 4096 -f $HOME/.ssh/id_rsa -N ""
    cat ~/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

```
fi
```

> **Note**
>
> Las instancias deben usar un grupo de seguridad que permita las conexiones SSH entre los nodos del clúster.

Reservas de capacidad en AWS PCS

Puede reservar la EC2 capacidad de Amazon en una zona de disponibilidad específica y durante un período específico mediante reservas de capacidad bajo demanda o bloques de EC2 capacidad de Amazon para aprendizaje automático para asegurarse de que tiene la capacidad informática necesaria disponible cuando la necesite.

Las reservas de capacidad bajo demanda (ODCRs) le permiten reservar capacidad de cómputo para sus EC2 instancias de Amazon en una zona de disponibilidad específica durante cualquier período. Puede crear y cancelar reservas en cualquier momento, sin compromisos a largo plazo ni pagos por adelantado. ODCRs son ideales cuando necesita reservas de capacidad flexibles que pueda modificar a medida que cambien sus requisitos. Para obtener más información, consulte Reservas de capacidad bajo demanda en la Guía del usuario de Amazon Elastic Compute Cloud.

Amazon EC2 Capacity Blocks for ML le permite reservar instancias de computación acelerada basadas en GPU para usarlas en el futuro, con hasta 8 semanas de antelación. Puede reservar bloques de 1 a 64 instancias con una duración de 1 día a 6 meses. Los bloques de capacidad son ideales para las cargas de trabajo de aprendizaje automático que requieren un acceso garantizado a la capacidad de la GPU en momentos específicos. Para obtener más información, consulte Capacity Blocks for ML en la Guía del usuario de Amazon Elastic Compute Cloud.

Temas

- Uso ODCRs con AWS PCS
- Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS

Uso ODCRs con AWS PCS

Puede elegir la forma en que AWS PCS consume las instancias reservadas. Si crea una ODCR abierta, cualquier instancia coincidente lanzada por AWS PCS u otros procesos de su cuenta

se descontará de la reserva. Con una ODCR segmentada, solo las instancias lanzadas con el identificador de reserva específico se tienen en cuenta para la reserva. En el caso de las cargas de trabajo urgentes, las segmentadas ODCRs son más habituales.

Puede configurar un grupo de nodos de cómputo de AWS PCS para que utilice un ODCR de destino agregándolo a una plantilla de lanzamiento. Estos son los pasos para hacerlo:

1. Cree una reserva de capacidad bajo demanda (ODCR) específica mediante la Guía del usuario para crear una reserva de capacidad de Amazon EC2.
2. Asocie la ODCR a una plantilla de lanzamiento. Hay dos maneras de hacerlo:

    a. Asociación directa de ODCR: haga referencia al ID de ODCR directamente en la plantilla de lanzamiento. Este enfoque proporciona un control estricto de la capacidad y no admite la reposición de instancias (si el grupo de nodos de procesamiento solicita más instancias de las disponibles en la ODCR, no se lanzará ninguna instancia adicional).

    b. Asociación de grupos de reserva de capacidad: añada el ODCR a un grupo de reserva de capacidad y haga referencia al grupo en la plantilla de lanzamiento. Este enfoque permite la reposición de instancias, lo que permite a AWS PCS lanzar instancias adicionales bajo demanda si se supera la capacidad de reserva.
3. Cree o actualice un grupo de nodos de cómputo de AWS PCS para usar la plantilla de lanzamiento. Para obtener más información, consulte la Guía del usuario de los grupos de nodos de cómputo de AWS PCS.

    - Establezca el grupo `purchaseOption` de nodos de cómputo en`ONDEMAND`.

Ejemplo: reserve y use instancias hpc6a.48xlarge con un ODCR de destino

Este comando de ejemplo crea un ODCR de destino para 32 instancias de hpc6a.48xlarge. Para lanzar las instancias reservadas en un grupo de ubicación, agréguelas al comando. `--placement-group-arn` Puede definir una fecha de finalización con `--end-date` y`--end-date-type`, de lo contrario, la reserva continuará hasta que se finalice manualmente.

```
aws ec2 create-capacity-reservation \
    --instance-type hpc6a.48xlarge \
    --instance-platform Linux/UNIX \
    --availability-zone us-east-2a \
    --instance-count 32 \
    --instance-match-criteria targeted
```

El resultado de este comando será un ARN para el nuevo ODCR. El ID de ODCR se puede recuperar del `"arn:aws:ec2:us-east-2:123456789012:capacity-reservation/ODCR-ID"` ARN o mediante Amazon EC2. DescribeCapacityReservations

Asociación ODCR directa: añada el ID de ODCR a la plantilla de lanzamiento. A continuación, se muestra un ejemplo de plantilla de lanzamiento que hace referencia al ID de la ODCR.

```
{
  "CapacityReservationSpecification": {
    "CapacityReservationTarget": {
      "CapacityReservationId": "cr-1234567890abcdef1"
    }
  }
}
```

Asociación de grupos de reserva de capacidad: cree un grupo de reserva de capacidad y añada el grupo a la plantilla de lanzamiento. El siguiente comando crea un grupo de reserva de capacidad denominado`EXAMPLE-CR-GROUP`.

```
aws resource-groups create-group \
    --name EXAMPLE-CR-GROUP \
    --configuration \
        '{"Type": "AWS::EC2::CapacityReservationPool"}' \
        '{"Type": "AWS::ResourceGroups::Generic", "Parameters": [{"Name": "allowed-
resource-types", "Values": ["AWS::EC2::CapacityReservation"]}]}'
```

El siguiente comando agrega el ODCR al grupo de reserva de capacidad.

```
aws resource-groups group-resources --group EXAMPLE-CR-GROUP \
    --resource-arns arn:aws:ec2:us-east-2:123456789012:capacity-reservation/
cr-1234567890abcdef1
```

Con el ODCR creado y agregado a un grupo de reserva de capacidad, ahora se puede conectar a un grupo de nodos de cómputo del AWS PCS agregándolo a una plantilla de lanzamiento. A continuación, se muestra un ejemplo de plantilla de lanzamiento que hace referencia al grupo de reserva de capacidad.

```
{
  "CapacityReservationSpecification": {
```

```
    "CapacityReservationResourceGroupArn": "arn:aws:resource-groups:us-
east-2:123456789012:group/EXAMPLE-CR-GROUP"
  }
}
```

Por último, cree o actualice un grupo de nodos de cómputo de AWS PCS para utilizar instancias hpc6a.48xlarge y utilice la plantilla de lanzamiento que hace referencia a la ODCR. Para un grupo de nodos estático, defina las instancias mínimas y máximas según el tamaño de la reserva (32). Para un grupo de nodos dinámico, establece el número mínimo de instancias en 0 y el máximo en el tamaño de instancia deseado.

Este ejemplo es una implementación simple de un ODCR único que se aprovisiona para un grupo de nodos de cómputo. Sin embargo, AWS PCS admite muchos otros diseños. Por ejemplo, puede subdividir un grupo grande de ODCR o de reserva de capacidad entre varios grupos de nodos de cómputo. O bien, puede usar la ODCRs que otra cuenta de AWS haya creado y compartido con la suya.

Para obtener más información, consulte Reservas de capacidad bajo demanda y bloques de capacidad para aprendizaje automático en la Guía del usuario de Amazon Elastic Compute Cloud.

Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS

Amazon EC2 Capacity Blocks for ML es una opción de compra de Amazon EC2 que le permite pagar por adelantado la reserva de instancias de computación acelerada basadas en GPU dentro de un intervalo de fechas y horas específico para soportar cargas de trabajo de corta duración. Las instancias que se ejecutan dentro de un bloque de capacidad se colocan automáticamente juntas dentro de Amazon EC2 UltraClusters, para una red de baja latencia, escala de petabits y sin bloqueos. Para obtener más información, consulte Capacity Blocks for ML en la Guía del usuario de Amazon Elastic Compute Cloud.

Puede usar una plantilla de lanzamiento para que AWS PCS utilice un bloque de capacidad cuando lance instancias para un grupo de nodos de cómputo.

> **Note**
>
> AWS PCS introdujo la compatibilidad con los bloques de capacidad desde la versión 24.05 de Slurm.

Limitaciones

- AWS PCS solo admite bloques de capacidad con las familias de instancias P5en, P5e, P5 y P4d.
- Solo puede asociar un grupo de nodos de cómputo a un bloque de capacidad a la vez.
- No puede asociar un grupo de nodos de cómputo a un grupo de reserva de capacidad que combine varios bloques de capacidad.
- Los bloques de capacidad deben estar en un `active` estado `scheduled` o estado para poder usarse con AWS PCS. No puedes usar bloques de capacidad en otros estados, como`payment-failed`. Para obtener más información, consulte Ver bloques de capacidad en la Guía del usuario de Amazon Elastic Compute Cloud.

Caducidad del bloque de capacidad

Los bloques de capacidad están limitados a un intervalo de fecha y hora específicos. Cuando caduca un bloque de capacidad:

- El grupo de nodos de cómputo asociado a ese bloque de capacidad sigue existiendo y sigue asociado a las mismas colas.
- Todas las instancias del grupo de nodos de cómputo están cerradas y es posible que los trabajos activos fallen, según la configuración de Slurm.
- AWS PCS no puede lanzar nuevas instancias en el grupo de nodos de cómputo.
- Todos los trabajos en cola o recién enviados permanecen pendientes hasta que se adjunte otro grupo de nodos de cómputo a la cola o hasta que se actualice el grupo de nodos de cómputo para usar una nueva plantilla de lanzamiento que especifique un nuevo bloque de capacidad.

Configurar un grupo de nodos de cómputo de AWS PCS para usar un bloque de capacidad

Para asociar un bloque de capacidad a un grupo de nodos de cómputo

1. Crea una plantilla de EC2 lanzamiento de Amazon para AWS PCS que especifique tu bloque de capacidad. Para obtener más información sobre la creación de una plantilla de lanzamiento para AWS PCS, consulteUso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

   La plantilla de lanzamiento debe incluir:

- El valor `MarketType` de `InstanceMarketOptions` debe estar establecido en`capacity-block`.
- A `CapacityReservationSpecification` con un valor válido `CapacityReservationId`
- Una válida `InstanceType` que coincida con el tipo de instancia del bloque de capacidad que compraste.

2. Cree un grupo de nodos de cómputo que utilice la plantilla de lanzamiento. Para obtener más información, consulte Creación de un grupo de nodos de cómputo en AWS PCS. También puede actualizar un grupo de nodos de procesamiento existente para usar la plantilla de lanzamiento. Para obtener más información, consulte Actualización de un grupo de nodos de cómputo AWS PCS.

   Al crear o actualizar el grupo de nodos de cómputo:

   - La identidad de IAM que utilices para crear o actualizar el grupo de nodos de procesamiento debe tener el siguiente permiso:

     ```
     ec2:DescribeCapacityReservations
     ```

     Para obtener más información, consulte Permisos mínimos para PCS AWS.
   - El bloque de capacidad debe estar en un `active` estado `scheduled` o.
   - Defina el grupo `purchaseOption` de nodos de cómputo en`CAPACITY_BLOCK`.
   - El tamaño `maxInstanceCount` del grupo de nodos de cómputo no debe superar el tamaño del bloque de capacidad.
   - La zona de disponibilidad del grupo de nodos de cómputo debe coincidir con una de las zonas de disponibilidad de subred del grupo de nodos de cómputo.

> **Important**
>
> No puedes cambiar el tipo de instancia de un grupo de nodos de cómputo al actualizarlo. Solo puedes usar un bloque de capacidad con el mismo tipo de instancia que el grupo de nodos de cómputo. Si quieres usar un bloque de capacidad con un tipo de instancia diferente, debes crear un nuevo grupo de nodos de cómputo.

Preguntas frecuentes sobre el uso de bloques de capacidad con AWS PCS

Acabo de pagar un bloque de capacidad e inmediatamente intenté usarlo con AWS PCS, pero no se pudo crear el grupo de nodos de cómputo. ¿Qué ha pasado?

> Es posible que su bloque de capacidad no esté en un `active` estado `scheduled` o. Vuelva a intentarlo cuando el bloque de capacidad esté `scheduled` o`active`.

Estoy utilizando un bloque de capacidad en AWS PCS y he comprado una extensión antes de que caducara. ¿Cómo puedo seguir utilizándolo en AWS PCS?

> No tiene que hacer nada para seguir utilizando el bloque de capacidad en AWS PCS. La fecha de finalización de tu bloque de capacidad se actualiza una vez que se haya realizado el pago de la extensión. Mientras el bloque de capacidad no caduque, el grupo de nodos de cómputo seguirá funcionando. Si no se realiza el pago de la extensión, el bloque de capacidad permanece `active` y el grupo de nodos de cómputo funciona hasta que el bloque de capacidad venza en su fecha de finalización original.

¿Qué ocurre con mis trabajos en cola y en ejecución si mi bloque de capacidad caduca?

> Los trabajos en cola que no se iniciaron antes de que expirara el bloque de capacidad permanecen pendientes hasta que asocie otro grupo de nodos de cómputo a la cola o actualice el grupo de nodos de cómputo con un nuevo bloque de capacidad. Aún puede enviar trabajos a la cola. La configuración de Slurm afecta a los trabajos activos. De forma predeterminada, los trabajos activos se vuelven a poner en cola automáticamente, pero pueden tener errores o fallar.

Mi bloque de capacidad ha caducado. ¿Debo hacer algo?

> No tienes que hacer nada. Puede comprobar el estado de sus reservas de capacidad de EC2 en la consola Amazon EC2. Cuando un bloque de capacidad caduca, el grupo de nodos de cómputo asociado a ese bloque de capacidad sigue existiendo y gestionando las mismas colas. El grupo de nodos de cómputo no tiene instancias para ejecutar trabajos. Puedes eliminar el grupo de nodos de cómputo o desasociarlo de las colas para evitar que los usuarios envíen trabajos que no se ejecutarán.

Quiero usar un nuevo bloque de capacidad con mi grupo de nodos de cómputo de AWS PCS. ¿Qué tengo que hacer?

> Le recomendamos que cree un nuevo grupo de nodos de cómputo para usar el nuevo bloque de capacidad. Para obtener más información, consulte Configurar un grupo de nodos de cómputo de AWS PCS para usar un bloque de capacidad.

¿Cómo puedo compartir 1 bloque de capacidad entre clústeres y servicios?

Puede dividir un bloque de capacidad en varios clústeres y servicios. Por ejemplo, para dividir un bloque de capacidad con 64 `p5.48xlarge` instancias con 20 nodos en el PCS-Cluster-1, 16 nodos en el PCS-Cluster-2 y los nodos restantes para otros servicios, defina ambos `minInstanceCount` nodos en 20 para el PCS-Cluster-1 y 16 para el PCS-Cluster-2. `maxInstanceCount`

¿Puedo usar más de un bloque de capacidad o una capacidad combinada con un grupo de nodos de cómputo?

No. Solo se puede asociar un bloque de capacidad a un único grupo de nodos de procesamiento. AWS PCS no admite grupos de reserva de capacidad que combinen varios bloques de capacidad.

¿Cómo sé cuándo comienzan o caducan mis bloques de capacidad?

Independientemente del AWS PCS, Amazon EC2 envía un `Capacity Block Reservation Delivered` evento EventBridge cuando se inicia una reserva de bloque de capacidad y un `Capacity Block Reservation Expiration Warning` evento 40 minutos antes de que caduque la reserva de bloque de capacidad. Para obtener más información, consulte Supervisar los bloques de capacidad EventBridge en la Guía del usuario de Amazon Elastic Compute Cloud.

¿Cómo rastrea Slurm el estado de mi bloque de capacidad?

Puede correr `sinfo` para entender cómo AWS PCS utiliza el bloque de capacidad. En el siguiente resultado de ejemplo, se asocia una cola a un grupo de nodos de cómputo que ejecuta 4 instancias desde un bloque de `active` capacidad. Los nodos están en el estado `idle` Slurm (están disponibles para su uso y aún no están asignados a ningún trabajo).

```
$ sinfo
PARTITION AVAIL TIMELIMIT NODES STATE NODELIST
fanout up infinite 4 idle node-fanout-[1-4]
```

Si, por el contrario, los nodos están en `maint` estado, puedes ir `scontrol show res` a ver los detalles sobre la reserva de Slurm que controla este estado. En el siguiente ejemplo de salida, el bloque de capacidad tiene `scheduled` una fecha de inicio futura.

```
$ scontrol show res

ReservationName=node-fanout-scheduled StartTime=2025-10-14T13:09:17
 EndTime=2025-10-14T13:11:17 Duration=00:02:00
```

```
   Nodes=node-fanout-[1-4] NodeCnt=4 CoreCnt=16 Features=(null) PartitionName=(null)
 Flags=MAINT,SPEC_NODES
   TRES=cpu=16

   Users=root Groups=(null) Accounts=(null) Licenses=(null) State=ACTIVE
 BurstBuffer=(null)
   MaxStartDelay=(null)

   Comment=node-fanout Scheduled
```

¿Cómo puedo saber si los errores que recibo al lanzar la capacidad se deben a que mi bloque de capacidad está compartido?

Compruebe las reservas de capacidad en la consola de Amazon EC2 para averiguar cuántas instancias del bloque de capacidad están aprovisionadas activamente. Compruebe las etiquetas de cada instancia para saber qué servicio o clúster la utiliza. Por ejemplo, todas las instancias de AWS PCS tienen etiquetas de AWS PCS, como las `aws:pcs:cluster-id = pcs_l0mizqyk5o | aws:pcs:compute-node-group-id = pcs_ic7onkmfqk` que indican a qué clústeres y grupos de nodos de cómputo pertenece la instancia. A continuación, puede comprobar si el bloque de capacidad está al máximo de su capacidad.

`scontrol show nodes`Para comprobar si un nodo de bloque de capacidad de un clúster de AWS PCS se está activando`ReservationCapacityExceeded`:

```
[root@ip-172-16-10-54 ~]# scontrol show nodes test-node-8-gamma-cb-2
NodeName=test-8-gamma-cb-2 CoresPerSocket=1
   CPUAlloc=0 CPUEfctv=8 CPUTot=8 CPULoad=0.00
   AvailableFeatures=test-8-gamma-cb,gpu
   ActiveFeatures=test-8-gamma-cb,gpu
   Gres=gpu:H100:1
   NodeAddr=test-8-gamma-cb-2 NodeHostName=test-8-gamma-cb-2
   RealMemory=249036 AllocMem=0 FreeMem=N/A Sockets=8 Boards=1
   State=IDLE+CLOUD+POWERING_DOWN ThreadsPerCore=1 TmpDisk=0 Weight=1 Owner=N/A
 MCS_label=N/A
   Partitions=my-q
   BootTime=None SlurmdStartTime=None
   LastBusyTime=Unknown ResumeAfterTime=None
   CfgTRES=cpu=8,mem=249036M,billing=8
   AllocTRES=
   CurrentWatts=0 AveWatts=0
   Reason=Failed to launch backing instance (Error Code:
 ReservationCapacityExceeded) [root@2025-08-28T15:15:33]
```

Cuando hay varios grupos de nodos de cómputo conectados a la misma cola, ¿cómo puedo forzar la ejecución de un trabajo en instancias respaldadas por Capacity Block?

Puedes usar las funciones y restricciones de Slurm para bloquear un trabajo en un determinado conjunto de nodos. Le recomendamos que no establezca ponderaciones de Slurm para cada grupo de nodos de cómputo, ya que eso solo funciona con los nodos que no están en ese estado. `maint`

Parámetros útiles de la plantilla de lanzamiento

En esta sección se describen algunos parámetros de la plantilla de lanzamiento que pueden resultar muy útiles con el AWS PCS.

Active la CloudWatch supervisión detallada

Puede habilitar la recopilación de CloudWatch métricas en un intervalo más corto mediante un parámetro de plantilla de lanzamiento.

Consola de administración de AWS

En las páginas de la consola para crear o editar plantillas de lanzamiento, esta opción se encuentra en la sección de detalles avanzados. Configure la CloudWatch supervisión detallada en Activar.

YAML

```
Monitoring:
    Enabled: True
```

JSON

```
{"Monitoring": {"Enabled": "True"}}
```

Para obtener más información, consulte Habilitar o desactivar la supervisión detallada de sus instancias en la Guía del usuario de Amazon Elastic Compute Cloud para instancias de Linux.

Instance Metadata Service, versión 2 (IMDS v2)

El uso de IMDS v2 con instancias EC2 ofrece importantes mejoras de seguridad y ayuda a mitigar los posibles riesgos asociados al acceso a los metadatos de las instancias en los entornos. AWS

Consola de administración de AWS

En las páginas de la consola para crear o editar plantillas de lanzamiento, esta opción se encuentra en la sección de detalles avanzados. Configura los metadatos accesibles en Habilitados, la versión de metadatos en V2 únicamente (se requiere un token) y el límite de saltos de respuesta de los metadatos en 4.

YAML

```
MetadataOptions:
  HttpEndpoint: enabled
  HttpTokens: required
  HttpPutResponseHopLimit: 4
```

JSON

```
{
    "MetadataOptions": {
        "HttpEndpoint": "enabled",
        "HttpPutResponseHopLimit": 4,
        "HttpTokens": "required"
    }
}
```

.

AWS Colas PCS

Una cola de AWS PCS es una abstracción ligera de la implementación nativa de una cola de trabajos por parte del programador. En el caso de Slurm, una cola AWS PCS equivale a una partición Slurm.

Los usuarios envían los trabajos a una cola en la que residen hasta que puedan programarse para que se ejecuten en los nodos proporcionados por uno o más grupos de nodos de procesamiento. Un clúster de AWS PCS puede tener varias colas de trabajos. Por ejemplo, puede crear una cola que utilice Amazon EC2 On-Demand Instances para los trabajos de alta prioridad y otra cola que utilice Amazon EC2 Spot Instances para los trabajos de baja prioridad.

Temas

Creación de una cola en PCS AWS

En este tema se proporciona una visión general de las opciones disponibles y se describe lo que se debe tener en cuenta al crear una cola en AWS PCS.

> **Note**
>
> Puede configurar los ajustes de Slurm personalizados en las colas para implementar políticas de programación específicas de las particiones y la administración de recursos. Para obtener más información, consulte Configuración de ajustes de Slurm personalizados en PCS AWS.

Requisitos previos

- Un clúster de AWS PCS: las colas solo se pueden crear en asociación con un clúster de PCS específico. AWS
- Uno o más grupos de nodos de cómputo de AWS PCS: una cola debe estar asociada a al menos un grupo de nodos de cómputo de AWS PCS.

Para crear una cola en PCS AWS

Puede crear una cola utilizando el Consola de administración de AWS o el. AWS CLI

Consola de administración de AWS

Para crear una cola mediante la consola

1. Abra la consola AWS PCS.
2. Seleccione el clúster para la cola. Navegue hasta Colas y elija Crear cola.
3. En la sección de configuración de colas, proporciona los siguientes valores:
   a. Nombre de la cola: nombre de la cola. El nombre solo puede contener caracteres alfanuméricos (con distinción de mayúsculas y minúsculas) y guiones. Debe empezar por un carácter alfabético y no puede tener más de 25 caracteres. El nombre debe ser único en el clúster.
   b. Grupos de nodos de cómputo: seleccione uno o más grupos de nodos de cómputo para dar servicio a esta cola. Un grupo de nodos de cómputo se puede asociar a más de una cola.
4. (Opcional) En la sección Configuración adicional del programador, puede añadir pares de nombre y valor del parámetro para configurar ajustes adicionales de Slurm. Para obtener una lista completa de los parámetros compatibles, consulte. Configuración de Slurm personalizada para colas de PCS AWS
5. (Opcional) En Etiquetas, añada cualquier etiqueta a su cola de AWS PCS
6. Elige Crear cola. El campo Estado mostrará la opción Crear mientras AWS PCS crea la cola. La creación de la cola puede tardar varios minutos.

Siguiente paso recomendado

- Envía un trabajo a tu nueva lista.

AWS CLI

Para crear una cola mediante AWS CLI

Utilice el siguiente comando para crear la cola. Realice las siguientes sustituciones:

1. Reemplace *region-code* por la AWS región del clúster. Por ejemplo, `us-east-1`.

2. *my-queue*Sustitúyalo por el nombre de la cola. El nombre solo puede contener caracteres alfanuméricos (con distinción de mayúsculas y minúsculas) y guiones. Debe empezar por un carácter alfabético y no puede tener más de 25 caracteres. El nombre debe ser único en el clúster.
3. *my-cluster*Sustitúyalo por el nombre o el ID del clúster.
4. *compute-node-group-id*Sustitúyalo por el ID del grupo de nodos de procesamiento para dar servicio a la cola. Por ejemplo, `pcs_abcdef12345`.

> **Note**
>
> Al crear una cola, debe proporcionar el ID del grupo de nodos de procesamiento y no su nombre.

```
aws pcs create-queue --region region-code \
    --queue-name my-queue \
    --cluster-identifier my-cluster \
    --compute-node-group-configurations \
    computeNodeGroupId=compute-node-group-id
```

Example— Crear una cola con una configuración de Slurm personalizada

```
aws pcs create-queue --region region-code \
    --queue-name my-queue \
    --cluster-identifier my-cluster \
    --compute-node-group-configurations \
    computeNodeGroupId=compute-node-group-id \
    --slurm-configuration \
    'slurmCustomSettings=[{parameterName=Default,parameterValue=YES}]'
```

Para obtener más información, consulte Configuración de Slurm personalizada para colas de PCS AWS.

La creación de la cola puede tardar varios minutos. Puede consultar el estado de la cola con el siguiente comando. No podrá enviar trabajos a la cola hasta que alcance su estado. `ACTIVE`

```
aws pcs get-queue --region region-code \
    --cluster-identifier my-cluster \
    --queue-identifier my-queue
```

Siguiente paso recomendado

- Envía un trabajo a tu nueva lista

Actualización de una cola de AWS PCS

En este tema se proporciona una visión general de las opciones disponibles y se describe lo que se debe tener en cuenta al actualizar una cola de AWS PCS. Para obtener información sobre la configuración personalizada de Slurm, consulte. Configuración de Slurm personalizada para colas de PCS AWS

Consideraciones a la hora de actualizar una cola de PCS AWS

Las actualizaciones de las colas no afectarán a los trabajos en ejecución, pero es posible que el clúster no pueda aceptar nuevos trabajos mientras se actualiza la cola.

Para actualizar una cola de PCS AWS

Puede usar Consola de administración de AWS o AWS CLI para actualizar una cola.

Consola de administración de AWS

Para actualizar una cola

1. Abra la consola AWS PCS en `https://console.aws.amazon.com/pcs/home#/clusters`
2. Seleccione el clúster en el que desee actualizar una cola.
3. Ve a Colas, ve a la cola que deseas actualizar y, a continuación, selecciona Editar.
4. En la sección de configuración de colas, actualiza cualquiera de los siguientes valores:
   - Grupos de nodos: agregue o elimine grupos de nodos de cómputo de la asociación con la cola.
   - Configuración adicional del programador: añada, modifique o elimine la configuración de Slurm personalizada para la cola. Para obtener más información, consulte Configuración de Slurm personalizada para colas de PCS AWS.
   - Etiquetas: agrega o elimina etiquetas para la cola.
5. Elija Actualizar. El campo Estado mostrará la actualización mientras se aplican los cambios.

> **Important**
>
> Las actualizaciones de las colas pueden tardar varios minutos.

AWS CLI

Para actualizar una cola

1. Actualice la cola con el siguiente comando. Antes de ejecutar el comando, realice los siguientes reemplazos:

   a. *region-code*Sustitúyala por la Región de AWS que deseas crear tu clúster.

   b. *my-queue*Sustitúyalo por el nombre o `computeNodeGroupId` por el de tu cola.

   c. *my-cluster*Sustitúyalo por el nombre o `clusterId` el de tu clúster.

   d. Para cambiar las asociaciones de grupos de nodos de procesamiento, proporciona una lista actualizada de`--compute-node-group-configurations`.

      - Por ejemplo, para añadir un segundo grupo de nodos de procesamiento`computeNodeGroupExampleID2`:

        ```
        --compute-node-group-configurations
         computeNodeGroupId=computeNodeGroupExampleID1,computeNodeGroupId=computeNodeGro
        ```

```
aws pcs update-queue --region region-code \
    --queue-identifier my-queue \
    --cluster-identifier my-cluster \
    --compute-node-group-configurations \
    computeNodeGroupId=computeNodeGroupExampleID1
```

Example— Actualizar una cola con una configuración de Slurm personalizada

```
aws pcs update-queue --region region-code \
    --queue-identifier my-queue \
    --cluster-identifier my-cluster \
    --slurm-configuration \
    'slurmCustomSettings=[{parameterName=Default,parameterValue=YES}]'
```

Para obtener más información, consulte Configuración de Slurm personalizada para colas de PCS AWS.

2. La actualización de la cola puede tardar varios minutos. Puede consultar el estado de la cola con el siguiente comando. No podrá enviar trabajos a la cola hasta que alcance su estado. `ACTIVE`

```
aws pcs get-queue --region region-code \
    --cluster-identifier my-cluster \
    --queue-identifier my-queue
```

Próximos pasos recomendados

- Envía un trabajo a tu lista actualizada.

Eliminar una cola en PCS AWS

En este tema se proporciona una descripción general de cómo eliminar una cola en AWS PCS.

Consideraciones a la hora de eliminar una cola

- Si hay trabajos en ejecución en la cola, el programador los finalizará cuando se elimine la cola. Los trabajos pendientes de la cola se cancelarán. Considere la posibilidad de esperar a que terminen los trabajos de la cola o de forma manual mediante stop/cancel los comandos nativos del programador (como `scancel` los de Slurm).

Eliminación de la cola

Puede usar Consola de administración de AWS o AWS CLI para eliminar una cola.

Consola de administración de AWS

Para eliminar una cola

1. Abra la consola AWS PCS.
2. Seleccione el grupo de la cola.
3. Navegue hasta Colas y seleccione la cola que desee eliminar.

4. Elija Eliminar.
5. Aparece el campo Estado. `Deleting` Puede tardar varios minutos en completarse.

> **Note**
>
> Puede usar los comandos nativos de su programador para confirmar que la cola se ha eliminado. Por ejemplo, usa `sinfo` o `squeue` para Slurm.

AWS CLI

Para eliminar una cola

- Utilice el siguiente comando para eliminar una cola, con estas sustituciones:
  - *region-code*Sustitúyala por la que se encuentra Región de AWS tu clúster.
  - *my-queue*Sustitúyalo por el nombre o el ID de la cola.
  - *my-cluster*Sustitúyalo por el nombre o el ID de tu clúster.

```
aws pcs delete-queue --region region-code \
       --queue-identifier my-queue \
       --cluster-identifier my-cluster
```

Eliminar la cola puede tardar varios minutos.

> **Note**
>
> Puede usar los comandos nativos de su programador para confirmar que la cola se ha eliminado. Por ejemplo, usa `sinfo` o `squeue` para Slurm.

AWS nodos de inicio de sesión PCS

Por lo general, un clúster de AWS PCS necesita al menos un nodo de inicio de sesión para admitir el acceso interactivo y la administración de tareas. Una forma de lograrlo es con un grupo de nodos de cómputo AWS PCS estáticos configurado para funcionar con nodos de inicio de sesión. También puede configurar una instancia EC2 independiente para que actúe como nodo de inicio de sesión.

Temas

Uso de un grupo de nodos de cómputo de AWS PCS para proporcionar nodos de inicio de sesión

En este tema se proporciona una descripción general de las opciones de configuración sugeridas y se describe qué se debe tener en cuenta al utilizar un grupo de nodos de cómputo de AWS PCS para proporcionar un acceso persistente e interactivo a su clúster.

Crear un grupo de nodos de cómputo de AWS PCS para los nodos de inicio de sesión

Operacionalmente, esto no es muy diferente de crear un grupo de nodos de cómputo normal. Sin embargo, hay que tomar algunas decisiones de configuración clave:

- Establezca una configuración de escalado estático de al menos una instancia EC2 del grupo de nodos de procesamiento.
- Elija la opción de compra bajo demanda para evitar la recuperación de sus instancias.
- Elija un nombre informativo para el grupo de nodos de cómputo, como el inicio de sesión.
- Si desea que se pueda acceder a las instancias del nodo de inicio de sesión desde fuera de su VPC, considere la posibilidad de utilizar una subred pública.
- Si pretende permitir el acceso a SSH, la plantilla de lanzamiento necesitará tener un grupo de seguridad que exponga el puerto SSH a las direcciones IP que elija.

- El perfil de instancia de IAM debe tener solo los permisos de AWS que desee que tengan sus usuarios finales. Para obtener más información, consulte Perfiles de instancia de IAM para AWS Parallel Computing Service.
- Considere permitir que AWS Systems Manager Session Manager administre sus instancias de inicio de sesión.
- Considere restringir el acceso a las credenciales de AWS de la instancia solo a los usuarios administrativos
- Seleccione tipos de instancias menos costosos que los de los grupos de nodos de cómputo normales, ya que los nodos de inicio de sesión se ejecutarán de forma continua.
- Use la misma AMI (o una derivada) que para sus otros grupos de nodos de cómputo para garantizar que todas las instancias tengan instalado el mismo software. Para obtener más información sobre la personalización AMIs, consulte Amazon Machine Images (AMIs) para AWS PCS
- Configure el mismo sistema de archivos de red (Amazon EFS, Amazon FSx for Lustre, etc.) para montarlo en los nodos de inicio de sesión que en las instancias informáticas. Para obtener más información, consulte Uso de sistemas de archivos de red con AWS PCS.

Acceda a sus nodos de inicio de sesión

Cuando su nuevo grupo de nodos de cómputo alcance el estado ACTIVO, podrá encontrar las instancias EC2 que ha creado e iniciar sesión en ellas. Para obtener más información, consulte Búsqueda de instancias de grupos de nodos de cómputo en AWS PCS.

Actualización de un grupo de nodos de procesamiento de AWS PCS para los nodos de inicio de sesión

Puede actualizar un grupo de nodos de inicio de sesión mediante UpdateComputeNodeGroup. Como parte del proceso de actualización del grupo de nodos, se sustituirán las instancias en ejecución. Ten en cuenta que esto interrumpirá cualquier sesión de usuario o proceso activo en la instancia. Los trabajos de Slurm en ejecución o en cola no se verán afectados. Para obtener más información, consulte Actualización de un grupo de nodos de cómputo AWS PCS.

También puede editar la plantilla de lanzamiento utilizada por su grupo de nodos de cómputo. Debe utilizarla UpdateComputeNodeGroup para aplicar la plantilla de lanzamiento actualizada al grupo de nodos de procesamiento. Las nuevas instancias de EC2 lanzadas en el grupo de nodos

de procesamiento utilizan la plantilla de lanzamiento actualizada. Para obtener más información, consulte Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

Eliminar un grupo de nodos de cómputo de AWS PCS para los nodos de inicio de sesión

Puede actualizar un grupo de nodos de inicio de sesión mediante el mecanismo de eliminación del grupo de nodos de cálculo de AWS PCS. Las instancias en ejecución se cancelarán como parte de la eliminación del grupo de nodos. Ten en cuenta que esto interrumpirá cualquier sesión de usuario o proceso activo en la instancia. Los trabajos de Slurm en ejecución o en cola no se verán afectados. Para obtener más información, consulte Eliminar un grupo de nodos de cómputo en AWS PCS.

Uso de instancias independientes como nodos de inicio de sesión de AWS PCS

Puede configurar instancias EC2 independientes para que interactúen con el programador Slurm de un clúster de AWS PCS. Esto resulta útil para crear nodos de inicio de sesión, estaciones de trabajo o hosts de administración de flujos de trabajo dedicados que funcionen con clústeres de PCS pero que operen fuera de la administración de AWS PCS. AWS Para ello, cada instancia independiente debe:

1. Tener instalada una versión de software Slurm compatible.
2. Podrá conectarse al punto final Slurmctld del clúster AWS PCS.
3. Configure correctamente el Slurm Auth and Cred Kiosk Daemon (`sackd`) con el punto final y el secreto del clúster de PCS. AWS Para obtener más información, consulte sackd en la documentación de Slurm.

Este tutorial le ayuda a configurar una instancia independiente que se conecta a un clúster de PCS. AWS

Contenido

- Paso 5: Configurar la conexión al clúster de PCS AWS
- Paso 6: (opcional) Pruebe la conexión

Paso 1: Recupere la dirección y el secreto del clúster de AWS PCS de destino

Recupere los detalles sobre el clúster AWS PCS AWS CLI de destino mediante el comando siguiente. Antes de ejecutar el comando, realice los siguientes reemplazos:

- *region-code*Sustitúyalo por el Región de AWS lugar en el que se ejecuta el clúster de destino.
- *cluster-ident*Sustitúyalo por el nombre o identificador del clúster de destino

```
aws pcs get-cluster --region region-code --cluster-identifier cluster-ident
```

El comando devolverá un resultado similar al de este ejemplo.

```
{
    "cluster": {
        "name": "get-started",
        "id": "pcs_123456abcd",
        "arn": "arn:aws:pcs:us-east-1:111122223333:cluster/pcs_123456abcd",
        "status": "ACTIVE",
        "createdAt": "2024-12-17T21:03:52+00:00",
        "modifiedAt": "2024-12-17T21:03:52+00:00",
        "scheduler": {
            "type": "SLURM",
            "version": "25.05"
        },
        "size": "SMALL",
        "slurmConfiguration": {
            "authKey": {
                "secretArn": "arn:aws:secretsmanager:us-east-1:111122223333:secret:pcs!
slurm-secret-pcs_123456abcd-a12ABC",
                "secretVersion": "ef232370-d3e7-434c-9a87-ec35c1987f75"
            }
        },
        "networking": {
            "subnetIds": [
                "subnet-0123456789abcdef0"
```

```
            ],
            "securityGroupIds": [
                "sg-0123456789abcdef0"
            ]
        },
        "endpoints": [
            {
                "type": "SLURMCTLD",
                "privateIpAddress": "10.3.149.220",
                "port": "6817"
            }
        ]
    }
}
```

En este ejemplo, el punto final del controlador Slurm del clúster tiene una dirección IP de `10.3.149.220` y se ejecuta en el puerto. `6817` Se `secretArn` usará en pasos posteriores para recuperar el secreto del clúster. La dirección IP y el puerto se utilizarán en pasos posteriores para configurar el `sackd` servicio.

Paso 2: lanzar una instancia EC2

Para iniciar una instancia de EC2

1. Abra la consola de Amazon EC2.
2. En el panel de navegación, elija Instancias y, a continuación, Iniciar instancias para abrir el nuevo asistente de inicialización de instancias.
3. (Opcional) En la sección Nombre y etiquetas, proporciona un nombre para la instancia, por ejemplo. `PCS-LoginNode` El nombre se asigna a la instancia como etiqueta de recurso (`Name=PCS-LoginNode`).
4. En la sección Imágenes de aplicaciones y sistemas operativos, seleccione una AMI para uno de los sistemas operativos compatibles con AWS PCS. Para obtener más información, consulte Sistemas operativos compatibles.
5. En la sección Tipo de instancia, seleccione un tipo de instancia compatible. Para obtener más información, consulte Tipos de instancias admitidas.
6. En la sección Par de claves, selecciona el par de claves SSH que quieres usar en la instancia.
7. En la sección Configuración de red:

    - Elija Edit (Edición de).

i. Seleccione la VPC de su clúster de AWS PCS.

ii. En Firewall (grupos de seguridad), elija Seleccionar un grupo de seguridad existente.

A. Seleccione un grupo de seguridad que permita el tráfico entre la instancia y el controlador Slurm del clúster AWS PCS de destino. Para obtener más información, consulte Requisitos y consideraciones sobre los grupos de seguridad.

B. (Opcional) Seleccione un grupo de seguridad que permita el acceso SSH entrante a su instancia.

8. En la sección Almacenamiento, configura los volúmenes de almacenamiento según sea necesario. Asegúrese de configurar suficiente espacio para instalar aplicaciones y bibliotecas a fin de habilitar su caso de uso.
9. En Avanzado, elija un rol de IAM que permita el acceso al secreto del clúster. Para obtener más información, consulte Obtén el secreto del cúmulo de Slurm.
10. En el panel de resumen, elija Launch instance.

Paso 3: Instala Slurm en la instancia

Cuando la instancia se haya lanzado y se active, conéctese a ella mediante el mecanismo que prefiera. Use el instalador de Slurm proporcionado por AWS para instalar Slurm en la instancia. Para obtener más información, consulte Instalador de Slurm.

Descarga el instalador de Slurm, descomprímelo y usa el script para instalar Slurm. `installer.sh` Para obtener más información, consulte Paso 3: Instalar Slurm.

Paso 4: Recupere y almacene el secreto del clúster

Estas instrucciones requieren la AWS CLI. Para obtener más información, consulte Instalar o actualizar a la última versión de AWS CLI en la Guía del AWS Command Line Interface usuario de la versión 2.

Guarde el secreto del clúster con los siguientes comandos.

- Cree el directorio de configuración de Slurm.

```
sudo mkdir -p /etc/slurm
```

- Recupere, decodifique y almacene el secreto del clúster. Antes de ejecutar este comando, *region-code* sustitúyalo por la región en la que se ejecuta el clúster de destino y *secret-arn* sustitúyalo por el valor `secretArn` obtenido en el paso 1.

```
aws secretsmanager get-secret-value \
 --region region-code \
 --secret-id 'secret-arn' \
 --version-stage AWSCURRENT \
 --query 'SecretString' \
 --output text | base64 -d | sudo tee /etc/slurm/slurm.key
```

> **Warning**
>
> En un entorno multiusuario, cualquier usuario con acceso a la instancia podría obtener el secreto del clúster si puede acceder al servicio de metadatos de la instancia (IMDS). Esto, a su vez, podría permitirles hacerse pasar por otros usuarios. Considere la posibilidad de restringir el acceso al IMDS únicamente a los usuarios root o administrativos. Como alternativa, considere la posibilidad de utilizar un mecanismo diferente que no dependa del perfil de la instancia para obtener y configurar el secreto.

- Configura la propiedad y los permisos en el archivo de claves de Slurm.

```
sudo chmod 0600 /etc/slurm/slurm.key
sudo chown slurm:slurm /etc/slurm/slurm.key
```

> **Note**
>
> La clave Slurm debe ser propiedad del usuario y del grupo en los que se ejecuta el `sackd` servicio.

Paso 5: Configurar la conexión al clúster de PCS AWS

Para establecer una conexión con el clúster de AWS PCS, ejecútelo `sackd` como un servicio del sistema siguiendo estos pasos.

> **Note**
>
> Si usa Slurm 25.05 o una versión posterior, puede usar un script para configurar su nodo de inicio de sesión para que se conecte a varios clústeres. Para obtener más información, consulte Conexión de un nodo de inicio de sesión independiente a varios clústeres en PCS AWS.

1. Configure el archivo de entorno del `sackd` servicio con el siguiente comando. Antes de ejecutar el comando, sustituya *`ip-address`* y por *`port`* los valores recuperados de los puntos finales en el paso 1.

```
sudo echo "SACKD_OPTIONS='--conf-server=ip-address:port'" > /etc/sysconfig/sackd
```

2. Cree un archivo `systemd` de servicio para gestionar el `sackd` proceso.

```
sudo cat << EOF > /etc/systemd/system/sackd.service
[Unit]
Description=Slurm auth and cred kiosk daemon
After=network-online.target remote-fs.target
Wants=network-online.target
ConditionPathExists=/etc/sysconfig/sackd

[Service]
Type=notify
EnvironmentFile=/etc/sysconfig/sackd
User=slurm
Group=slurm
RuntimeDirectory=slurm
RuntimeDirectoryMode=0755
ExecStart=/opt/aws/pcs/scheduler/slurm-25.05/sbin/sackd --systemd \$SACKD_OPTIONS
ExecReload=/bin/kill -HUP \$MAINPID
KillMode=process
LimitNOFILE=131072
LimitMEMLOCK=infinity
LimitSTACK=infinity

[Install]
WantedBy=multi-user.target
EOF
```

3. Establezca la propiedad del archivo `sackd` de servicio.

```
sudo chown root:root /etc/systemd/system/sackd.service && \
    sudo chmod 0644 /etc/systemd/system/sackd.service
```

4. Habilite el `sackd` servicio.

```
sudo systemctl daemon-reload && sudo systemctl enable sackd
```

5. Inicie el servicio `sackd`.

```
sudo systemctl start sackd
```

Paso 6: (opcional) Pruebe la conexión

Confirme que el `sackd` servicio se esté ejecutando. A continuación, se muestra un resultado de ejemplo. Si hay errores, por lo general aparecen aquí.

```
[root@ip-10-3-27-112 ~]# systemctl status sackd
[x] sackd.service - Slurm auth and cred kiosk daemon
   Loaded: loaded (/etc/systemd/system/sackd.service; enabled; vendor preset: disabled)
   Active: active (running) since Tue 2024-12-17 16:34:55 UTC; 8s ago
 Main PID: 9985 (sackd)
   CGroup: /system.slice/sackd.service
           ##9985 /opt/aws/pcs/scheduler/slurm-25.05/sbin/sackd --systemd --conf-
server=10.3.149.220:6817

Dec 17 16:34:55 ip-10-3-27-112.ec2.internal systemd[1]: Starting Slurm auth and cred
 kiosk daemon...
Dec 17 16:34:55 ip-10-3-27-112.ec2.internal systemd[1]: Started Slurm auth and cred
 kiosk daemon.
Dec 17 16:34:55 ip-10-3-27-112.ec2.internal sackd[9985]: sackd: running
```

Confirme que las conexiones al clúster funcionan mediante comandos del cliente de Slurm como `sinfo` y. `squeue` A continuación, se muestra un ejemplo de la salida de. `sinfo`

```
[root@ip-10-3-27-112 ~]# /opt/aws/pcs/scheduler/slurm-25.05/bin/sinfo
PARTITION AVAIL TIMELIMIT NODES STATE NODELIST
all up infinite 4 idle~ compute-[1-4]
```

También deberías poder enviar trabajos. Por ejemplo, un comando similar a este ejemplo lanzaría un trabajo interactivo en un nodo del clúster.

```
/opt/aws/pcs/scheduler/slurm-25.05/bin/srun --nodes=1 -p all --pty bash -i
```

Conexión de un nodo de inicio de sesión independiente a varios clústeres en PCS AWS

El `pcs-multi-cluster-login-configure.sh` script proporciona una forma automática de configurar varios `sackd` daemons de Slurm en un único nodo de inicio de sesión independiente. Permite que el nodo de inicio de sesión se comunique con varios clústeres. El script automatiza las siguientes operaciones:

- Utiliza las acciones de la API de AWS PCS para obtener información del clúster
- Solicita la clave de autenticación Slurm codificada en base64
- Crea un archivo JWKS de Slurm con la clave de autenticación del clúster
- Configura el `sackd` servicio con puntos finales y puertos del clúster
- Crea un archivo de `systemd` servicio para un daemon específico de un clúster `sackd`
- Genera un script de activación para la configuración del entorno de clústeres
- Activa e inicia el `sackd` servicio

> **Note**
>
> Este script requiere la versión 25.05 o posterior de Slurm.

Slurm ya debe estar instalado en la instancia (lo que equivale al paso 3 del proceso manual). La instancia debe poder llegar a los puntos finales del clúster de destino. El script realiza las operaciones equivalentes a las de los pasos 4 y 5 del proceso de configuración manual. Obtiene automáticamente la información del clúster, configura el `sackd` servicio, crea los archivos de `systemd` servicio necesarios y crea un script de activación que los usuarios pueden usar para configurar su entorno de shell para la interacción del clúster.

Temas

Requisitos previos para el script de configuración del nodo de inicio de sesión multiclúster de AWS PCS

Requisitos del sistema

- Sistema operativo Linux con soporte `systemd`
- Privilegios de root para la configuración del sistema

Comandos y paquetes necesarios

- `bash`— Intérprete de shell (versión 4.0+)
- `curl`— Para la recuperación de AWS metadatos de IMDS v2
- `jq`— Procesador JSON para analizar las respuestas de la API AWS
- `aws`— AWS CLI v2 para ejecutar acciones de la API de AWS PCS y para acceder a Secrets Manager
- `systemctl`— gestión `systemd` de servicios
- `find`— Utilidad de búsqueda del sistema de archivos
- `grep`— Coincidencia de patrones de texto
- `sed`— Editor de secuencias para la manipulación de texto
- `sort`— Utilidad de clasificación de texto
- `tail`— Muestra las últimas líneas de un archivo
- `mkdir`— Creación de directorios
- `chmod`— Cambia los permisos de los archivos
- `chown`— Cambia la propiedad del archivo
- `ldconfig`— Configuración dinámica del enlazador

AWS requisitos

- Un clúster de AWS PCS que ejecute la versión 25.05 o posterior de Slurm
- AWS credenciales configuradas (mediante un rol de IAM, un archivo de credenciales o variables de entorno)
- Permisos para:
  - `pcs:GetCluster`
  - `secretsmanager:GetSecretValue`(si utilizas un secreto alternativo)

Usuarios y grupos del sistema

- El `slurm` usuario y el grupo deben existir en el sistema

Instalación de Slurm

- Slurm debe instalarse en la misma ubicación que los paquetes de instalación de AWS PCS Slurm:

```
/opt/aws/pcs/scheduler/slurm-version
```

AWS Código de script de configuración del nodo de inicio de sesión multiclúster PCS

Guarde el siguiente código fuente en un archivo con el siguiente nombre:

```
pcs-multi-cluster-login-configure.sh
```

Código fuente del script

```
#!/bin/bash
# Copyright Amazon.com, Inc. or its affiliates. All Rights Reserved.

# AWS PCS Multi-Cluster Standalone Login Node Configuration Script
#
# This script configures AWS Parallel Computing Service (PCS) multi-cluster stand alone
 login nodes
# by setting up the Slurm authentication and credential kiosk daemon (sackd)
```

```
# for connecting to remote PCS clusters.
#
# Prerequisites:
# - AWS CLI configured with appropriate permissions
# - Slurm version 25.05 or later
# - Root privileges for system configuration
# - Network connectivity to AWS PCS endpoints


set -eo pipefail

# Function to display usage
usage() {
    echo "Usage: $0 --cluster-identifier <cluster-identifier> [--endpoint-url
 <endpoint-url>]"
    echo "       $0 -h|--help"
}

# Function to display help
help() {
    echo "AWS PCS Multi-Cluster Standalone Login Node Configuration Script"
    echo "================================================"
    echo
    echo "This script configures multi-cluster standalone login node for AWS Parallel
 Computing Service (PCS)"
    echo "by setting up the Slurm authentication and credential kiosk daemon (sackd)."
    echo
    usage
    echo
    echo "Options:"
    echo "  --cluster-identifier <id>    AWS PCS cluster identifier (required)"
    echo "  --endpoint-url <url>         Custom PCS endpoint URL (optional)"
    echo "  -h, --help                   Show this help message"
    echo
    echo "Examples:"
    echo "  $0 --cluster-identifier my-pcs-cluster"
    echo
    echo "Note: This script requires root privileges and Slurm version 25.05 or later."
}

# Function to retrieve authentication key
get_auth_key() {
    if [ "$ALTERNATE_SECRET_RETRIEVAL" = "true" ]; then
        echo "Retrieving authentication key from AWS Secrets Manager..." >&2
```

```
        local auth_key_arn=$(echo "$CLUSTER_INFO" | jq -r
 '.cluster.slurmConfiguration.authKey.secretArn')
        local auth_key_version=$(echo "$CLUSTER_INFO" | jq -r
 '.cluster.slurmConfiguration.authKey.secretVersion')

        if [ "$auth_key_arn" = "null" ] || [ "$auth_key_version" = "null" ]; then
            echo "Error: Auth key information not found in cluster configuration" >&2
            exit 1
        fi

        if ! aws secretsmanager get-secret-value --secret-id "$auth_key_arn" --version-
id "$auth_key_version" --query SecretString --output text --region "$REGION" 2>/dev/
null; then
            echo "Error: Failed to retrieve auth key from Secrets Manager" >&2
            exit 1
        fi
    else
        echo "Please enter the base64-encoded Slurm authentication key:" >&2
        echo -n "Base64 of the Slurm secret key: " >&2
        local key
        read -rs key
        echo >&2
        echo "$key"
    fi
}

# Function to get next available SACKD port
get_next_sackd_port() {
    local exclude_file="$1"
    local port=6918
    local used_ports=()

    # Get all currently used SACKD ports into an array
    while IFS= read -r line; do
        used_ports+=("$line")
    done < <(find /etc/sysconfig -name "sackd-pcs-*" ! -path "$exclude_file" \
             -exec grep SACKD_PORT= '{}' ';' 2>/dev/null | \
             sed 's/.*SACKD_PORT=//' | sort -n)

    # Loop through used ports to find first available port
    for used_port in "${used_ports[@]}"; do
        if [ "$port" -lt "$used_port" ]; then
            break
        elif [ "$port" -eq "$used_port" ]; then
```

```
            ((port++))
        fi
    done

    echo "$port"
}

# Function to configure cluster
configure_cluster() {
    mkdir -p /etc/slurm
    SLURM_JWKS_FILE="/etc/slurm/slurm-${CLUSTER_NAME}.jwks"
    echo '{"keys":
[{"alg":"HS256","kty":"oct","kid":"key-'"${CLUSTER_ID}"'","k":"'"${BASE64_SLURM_KEY}"'"}]}'
 | jq -c '.' > "${SLURM_JWKS_FILE}"

    chmod 0600 "$SLURM_JWKS_FILE"
    chown slurm:slurm "$SLURM_JWKS_FILE"

    SLURM_INSTALL_PATH="/opt/aws/pcs/scheduler/slurm-${SLURM_VERSION}"

    SACKD_RUNTIME_DIRECTORY="/run/slurm-${CLUSTER_NAME}"
    mkdir -p "${SACKD_RUNTIME_DIRECTORY}"
    chown slurm:slurm "${SACKD_RUNTIME_DIRECTORY}"

    mkdir -p /etc/sysconfig
    SACKD_SERVICE_NAME="sackd-pcs-${CLUSTER_NAME}"
    SACKD_SERVICE_ENV="/etc/sysconfig/${SACKD_SERVICE_NAME}"
    SACKD_PORT=$(get_next_sackd_port "$SACKD_SERVICE_ENV")
    cat > "${SACKD_SERVICE_ENV}" << EOF
SACKD_OPTIONS='--conf-server=$ENDPOINTS'
SLURM_SACK_JWKS='$SLURM_JWKS_FILE'
RUNTIME_DIRECTORY='$SACKD_RUNTIME_DIRECTORY'
SACKD_PORT=$SACKD_PORT
EOF

    SACKD_SERVICE_PATH="/etc/systemd/system/${SACKD_SERVICE_NAME}.service"

    cat << EOF > "$SACKD_SERVICE_PATH"
[Unit]
Description=Slurm auth and cred kiosk daemon
After=network-online.target remote-fs.target
Wants=network-online.target
ConditionPathExists=${SACKD_SERVICE_ENV}
```

```
[Service]
Type=notify
EnvironmentFile=${SACKD_SERVICE_ENV}
User=slurm
Group=slurm
RuntimeDirectory=slurm-${CLUSTER_NAME}
RuntimeDirectoryMode=0755
ExecStart=${SLURM_INSTALL_PATH}/sbin/sackd --systemd \$SACKD_OPTIONS
ExecReload=/bin/kill -HUP \$MAINPID
KillMode=process
LimitNOFILE=131072
LimitMEMLOCK=infinity
LimitSTACK=infinity

[Install]
WantedBy=multi-user.target
EOF

    chown root:root "$SACKD_SERVICE_PATH"
    chmod 0644 "$SACKD_SERVICE_PATH"
    systemctl daemon-reload && systemctl enable "$SACKD_SERVICE_NAME"
    systemctl restart "$SACKD_SERVICE_NAME"

    ACTIVATE_SCRIPT="activate-pcs-${CLUSTER_NAME}"
    cat > "$ACTIVATE_SCRIPT" << EOF
# Activate script for Slurm cluster ${CLUSTER_NAME}

# Add Slurm paths
export PATH="${SLURM_INSTALL_PATH}/bin:\$PATH"
export MANPATH="${SLURM_INSTALL_PATH}/share/man:\$MANPATH"
export LD_LIBRARY_PATH="${SLURM_INSTALL_PATH}/lib:\$LD_LIBRARY_PATH"
ldconfig

# Set Slurm configuration
export SLURM_CONF="/run/slurm-${CLUSTER_NAME}/conf/slurm.conf"
export PCS_CLUSTER_NAME="${CLUSTER_NAME}"
export PCS_CLUSTER_IDENTIFIER="${CLUSTER_IDENTIFIER}"
export PCS_CLUSTER_ID="${CLUSTER_ID}"

echo "Activated PCS cluster environment: ${CLUSTER_NAME}"

# Deactivate function
function deactivate-pcs-${CLUSTER_NAME}() {
```

```
    export PATH="\$(echo "\$PATH" | sed -e "s|${SLURM_INSTALL_PATH}/bin:||g" -e "s|:
${SLURM_INSTALL_PATH}/bin||g" -e "s|^${SLURM_INSTALL_PATH}/bin\$||")"
    export MANPATH="\$(echo "\$MANPATH" | sed -e "s|${SLURM_INSTALL_PATH}/share/man:||
g" -e "s|:${SLURM_INSTALL_PATH}/share/man||g" -e "s|^${SLURM_INSTALL_PATH}/share/man\
$||")"
    export LD_LIBRARY_PATH="\$(echo "\$LD_LIBRARY_PATH" | sed -e "s|
${SLURM_INSTALL_PATH}/lib:||g" -e "s|:${SLURM_INSTALL_PATH}/lib||g" -e "s|^
${SLURM_INSTALL_PATH}/lib\$||")"
    unset SLURM_CONF
    unset PCS_CLUSTER_NAME
    unset PCS_CLUSTER_IDENTIFIER
    unset PCS_CLUSTER_ID
    unset -f deactivate-pcs-${CLUSTER_NAME}
    ldconfig
    echo "Deactivated PCS cluster environment: ${CLUSTER_NAME}"
}

export -f deactivate-pcs-${CLUSTER_NAME}

EOF
}

# Main function
main() {
    # Parse arguments
    CLUSTER_IDENTIFIER=""
    PCS_ENDPOINT_URL=""

    while [ "$1" != "" ]; do
        case $1 in
            --cluster-identifier)
                shift
                CLUSTER_IDENTIFIER="$1"
                ;;
            --endpoint-url)
                shift
                PCS_ENDPOINT_URL="--endpoint-url $1"
                ;;
            -h|--help)
                help
                exit 0
                ;;
            *)
                echo "Invalid argument: $1" >&2
```

```
                usage >&2
                exit 1
                ;;
        esac
        shift
    done

    # Validate required arguments
    if [ -z "$CLUSTER_IDENTIFIER" ]; then
        echo "Error: --cluster-identifier is required" >&2
        usage >&2
        exit 1
    fi

    # Validate running as root
    if [ "$EUID" -ne 0 ]; then
        echo "Error: This script must be run as root" >&2
        exit 1
    fi

    # Validate required commands are available
    for cmd in aws jq curl; do
        if ! command -v "$cmd" &> /dev/null; then
            echo "Error: Required command '$cmd' not found" >&2
            exit 1
        fi
    done

    # Get the region name from IMDS v2 with error handling (try IPv6 first, fallback to
 IPv4)
    echo "Retrieving AWS region from instance metadata..."
    # Try IPv6 IMDS endpoint first (fd00:ec2::254) with fast timeout (1s connect, 2s
 total)
    # If IPv6 fails, fallback to IPv4 IMDS endpoint (169.254.169.254)
    IMDS_ENDPOINT="http://[fd00:ec2::254]"
    if ! TOKEN=$(curl -s -X PUT "${IMDS_ENDPOINT}/latest/api/token" -H "X-aws-ec2-
metadata-token-ttl-seconds: 21600" --connect-timeout 1 --max-time 2 2>/dev/null); then
        IMDS_ENDPOINT="http://169.254.169.254"
        if ! TOKEN=$(curl -s -X PUT "${IMDS_ENDPOINT}/latest/api/token" -H "X-aws-ec2-
metadata-token-ttl-seconds: 21600" --max-time 5); then
            echo "Error: Failed to retrieve IMDS token. Ensure this script is running
 on an EC2 instance." >&2
            exit 1
        fi
```

```
    fi

    if ! REGION=$(curl -s -H "X-aws-ec2-metadata-token: $TOKEN" "${IMDS_ENDPOINT}/
latest/dynamic/instance-identity/document" --max-time 5 | jq -r '.region'); then
        echo "Error: Failed to retrieve AWS region from instance metadata" >&2
        exit 1
    fi

    echo "Detected AWS region: $REGION"

    # Retrieve cluster information from AWS PCS
    echo "Retrieving cluster information for: $CLUSTER_IDENTIFIER"
    # shellcheck disable=SC2086
    if ! CLUSTER_INFO=$(aws pcs get-cluster --region "$REGION" --cluster-identifier
 "$CLUSTER_IDENTIFIER" $PCS_ENDPOINT_URL 2>/dev/null); then
        echo "Error: Failed to retrieve cluster information. Check cluster identifier
 and AWS permissions." >&2
        exit 1
    fi

    CLUSTER_ID=$(echo "$CLUSTER_INFO" | jq -r '.cluster.id')
    CLUSTER_NAME="$(echo "$CLUSTER_INFO" | jq -r '.cluster.name')"
    SLURM_VERSION=$(echo "$CLUSTER_INFO" | jq -r '.cluster.scheduler.version')
    SLURM_VERSION=${SLURM_VERSION#Slurm_}

    # Check if Slurm version is >= 25.05
    # shellcheck disable=SC2072
    if [[ "$SLURM_VERSION" < "25.05" ]]; then
        echo "Error: This script requires Slurm version 25.05 or later. Found version:
 $SLURM_VERSION" >&2
        exit 1
    fi

    ENDPOINTS=$(echo "$CLUSTER_INFO" | jq -r '.cluster.endpoints[] | select(.type
 == "SLURMCTLD") | (if .privateIpAddress != "" then .privateIpAddress else "["
 + .ipv6Address + "]" end) + ":" + .port' | tr '\n' ',' | sed 's/,$//')

    # Get BASE64_SLURM_KEY
    BASE64_SLURM_KEY=$(get_auth_key)

    if [ -z "$BASE64_SLURM_KEY" ]; then
        echo "Error: base64 Slurm key cannot be empty" >&2
        exit 1
    fi
```

```
    configure_cluster

    # Final configuration summary
    echo "========================================"
    echo "Configuration completed successfully!"
    echo "========================================"
    echo "Cluster Name: $CLUSTER_NAME"
    echo "Cluster ID: $CLUSTER_ID"
    echo "Slurm Version: $SLURM_VERSION"
    echo "Service Name: $SACKD_SERVICE_NAME"
    echo "SACKD Port: $SACKD_PORT"
    echo
    echo "To activate this cluster environment, run:"
    echo "  source ./$ACTIVATE_SCRIPT"
    echo
    echo "To deactivate this cluster environment, run:"
    echo "  deactivate-pcs-${CLUSTER_NAME}"
    echo
    echo "To check service status:"
    echo "  systemctl status $SACKD_SERVICE_NAME"
    echo
    echo "To view service logs:"
    echo "  journalctl -u $SACKD_SERVICE_NAME -f"
}

# Exit if being sourced for testing
[[ "${BASH_SOURCE[0]}" != "${0}" ]] && return

# Execute main function
main "$@"
```

Uso del script de configuración del nodo de inicio de sesión multiclúster AWS PCS

Ejecución del script

Ejecución del script de configuración

1. Guarde el contenido del script en un archivo denominado:

```
pcs-multi-cluster-login-configure.sh
```

2. Hágalo ejecutable:

```
chmod +x pcs-multi-cluster-login-configure.sh
```

3. Ejecute el script :

```
./pcs-multi-cluster-login-configure.sh --cluster-identifier cluster-name
```

Entornos de interacción en clúster

Tras una configuración correcta, el script genera un script de activación específico del clúster en el directorio actual. El script tiene el nombre. `activate-pcs-`*`cluster-name`* El script de activación configura las variables de entorno y las rutas necesarias para interactuar con el clúster de destino.

Para activar un entorno de clúster

- Utilice el `source` comando para ejecutar el script de activación

```
source ./activate-pcs-cluster-name
```

Example

```
# Activate cluster environment for cluster 'my-cluster'
source ./activate-pcs-my-cluster

# Now you can use Slurm commands
sinfo
squeue
sbatch my-job.sh
```

Qué hace el script de activación

- Establece la variable de `SLURM_CONF` entorno para que apunte a la configuración del clúster.
- Actualiza el `PATH` para incluir los binarios de Slurm del clúster.
- Configura otras variables de entorno de Slurm necesarias (,). `MANPATH` `LD_LIBRARY_PATH`
- Establece las variables de AWS identificación del clúster de PCS.
- Permite una interacción fluida con el clúster de AWS PCS de destino.

Para desactivar un entorno de clúster

- Ejecute el comando de desactivación.

```
deactivate-pcs-cluster-name
```

Example

```
# After activating a cluster
source ./activate-pcs-my-cluster

# Work with the cluster
sinfo

# Deactivate when done
deactivate-pcs-my-cluster
```

Qué hace el comando de desactivación

- Restaura la variable de PATH entorno original.
- Desactiva las variables de entorno de Slurm específicas del clúster.
- Devuelve el entorno del shell a su estado previo a la activación.

> **Note**
>
> La activación es específica de la sesión y debe originarse en la sesión de shell en la que desee interactuar con el clúster.

AWS Redes PCS

El clúster de AWS PCS se crea en una Amazon VPC. En este capítulo se incluyen los siguientes temas sobre las redes para el programador y los nodos del clúster.

A excepción de elegir una subred para lanzar instancias, debe usar plantillas de EC2 lanzamiento para configurar las redes de los grupos de nodos de cómputo de AWS PCS. Para obtener más información acerca de las plantillas de inicialización, consulte Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

Temas

AWS Requisitos y consideraciones sobre la VPC y la subred

Al crear un clúster de AWS PCS, se especifica una VPC, una subred en esa VPC. En este tema se proporciona una descripción general de los requisitos y consideraciones específicos del AWS PCS para la VPC y las subredes que se utilizan con el clúster. Si no tiene una VPC para usarla con AWS PCS, puede crear una mediante una plantilla proporcionada AWS. CloudFormation Para obtener más información VPCs, consulte Nubes privadas virtuales (VPC) en la Guía del usuario de Amazon VPC.

Requisitos y consideraciones de la VPC

Al crear un clúster, la VPC que especifique debe cumplir los siguientes requisitos y consideraciones:

- La VPC debe tener un número suficiente de direcciones IP disponibles para el clúster, los nodos y otros recursos del clúster que desee crear. Para obtener más información, consulte el direccionamiento IP de su red VPCs y de sus subredes en la Guía del usuario de Amazon VPC.
- Si su clúster usa: IPv6

- Asocie un bloque IPv6 CIDR a su VPC. Para obtener más información, consulte Crear una VPC en la Guía del usuario de Amazon VPC.

  > **Important**
  >
  > Si bien puede configurar su VPC con ambos IPv4 tipos de red IPv6, solo puede elegir un tipo de red para el clúster.

- Habilite la asignación automática de IPv6 direcciones para sus subredes.
- Para obtener más información, consulte lo siguiente:
  - IPv6 activado AWS
  - Comprender el IPv6 direccionamiento en AWS y diseñar un plan de direccionamiento escalable

- La VPC debe tener un nombre de host DNS y ser compatible con la resolución de DNS. De lo contrario, los nodos no podrán registrar el clúster de clientes. Para obtener más información, consulte Atributos de DNS para su VPC en la Guía del usuario de Amazon VPC.
- Es posible que la VPC requiera el uso de puntos finales de la VPC AWS PrivateLink para poder contactar con la API de PCS. AWS Para obtener más información, consulte Conectar la VPC a los servicios mediante la Guía del AWS PrivateLink usuario de Amazon VPC.

> **Important**
>
> AWS PCS no admite una VPC con arrendamiento de instancias dedicado. La VPC que utilice para AWS PCS debe utilizar la tenencia de `default` instancias. Puede cambiar la tenencia de la instancia de una VPC existente. Para obtener más información, consulte Cambiar la tenencia de una instancia de una VPC en la Guía del usuario de Amazon Elastic Compute Cloud.

Requisitos y consideraciones de la subred

Al crear un clúster de Slurm, AWS PCS crea una interfaz de red elástica (ENI) en la subred que especificó. Esta interfaz de red permite la comunicación entre el controlador del planificador y la VPC del cliente. La interfaz de red también permite a Slurm comunicarse con los componentes implementados en su cuenta. Solo puede especificar la subred de un clúster en el momento de la creación.

Requisitos de la subred para los clústeres

La subred que especifique al crear un clúster debe cumplir los siguientes requisitos:

- La subred debe tener al menos una dirección IP para que la utilice PCS. AWS
- Si su clúster la usa IPv6, deben usarla todas las subredes del clúster. IPv6

> **Important**
>
> Los grupos de nodos de cómputo configurados con una muestra de AWS PCS AMIs y varias interfaces de red no funcionarán actualmente si las subredes solo están configuradas para su uso. IPv6 En su lugar, utilice subredes de doble pila (IPv4 y IPv6) o IPv4 solo subredes. Para obtener más información, consulte Uso de Amazon Machine Images (AMIs) de muestra con AWS PCS.

- La subred no puede residir en AWS Outposts, AWS Wavelength o en una zona local. AWS
- La subred puede ser pública o privada. Le recomendamos que especifique una subred privada, si es posible. Una subred pública es una subred con una tabla de enrutamiento que incluye una ruta a una puerta de enlace a Internet; una subred privada es una subred con una tabla de enrutamiento que no incluye una ruta a una puerta de enlace a Internet.

Requisitos de la subred para los nodos

Puede implementar nodos y otros recursos de clúster en la subred que especifique al crear el clúster de AWS PCS y en otras subredes de la misma VPC.

Cualquier subred en la que implementes nodos y recursos de clúster debe cumplir los siguientes requisitos:

- Debe asegurarse de que la subred tenga suficientes direcciones IP disponibles para implementar todos los nodos y los recursos del clúster.
- Si tu clúster usa nodos en una subred pública IPv4 y planeas implementarlos, esa subred debe asignar automáticamente IPv4 direcciones públicas.

> **Note**
>
> Las instancias de una subred pública deben usar un grupo de seguridad con reglas de entrada que permitan el tráfico desde direcciones IP públicas. A menos que tenga restricciones de dirección de origen específicas, se trata de una dirección de IPv4 origen de 0.0.0.0/0 o una IPv6 dirección de origen de: :/0.

- Si la subred en la que despliega los nodos es una subred privada y su tabla de enrutamiento no incluye una ruta a un dispositivo de traducción de direcciones de red (NAT) (IPv4), añada puntos de enlace de VPC a AWS PrivateLink la VPC del cliente. Los puntos finales de VPC son necesarios para todos los AWS servicios con los que contactan los nodos. El único punto final necesario es que el AWS PCS permita que el nodo inicie la acción de la `RegisterComputeNodeGroupInstance` API. Para obtener más información, consulte la referencia RegisterComputeNodeGroupInstancede la API de AWS PCS.
- El estado de la subred pública o privada no afecta al AWS PCS; los puntos finales necesarios deben estar accesibles.

Creación de una VPC para su AWS clúster de PCS

Puede crear una Amazon Virtual Private Cloud (Amazon VPC) para sus clústeres dentro de AWS Parallel Computing Service (AWS PCS).

Utilice Amazon VPC para lanzar recursos de VPC en una red virtual que haya definido. Esta red virtual es prácticamente idéntica a una red tradicional que podría operar en su propio centro de datos. Sin embargo, incluye los beneficios que supone utilizar la infraestructura escalable de Amazon Web Services. Le recomendamos que conozca a fondo el servicio Amazon VPC antes de implementar clústeres de VPC de producción. Para obtener más información, consulte ¿Qué es Amazon VPC? en el modo visual de autor. Guía del usuario de Amazon VPC.

En su Amazon VPC se implementan un clúster de PCS, nodos y recursos de soporte (como sistemas de archivos y servicios de directorio). Si desea utilizar una Amazon VPC existente con PCS, debe cumplir los requisitos descritos en. AWS Requisitos y consideraciones sobre la VPC y la subred En este tema se describe cómo crear una VPC que cumpla con los requisitos de PCS mediante una plantilla proporcionada AWS. CloudFormation Una vez que haya implementado una plantilla, podrá ver los recursos creados por la plantilla para saber exactamente qué recursos creó y la configuración de esos recursos.

Requisitos previos

Para crear una Amazon VPC para PCS, debe tener los permisos de IAM necesarios para crear recursos de Amazon VPC. Estos recursos son subredes VPCs, grupos de seguridad, tablas y rutas de enrutamiento y puertas de enlace de Internet y NAT. Para obtener más información, consulte Crear una VPC con una subred pública en la Guía del usuario de Amazon VPC. Para revisar la lista completa de Amazon EC2, consulte Acciones, recursos y claves de condición de Amazon EC2 en la Referencia de autorización de servicio.

Creación de una Amazon VPC

Cree una VPC copiando y pegando la URL adecuada para el lugar en el Región de AWS que utilizará la PCS. También puede descargar la CloudFormation plantilla y subirla usted mismo a la CloudFormation consola.

- EE.UU. Este (Virginia) (us-east-1)

```
https://console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/
create/review?stackName=hpc-networking&templateURL=https://aws-hpc-recipes.s3.us-
east-1.amazonaws.com/main/recipes/net/hpc_large_scale/assets/main.yaml
```

- EE.UU. Este (Ohio) (us-east-2)

```
https://console.aws.amazon.com/cloudformation/home?region=us-east-2#/stacks/
create/review?stackName=hpc-networking&templateURL=https://aws-hpc-recipes.s3.us-
east-1.amazonaws.com/main/recipes/net/hpc_large_scale/assets/main.yaml
```

- EE.UU. Oeste (Oregón) (us-west-2)

```
https://console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/
create/review?stackName=hpc-networking&templateURL=https://aws-hpc-recipes.s3.us-
east-1.amazonaws.com/main/recipes/net/hpc_large_scale/assets/main.yaml
```

- Solo plantilla

```
https://aws-hpc-recipes.s3.us-east-1.amazonaws.com/main/recipes/net/hpc_large_scale/
assets/main.yaml
```

Para crear una Amazon VPC para PCS

1. Abra la plantilla en la CloudFormation consola.

   > **Note**
   >
   > Se rellenan previamente en la plantilla, por lo que puede simplemente dejarlos como valores predeterminados.

2. En Proporcione un nombre de pila y, a continuación, en Nombre de pila, introduzcahpc-networking.
3. En Parámetros, introduce los siguientes detalles:
   a. A continuación, en VPC, introduzca CidrBlock10.3.0.0/16
   b. En las subredes A:
      i. Luego CidrPublicSubnetA, introduzca 10.3.0.0/20
      ii. Luego CidrPrivateSubnetA, ingresa 10.3.128.0/20
   c. En las subredes B:
      i. A continuación, CidrPublicSubnetB, introduzca 10.3.16.0/20
      ii. Luego CidrPrivateSubnetA, ingresa 10.3.144.0/20
   d. En las subredes C:
      i. Para ProvisionSubnetsC, seleccioneTrue.

      > **Note**
      >
      > Si va a crear una VPC en una región que tiene menos de tres zonas de disponibilidad, esta opción se ignorará si se establece en. True

      ii. A continuación, CidrPublicSubnetB, introduzca 10.3.32.0/20
      iii. Luego CidrPrivateSubnetA, ingresa 10.3.160.0/20
4. En Capacidades, marca la casilla Acepto que AWS CloudFormation podría crear recursos de IAM.

Supervise el estado de la CloudFormation pila. Cuando llegue`CREATE_COMPLETE`, el recurso de VPC estará listo para su uso.

> **Note**
>
> Para ver todos los recursos que creó la CloudFormation plantilla, abre la CloudFormation consola. Elija la pila `hpc-networking` y, a continuación, elija la pestaña Resources (Recursos).

Grupos de seguridad en AWS PCS

Los grupos de seguridad de Amazon EC2 actúan como firewalls virtuales para controlar el tráfico entrante y saliente a las instancias. Utilice una plantilla de lanzamiento para un grupo de nodos de cómputo de AWS PCS para añadir o eliminar grupos de seguridad en sus instancias. Si la plantilla de lanzamiento no contiene ninguna interfaz de red, utilícela `SecurityGroupIds` para proporcionar una lista de grupos de seguridad. Si la plantilla de lanzamiento define las interfaces de red, debe usar el `Groups` parámetro para asignar grupos de seguridad a cada interfaz de red. Para obtener más información acerca de las plantillas de inicialización, consulte Uso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

> **Note**
>
> Los cambios en la configuración del grupo de seguridad en la plantilla de lanzamiento solo afectan a las nuevas instancias lanzadas una vez actualizado el grupo de nodos de procesamiento.

Requisitos y consideraciones sobre los grupos de seguridad

AWS PCS crea una interfaz de red elástica (ENI) entre cuentas en la subred que especifique al crear un clúster. Esto proporciona al programador de HPC, que se ejecuta en una cuenta administrada por AWS, una ruta para comunicarse con las instancias EC2 lanzadas por PCS. AWS Debe proporcionar un grupo de seguridad para ese ENI que permita la comunicación bidireccional entre el ENI del programador y las instancias EC2 del clúster.

Una forma sencilla de lograrlo consiste en crear un grupo de seguridad autorreferenciado permisivo que permita el TCP/IP tráfico en todos los puertos entre todos los miembros del grupo. Puede adjuntarlo a las instancias EC2 del clúster y del grupo de nodos.

Ejemplo de configuración de grupo de seguridad permisiva

IPv4

Tipo de regla	Protocolos	Puertos	Origen	Destino
Entrada	Todos	Todos	Auto	
Salida	Todos	Todos		0.0.0.0/0
Salida	Todos	Todos		Auto

IPv6

Tipo de regla	Protocolos	Puertos	Origen	Destino
Entrada	Todos	Todos	Auto	
Salida	Todos	Todos		::/0
Salida	Todos	Todos		Auto

Estas reglas permiten que todo el tráfico fluya libremente entre el controlador Slurm y los nodos, permiten que todo el tráfico saliente se dirija a cualquier destino y habilitan el tráfico EFA.

Ejemplo de configuración restrictiva de un grupo de seguridad

También puede limitar los puertos abiertos entre el clúster y sus nodos de procesamiento. En el caso del programador Slurm, el grupo de seguridad adjunto al clúster debe permitir los siguientes puertos:

- 6817: habilita las conexiones entrantes desde instancias EC2 `slurmctld`
- 6818: habilita las conexiones salientes desde `slurmctld` y hasta que se ejecuten en las instancias EC2 `slurmd`

El grupo de seguridad adjunto a sus nodos de cómputo debe permitir los siguientes puertos:

- 6817: habilita las conexiones salientes `slurmctld` desde instancias EC2.
- 6818: habilita las conexiones entrantes y salientes desde y hacia las instancias `slurmd` del grupo de nodos `slurmctld` `slurmd`
- 60001—63000: conexiones entrantes y salientes entre instancias de grupos de nodos para admitir `srun`
- Tráfico de EFA entre instancias de grupos de nodos. Para obtener más información, consulte Preparar un grupo de seguridad habilitado para EFA en la Guía del usuario para instancias de Linux
- Cualquier otro tráfico entre nodos que requiera su carga de trabajo

Múltiples interfaces de red en AWS PCS

Algunas instancias EC2 tienen varias tarjetas de red. Esto les permite ofrecer un mayor rendimiento de la red, incluidas capacidades de ancho de banda superiores a 100 Gbps y una mejor gestión de paquetes. Para obtener más información sobre las instancias con varias tarjetas de red, consulte Interfaces de red elásticas en la Guía del usuario de Amazon Elastic Compute Cloud.

Configure tarjetas de red adicionales para las instancias de un grupo de nodos de cómputo de AWS PCS añadiendo interfaces de red a su plantilla de lanzamiento de EC2. A continuación, se muestra un ejemplo de plantilla de lanzamiento que habilita dos tarjetas de red, como las que se encuentran en una `hpc7a.96xlarge` instancia. Tenga en cuenta la siguiente información:

- La subred de cada interfaz de red debe ser la misma que la elegida al configurar el grupo de nodos de cómputo del AWS PCS que utilizará la plantilla de lanzamiento.
- El dispositivo de red principal, donde se producirá la comunicación de red rutinaria, como el tráfico SSH y HTTPS, se establece configurando una `DeviceIndex` de. `0` Otras interfaces de red tienen un `DeviceIndex` de`1`. Solo puede haber una interfaz de red principal; todas las demás interfaces son secundarias.
- Todas las interfaces de red deben tener una única. `NetworkCardIndex` Una práctica recomendada es numerarlas secuencialmente tal como se definen en la plantilla de lanzamiento.
- Los grupos de seguridad para cada interfaz de red se configuran mediante`Groups`. En este ejemplo, se agrega un grupo de seguridad SSH entrante (`sg-`*`SshSecurityGroupId`*) a la interfaz de red principal, así como un grupo de seguridad que permite las comunicaciones dentro

del clúster (). `sg-ClusterSecurityGroupId` Por último, se agrega un grupo de seguridad que permite las conexiones salientes a Internet (`sg-InternetOutboundSecurityGroupId`) a las interfaces principal y secundaria.

```
{
    "NetworkInterfaces": [
        {
            "DeviceIndex": 0,
            "NetworkCardIndex": 0,
            "SubnetId": "subnet-SubnetId",
            "Groups": [
               "sg-SshSecurityGroupId",
               "sg-ClusterSecurityGroupId",
               "sg-InternetOutboundSecurityGroupId"
            ]
        },
        {
            "DeviceIndex": 1,
            "NetworkCardIndex": 1,
            "SubnetId": "subnet-SubnetId",
            "Groups": ["sg-InternetOutboundSecurityGroupId"]
        }
    ]
}
```

Grupos de ubicación para instancias EC2 en PCS AWS

Puede utilizar un grupo de ubicación para influir en la ubicación de las instancias de EC2 y adaptarlas a las necesidades de la carga de trabajo que se ejecuta en ellas.

Tipos de grupos de ubicación

- Clúster: agrupa las instancias juntas en una zona de disponibilidad para optimizar la comunicación de baja latencia.
- Partición: distribuye las instancias entre las particiones lógicas para ayudar a maximizar la resiliencia.
- Distribución: exige estrictamente que un número reducido de instancias se lance en un hardware distinto, lo que también contribuye a aumentar la resiliencia.

Para obtener más información, consulte Grupos de ubicación para sus instancias de Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud.

Se recomienda incluir un grupo de ubicación de clústeres al configurar un grupo de nodos de procesamiento de AWS PCS para usar Elastic Fabric Adapter (EFA).

Para crear un grupo de ubicación en clústeres que funcione con EFA

1. Cree un grupo de ubicación con el tipo cluster para el grupo de nodos de procesamiento.
   - Utilice el siguiente AWS CLI comando:

     ```
     aws ec2 create-placement-group --strategy cluster --group-name PLACEMENT-GROUP-NAME
     ```

   - También puede utilizar una CloudFormation plantilla para crear un grupo de ubicaciones. Para obtener más información, consulte Trabajar con CloudFormation plantillas en la Guía del AWS CloudFormation usuario. Descargue la plantilla desde la siguiente URL y cárguela en la CloudFormation consola.

     ```
     https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/enable_efa/assets/efa-placement-group.yaml
     ```

2. Incluya el grupo de ubicación en la plantilla de lanzamiento de EC2 para el grupo de nodos de cómputo de AWS PCS.

Uso del Elastic Fabric Adapter (EFA) con PCS AWS

El Elastic Fabric Adapter (EFA) es una interconexión de red avanzada de alto rendimiento AWS que puede conectar a su instancia EC2 para acelerar las aplicaciones de computación de alto rendimiento (HPC) y aprendizaje automático. Para permitir que sus aplicaciones se ejecuten en un clúster de AWS PCS con EFA, es necesario configurar las instancias del grupo de nodos de cómputo del AWS PCS para que utilicen EFA de la siguiente manera.

> **Note**
>
> Instalación de EFA en una AMI AWS compatible con PCS: la AMI utilizada en AWS el grupo de nodos de cómputo de PCS debe tener el controlador EFA instalado y cargado. Para obtener información sobre cómo crear una AMI personalizada con el software EFA instalado, consulteImágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Contenido

Identifique las instancias EC2 habilitadas para EFA

Para usar EFA, todos los tipos de instancias permitidos para un grupo de cómputo de AWS PCS deben admitir EFA y deben tener el mismo número de v CPUs (y si corresponde). GPUs Para obtener una lista de instancias habilitadas para EFA, consulte Elastic Fabric Adapter para cargas de trabajo de HPC y ML en Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud. También puede utilizarla AWS CLI para ver una lista de los tipos de instancias que admiten la EFA. *region-code*Sustitúyala por la Región de AWS que utilices AWS PCS, por ejemplo`us-east-1`.

```
aws ec2 describe-instance-types \
   --region region-code \
   --filters Name=network-info.efa-supported,Values=true \
   --query "InstanceTypes[*].[InstanceType]" \
   --output text | sort
```

> **Note**
>
> Determine cuántas interfaces de red están disponibles: algunas instancias EC2 tienen varias tarjetas de red. Esto les permite tener varias EFAs. Para obtener más información, consulte Múltiples interfaces de red en AWS PCS.

Cree un grupo de seguridad para respaldar las comunicaciones de la EFA

AWS CLI

Puede usar el siguiente AWS CLI comando para crear un grupo de seguridad que admita la EFA. El comando genera un ID de grupo de seguridad. Realice las siguientes sustituciones:

- *region-code*— Especifique Región de AWS dónde va a utilizar el AWS PCS, por ejemplo. `us-east-1`
- *vpc-id*— Especifique el ID de la VPC que utiliza para AWS PCS.
- *efa-group-name*— Proporcione el nombre que elija para el grupo de seguridad.

```
aws ec2 create-security-group \
    --group-name efa-group-name \
    --description "Security group to enable EFA traffic" \
    --vpc-id vpc-id \
    --region region-code
```

Utilice los siguientes comandos para adjuntar las reglas de los grupos de seguridad entrantes y salientes. Realice la siguiente sustitución:

- *efa-secgroup-id*— Proporcione el ID del grupo de seguridad de EFA que acaba de crear.

```
aws ec2 authorize-security-group-ingress \
    --group-id efa-secgroup-id \
    --protocol -1 \
    --source-group efa-secgroup-id

aws ec2 authorize-security-group-egress \
    --group-id efa-secgroup-id \
    --protocol -1 \
    --source-group efa-secgroup-id
```

CloudFormation template

Puede usar una CloudFormation plantilla para crear un grupo de seguridad que admita la EFA. Descargue la plantilla desde la siguiente URL y cárguela en la AWS CloudFormation consola.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/enable_efa/assets/efa-sg.yaml
```

Con la plantilla abierta en la AWS CloudFormation consola, introduce las siguientes opciones.

- En Proporcione un nombre de pila
  - En Nombre de pila, introduce un nombre como`efa-sg-stack`.
- En Parámetros
  - En SecurityGroupName, introduzca un nombre como`efa-sg`.
  - En VPC, seleccione la VPC en la que utilizará PCS. AWS

Termine de crear la CloudFormation pila y supervise su estado. Cuando llegue al EFA, `CREATE_COMPLETE` el grupo de seguridad estará listo para su uso.

(Opcional) Cree un grupo de ubicación

Se recomienda lanzar todas las instancias que utilizan la EFA en un grupo de ubicación en clústeres para minimizar la distancia física entre ellas. Cree un grupo de ubicación para cada grupo de nodos de cómputo en el que vaya a usar EFA. Consulte esta Grupos de ubicación para instancias EC2 en PCS AWS sección para crear un grupo de ubicación para su grupo de nodos de cómputo.

Cree o actualice una plantilla de lanzamiento de EC2

Las interfaces de red EFA se configuran en la plantilla de lanzamiento de EC2 para un grupo de nodos de cómputo del AWS PCS. Si hay varias tarjetas de red, se EFAs pueden configurar varias. El grupo de seguridad EFA y el grupo de ubicación opcional también se incluyen en la plantilla de lanzamiento.

A continuación, se muestra un ejemplo de plantilla de lanzamiento para instancias con dos tarjetas de red, como hpc7a.96xlarge. Las instancias se lanzarán en un grupo de ubicación en clústeres. `subnet-`*`SubnetID1`* `pg-`*`PlacementGroupId1`*

Los grupos de seguridad se deben agregar específicamente a cada interfaz EFA. Cada EFA necesita el grupo de seguridad que habilita el tráfico EFA (). `sg-`*`EfaSecGroupId`* Otros grupos de seguridad, especialmente los que gestionan tráfico normal, como SSH o HTTPS, solo necesitan estar conectados a la interfaz de red principal (designada con un `DeviceIndex` de). `0` Las plantillas de

lanzamiento en las que se definen las interfaces de red no admiten la configuración de grupos de seguridad mediante el `SecurityGroupIds` parámetro; debe establecer un valor para `Groups` cada interfaz de red que configure.

```
{
    "Placement": {
        "GroupId": "pg-PlacementGroupId1"
    },
    "NetworkInterfaces": [
        {
            "DeviceIndex": 0,
            "InterfaceType": "efa",
            "NetworkCardIndex": 0,
            "SubnetId": "subnet-SubnetId1",
            "Groups": [
                "sg-SecurityGroupId1",
                "sg-EfaSecGroupId"
            ]
        },
        {
            "DeviceIndex": 1,
            "InterfaceType": "efa",
            "NetworkCardIndex": 1,
            "SubnetId": "subnet-SubnetId1"
            "Groups": ["sg-EfaSecGroupId"]
        }
    ]
}
```

Cree o actualice grupos de nodos de cómputo para EFA

Los grupos de nodos de cómputo de AWS PCS deben contener instancias que tengan el mismo número de vCPUs, arquitectura de procesador y compatibilidad con EFA. Configure el grupo de nodos de procesamiento para usar la AMI con el software EFA instalado y para usar la plantilla de lanzamiento que configura las interfaces de red habilitadas para EFA.

(Opcional) Pruebe la EFA

Para demostrar la comunicación habilitada por EFA entre dos nodos de un grupo de nodos de cómputo, ejecute el `fi_pingpong` programa, que se incluye en la instalación del software EFA. Si esta prueba se realiza correctamente, es probable que el EFA esté configurado correctamente.

Para empezar, necesita dos instancias en ejecución en el grupo de nodos de cómputo. Si tu grupo de nodos de cómputo usa capacidad estática, ya debería haber instancias disponibles. En el caso de un grupo de nodos de cómputo que utilice capacidad dinámica, puede lanzar dos nodos mediante el `salloc` comando. A continuación, se muestra un ejemplo de un clúster con un nombre de grupo de nodos dinámico `hpc7g` asociado a una cola denominada`all`.

```
% salloc --nodes 2 -p all
salloc: Granted job allocation 6
salloc: Waiting for resource configuration
... a few minutes pass ...
salloc: Nodes hpc7g-[1-2] are ready for job
```

Averigüe la dirección IP de los dos nodos asignados mediante`scontrol`. En el siguiente ejemplo, las direcciones son `10.3.140.69` para `hpc7g-1` y `10.3.132.211` para`hpc7g-2`.

```
% scontrol show nodes hpc7g-[1-2]
NodeName=hpc7g-1 Arch=aarch64 CoresPerSocket=1
   CPUAlloc=0 CPUEfctv=64 CPUTot=64 CPULoad=0.00
   AvailableFeatures=hpc7g
   ActiveFeatures=hpc7g
   Gres=(null)
   NodeAddr=10.3.140.69 NodeHostName=ip-10-3-140-69 Version=25.05.4
   OS=Linux 5.10.218-208.862.amzn2.aarch64 #1 SMP Tue Jun 4 16:52:10 UTC 2024
   RealMemory=124518 AllocMem=0 FreeMem=110763 Sockets=64 Boards=1
   State=IDLE+CLOUD ThreadsPerCore=1 TmpDisk=0 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=efa
   BootTime=2024-07-02T19:00:09 SlurmdStartTime=2024-07-08T19:33:25
   LastBusyTime=2024-07-08T19:33:25 ResumeAfterTime=None
   CfgTRES=cpu=64,mem=124518M,billing=64
   AllocTRES=
   CapWatts=n/a
   CurrentWatts=0 AveWatts=0
   ExtSensorsJoules=n/a ExtSensorsWatts=0 ExtSensorsTemp=n/a
   Reason=Maintain Minimum Number Of Instances [root@2024-07-02T18:59:00]
   InstanceId=i-04927897a9ce3c143 InstanceType=hpc7g.16xlarge

NodeName=hpc7g-2 Arch=aarch64 CoresPerSocket=1
   CPUAlloc=0 CPUEfctv=64 CPUTot=64 CPULoad=0.00
   AvailableFeatures=hpc7g
   ActiveFeatures=hpc7g
   Gres=(null)
   NodeAddr=10.3.132.211 NodeHostName=ip-10-3-132-211 Version=25.05.4
```

```
   OS=Linux 5.10.218-208.862.amzn2.aarch64 #1 SMP Tue Jun 4 16:52:10 UTC 2024
   RealMemory=124518 AllocMem=0 FreeMem=110759 Sockets=64 Boards=1
   State=IDLE+CLOUD ThreadsPerCore=1 TmpDisk=0 Weight=1 Owner=N/A MCS_label=N/A
   Partitions=efa
   BootTime=2024-07-02T19:00:09 SlurmdStartTime=2024-07-08T19:33:25
   LastBusyTime=2024-07-08T19:33:25 ResumeAfterTime=None
   CfgTRES=cpu=64,mem=124518M,billing=64
   AllocTRES=
   CapWatts=n/a
   CurrentWatts=0 AveWatts=0
   ExtSensorsJoules=n/a ExtSensorsWatts=0 ExtSensorsTemp=n/a
   Reason=Maintain Minimum Number Of Instances [root@2024-07-02T18:59:00]
   InstanceId=i-0a2c82623cb1393a7 InstanceType=hpc7g.16xlarge
```

Conéctese a uno de los nodos (en este caso de ejemplo`hpc7g-1`) mediante SSH (o SSM). Tenga en cuenta que se trata de una dirección IP interna, por lo que puede que necesite conectarse desde uno de sus nodos de inicio de sesión si utiliza SSH. Ten en cuenta también que la instancia debe configurarse con una clave SSH mediante la plantilla de lanzamiento del grupo de nodos de cómputo.

```
% ssh ec2-user@10.3.140.69
```

Ahora, `fi_pingpong` ejecútala en modo servidor.

```
/opt/amazon/efa/bin/fi_pingpong -p efa
```

Conéctese a la segunda instancia (`hpc7g-2`).

```
% ssh ec2-user@10.3.132.211
```

Se ejecuta `fi_pingpong` en modo cliente, conectándose al servidor activado`hpc7g-1`. Debería ver un resultado parecido al del ejemplo siguiente.

```
% /opt/amazon/efa/bin/fi_pingpong -p efa 10.3.140.69

bytes   #sent   #ack     total       time     MB/sec    usec/xfer   Mxfers/sec
64      10      =10      1.2k        0.00s      3.08      20.75       0.05
256     10      =10      5k          0.00s     21.24      12.05       0.08
1k      10      =10      20k         0.00s     82.91      12.35       0.08
4k      10      =10      80k         0.00s    311.48      13.15       0.08
```

```
[error] util/pingpong.c:1876: fi_close (-22) fid 0
```

(Opcional) Usa una CloudFormation plantilla para crear una plantilla de lanzamiento compatible con la EFA

Como hay varias dependencias para configurar la EFA, se ha proporcionado una CloudFormation plantilla que puede utilizar para configurar un grupo de nodos de cálculo. Admite instancias con hasta cuatro tarjetas de red. Para obtener más información sobre las instancias con varias tarjetas de red, consulte las interfaces de red elásticas en la Guía del usuario de Amazon Elastic Compute Cloud.

Descargue la CloudFormation plantilla desde la siguiente URL y, a continuación, cárguela en la CloudFormation consola en la Región de AWS que utilice el AWS PCS.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/enable_efa/assets/pcs-lt-efa.yaml
```

Con la plantilla abierta en la CloudFormation consola, introduzca los siguientes valores. Tenga en cuenta que la plantilla proporcionará algunos valores de parámetros predeterminados; puede dejarlos como valores predeterminados.

- En Indique un nombre de pila
  - En Nombre de pila, introduce un nombre descriptivo. Le recomendamos que incorpore el nombre que elija para su grupo de nodos de cómputo de AWS PCS, por ejemplo*NODEGROUPNAME*`-efa-lt`.
- En Parámetros
  - En NumberOfNetworkCards, selecciona el número de tarjetas de red en las instancias que estarán en tu grupo de nodos.
  - En VpcId, elija la VPC en la que se implementa el clúster de AWS PCS.
  - En NodeGroupSubnetId, elija la subred de la VPC de su clúster donde se lanzarán las instancias habilitadas para EFA.
  - En PlacementGroupName, deje el campo en blanco para crear un nuevo grupo de ubicación de clústeres para el grupo de nodos. Si ya tiene un grupo de ubicación que quiere usar, introduzca su nombre aquí.
  - En ClusterSecurityGroupId, elija el grupo de seguridad que va a utilizar para permitir el acceso a otras instancias del clúster y a la API de AWS PCS. Muchos clientes eligen el grupo de seguridad predeterminado de su VPC de clúster.

  - En la sección SshSecurityGroupId, indique el ID del grupo de seguridad que esté utilizando para permitir el acceso SSH entrante a los nodos del clúster.
  - Para SshKeyName, seleccione el par de claves SSH para acceder a los nodos de su clúster.
  - Para LaunchTemplateName, introduzca un nombre descriptivo para la plantilla de lanzamiento, por ejemplo. *NODEGROUPNAME*`-efa-lt` El nombre debe ser exclusivo del usuario Cuenta de AWS en el Región de AWS lugar en el que vaya a utilizar AWS PCS.
- En Capacidades
  - Marque la casilla Reconozco que AWS CloudFormation podría crear recursos de IAM.

Supervisa el estado de la CloudFormation pila. Cuando llegue a `CREATE_COMPLETE` la plantilla de lanzamiento estará lista para ser utilizada. Úselo con un grupo de nodos de cómputo del AWS PCS, tal y como se describe anteriormente enCree o actualice grupos de nodos de cómputo para EFA.

Uso de sistemas de archivos de red con AWS PCS

Puede conectar los sistemas de archivos de red a los nodos lanzados en un grupo de nodos de cómputo del Servicio de Computación AWS Paralela (AWS PCS) para proporcionar una ubicación persistente en la que se puedan escribir los datos y los archivos y acceder a ellos. Puede utilizar los sistemas de archivos proporcionados por AWS los servicios, como Amazon Elastic File System (Amazon EFS), Amazon FSx for Lustre, Amazon FSx for NetApp ONTAP, Amazon FSx for OpenZFS y Amazon File Cache. También puede utilizar sistemas de archivos autogestionados, como servidores NFS.

En este tema se describen algunas consideraciones y ejemplos del uso de sistemas de archivos de red con AWS PCS.

Consideraciones sobre el uso de sistemas de archivos de red

Los detalles de implementación de los distintos sistemas de archivos son diferentes, pero hay algunas consideraciones comunes.

- El software del sistema de archivos correspondiente debe estar instalado en la instancia. Por ejemplo, para usar Amazon FSx for Lustre, debe estar presente el Lustre paquete adecuado. Esto se puede lograr incluyéndolo en la AMI del grupo de nodos de cómputo o utilizando un script que se ejecute al arrancar la instancia.
- Debe haber una ruta de red entre el sistema de archivos de red compartido y las instancias del grupo de nodos de procesamiento.
- Las reglas del grupo de seguridad tanto para el sistema de archivos de red compartido como para las instancias del grupo de nodos de procesamiento deben permitir las conexiones a los puertos correspondientes.
- Debe mantener un espacio de nombres de POSIX usuarios y grupos coherente en todos los recursos que acceden a los sistemas de archivos. De lo contrario, los trabajos y los procesos interactivos que se ejecutan en el clúster de PCS podrían producir errores de permisos.
- Los montajes del sistema de archivos se realizan mediante plantillas de EC2 lanzamiento. Los errores o los tiempos de espera al montar un sistema de archivos de red pueden impedir que las instancias estén disponibles para ejecutar tareas. Esto, a su vez, puede generar costes inesperados. Para obtener más información sobre la depuración de plantillas de lanzamiento, consulteUso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

Ejemplos de montajes de red

Puede crear sistemas de archivos con Amazon EFS, Amazon FSx for Lustre, Amazon FSx for NetApp ONTAP, Amazon FSx for OpenZFS y Amazon File Cache. Amplíe la sección correspondiente a continuación para ver un ejemplo de cada montaje de red.

Amazon EFS

Configuración del sistema de archivos

Crear un sistema de archivos de Amazon EFS. Asegúrese de que tenga un objetivo de montaje en cada zona de disponibilidad en la que vaya a lanzar las instancias del grupo de nodos de cómputo de PCS. Asegúrese también de que cada destino de montaje esté asociado a un grupo de seguridad que permita el acceso entrante y saliente desde las instancias del grupo de nodos de cómputo del PCS. Para obtener más información, consulte Montar objetivos y grupos de seguridad en la Guía del usuario de Amazon Elastic File System.

Plantilla de lanzamiento

Añada los grupos de seguridad de la configuración del sistema de archivos a la plantilla de lanzamiento que utilizará para el grupo de nodos de cómputo.

Incluya datos de usuario que utilicen `cloud-config` un mecanismo para montar el sistema de archivos Amazon EFS. Sustituya los siguientes valores de este script por sus propios detalles:

- *`mount-point-directory`*— La ruta en cada instancia en la que va a montar Amazon EFS
- *`filesystem-id`*— El ID del sistema de archivos del sistema de archivos EFS

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
  - amazon-efs-utils

runcmd:
  - mkdir -p /mount-point-directory
  - echo "filesystem-id:/ /mount-point-directory efs tls,_netdev" >> /etc/fstab
```

```
  - mount -a -t efs defaults

--==MYBOUNDARY==--
```

Amazon FSx para Lustre

Configuración del sistema de archivos

Cree un sistema de archivos FSx para Lustre en la VPC en el que utilizará AWS PCS. Para minimizar las transferencias entre zonas, despliéguelo en una subred de la misma zona de disponibilidad donde lanzará la mayoría de las instancias del grupo de nodos de cómputo de PCS. Asegúrese de que el sistema de archivos esté asociado a un grupo de seguridad que permita el acceso entrante y saliente desde las instancias del grupo de nodos de cómputo del PCS. Para obtener más información sobre los grupos de seguridad, consulte Control de acceso al sistema de archivos con Amazon VPC en la Guía del usuario de Amazon FSx for Lustre.

Plantilla de lanzamiento

Incluya los datos de usuario que se utilizan `cloud-config` para montar el sistema de FSx archivos de Lustre. Sustituya los siguientes valores de este script por sus propios detalles:

- *`mount-point-directory`*— La ruta de la instancia en la que quieres montarla FSx para Lustre
- *`filesystem-id`*— El ID del sistema de archivos del sistema de archivos FSx de Lustre
- *`mount-name`*— El nombre de montaje del sistema de FSx archivos para Lustre
- *`region-code`*— El Región de AWS lugar donde se implementa el sistema de archivos FSx for Lustre (debe ser el mismo que su sistema AWS PCS)
- (Opcional)*`latest`*: cualquier versión de FSx for Lustre Lustre compatible

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- amazon-linux-extras install -y lustre=latest
- mkdir -p /mount-point-directory
- mount -t lustre filesystem-id.fsx.region-code.amazonaws.com@tcp:/mount-name /mount-
point-directory
```

```
--==MYBOUNDARY==
```

Amazon FSx para NetApp ONTAP

Configuración del sistema de archivos

Cree un sistema de archivos Amazon FSx for NetApp ONTAP en la VPC en la que utilizará AWS PCS. Para minimizar las transferencias entre zonas, despliéguelo en una subred de la misma zona de disponibilidad donde lanzará la mayoría de las instancias del grupo de nodos de cómputo de AWS PCS. Asegúrese de que el sistema de archivos esté asociado a un grupo de seguridad que permita el acceso entrante y saliente desde las instancias del grupo de nodos de cómputo de AWS PCS. Para obtener más información sobre los grupos de seguridad, consulte Control de acceso al sistema de archivos con Amazon VPC en la Guía del usuario FSx de ONTAP.

Plantilla de lanzamiento

Incluya los datos de usuario que se utilizan `cloud-config` para montar el volumen raíz de un sistema de archivos FSx de ONTAP. Sustituya los siguientes valores de este script por sus propios detalles:

- *`mount-point-directory`*— La ruta de la instancia en la que desea montar su volumen FSx para ONTAP
- *`svm-id`*— El ID de SVM del sistema de archivos FSx ONTAP
- *`filesystem-id`*— El ID del sistema de archivos del sistema de archivos FSx de ONTAP
- *`region-code`*— El Región de AWS lugar donde se implementa el FSx sistema de archivos de ONTAP (debe ser el mismo que el de su sistema AWS PCS)
- *`volume-name`*— El nombre del FSx volumen de ONTAP

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- mkdir -p /mount-point-directory
```

```
- mount -t nfs svm-id.filesystem-id.fsx.region-code.amazonaws.com:/volume-name /mount-point-directory

--==MYBOUNDARY==
```

Amazon FSx para OpenZFS

Configuración del sistema de archivos

Cree un sistema de archivos FSx para OpenZFS en la VPC en el que utilizará PCS. AWS Para minimizar las transferencias entre zonas, impleméntelo en una subred de la misma zona de disponibilidad donde lanzará la mayoría de las instancias del grupo de nodos de cómputo de AWS PCS. Asegúrese de que el sistema de archivos esté asociado a un grupo de seguridad que permita el acceso entrante y saliente desde las instancias del grupo de nodos de cómputo de AWS PCS. Para obtener más información sobre los grupos de seguridad, consulte Administrar el acceso al sistema de archivos con Amazon VPC en la Guía del usuario FSx de OpenZFS.

Plantilla de lanzamiento

Incluya los datos de usuario que se utilizan `cloud-config` para montar el volumen raíz de un sistema de archivos FSx para OpenZFS. Sustituya los siguientes valores de este script por sus propios detalles:

- *`mount-point-directory`*— La ruta de una instancia en la que quieres montar tu recurso compartido FSx para OpenZFS
- *`filesystem-id`*— El ID del sistema de archivos del sistema de archivos FSx de OpenZFS
- *`region-code`*— El Región de AWS lugar donde se implementa el sistema de archivos FSx para OpenZFS (debe ser el mismo que el de su sistema PCS) AWS

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- mkdir -p /mount-point-directory
- mount -t nfs -o noatime,nfsvers=4.2,sync,rsize=1048576,wsize=1048576 filesystem-id.fsx.region-code.amazonaws.com:/fsx/ /mount-point-directory
```

```
--==MYBOUNDARY==
```

Amazon File Cache

Configuración del sistema de archivos

Cree una caché de archivos de Amazon en la VPC en la que utilizará AWS PCS. Para minimizar las transferencias entre zonas, elija una subred en la misma zona de disponibilidad en la que lanzará la mayoría de las instancias del grupo de nodos de cómputo de PCS. Asegúrese de que la caché de archivos esté asociada a un grupo de seguridad que permita el tráfico entrante y saliente en el puerto 988 entre las instancias de PCS y la caché de archivos. Para obtener más información sobre los grupos de seguridad, consulte Control de acceso a caché con Amazon VPC en la Guía del usuario de Amazon File Cache.

Plantilla de lanzamiento

Añada los grupos de seguridad de la configuración de su sistema de archivos a la plantilla de lanzamiento que utilizará para el grupo de nodos de cómputo.

Incluye los datos de usuario que se utilizan `cloud-config` para montar la caché de archivos de Amazon. Sustituya los siguientes valores de este script por sus propios detalles:

- *`mount-point-directory`*— La ruta de la instancia en la que quieres montarla FSx para Lustre
- *`cache-dns-name`*— El nombre del sistema de nombres de dominio (DNS) de la caché de archivos
- *`mount-name`*— El nombre de montaje de la caché de archivos

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- amazon-linux-extras install -y lustre=2.12
- mkdir -p /mount-point-directory
- mount -t lustre -o relatime,flock cache-dns-name@tcp:/mount-name /mount-point-
directory
```

```
--==MYBOUNDARY==
```

Amazon Machine Images (AMIs) para AWS PCS

AWS PCS trabaja con lo AMIs que usted proporciona, lo que ofrece una gran flexibilidad en el software y la configuración que se encuentran en los nodos de su clúster. Si está probando AWS PCS, puede utilizar un ejemplo de AMI proporcionado y mantenido por AWS. Si utiliza AWS PCS en producción, le recomendamos que construya el suyo propio AMIs. En este tema se explica cómo descubrir y utilizar el ejemplo AMIs, así como cómo crear y utilizar el suyo propio personalizado AMIs.

Temas

Uso de Amazon Machine Images (AMIs) de muestra con AWS PCS

AWS proporciona un ejemplo AMIs que puede utilizar como punto de partida para trabajar con AWS PCS.

> **Important**
>
> AMIs Los ejemplos son para fines de demostración y no se recomiendan para cargas de trabajo de producción.

> **Important**
>
> Los grupos de nodos de cómputo configurados con una muestra de AWS PCS AMIs y varias interfaces de red no funcionarán actualmente si las subredes solo están configuradas para su uso. IPv6 En su lugar, utilice subredes de doble pila (IPv4 y IPv6) o IPv4 solo subredes.

Encuentre un ejemplo de PCS actual AWS AMIs

Consola de administración de AWS

AMIs Los ejemplos de PCS de AWS tienen la siguiente convención de nomenclatura:

```
aws-pcs-sample_ami-OS-architecture-scheduler-scheduler-major-version
```

Valores aceptados

- *OS* – amzn2
- *architecture* – x86_64 o arm64
- *scheduler* – slurm
- *scheduler-major-version* – 25.05

Para encontrar un ejemplo de AWS PCS AMIs

1. Abre la EC2 consola de Amazon.
2. Vaya a AMIs.
3. Seleccione Imágenes públicas.
4. En Buscar AMI por atributo o etiqueta, busque una AMI con el nombre de la plantilla.

   Ejemplos

   - Ejemplo de AMI para Slurm 25.05 en instancias Arm64

     ```
     aws-pcs-sample_ami-amzn2-arm64-slurm-25.05
     ```

   - Ejemplo de AMI para Slurm 25.05 en instancias x86

     ```
     aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05
     ```

   > **Note**
   >
   > Si hay varios AMIs, utilice la AMI con la marca de tiempo más reciente.

5. Use el ID de AMI cuando cree o actualice un grupo de nodos de cómputo.

AWS CLI

Puede encontrar el ejemplo más reciente de AMI de AWS PCS con los siguientes comandos. *region-code*Sustitúyala por la Región de AWS que utilices AWS PCS, por ejemplo`us-east-1`.

- x86_64

```
aws ec2 describe-images --region region-code --owners amazon \
--filters 'Name=name,Values=aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05*' \
            'Name=state,Values=available' \
--query 'sort_by(Images, &CreationDate)[-1].[Name,ImageId]' --output text
```

- Arm64

```
aws ec2 describe-images --region region-code --owners amazon \
--filters 'Name=name,Values=aws-pcs-sample_ami-amzn2-arm64-slurm-25.05*' \
            'Name=state,Values=available' \
--query 'sort_by(Images, &CreationDate)[-1].[Name,ImageId]' --output text
```

Use el ID de AMI cuando cree o actualice un grupo de nodos de cómputo.

Obtenga más información sobre el ejemplo de AWS PCS AMIs

Para ver el contenido y los detalles de configuración de las versiones actuales y anteriores del ejemplo de AWS PCS AMIs, consulteNotas de publicación para un ejemplo de AWS PCS AMIs.

Cree la suya propia AMIs compatible con AWS PCS

Para obtener información sobre cómo crear uno propio AMIs que funcione con AWS PCS, consulteImágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS

AWS El PCS está diseñado para funcionar con Amazon Machine Images (AMI) que usted incorpore al servicio. AMIs Pueden tener instalados software y configuraciones arbitrarios, siempre que tengan el agente AWS PCS y una versión compatible de Slurm instalados y configurados correctamente. Debe utilizar los AWS instaladores proporcionados para instalar el software AWS PCS en su AMI

personalizada. Le recomendamos que utilice los AWS instaladores proporcionados para instalar Slurm en su AMI personalizada, pero puede instalar Slurm por su cuenta si lo prefiere (no se recomienda).

> **Note**
>
> Si quiere probar AWS PCS sin crear una AMI personalizada, puede utilizar un ejemplo de AMI proporcionado por AWS. Para obtener más información, consulte Uso de Amazon Machine Images (AMIs) de muestra con AWS PCS.

> **Important**
>
> AWS Actualmente, el PCS requiere un núcleo IPv4 compatible con la comunicación entre nodos locales, incluso cuando se utiliza el AWS PCS IPv6 solo en una red.

Este tutorial le ayuda a crear una AMI que se pueda usar con grupos de nodos de cómputo de PCS para impulsar su HPC y sus cargas de AI/ML trabajo.

Temas

Paso 1: lanzar una instancia temporal

Lance una instancia temporal que pueda usar para instalar y configurar el software AWS PCS y el programador Slurm. Esta instancia se utiliza para crear una AMI compatible con AWS PCS.

Para iniciar una instancia temporal

1. Abra la consola de Amazon EC2.

2. En el panel de navegación, elija Instances y, a continuación, elija Launch instances para abrir el asistente de nueva instancia de lanzamiento.
3. (Opcional) En la sección Nombre y etiquetas, proporciona un nombre para la instancia, como`PCS-AMI-instance`. El nombre se asigna a la instancia como etiqueta de recurso (`Name=PCS-AMI-instance`).
4. En la sección Application and OS Images (Imágenes de aplicaciones y sistema operativo), seleccione una AMI para uno de los sistemas operativos compatibles.
5. En la sección Tipo de instancia, seleccione el tipo de instancia admitida.
6. En la sección Par de claves, seleccione el par de claves que desea utilizar en la instancia.
7. En la sección Configuración de red:
   - En Firewall (grupos de seguridad), selecciona Seleccionar un grupo de seguridad existente y, a continuación, selecciona un grupo de seguridad que permita el acceso SSH entrante a la instancia.
8. En la sección Almacenamiento, configure los volúmenes según sea necesario. Asegúrese de configurar suficiente espacio para instalar sus propias aplicaciones y bibliotecas.
9. En el panel Resumen, elija Iniciar instancia.

Paso 2: Instalar el agente AWS PCS

Instale el agente que configura las instancias lanzadas por AWS PCS para su uso con Slurm. Para obtener más información sobre el agente AWS PCS, consulte. AWS Versiones del agente PCS

Para instalar el agente AWS PCS

1. Conéctese a la instancia que lanzó. Para obtener más información, consulte Conexión con la instancia de Linux.
2. (Opcional) Para asegurarse de que todos los paquetes de software estén actualizados, realice una actualización rápida del software de la instancia. Este proceso puede demorar unos minutos.
   - Amazon Linux 2, Amazon Linux 2023, RHEL 9, RHEL 8, Rocky Linux 9 y Rocky Linux 8

     ```
     sudo yum update -y
     ```
   - Ubuntu 22.04 y Ubuntu 24.04

```
sudo apt-get update && sudo apt-get upgrade -y
```

3. Reinicie la instancia y vuelva a conectarse a ella.
4. Descargue los archivos de instalación del agente AWS PCS. Los archivos de instalación se empaquetan en un archivo tarball (`.tar.gz`) comprimido. Para descargar la última versión estable, utilice el comando siguiente. *region*Sustitúyala por la ubicación Región de AWS en la que lanzaste la instancia temporal, por ejemplo. `us-east-1`

   ```
   curl https://aws-pcs-repo-region.s3.region.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.3.2-1.tar.gz -o aws-pcs-agent-v1.3.2-1.tar.gz
   ```

   También puede obtener la última versión sustituyendo el número de versión por `latest` el del comando anterior (por ejemplo:`aws-pcs-agent-v1-latest.tar.gz`).

   > **Note**
   >
   > Esto podría cambiar en futuras versiones del software del agente AWS PCS.

5. (Opcional) Compruebe la autenticidad e integridad del archivo tar del software del AWS PCS. Le recomendamos que lo haga para verificar la identidad del editor de software y para verificar que el archivo no se haya modificado ni dañado desde que se publicó.

   a. Descarga la clave GPG pública para AWS PCS e impórtala a tu conjunto de claves. *region*Sustitúyala por la Región de AWS ubicación en la que lanzaste la instancia temporal. El comando debe devolver un valor de clave. Registre el valor clave y utilícelo en el siguiente paso.

      ```
      wget https://aws-pcs-repo-public-keys-region.s3.region.amazonaws.com/aws-pcs-public-key.pub && \
          gpg --import aws-pcs-public-key.pub
      ```

   b. Ejecuta el siguiente comando para verificar la huella digital de la clave GPG.

      ```
      gpg --fingerprint 7EEF030EDDF5C21C
      ```

      El comando debería devolver una huella digital idéntica a la siguiente:

```
1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
```

> **Important**
>
> No ejecute el script de instalación del agente AWS PCS si la huella digital no coincide. Contacte con AWS Support.

c. Descargue el archivo de firma y verifique la firma del archivo tar del software AWS PCS. *region*Sustitúyalo por el Región de AWS lugar donde lanzaste la instancia temporal, por ejemplo. `us-east-1`

```
wget https://aws-pcs-repo-region.s3.region.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.3.2-1.tar.gz.sig && \
    gpg --verify ./aws-pcs-agent-v1.3.2-1.tar.gz.sig
```

El resultado debería ser similar al siguiente:

```
gpg: assuming signed data in './aws-pcs-agent-v1.3.2-1.tar.gz'
gpg: Signature made Thu 06 Nov 2025 11:10:36 AM CET using RSA key ID ECC0AE5C
gpg: Good signature from "AWS PCS Packages (AWS PCS Packages)"
gpg: WARNING: This key is not certified with a trusted signature!
gpg:          There is no indication that the signature belongs to the owner.
Primary key fingerprint: 1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
   Subkey fingerprint: B7E1 8788 3517 6A74 C3D5  EAF5 6088 136D ECC0 AE5C
```

Si el resultado incluye `Good signature` y la huella digital coincide con la huella digital devuelta en el paso anterior, continúe con el paso siguiente.

> **Important**
>
> No ejecute el script de instalación del software AWS PCS si la huella digital no coincide. Contacte con AWS Support.

6. Extraiga los archivos del `.tar.gz` archivo comprimido y navegue hasta el directorio extraído.

```
tar -xf aws-pcs-agent-v1.3.2-1.tar.gz && \
    cd aws-pcs-agent
```

7. Instale el software AWS PCS.

```
sudo ./installer.sh
```

8. Compruebe el archivo de versión del software AWS PCS para confirmar que la instalación se ha realizado correctamente.

```
cat /opt/aws/pcs/version
```

El resultado debería ser similar al siguiente:

```
AGENT_INSTALL_DATE='Fri Dec 13 12:28:43 UTC 2024'
AGENT_VERSION='1.3.2'
AGENT_RELEASE='1'
```

Paso 3: Instalar Slurm

Instale una versión de Slurm que sea compatible con el PCS. AWS Para obtener más información, consulte Versiones de Slurm en PCS AWS.

> **Note**
>
> Si tiene una AMI con una versión anterior del software Slurm instalada, debe realizar los siguientes pasos para instalar la nueva versión de Slurm. El agente AWS PCS habilita la versión correcta de los binarios de Slurm en tiempo de ejecución, de acuerdo con la versión de Slurm configurada en el momento de la creación del clúster.

Para instalar Slurm

1. Conéctese a la misma instancia temporal en la que instaló el software AWS PCS.
2. Descargue el software de instalación de Slurm. El instalador de Slurm está empaquetado en un archivo tarball () comprimido. `.tar.gz` Para descargar la última versión estable, utilice el comando siguiente. *`region`*Sustitúyalo por el Región de AWS de su instancia temporal, por ejemplo. `us-east-1`

```
curl https://aws-pcs-repo-region.s3.region.amazonaws.com/aws-pcs-slurm/aws-pcs-
slurm-25.05-installer-25.05.4-1.tar.gz \
```

```
    -o aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz
```

También puede obtener la última versión sustituyendo el número de versión por `latest` el del comando anterior (por ejemplo:`aws-pcs-slurm-25.05-installer-latest.tar.gz`). Para obtener una lista completa de las versiones disponibles con sumas de verificación, consulte[Versiones de Slurm en PCS AWS](#).

> **Note**
>
> Esto podría cambiar en futuras versiones del software de instalación de Slurm.

3. (Opcional) Compruebe la autenticidad e integridad del archivo tar del instalador de Slurm. Le recomendamos que lo haga para verificar la identidad del editor de software y para verificar que el archivo no se haya modificado ni dañado desde que se publicó.

   a. Descarga la clave GPG pública para AWS PCS e impórtala a tu conjunto de claves. *`region`*Sustitúyala por la Región de AWS ubicación en la que lanzaste la instancia temporal. El comando debe devolver un valor de clave. Registre el valor clave y utilícelo en el siguiente paso.

   ```
   wget https://aws-pcs-repo-public-keys-region.s3.region.amazonaws.com/aws-pcs-public-key.pub && \
       gpg --import aws-pcs-public-key.pub
   ```

   b. Ejecuta el siguiente comando para verificar la huella digital de la clave GPG.

   ```
   gpg --fingerprint 7EEF030EDDF5C21C
   ```

   El comando debería devolver una huella digital idéntica a la siguiente:

   ```
   1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
   ```

   > **Important**
   >
   > No ejecute el script de instalación de Slurm si la huella digital no coincide. Contacte con AWS Support.

c. Descargue el archivo de firma y verifique la firma del archivo tarball del instalador de Slurm. *region*Sustitúyalo por el Región de AWS lugar en el que lanzaste la instancia temporal, por ejemplo. `us-east-1`

```
wget https://aws-pcs-repo-region.s3.region.amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz.sig && \
     gpg --verify ./aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz.sig
```

El resultado debería ser similar al siguiente:

```
gpg: assuming signed data in './aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz'
gpg: Signature made Fri 24 Oct 2025 05:05:11 PM UTC using RSA key ID ECC0AE5C
gpg: Good signature from "AWS PCS Packages (AWS PCS Packages)"
gpg: WARNING: This key is not certified with a trusted signature!
gpg:          There is no indication that the signature belongs to the owner.
Primary key fingerprint: 1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
     Subkey fingerprint: B7E1 8788 3517 6A74 C3D5  EAF5 6088 136D ECC0 AE5C
```

Si el resultado incluye `Good signature` y la huella digital coincide con la huella digital devuelta en el paso anterior, continúe con el paso siguiente.

> **Important**
>
> No ejecute el script de instalación de Slurm si la huella digital no coincide. Contacte con AWS Support.

4. Extraiga los archivos desde el archivo `.tar.gz` comprimido y acceda al directorio extraído.

```
tar -xf aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz && \
    cd aws-pcs-slurm-25.05-installer
```

5. Instala Slurm. El instalador descarga, compila e instala Slurm y sus dependencias. Tarda varios minutos, según las especificaciones de la instancia temporal que haya seleccionado.

```
sudo ./installer.sh -y
```

6. Compruebe el archivo de versión del programador para confirmar la instalación.

```
cat /opt/aws/pcs/scheduler/slurm-25.05/version
```

El resultado debería ser similar al siguiente:

```
SLURM_INSTALL_DATE='Mon Nov 3 14:23:38 UTC 2025'
SLURM_VERSION='25.05.4'
PCS_SLURM_RELEASE='1'
```

Paso 4: (opcional) Instalar controladores, bibliotecas y software de aplicación adicionales

Instale controladores, bibliotecas y software de aplicación adicionales en la instancia temporal. Los procedimientos de instalación variarán en función de las aplicaciones y bibliotecas específicas. Si no ha creado una AMI personalizada para AWS PCS anteriormente, le recomendamos que primero cree y pruebe una AMI solo con el software de AWS PCS y Slurm instalados y, a continuación, añada gradualmente su propio software y configuraciones una vez que haya confirmado el éxito inicial.

Ejemplos

- Software Elastic Fabric Adapter (EFA). Para obtener más información, consulte Introducción a EFA y MPI para cargas de trabajo de HPC en Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud.
- Cliente Amazon Elastic File System (Amazon EFS). Para obtener más información, consulte Instalación manual del cliente Amazon EFS en la Guía del usuario de Amazon Elastic File System.
- Cliente Lustre, para usar Amazon FSx for Lustre y Amazon File Cache. Para obtener más información, consulte Instalación del cliente de Lustre en la Guía del usuario FSx de Lustre.
- CloudWatch Agente de Amazon, para usar CloudWatch registros y métricas. Para obtener más información, consulte Instalar el CloudWatch agente en la Guía del CloudWatch usuario de Amazon.
- AWS Neuron, para usar los tipos de instancia trn* e inf*. Para obtener más información, consulte la documentación de Neuron.AWS
- Controlador NVIDIA, CUDA y DCGM para usar los tipos de instancia p* o g*.

Paso 5: Crear una AMI compatible con AWS PCS

Una vez instalados los componentes de software necesarios, crea una AMI que puede reutilizar para lanzar instancias en grupos de nodos de cómputo de AWS PCS.

> **Important**
>
> AWS Actualmente, el PCS requiere un núcleo IPv4 compatible con la comunicación entre nodos locales, incluso cuando se utiliza el AWS PCS en una red IPv6 exclusiva.

Para crear una AMI desde la instancia temporal

1. Abra la consola de Amazon EC2.
2. En el panel de navegación, seleccione Instances (Instancias).
3. Seleccione la instancia temporal que ha creado. Selecciona Acciones, Imagen y Crear imagen.
4. En Crear imagen, realice lo siguiente:
   a. En Nombre de imagen, ingrese un nombre descriptivo para la AMI.
   b. (Opcional) En Descripción de imagen, ingrese una breve descripción del propósito la AMI.
   c. Elija Crear imagen.
5. En el panel de navegación, elija AMIs.
6. Busque la AMI que creó en la lista. Espere a que su estado cambie de Pendiente a Disponible y, a continuación, utilícela con un grupo de nodos de cómputo de AWS PCS.

Paso 6: Utilice la AMI personalizada con un grupo de nodos de cómputo de AWS PCS

Puede usar su AMI personalizada con un grupo de nodos de cómputo de AWS PCS nuevo o existente.

> **Important**
>
> AWS Actualmente, el PCS requiere un núcleo IPv4 compatible con la comunicación entre nodos locales, incluso cuando se utiliza el AWS PCS en una red IPv6 exclusiva.

New compute node group

Para usar la AMI personalizada

1. Abra la consola AWS PCS.

2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Elija el clúster en el que utilizará la AMI personalizada y, a continuación, seleccione Compute node groups.
4. Cree un nuevo grupo de nodos de procesamiento. Para obtener más información, consulte Creación de un grupo de nodos de cómputo en AWS PCS. En ID de AMI, busque el nombre o la ID de la AMI personalizada que desee usar. Termine de configurar el grupo de nodos de cómputo y, a continuación, seleccione Crear grupo de nodos de cómputo.
5. (Opcional) Confirme que la AMI admite el lanzamiento de instancias. Lanza una instancia en el grupo de nodos de cómputo. Para ello, configura el grupo de nodos de cómputo para que tenga una única instancia estática, o puedes enviar un trabajo a una cola que utilice el grupo de nodos de cómputo.

    a. Compruebe la consola Amazon EC2 hasta que aparezca una instancia etiquetada con el nuevo ID de grupo de nodos de cómputo. Para obtener más información al respecto, consulte.. Búsqueda de instancias de grupos de nodos de cómputo en AWS PCS

    b. Cuando veas que una instancia se lanza y completa su proceso de arranque, confirma que usa la AMI esperada. Para ello, selecciona la instancia y, a continuación, inspecciona el ID de AMI en Detalles. Debe coincidir con la AMI que configuró en la configuración del grupo de nodos de cómputo.

    c. (Opcional) Actualice la configuración de escalado del grupo de nodos de cómputo a sus valores preferidos.

Existing compute node group

Para usar la AMI personalizada

1. Abra la consola AWS PCS.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Elija el clúster en el que utilizará la AMI personalizada y, a continuación, seleccione Compute node groups.
4. Seleccione el grupo de nodos que desee configurar y elija Editar. En ID de AMI, busque el nombre o la ID de la AMI personalizada que desee usar. Termine de configurar el grupo de nodos de procesamiento y, a continuación, seleccione Actualizar. Las nuevas instancias lanzadas en el grupo de nodos de procesamiento utilizarán el ID de AMI actualizado. Las instancias existentes seguirán utilizando la AMI anterior hasta que AWS PCS las sustituya.

Para obtener más información, consulte Actualización de un grupo de nodos de cómputo AWS PCS.

5. (Opcional) Confirme que la AMI admite el lanzamiento de instancias. Lanza una instancia en el grupo de nodos de cómputo. Para ello, configura el grupo de nodos de cómputo para que tenga una única instancia estática, o puedes enviar un trabajo a una cola que utilice el grupo de nodos de cómputo.

   a. Compruebe la consola Amazon EC2 hasta que aparezca una instancia etiquetada con el nuevo ID de grupo de nodos de cómputo. Para obtener más información al respecto, consulte.. Búsqueda de instancias de grupos de nodos de cómputo en AWS PCS

   b. Cuando veas que una instancia se lanza y completa su proceso de arranque, confirma que usa la AMI esperada. Para ello, selecciona la instancia y, a continuación, inspecciona el ID de AMI en Detalles. Debe coincidir con la AMI que configuró en la configuración del grupo de nodos de cómputo.

   c. (Opcional) Actualice la configuración de escalado del grupo de nodos de cómputo a sus valores preferidos.

Paso 7: Finalizar la instancia temporal

Una vez que haya confirmado que su AMI funciona según lo previsto con AWS PCS, puede cancelar la instancia temporal para dejar de incurrir en cargos por ella.

Para finalizar la instancia temporal

1. Abra la consola de Amazon EC2.
2. En el panel de navegación, seleccione Instances (Instancias).
3. Seleccione la instancia temporal que creó y elija Acciones, Estado de la instancia y Finalizar instancia.
4. Cuando se le pida que confirme, elija Finalizar.

Instaladores de software para crear PCS personalizados AMIs AWS

AWS proporciona un archivo descargable que puede instalar el software AWS PCS en una instancia. AWS también proporciona software que puede descargar, compilar e instalar las versiones

pertinentes de Slurm y sus dependencias. Puede usar estas instrucciones para crear una versión personalizada AMIs para su uso con AWS PCS o puede usar sus propios métodos.

Contenido

AWS Instalador del software PCS Agent

El instalador del software del agente AWS PCS configura una instancia para que funcione con el AWS PCS durante el proceso de arranque de la instancia. Debe utilizar los AWS instaladores proporcionados para instalar el agente AWS PCS en la AMI personalizada.

Para obtener más información sobre el software del agente AWS PCS, consulte. AWS Versiones del agente PCS

Instalador de Slurm

El instalador de Slurm descarga, compila e instala las versiones relevantes de Slurm y sus dependencias. Puede usar el instalador de Slurm para crear versiones personalizadas para PCS. AMIs AWS También puede utilizar sus propios mecanismos si son coherentes con la configuración de software que proporciona el instalador de Slurm. Para obtener más información sobre la compatibilidad de AWS PCS con Slurm, consulte. Versiones de Slurm en PCS AWS

El software AWS suministrado instala lo siguiente:

- Utilice la versión principal y de mantenimiento solicitada (actualmente la versión 25.05.x): licencia GPL 2
  - Slurm está construido con un conjunto de `--sysconfdir /etc/slurm`
  - Slurm está diseñado con la opción y `--enable-pam --without-munge`
  - Slurm se construye con la opción `--sharedstatedir=/run/slurm/`
  - Slurm está construido con soporte para PMIX y JWT

  - Slurm está instalado en `/opt/aws/pcs/schedulers/slurm-25.05`
- OpenPMix (versión 4.2.6) — Licencia
  - OpenPMIX se instala como un subdirectorio de `/opt/aws/pcs/scheduler/`
- libjwt (versión 1.17.0) — Licencia MPL-2.0
  - libjwt se instala como un subdirectorio de `/opt/aws/pcs/scheduler/`

El software AWS suministrado cambia la configuración del sistema de la siguiente manera:

- El `systemd` archivo Slurm creado por la compilación se copia `/etc/systemd/system/` con el nombre del archivo. `slurmd-25.05.service`
- Si no existen, se crean un usuario y un grupo de Slurm (`slurm:slurm`) con of. UID/GID `401`
- `/etc/aws/pcs/scheduler/slurm-25.05/plugstack.conf.d/`Se crea la carpeta para almacenar la configuración. Amplíe la funcionalidad de Slurm en los PCS con los complementos de AWS SPANK
- En Amazon Linux 2 y Rocky Linux 9, la instalación añade el repositorio EPEL para instalar el software necesario para compilar Slurm o sus dependencias.
- Durante RHEL9 la instalación, se habilitará `codeready-builder-for-rhel-9-rhui-rpms` y `epel-release-latest-9` se instalará el software necesario `fedoraproject` para compilar Slurm o sus dependencias.

Sistemas operativos compatibles

Consulte Sistemas operativos compatibles en AWS PCS.

> **Note**
>
> AWS Deep Learning AMIs Las versiones (DLAMI) basadas en Amazon Linux 2 y Ubuntu 22.04 deben ser compatibles con el software PCS y los instaladores de AWS Slurm. Para obtener más información, consulte Cómo elegir su DLAMI en AWS Deep Learning AMIs la Guía para desarrolladores.

Tipos de instancias admitidas

AWS El software PCS y los instaladores de Slurm admiten cualquier tipo de instancia x86_64 o arm64 que pueda ejecutar uno de los sistemas operativos compatibles.

Versiones de Slurm compatibles

Consulte Versiones de Slurm en PCS AWS.

Verifique los instaladores mediante una suma de verificación

Puede utilizar SHA256 sumas de comprobación para comprobar los archivos tar (.tar.gz) del instalador. Le recomendamos que lo haga para verificar la identidad del editor de software y para comprobar que la aplicación no se ha modificado ni dañado desde que se publicó.

Para verificar un tarball

Utilice la utilidad sha256sum para la suma de SHA256 comprobación y especifique el nombre del archivo tarball. Debe ejecutar el comando desde el directorio en el que guardó el archivo tarball.

- SHA256

```
$ sha256sum tarball_filename.tar.gz
```

El comando debe devolver un valor de suma de comprobación con el siguiente formato.

```
checksum_value tarball_filename.tar.gz
```

Compare el valor de la suma de verificación devuelto por el comando con el valor de la suma de verificación que se proporciona en la siguiente tabla. Si las sumas de comprobación coinciden, es seguro ejecutar el script de instalación.

> **Important**
>
> Si las sumas de comprobación no coinciden, no ejecute el script de instalación. Ponte en contacto con Soporte.

Por ejemplo, el siguiente comando genera la SHA256 suma de comprobación del tarball Slurm 25.05.4-1.

```
$ sha256sum aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz
```

Ejemplo de código de salida:

```
3b0f93bce441d4f4f6935175f2c1e81cd961cb923adb416fa6689f5592047a7d aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz
```

En las tablas siguientes se muestran las sumas de comprobación de las versiones recientes de los instaladores. *us-east-1*Sustitúyalo por el Región de AWS lugar en el que utilice PCS AWS .

AWS Agente de PCS

Installer (Instalador)	Descargar URL	SHA256 suma de control
AWS agente PCS 1.3.2-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.3.2-1.tar.gz`	`06b32a952a1c849e3442e35c28ac2e4d6962b09286cad748f3c83d561b52ec6f`
AWS Agente PCS 1.3.1-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.3.1-1.tar.gz`	`5b7f1eb7b3a86bd2d331b5cb0138d868dc9452da34b480becd86af892c7e8d19`
AWS Agente PCS 1.3.0-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.3.0-1.tar.gz`	`eadc9b65c3db248bdde2a6c41814dfb1b97239f24ad55e03d8526dd9ab4a8d16`

Installer (Instalador)	Descargar URL	SHA256 suma de control
AWS Agente PCS 1.2.2-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.2.2-1.tar.gz`	`fd7b6ea5442db75d723fc4971781ce6ae511baa21b87c4286fc1df8127b282b8`
AWS Agente PCS 1.2.1-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.2.1-1.tar.gz`	`2b784643ca01ccca1baa64fbfb34bb41efe8bdca69470998b74ce3962bc271d4`
AWS Agente PCS 1.2.0-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.2.0-1.tar.gz`	`470db8c4fc9e50277b6317f98584b6b547e73523043e34f018eecae767846805`
AWS Agente PCS 1.1.1-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.1.1-1.tar.gz`	`bef078bf60a6d8ecde2e6c49cd34d088703f02550279e3bf483d57a235334dc6`
AWS Agente PCS 1.1.0-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.1.0-1.tar.gz`	`594c32194c71bccc5d66e5213213ae38dd2c6d2f9a950bb01accea0bbab0873a`

Installer (Instalador)	Descargar URL	SHA256 suma de control
AWS Agente PCS 1.0.1-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.0.1-1.tar.gz`	`04e22264019837e3f42d8346daf5886eaacecd21571742eb505ea8911786bcb2`
AWS Agente PCS 1.0.0-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-agent/aws-pcs-agent-v1.0.0-1.tar.gz`	`d2d3d68d00c685435c38af471d7e2492dde5ce9eb222d7b6ef0042144b134ce0`

Instalador Slurm

Installer (Instalador)	Descargar URL	SHA256 suma de verificación
Slurm 25.05.4-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz`	`3b0f93bce441d4f4f6935175f2c1e81cd961cb923adb416fa6689f5592047a7d`
Slurm 25.05.3-1	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-25.05-installer-25.05.3-1.tar.gz`	`851bb5815b6700ceb30cc4a3fda204ca8ce362c14528c339908983255a936cf0`
Slurm 24.11.6-2	`https://aws-pcs-repo-us-east-1.s3.us-east-1.amazonaws.com/aws-pcs-slurm/aws-pcs`	`f17cd78e0bc6b9c818b794d9d2685cceabdc`

Installer (Instalador)	Descargar URL	SHA256 suma de verificación
	-slurm-24.11-insta ller-24.11.6-2.tar.gz	73f4fbb12f7566ae5b 86a5abc32b
Slurm 24.11.6-1	https://aws-pcs-re po- *us-east-1* .s3.*us-east-1* .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.11-insta ller-24.11.6-1.tar.gz	225de9fc18206f5f65 f412effe1fd457614a c97ee9822b3ff804a4 52b0fae522
Slurm 24.11.5-1	https://aws-pcs-re po- *us-east-1* .s3.*us-east-1* .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.11-insta ller-24.11.5-1.tar.gz	593efe4d66bef2f3e4 6d5a382fb5a32f7a3c a2510bcf1b3c85739f 4f951810d5
Slurm 24.05.8-2	https://aws-pcs-re po- *us-east-1* .s3.*us-east-1* .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.05-insta ller-24.05.8-2.tar.gz	c494b0b55c319a4c2f 3faf668c759d46c32c 4c7aa94ae97d941283 28fe95364b
Slurm 24.05.8-1	https://aws-pcs-re po- *us-east-1* .s3.*us-east-1* .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.05-insta ller-24.05.8-1.tar.gz	210a43b376af082bba d640b2032655885790 c5dab0e6489cc327c7 310a375849

Installer (Instalador)	Descargar URL	SHA256 suma de verificación
Slurm 24.05.7-1	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-24.05-installer-24.05.7-1.tar.gz`	`0b5ed7c81195de2628c78f37c79e63fc4ae99132ca6b019b53a0d68792ee82c5`
Slurm 24.05.5-2	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-24.05-installer-24.05.5-2.tar.gz`	`7cc8d8294f2fbff95fe0602cf9e21e02003b5d96c0730e0a18c6aa04c7a4967b`
Slurm 23.11.10-4 (obsoleto)	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-23.11-installer-23.11.10-4.tar.gz`	`bb2d8c919c69dba38d14358f49c7f0427564c5dd4af85a1c9eca2c57ceeae29a`
Slurm 23.11.10-3 (obsoleto)	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-23.11-installer-23.11.10-3.tar.gz`	`488a10ee0fbd57ec0e0ff7ea708a9e3038fafdc025c6bb391c75c2e2a7852a00`

Installer (Instalador)	Descargar URL	SHA256 suma de verificación
Slurm 23.11.10-2 (obsoleto)	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-23.11-installer-23.11.10-2.tar.gz`	`0bbe85423305c05987931168caf98da08a34c25f9eec0690e8e74de0b7bc8752`
Slurm 23.11.10-1 (obsoleto)	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-23.11-installer-23.11.10-1.tar.gz`	`27e8faa9980e92cdfd8cfdc71f937777f0934552ce61e33dac4ecf5a20321e44`
Slurm 23.11.9-1 (obsoleto)	`https://aws-pcs-repo- us-east-1 .s3.us-east-1 .amazonaws.com/aws-pcs-slurm/aws-pcs-slurm-23.11-installer-23.11.9-1.tar.gz`	`1de7d919c8632fe8e2806611bed4fde1005a4fadc795412456e935c7bba2a9b8`

Notas de publicación para un ejemplo de AWS PCS AMIs

AMIs para las últimas versiones principales compatibles del programador, reciba actualizaciones de seguridad y correcciones de errores críticos. Estos parches de seguridad incrementales no se incluyen en las notas oficiales de la versión.

> **Important**
>
> Los ejemplos AMIs relacionados con las versiones antiguas del programador no son compatibles y no reciben actualizaciones.

> **Important**
>
> AMIs Los ejemplos son para fines de demostración y no se recomiendan para cargas de trabajo de producción.

Contenido

AWS Ejemplo de PCS AMIs para x86_64 (Amazon Linux 2)

Slurm 25.05

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05`

Instancias de EC2 admitidas

- Todas las instancias con un procesador x86 de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque Architectures=x86_64.

Contenido de AMI

- Servicio de AWS compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura informática: x86_64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.43.1
- GDRCopy: 2.5.1
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.11

> **Note**
>
> AWS PCS admite la contabilidad de Slurm 24.11 y versiones posteriores. Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.11`

Instancias de EC2 admitidas

- Todas las instancias con un procesador x86 de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=x86_64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: x86_64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.05

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.05`

Instancias de EC2 admitidas

- Todas las instancias con un procesador x86 de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=x86_64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: x86_64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 23.11

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-23.11`

Instancias de EC2 admitidas

- Todas las instancias con un procesador x86 de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=x86_64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: x86_64
- Tipo de volumen de EBS: gp2

- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

AWS Ejemplo de PCS AMIs para Arm64 (Amazon Linux 2)

Slurm 25.05

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-25.05`

Instancias de EC2 admitidas

- Todas las instancias cuentan con un procesador Arm de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque Architectures=arm64.

Contenido de AMI

- Servicio de AWS compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura informática: arm64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.43.1
- GDRCopy: 2.5.1
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.11

> **Note**
>
> AWS PCS admite la contabilidad de Slurm 24.11 y versiones posteriores. Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-24.11`

Instancias de EC2 admitidas

- Todas las instancias cuentan con un procesador Arm de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=arm64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: arm64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.05

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-24.05`

Instancias de EC2 admitidas

- Todas las instancias cuentan con un procesador Arm de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=arm64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: arm64
- Tipo de volumen de EBS: gp2
- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 23.11

Nombre de la AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-23.11`

Instancias de EC2 admitidas

- Todas las instancias cuentan con un procesador Arm de 64 bits. Para buscar instancias compatibles, vaya a la consola Amazon EC2. Elija Tipos de instancia y, a continuación, busque`Architectures=arm64`.

Contenido de AMI

- AWS Servicio compatible: AWS PCS
- Sistema operativo: Amazon Linux 2
- Arquitectura de cómputo: arm64
- Tipo de volumen de EBS: gp2

- Instalador de EFA: 1.33.0
- GDRCopy: 2.4
- Controlador NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Sistemas operativos compatibles en AWS PCS

AWS PCS usa la Amazon Machine Image (AMI) configurada para un grupo de nodos de cómputo para lanzar EC2 instancias en ese grupo de nodos de cómputo. La AMI determina el sistema operativo que utilizan las EC2 instancias. No se puede cambiar el sistema operativo en el ejemplo de AWS PCS AMIs. Debe crear una AMI personalizada si desea utilizar un sistema operativo diferente. Para obtener más información, consulte Amazon Machine Images (AMIs) para AWS PCS.

Sistemas operativos compatibles

- Amazon Linux 2

  Este es el sistema operativo del ejemplo de AWS PCS AMIs.

  > **Important**
  >
  > AMIs Los ejemplos son para fines de demostración y no se recomiendan para cargas de trabajo de producción. Debe crear y usar una AMI personalizada para las cargas de trabajo de producción, incluso si piensa usar Amazon Linux 2.

- Amazon Linux 2023
- RedHat Linux empresarial 9 (RHEL 9)

  El coste bajo demanda de cualquier tipo de instancia de RHEL es superior al de otros sistemas operativos compatibles. Para obtener más información sobre los precios, consulte Precios bajo demanda y ¿Cómo se ofrece y se cotiza Red Hat Enterprise Linux en Amazon Elastic Compute Cloud?.
- RedHat Linux empresarial 8 (RHEL 8)
- Rocky Linux 9

  Puedes usar el Rocky Linux 9 oficial AMIs como base para una AMI personalizada. La compilación de la AMI personalizada podría fallar si la AMI base no tiene el núcleo más reciente.

  Para actualizar el núcleo

  1. Lance una instancia con un ID de AMI de rocky9 desde aquí: https://rockylinux.org/cloud-images/
  2. Inicie sesión en la instancia y ejecute el comando siguiente:

```
sudo yum -y update
```

3. Cree una imagen a partir de la instancia. Especifique esta imagen como imagen `ParentImage` para su AMI personalizada.

- Rocky Linux 8
- Ubuntu 22.04

  Ubuntu 22.04 requiere claves más seguras para SSH y no admite claves RSA de forma predeterminada. Te recomendamos que generes y utilices una ED25519 clave en su lugar.
- Ubuntu 24.04

AWS Versiones del agente PCS

El software del agente AWS PCS configura las instancias EC2 que AWS PCS lanza para usarlas con Slurm. El agente se incluye en una Amazon Machine Images (AMI) que se especifica al crear grupos de nodos de procesamiento para el clúster. Las instancias EC2 lanzadas en esos grupos de nodos de cómputo utilizan la AMI especificada y el software de agente AWS PCS incluido. El agente AWS PCS permite que una instancia EC2 se registre como parte del clúster. Para utilizar el software de agente de AWS PCS más reciente, debe actualizar su versión personalizada AMIs. Para obtener más información, consulta Paso 2: Instalar el agente AWS PCS en Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

AWS Versión del agente de PCS	Fecha de publicación	Notas de la versión
v1.3.2-1	10 de marzo de 2026	• Se solucionó un problema que provocaba que los nodos de cómputo que ejecutaban RHEL 8.10 o Rocky Linux 8.10 no se pudieran arrancar debido a una falla en el backport de `curl` SiGv4 en esos sistemas operativos.
v1.3.1-1	7 de noviembre de 2025	• Se ha mejorado la desactivación del hiperproceso mediante el uso del parámetro sysfs `smt/control` cuando está disponible. • Se ha corregido una posible condición de carrera que provocaba que la CPU se bloqueara durante el arranque mientras el agente PCS intentaba desactivar el hiperproceso.

AWS Versión del agente de PCS	Fecha de publicación	Notas de la versión
		• Se ha corregido un error que provocaba que `InstanceType` los campos `InstanceId` y de los nodos de cálculo de Slurm se rellenaran con una marca de tiempo y un guión, respectivamente.
v1.3.0-1	3 de noviembre de 2025	• Se agregó soporte para nuevos sistemas operativos: Amazon Linux 2023, Ubuntu 24, RHEL 8, Rocky 8.
v1.2.2-1	16 de octubre de 2025	• Se permiten consultas de metadatos de instancias a un IPv6 punto IPv4 final si este no está disponible. • Se ha corregido un problema que impedía inhabilitar el hiperproc eso si el núcleo devolvía subprocesos hermanos como rangos de ID de CPU. • Se ha corregido un problema que producía falsos mensajes de error en los registros cuando el hiperproceso se desactiva ba correctamente.

AWS Versión del agente de PCS	Fecha de publicación	Notas de la versión
v1.2.1-1	19 de junio de 2025	• El agente de AWS PCS ahora intenta iniciar slurmd durante un máximo de 30 minutos si el mando no está disponible. • Se ha corregido un problema que provocaba una configuración de slurmd incorrecta si la respuesta RegisterComputeNod eGroupInstance contenía un punto final de SLURMDBD.
v1.2.0-1	7 de marzo de 2025	• Soporte habilitado para in. IPv6 `slurmd.conf`
v1.1.1-1	13 de diciembre de 2024	• Se ha corregido un error que provocaba que en la llamada a se informara de una versión incorrect a de Slurm. RegisterC omputeNodeGroupInstance • Se ha corregido un problema por el que los metadatos de la instancia no se recuperab an correctamente si se ejecutaba un script personalizado. `/opt/aws/pcs/etc/bootstrap_hooks/`

AWS Versión del agente de PCS	Fecha de publicación	Notas de la versión
v1.1.0-1	6 de diciembre de 2024	• Se habilitó la ejecución de scripts personalizados antes de los pasos de arranque. `/opt/aws/pcs/etc/bootstrap_hooks/`
v1.0.1-1	22 de octubre de 2024	• Se ha corregido un problema que provocaba que los dispositivos NVIDIA no funcionaran cuando se `slurmd` iniciaban en instancias con GPU habilitada.
v1.0.0-1	28 de agosto de 2024	• Versión inicial.

programador Slurm en PCS AWS

Slurm es un administrador de cargas de trabajo de código abierto diseñado para clústeres de Linux que proporciona capacidades de programación de tareas, asignación de recursos y supervisión de tareas para las cargas de trabajo de HPC. AWS PCS es compatible con el programador Slurm para gestionar las cargas de trabajo de los clústeres.

Temas

Versiones de Slurm en PCS AWS

SchedMD mejora continuamente Slurm con nuevas capacidades, optimizaciones y parches de seguridad. SchedMD lanza una nueva versión principal a intervalos regulares y planea admitir hasta 3 versiones en un momento dado. AWS El PCS está diseñado para actualizar automáticamente el controlador Slurm con versiones de parches.

Cuando SchedMD finaliza el soporte para una versión principal en particular, AWS PCS designa esa versión como End of Life (EOL). Tras la finalización de la vida útil, no se pueden crear nuevos clústeres con esa versión, aunque los clústeres existentes pueden seguir funcionando durante un máximo de 12 meses sin soporte garantizado. AWS PCS envía un aviso anticipado si una versión principal de Slurm está próxima a la EOL, para que los clientes sepan cuándo actualizar sus clústeres a una versión compatible más reciente.

Le recomendamos que utilice la última versión compatible de Slurm para implementar su clúster y acceder a los avances y mejoras más recientes.

Versiones de Slurm compatibles en PCS AWS

La siguiente tabla muestra las versiones de Slurm compatibles y las fechas e información importantes de cada versión.

Versión Slurm	Fecha de lanzamiento de SchedMD	AWS Fecha de lanzamiento de PCS	AWS Fecha de EOL de PCS	Versión mínima compatible del agente AWS PCS	Ejemplo de AWS PCS compatible AMIs
25.05	29 de mayo de 2025	16/10/2025	31/05/2027	1,00-1	• `aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05` • `aws-pcs-sample_ami-amzn2-arm64-slurm-25.05`
24,11	29/11/2024	14/05/2025	31/05/2026	1.0.0-1	• `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.11` • `aws-pcs-sample_ami`

Versión Slurm	Fecha de lanzamiento de SchedMD	AWS Fecha de lanzamiento de PCS	AWS Fecha de EOL de PCS	Versión mínima compatible del agente AWS PCS	Ejemplo de AWS PCS compatible AMIs
					`-amzn2-arm64-slurm-24.11`

Versiones de Slurm no compatibles en PCS AWS

La siguiente tabla muestra las versiones de Slurm que no son compatibles con el PCS. AWS

Versión Slurm	Fecha de lanzamiento de SchedMD	AWS Fecha de lanzamiento de PCS	AWS Fecha de EOL de PCS		
24.05	30/05/2024	18 de diciembre de 2024	30/11/2025		
23,11	21/11/2023	28 de agosto de 2024	31/05/2025		

Notas de publicación de las versiones de Slurm en PCS AWS

En este tema se describen los cambios importantes de cada versión de Slurm actualmente compatible con el PCS. AWS Le recomendamos que revise los cambios entre la versión antigua y la nueva cuando actualice el clúster.

Slurm 25.05

Cambios implementados en el PCS AWS

- El Slurm SchedulerParameter requeue_on_resume_failure ahora está activado de forma predeterminada.
- Se eliminó «stderr» como opción, ya que estaba deshabilitado en Slurm 25.05. LogTimeFormat
- AWS El PCS admite la configuración de paquetes de varios clústeres: el nodo de inicio de sesión puede acceder a varios clústeres.

Para obtener más información sobre Slurm 25.05, consulte las siguientes publicaciones:

- Anuncio de lanzamiento de SchedMD: https://www.schedmd.com/slurm-version-25-05-0-is-now-available/
- Notas de publicación de SchedMD: _Notes.md https://github.com/SchedMD/ slurm/blob/ slurm-25-05-0-1/RELEASE

Slurm 24.11

Cambios implementados en el PCS AWS

- AWS PCS admite la contabilidad de Slurm. Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.

Para obtener más información sobre Slurm 24.11, consulte las siguientes publicaciones:

- Anuncio de lanzamiento de SchedMD
- Notas de lanzamiento de SchedMD

Slurm 24.05

Cambios implementados en el PCS AWS

- El nuevo módulo Slurm Step Manager ahora está activado por defecto en AWS el PCS. Este módulo ofrece importantes ventajas al transferir la gestión por pasos del controlador central a los nodos de cómputo, lo que mejora sustancialmente la simultaneidad del sistema en entornos con un

uso intensivo de pasos. Para admitir esta configuración y aislar `Prolog` y `Epilog` procesar mejor los procesos, están habilitados los nuevos indicadores de prólogo (`Contain`,`Alloc`).

- La comunicación jerárquica entre el controlador y los nodos de cómputo permite optimizar la comunicación entre los nodos de Slurm, lo que mejora la escalabilidad y el rendimiento. Además, la configuración de enrutamiento ahora usa listas de nodos de partición para las comunicaciones desde el controlador, en lugar del algoritmo de enrutamiento predeterminado del complemento, lo que mejora la resiliencia del sistema.
- Un nuevo complemento de hash `HashPlugin=hash/sha3` reemplaza al anterior`hash/k12` `plugin`. Ahora está activado de forma predeterminada en los clústeres de AWS PCS.
- Los registros del controlador Slurm ahora incluyen capacidades de auditoría mejoradas para todas las llamadas entrantes a procedimientos remotos (RPC). `slurmctld` Los registros incluyen la dirección de origen, el usuario autenticado y el tipo de RPC antes del procesamiento de la conexión.

Para obtener más información sobre Slurm 24.05, consulte las siguientes publicaciones:

- Anuncio de lanzamiento de SchedMD
- Notas de lanzamiento de SchedMD

Slurm 23.11

La configuración de Slurm se puede cambiar en PCS AWS

- El `SuspendTime` valor predeterminado es. `60` Utilice el parámetro de `scaleDownIdleTimeInSeconds` configuración AWS PCS para configurarlo. Para obtener más información, consulte el `scaleDownIdleTimeInSeconds`parámetro del tipo de `ClusterSlurmConfiguration` datos en la referencia de la API de AWS PCS.
- El `MaxJobCount` y `MaxArraySize` se basa en el tamaño que elija para el clúster. Para obtener más información, consulte el `size`parámetro de la acción de la `CreateCluster` API en la referencia de la API de AWS PCS.
- La configuración predeterminada de `SelectTypeParameters` Slurm es. `CR_CPU` Puede proporcionarlo como un valor para configurarlo `slurmCustomSettings` al crear un clúster. Para obtener más información, consulte el `slurmCustomSettings`parámetro de la acción de la `CreateCluster` API y SlurmCustomSettingen la Referencia de la API de AWS PCS.

- Puede configurar `Prolog` y `Epilog` a nivel de clúster. Puede proporcionarlo como un valor `slurmCustomSettings` para configurarlo al crear un clúster. Para obtener más información, consulte `CreateCluster`y SlurmCustomSettingen la referencia de la API de AWS PCS.
- Puede configurar `Weight` y `RealMemory` a nivel de grupo de nodos de cómputo. Puede proporcionarlo como un valor para configurarlo `slurmCustomSettings` al crear un grupo de nodos de procesamiento. Para obtener más información, consulta `CreateComputeNodeGroup`y consulta SlurmCustomSettingla referencia de la API de AWS PCS.

Preguntas frecuentes sobre las versiones de Slurm en PCS AWS

AWS El PCS mantiene el soporte para varias versiones de Slurm. Cuando se presenta una nueva versión de Slurm, AWS PCS proporciona soporte técnico y parches de seguridad hasta que SchedMD llegue a su fin de soporte (EOS). AWS Para mantener la coherencia con la terminología utilizada en PCS, una versión de Slurm indica el final de su vida útil (EOL) para mantener la coherencia con la terminología. AWS

¿Durante cuánto tiempo es compatible AWS PCS con una versión Slurm?

AWS El soporte de PCS para las versiones de Slurm se alinea con los ciclos de soporte de SchedMD para las versiones principales. AWS PCS admite la versión actual y las 2 versiones principales anteriores más recientes. Cuando SchedMD lanza una nueva versión principal, AWS PCS deja de dar soporte a la versión compatible más antigua. AWS PCS lanza las nuevas versiones principales de Slurm lo antes posible, pero es posible que haya un retraso entre el lanzamiento de SchedMD y su disponibilidad en PCS. AWS

¿Cómo obtienen mis clústeres las nuevas versiones de parches de Slurm?

Para corregir errores y corregir problemas de seguridad, el AWS PCS está diseñado para aplicar automáticamente los parches a los controladores de clúster que se ejecutan en las cuentas internas propiedad del servicio. Para instalar parches en sus instancias EC2 Cuenta de AWS, actualice la Amazon Machine Image (AMI) de sus grupos de nodos de cómputo y actualice los grupos de nodos de cómputo para usar la AMI actualizada. Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

> **Note**
>
> Los controladores Slurm no estarán disponibles mientras los actualizamos. Los trabajos en ejecución no se ven afectados. Los trabajos enviados antes de que el controlador del clúster dejara de estar disponible se retienen hasta que el controlador esté disponible.

¿Cómo me informan sobre un próximo evento de EOL de la versión de Slurm?

Te enviamos un mensaje de correo electrónico 6 meses antes de la fecha de fin de vida. Le enviamos un mensaje de correo electrónico cada mes antes de la EOL, y un último mensaje de correo electrónico una semana antes de la fecha de EOL. Después de la fecha de fin de vida, enviamos mensajes de correo electrónico mensuales durante 12 meses a los clientes que utilizan clústeres de AWS PCS con versiones de EOL Slurm. Podríamos suspender un clúster con una versión de EOL Slurm si se identifican vulnerabilidades de seguridad en esa versión.

¿Cómo puedo determinar si la versión de Slurm que utiliza mi clúster ejecuta una versión de Slurm de EOL?

Le enviamos un mensaje de correo electrónico para notificarle que tiene un clúster en ejecución con una versión de EOL Slurm. Publicamos una alerta en las Panel de AWS Health alertas que contiene los detalles de sus clústeres con versiones de EOL Slurm. También puede utilizar la consola AWS PCS para identificar los clústeres con las versiones de EOL Slurm.

¿Qué debo hacer si mi versión de Slurm se acerca o supera el EOL?

Cree un clúster nuevo con una versión compatible más reciente de Slurm y actualice la versión de Slurm en las AMI del grupo de nodos de cómputo. La versión de Slurm de las AMI y las instancias de EC2 en ejecución no pueden estar más de dos versiones por detrás de la versión de Slurm del clúster. Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

¿Qué pasará si no cambio a una versión más reciente de Slurm antes de la fecha de fin de vida?

No puedes crear nuevos clústeres con una versión de EOL Slurm. Los clústeres existentes pueden funcionar hasta 12 meses sin AWS soporte y no se requiere ninguna acción inmediata para mantener su funcionamiento. Después de la fecha de fin de vida, no se garantizan el soporte, las actualizaciones de seguridad ni la disponibilidad. Es posible que suspendamos un clúster por motivos de seguridad. Le recomendamos encarecidamente que utilice una versión de Slurm compatible para mantener la seguridad y el soporte de sus clústeres de AWS PCS.

¿Cuáles son los riesgos de operar un clúster con las versiones de EOL Slurm?

Los clústeres con versiones de EOL Slurm presentan importantes riesgos operativos y de seguridad. Sin la supervisión activa de SchedMD, es posible que las vulnerabilidades de seguridad pasen desapercibidas o no se aborden. Si se descubren vulnerabilidades críticas, es posible que suspendamos sus clústeres de inmediato.

¿Qué ocurre con mis trabajos, la computación del clúster, el almacenamiento y los recursos de red cuando se suspende mi clúster?

Se cancelan todos los recursos gestionados por AWS PCS. Esto incluye el controlador Slurm, los grupos de nodos de cómputo y las instancias EC2. Todos los trabajos que se ejecuten en las instancias de procesamiento finalizan inmediatamente y el clúster pasa a un estado suspendido. Los recursos administrados por el cliente, como los sistemas de archivos externos, permanecen intactos. Puede utilizar la consola AWS PCS y las acciones de la API para acceder a la configuración del clúster.

¿Puedo reiniciar un clúster suspendido para reanudar sus tareas restantes?

No, no puedes reiniciar un clúster suspendido. Puedes usar la configuración del clúster suspendido para crear un clúster nuevo con una versión de Slurm compatible. Puede ejecutar los trabajos restantes si los ha guardado en un sistema de archivos externo.

¿Puedo solicitar una prórroga más allá del período de gracia de 12 meses?

No, no puedes solicitar una extensión para ejecutar tu clúster más allá del período de gracia de 12 meses. Proporcionamos la prórroga para ayudarte a cambiar a una versión de Slurm compatible. Para evitar que se interrumpan las operaciones del clúster, le recomendamos que cambie antes de que la versión de Slurm llegue a su fin de vida.

Contabilidad de Slurm en PCS AWS

Puede habilitar la contabilidad en sus nuevos clústeres de AWS PCS para supervisar el uso de los clústeres, hacer cumplir los límites de recursos y gestionar un control de acceso detallado a colas o grupos de nodos de cómputo específicos. AWS PCS crea y administra la base de datos de contabilidad para su clúster, lo que elimina la necesidad de crear y administrar su propia base de datos de contabilidad independiente. AWS PCS utiliza la función de contabilidad de Slurm. Para obtener más información sobre la función de contabilidad de Slurm, consulte la documentación de Slurm en SchedMD.

Para utilizar la contabilidad, habilítela al crear un nuevo clúster y, si lo desea, defina los parámetros contables. Cuando el estado del clúster sea `Active` y tenga grupos de nodos de cálculo, puede conectarse al shell de Linux de un nodo de inicio de sesión para realizar funciones de contabilidad, como ver los datos del trabajo con el comando Slurm`sacct`.

> **Note**
>
> La contabilidad es compatible con Slurm 24.11 o versiones posteriores.

AWS PCS console

En la página Crear clúster, debe seleccionar una versión válida de Slurm (versión 24.11 o posterior). En la configuración del programador, habilite la contabilidad.

AWS PCS API

Proporciona la `accounting` configuración en tu llamada a la acción de la `CreateCluster` API. En el `accounting` objeto, `mode` defina el`STANDARD`. Para obtener más información, consulte CreateClusterla referencia sobre la contabilidad en la API de AWS PCS.

En el siguiente ejemplo, se utiliza AWS CLI para llamar a la acción `CreateCluster` de la API. La subcadena de valores del parámetro `accounting='{mode=STANDARD}'` permite la contabilidad.

```
aws pcs create-cluster --cluster-name cluster-name \
                       --scheduler type=SLURM,version=24.11 \
                       --size SMALL \
                       --networking subnetIds=cluster-subnet-
id,securityGroupIds=cluster-security-group-id \
                       --slurm-configuration
 scaleDownIdleTimeInSeconds=180,accounting='{mode=STANDARD}',slurmCustomSettings='[{paramete
```

> **Important**
>
> Obtendrá cargos de facturación adicionales si habilita la contabilidad. Para obtener más información, consulta la página de precios de AWS PCS.

Modificación de la configuración contable

Puede activar o desactivar la contabilidad en los clústeres existentes sin necesidad de reconstruir la infraestructura. Para obtener más información, consulte Actualización de un clúster en AWS PCS.

Al deshabilitar la contabilidad, la facturación de la función de contabilidad se detiene en cuanto el clúster entra en `UPDATING` estado. Al activar la contabilidad, la facturación comienza cuando el clúster vuelve correctamente al `ACTIVE` estado.

Conceptos clave para la contabilidad de Slurm en PCS AWS

Los siguientes conceptos son específicos de AWS PCS y controlan la forma AWS en que PCS implementa la contabilidad de Slurm.

Base de datos de contabilidad

AWS PCS almacena sus datos contables en una base de datos creada en una Cuenta de AWS entidad AWS propietaria. No tiene acceso al `slurmdbd.conf`.

Tiempo de purga predeterminado

Esta configuración de AWS PCS especifica el período de retención (en días) para todos los tipos de registros contables (trabajos, eventos, reservas, pasos, suspensiones, transacciones y datos de uso). Por ejemplo, si el valor es 30, AWS PCS conserva los registros contables durante 30 días. Este valor se proporciona al crear el clúster. Si no proporciona un valor, AWS PCS conserva los registros contables en la base de datos de forma indefinida.

AWS PCS console

El tiempo de purga predeterminado se especifica como parte de los pasos para crear un clúster. En la página Crear clúster, debe seleccionar una versión válida de Slurm (versión 24.11 o posterior) y activar la contabilidad. En la configuración del programador, proporcione un valor entero para el tiempo de purga predeterminado (días).

AWS PCS API

Especifica esto `defaultPurgeTimeInDays` como parte de la `accounting` información que proporciones en la llamada a la acción de la `CreateCluster` API. Para obtener más información, consulta CreateClusterla referencia sobre la contabilidad en la API de AWS PCS.

> **Note**
>
> Cuando utiliza la API de AWS PCS para crear un clúster, el valor predeterminado `defaultPurgeTimeInDays` es `-1` y `0` no es un valor válido.

Aplicación de la política contable

Esta configuración determina con qué rigor Slurm aplica las reglas de envío de trabajos, los límites de recursos y las políticas contables para su clúster. Esta configuración corresponde al `AccountingStorageEnforce` parámetro del archivo del clúster. `slurm.conf` Puede seleccionar cualquier combinación de opciones de aplicación. Si no selecciona ninguna opción, no se aplicarán restricciones contables a los trabajos del clúster. AWS PCS admite las siguientes opciones:

- asociaciones: job-to-account mapeo
- límites: restricciones de recursos
- QoS: requisitos de calidad de servicio
- modo seguro: finalización garantizada dentro de los límites
- nosteps: desactiva la contabilidad de pasos
- nojobs: desactiva la contabilidad de trabajos

Para obtener más información sobre estas opciones, consulte la documentación de Slurm en SchedMD.

AWS PCS console

Las opciones se configuran como parte de los pasos para crear un clúster. En la página Crear clúster, debe seleccionar una versión válida de Slurm (versión 24.11 o posterior) y activar la contabilidad. Seleccione las opciones que desee de la lista desplegable de aplicación de las políticas contables en la configuración del programador.

AWS PCS API

En Slurm, estas opciones se configuran en un archivo de clúster. `slurm.conf` No tiene acceso directo al clúster de `slurm.conf` su AWS PCS. En su lugar, usted proporciona `SlurmCustomSettings` a la `CreateCluster` API una acción al crear un clúster. Para obtener más información, consulta la referencia CreateClusterde la API de AWS PCS.

Obtenga la configuración de contabilidad de un clúster de AWS PCS existente

La configuración de contabilidad de Slurm se incluye en la configuración de Slurm de su clúster.

AWS PCS console

1. Seleccione Clústeres en el panel de navegación.
2. Elija el nombre del clúster de la lista.
3. En la pestaña Configuración, busque la configuración de cuentas en Configuración de Slurm

AWS PCS API

Utilice la acción de la `GetCluster` API para obtener la configuración del clúster. Puede encontrar la configuración de contabilidad en`slurmConfiguration`. La configuración `mode` y el valor de `defaultPurgeTimeInDays` son inferiores a`accounting`. Las opciones de aplicación de la política contable seleccionadas se muestran a continuación`slurmCustomSettings`. Para obtener más información, consulte la referencia GetClusterde la API de AWS PCS.

API REST de Slurm en PCS AWS

AWS PCS proporciona soporte gestionado para la API REST nativa de Slurm mediante una interfaz HTTP para la interacción `slurmrestd` programática entre clústeres. Puede enviar trabajos, supervisar el estado del clúster y administrar los recursos mediante solicitudes HTTP estándar sin necesidad de acceder directamente a su clúster desde el shell.

Casos de uso comunes

La API REST de Slurm admite varios escenarios de integración:

- Integración de aplicaciones web: cree interfaces y aplicaciones web personalizadas que envíen y administren trabajos directamente.
- Integración con Jupyter Notebook: permite a los investigadores enviar trabajos desde entornos portátiles sin abandonar su flujo de trabajo de desarrollo.
- Integración de soluciones de socios: conecte herramientas de HPC y administradores de flujo de trabajo de terceros a sus clústeres de AWS PCS.

- Gestión programática de clústeres: automatice los flujos de trabajo de envío de trabajos, supervisión y gestión de recursos.
- Flujos de trabajo de computación de investigación: Support entornos de investigación académicos y empresariales que requieren una gestión de trabajos basada en API.

Requisitos y limitaciones

Antes de usar la API REST de Slurm, revise estos detalles:

- Su clúster debe usar la versión 25.05 o superior de Slurm.
- Solo se podrá acceder al punto final de la API a través de una dirección IP privada dentro de la VPC del clúster.
- El grupo de seguridad del clúster debe permitir el tráfico HTTP en el puerto 6820.
- La autenticación requiere tokens JWT con declaraciones de identidad de usuario específicas.

Las limitaciones actuales incluyen:

- No se admiten `scontrol token` los tokens generados por.
- `X-SLURM-USER-NAME`la suplantación de encabezados no está disponible.
- Algunas funciones requieren que la contabilidad de Slurm esté habilitada.
- No es compatible con el mecanismo del complemento de filtro CLI de Slurm.
- Las conexiones al punto final de la API REST no se cifran con TLS.

Temas

Habilitación de la API REST de Slurm en PCS AWS

Habilite la API REST de Slurm para acceder a la interfaz HTTP de su clúster para gestionar y supervisar los trabajos mediante programación. Puede habilitar esta función durante la creación del clúster o actualizar un clúster existente que cumpla con los requisitos.

Requisitos previos

Antes de habilitar la API REST de Slurm, asegúrate de tener:

- Versión de clúster: Slurm, versión 25.05 o superior.
- Grupo de seguridad: reglas que permiten el tráfico HTTP en el puerto 6820 desde las fuentes deseadas.

Procedimiento

Para habilitar la API REST de Slurm en un clúster nuevo

Consola de administración de AWS

1. Abra la consola AWS PCS en. https://console.aws.amazon.com/pcs/
2. Elija Create cluster.
3. En Detalles del clúster, seleccione Slurm, versión 25.05 o superior.
4. Configure los demás ajustes del clúster según sea necesario.
5. En la sección de configuración del programador, establece la API REST en Habilitada.
6. Configure el grupo de seguridad del clúster para permitir el tráfico HTTP en el puerto 6820 desde las fuentes que desee.
7. Complete el proceso de creación del clúster.

AWS CLI

1. Agregue una configuración REST de Slurm al crear el clúster.

```
aws pcs create-cluster --region region \
    --cluster-name my-cluster \
    --scheduler type=SLURM, version=25.05 \
    --size SMALL \
    --networking subnetIds=subnet-ExampleId1,securityGroupIds=sg-ExampleId1 \
    --slurm-configuration slurmRest='{mode=STANDARD}'
```

2. Configure el grupo de seguridad del clúster para permitir el tráfico HTTP en el puerto 6820 desde las fuentes que desee.

Para habilitar la API REST de Slurm en un clúster existente

Consola de administración de AWS

1. Abra la consola AWS PCS en. https://console.aws.amazon.com/pcs/
2. Elija su clúster de la lista.
3. Compruebe que su clúster utilice la versión 25.05 o superior de Slurm en los detalles del clúster.
4. Seleccione Editar clúster.
5. En la sección de configuración del programador, establece la API REST en Habilitada.
6. Seleccione Actualizar clúster para aplicar los cambios.
7. Configure el grupo de seguridad del clúster para permitir el tráfico HTTP en el puerto 6820 desde las fuentes que desee.

AWS CLI

1. Actualice el clúster con una configuración REST de Slurm, como en este ejemplo.

```
aws pcs update-cluster --cluster-identifier my-cluster \
    --slurm-configuration 'slurmRest={mode=STANDARD}'
```

2. Configure el grupo de seguridad del clúster para permitir el tráfico HTTP en el puerto 6820 desde las fuentes que desee.

¿Qué sucede después de habilitarlo

Al habilitar la API REST, AWS PCS automáticamente:

- Genera una clave de firma JWT y la almacena en AWS Secrets Manager.
- Expone el punto final de la API `https://<clusterPrivateIpAddress>:6820` dentro de su VPC.
- Actualiza la configuración del clúster para mostrar los detalles del punto final de la API REST.

Ahora puedes autenticar y usar la API REST para la administración de trabajos y las operaciones de clúster.

Autenticación con la API REST de Slurm en PCS AWS

La API REST de Slurm en AWS PCS utiliza la autenticación por token web JSON (JWT) para garantizar el acceso seguro a los recursos del clúster. AWS PCS proporciona una clave de firma gestionada almacenada en AWS Secrets Manager, que se utiliza para generar tokens JWT que contienen declaraciones de identidad de usuario específicas.

Requisitos previos

Antes de autenticarse con la API REST de Slurm, asegúrese de tener lo siguiente:

- Configuración del clúster: clúster de AWS PCS con Slurm 25.05+ y la API REST habilitadas.
- Permisos de AWS: acceso a AWS Secrets Manager para obtener la clave de firma de JWT.
- Información de usuario: nombre de usuario, ID de usuario de POSIX y uno o más grupos de POSIX IDs para su cuenta de clúster.
- Acceso a la red: conectividad dentro de la VPC del clúster con un grupo de seguridad que permite el puerto 6820.

Procedimiento

Para recuperar la dirección de punto final de la API REST de Slurm

Consola de administración de AWS

1. Abra la consola AWS PCS en. https://console.aws.amazon.com/pcs/
2. Elija su clúster de la lista.
3. En los detalles de configuración del clúster, busque la sección Endpoints.
4. Anote la dirección IP privada y el puerto de la API REST de Slurm (slurmrestd).
5. Puede realizar llamadas a la API enviando solicitudes HTTP con el formato correcto a esta dirección.

AWS CLI

1. Consulta el estado de tu clúster con`aws pcs get-cluster`. Busca el `SLURMRESTD` punto final en el `endpoints` campo de la respuesta. A continuación se muestra un ejemplo:

```
"endpoints": [
```

```
    {
        "type": "SLURMCTLD",
        "privateIpAddress": "192.0.2.1",
        "port": "6817"
    },
    {
        "type": "SLURMRESTD",
        "privateIpAddress": "192.0.2.1",
        "port": "6820"
    }
]
```

2. Puede realizar llamadas a la API enviando solicitudes HTTP con el formato adecuado a `http://<privateIpAddress>:<port>/`

Para recuperar la clave de firma de JWT

1. Abra la consola AWS PCS en https://console.aws.amazon.com/pcs/.
2. Elija su clúster de la lista.
3. En los detalles de configuración del clúster, busque la sección de autenticación del programador.
4. Anote el ARN y la versión de la clave JSON Web Token (JWT).
5. Utilice el AWS CLI para recuperar la clave de firma de Secrets Manager:

```
aws secretsmanager get-secret-value --secret-
id arn:aws:secretsmanager:region:account:secret:name --version-id version
```

Para generar un token JWT

1. Cree un JWT con las siguientes afirmaciones obligatorias:
   - `exp`— Tiempo de caducidad en segundos desde 1970 para el JWT
   - `iat`— Tiempo actual en segundos desde 1970
   - `sun`— El nombre de usuario para la autenticación
   - `uid`— El seudónimo de POSIX
   - `gid`— El ID del grupo POSIX
   - `id`— Propiedades de identidad POSIX adicionales

- `gecos`— Campo de comentarios de usuario, que se utiliza a menudo para almacenar un nombre legible por humanos
- `dir`— Directorio de inicio del usuario
- `shell`— Consola predeterminada del usuario
- `gids`— Lista de grupos POSIX adicionales en los que se encuentra IDs el usuario

2. Firme el JWT con la clave de firma recuperada de Secrets Manager.
3. Establezca un tiempo de caducidad adecuado para el token.

> **Note**
>
> Como alternativa a la `sun` reclamación, puedes proporcionar cualquiera de los siguientes datos:
>
> - `username`
> - Un nombre de campo personalizado que se define mediante `userclaimfield` el `AuthAltParameters Slurm custom settings`
> - Un `name` campo dentro de la `id` reclamación

Para autenticar las solicitudes de API

1. Incluye el token JWT en tus solicitudes HTTP mediante uno de estos métodos:
   - Token de portador: agrega encabezado `Authorization: Bearer` *`<jwt>`*
   - Cabecera Slurm — Añadir cabecera `X-SLURM-USER-TOKEN:` *`<jwt>`*
2. Realice solicitudes HTTP al punto final de la API REST:

   Este es un ejemplo de cómo acceder a la `/ping` API mediante curl y el `Authorized: Bearer` encabezado.

   ```
   curl -X GET -H "Authorization: Bearer <jwt>" \
         http://<privateIpAddress>:6820/slurm/v0.0.43/ping
   ```

Ejemplo de generación de JWT

Obtenga la clave de firma JWT del clúster AWS PCS y guárdela como un archivo local. Sustituya los valores de aws-region, secret-arn y secret version por los valores adecuados para su clúster.

```
#!/bin/bash
SECRET_KEY=$(aws secretsmanager get-secret-value \
  --region aws-region \
  --secret-id secret-arn \
  --version-stage secret-version \
  --query 'SecretString' \
  --output text)
echo "$SECRET_KEY" | base64 --decode > jwt.key
```

Este ejemplo de Python ilustra cómo usar la clave de firma para generar un token JWT:

```
#!/usr/bin/env python3

import sys
import os
import pprint
import json
import time
from datetime import datetime, timedelta, timezone
from jwt import JWT
from jwt.jwa import HS256
from jwt.jwk import jwk_from_dict
from jwt.utils import b64decode,b64encode
if len(sys.argv) != 3:
    sys.exit("Usage: gen_jwt.py [jwt_key_file] [expiration_time_seconds]")
SIGNING_KEY = sys.argv[1]
EXPIRATION_TIME = int(sys.argv[2])
with open(SIGNING_KEY, "rb") as f:
    priv_key = f.read()
signing_key = jwk_from_dict({
    'kty': 'oct',
    'k': b64encode(priv_key)
})
message = {
    "exp": int(time.time() + EXPIRATION_TIME),
    "iat": int(time.time()),
    "sun": "ec2-user",
    "uid": 1000,
```

```
    "gid": 1000,
    "id": {
        "gecos": "EC2 User",
        "dir": "/home/ec2-user",
        "gids": [1000],
        "shell": "/bin/bash"
    }
}
a = JWT()
compact_jws = a.encode(message, signing_key, alg='HS256')
print(compact_jws)
```

El script imprimirá un JWT en la pantalla.

```
abcdefgtjwttoken...
```

Uso de la API REST de Slurm para la gestión de trabajos en PCS AWS

Descripción general de la API REST de Slurm

La API REST de Slurm proporciona acceso programático a las funciones de administración de clústeres a través de solicitudes HTTP. Comprender estas características clave le ayudará a utilizar la API con PCS de forma eficaz: AWS

- Protocolo de acceso: la API utiliza HTTP (no HTTPS) para la comunicación dentro de la red privada del clúster.
- Detalles de la conexión: accede a la API mediante la dirección IP privada del clúster y el `slurmrestd` puerto (normalmente el 6820). El formato de URL base completo es`http://<privateIpAddress>:6820`.
- Control de versiones de la API: la versión de la API corresponde a su instalación de Slurm. Para Slurm 25.05, utilice la versión v0.0.43. El número de versión cambia con cada versión de Slurm. Puede encontrar las versiones de API actualmente compatibles en las notas de la versión de Slurm.
- Estructura de URL: La estructura de URL de la API REST de Slurm es. `http://<privateIpAddress>:<port>/<api-version>/<endpoint>` Puede encontrar información detallada sobre el uso de los puntos finales de la API REST en la documentación de Slurm.

Requisitos previos

Antes de usar la API REST de Slurm, asegúrese de tener:

- Configuración del clúster: clúster de AWS PCS con Slurm 25.05+ y la API REST habilitadas.
- Autenticación: token JWT válido con las declaraciones de identidad de usuario adecuadas.
- Acceso a la red: conectividad dentro de la VPC del clúster con un grupo de seguridad que permita el puerto 6820.

Procedimiento

Para enviar un trabajo mediante la API REST

1. Cree una solicitud de envío de trabajos con los parámetros necesarios:

```
{
  "job": {
    "name": "my-job",
    "partition": "compute",
    "nodes": 1,
    "tasks": 1,
    "script": "#!/bin/bash\necho 'Hello from Slurm REST API'"
  }
}
```

2. Envíe el trabajo mediante una solicitud HTTP POST:

```
curl -X POST \
  -H "Authorization: Bearer <jwt>" \
  -H "Content-Type: application/json" \
  -d '<job-json>' \
  https://<privateIpAddress>:6820/slurm/v0.0.43/job/submit
```

3. Anote el ID de trabajo que aparece en la respuesta para fines de supervisión.

Para supervisar el estado del trabajo

1. Obtenga información sobre un trabajo específico:

```
curl -X GET -H "Authorization: Bearer <jwt>" \
```

```
    https://<privateIpAddress>:6820/slurm/v0.0.43/job/<job-id>
```

2. Enumere todos los trabajos del usuario autenticado:

```
curl -X GET -H "Authorization: Bearer <jwt>" \
    https://<privateIpAddress>:6820/slurm/v0.0.43/jobs
```

Para cancelar un trabajo

- Envíe una solicitud de ELIMINACIÓN para cancelar un trabajo específico:

```
curl -X DELETE -H "Authorization: Bearer <jwt>" \
    https://<privateIpAddress>:6820/slurm/v0.0.43/job/<job-id>
```

Preguntas frecuentes sobre la API REST de Slurm en PCS AWS

Esta sección responde a las preguntas frecuentes sobre la API REST de Slurm en AWS PCS.

¿Qué es la API REST de Slurm?

La API REST de Slurm es una interfaz HTTP que le permite interactuar con el administrador de cargas de trabajo de Slurm mediante programación. Puede usar métodos HTTP estándar, como GET, POST y DELETE, para enviar trabajos, monitorear el estado del clúster y administrar los recursos sin necesidad de acceder al clúster desde la línea de comandos.

¿Puedo usar los tokens generados por? **`scontrol token`**

No, la `scontrol token` salida estándar no es compatible con el AWS PCS. La API REST de PCS Slurm requiere tokens JWT enriquecidos que contengan declaraciones de identidad específicas que incluyan el nombre de usuario (`sun`), el ID de usuario POSIX (`uid`) y el grupo (). IDs `gids` Los tokens Slurm estándar carecen de estas afirmaciones obligatorias y la API los rechazará.

¿Puedo acceder a la API desde fuera de mi VPC?

No, solo se puede acceder al punto final de la API REST desde su VPC mediante la dirección IP privada del controlador Slurm. Para habilitar el acceso externo, implemente AWS servicios como Application Load Balancer with VPC Link o API Gateway, o establezca conexiones VPN o de emparejamiento de VPC para una conectividad segura.

¿Por qué la API usa HTTP en lugar de HTTPS?

La API REST de Slurm está pensada para ser un punto final interno dentro de la red privada del clúster. Para las implementaciones de producción que requieren cifrado, puedes implementar la SSL/TLS terminación en un nivel superior de tu arquitectura, por ejemplo, a través de una puerta de enlace de API, un equilibrador de carga o un proxy inverso.

¿Cómo puedo controlar el acceso a la API REST?

Configure las reglas del grupo de seguridad del clúster para restringir el acceso al puerto 6820 del controlador Slurm. Establezca reglas de entrada para permitir conexiones solo desde rangos de IP confiables o fuentes específicas dentro de su VPC, bloqueando el acceso no autorizado al punto final de la API.

¿Cómo se rota la clave de firma de JWT?

Pon tu clúster en modo de mantenimiento sin instancias activas y, a continuación, inicia la rotación de claves a través de AWS Secrets Manager. Cuando se complete la rotación, vuelva a activar las colas. Todos los tokens JWT existentes dejarán de ser válidos y deberán regenerarse con la nueva clave de firma de Secrets Manager.

¿Necesito activar la contabilidad de Slurm para usar la API REST?

No, la contabilidad de Slurm no es necesaria para las operaciones básicas de la API REST, como el envío y la supervisión de los trabajos. Sin embargo, todo el `/slurmdb` punto final requiere que la contabilidad esté activa.

¿Qué herramientas de terceros funcionan con la API REST de AWS PCS?

Muchos de los clientes de la API REST de Slurm existentes deberían funcionar con AWS PCS, incluido Slurm Exporter para Prometheus, y las aplicaciones personalizadas que siguen el formato estándar de la API REST de Slurm. SlurmWeb Sin embargo, las herramientas que se basan en la autenticación deberán modificarse `scontrol token` para que funcionen con los requisitos del PCS JWT. AWS

¿El uso de la API REST conlleva algún coste adicional?

No, no hay cargos adicionales por habilitar o usar la función de la API REST de Slurm. Como de costumbre, solo paga por los recursos del clúster subyacentes.

¿Cómo puedo solucionar los problemas de la API REST?

- Problemas de conectividad de red

Si no puede acceder al punto final de la API, verá que se agota el tiempo de espera de la conexión o se produce un error de «conexión rechazada» al realizar solicitudes HTTP al controlador de clúster.

Qué hacer: compruebe que su cliente esté en la misma VPC o que tenga el enrutamiento de red adecuado y confirme que su grupo de seguridad permita el tráfico HTTP en el puerto 6820 desde su IP o subred de origen.

- Problemas de autenticación REST de Slurm

  Si tu token JWT no es válido, ha caducado o está mal firmado, las solicitudes de API mostrarán el mensaje «Error de autenticación de protocolo» en el campo de errores de la respuesta.

  Ejemplos de mensajes de error:

```
{
"errors": [
    {
    "description": "Batch job submission failed",
    "error_number": 1007,
    "error": "Protocol authentication error",
    "source": "slurm_submit_batch_job()"
    }
  ]
}
```

  Qué hacer: Comprueba que tu token JWT esté formateado correctamente, no haya caducado y esté firmado con la clave correcta de Secrets Manager. Comprueba que el token esté correctamente formado e incluya las afirmaciones necesarias y que utilices el formato de encabezado de autenticación correcto.

- Job no se ejecuta después de enviarlo

  Si su token JWT es válido pero tiene una estructura interna o un contenido incorrectos, es posible que los trabajos hayan entrado en un estado de pausa (PD) con el código de motivo. `JobAdminHead` `scontrol show job <job-id>`Úselo para inspeccionar el trabajo; verá`JobState=PENDING, Reason=JobHeldAdmin`, y. `SystemComment=slurm_cred_create failure, holding job`

Qué hacer: La causa principal puede ser un error en los valores de JWT. Verifique que el token esté estructurado correctamente e incluya las afirmaciones requeridas según la documentación del PCS.

- Problemas con los permisos del directorio de trabajo

  Si la identidad de usuario especificada en su JWT carece de permisos de escritura en el directorio de trabajo del trabajo, el trabajo fallará y generará errores de permiso, algo similar a lo que ocurre `sbatch --chdir` con un directorio inaccesible.

  Qué hacer: Asegúrese de que el usuario especificado en su token de JWT tenga los permisos adecuados para el directorio de trabajo del trabajo.

Reiniciar los nodos de cómputo con Slurm en PCS AWS

AWS PCS admite el comando nativo de Slurm. `scontrol reboot` Utilice este comando para reiniciar los nodos de procesamiento sin reemplazar la instancia EC2. Otros métodos de reinicio (consola Amazon EC2 AWS CLI, parches automatizados o mantenimiento del sistema) hacen que el AWS PCS considere que la instancia EC2 está en mal estado y la sustituya.

Ventajas del reinicio de Slurm

El reinicio de Slurm ofrece varias ventajas para el mantenimiento del clúster:

- Conserve la capacidad: evite perder instancias EC2 con capacidad limitada a manos de otros clientes.
- Reduzca los costos: elimine los ciclos innecesarios de reemplazo de instancias y la facturación continua de los nodos inactivos.
- Recuperación más rápida: sin demoras en el aprovisionamiento en comparación con la sustitución de instancias.
- Flexibilidad operativa: elimine las pérdidas de memoria, elimine los archivos temporales y recupere los nodos de estados degradados.

Cuándo usar el reinicio de Slurm

Utilice el reinicio de Slurm para los escenarios comunes de mantenimiento operativo:

- Solución de problemas: resuelva los problemas de rendimiento o los procesos que no responden, especialmente en los nodos de la GPU.
- Limpieza de recursos: elimine las pérdidas de memoria, los archivos temporales o los `/tmp` procesos atascados que afectan al rendimiento laboral.
- Recuperación: recupere los nodos de estados bloqueados o degradados antes de tener que reemplazarlos por completo.

Limitaciones

- Solo los usuarios de Slurm Admin (usuarios root) pueden ejecutar comandos de reinicio.
- El soporte de reinicio está limitado a solo. `scontrol reboot`
- RebootProgram no se admite la configuración.
- Sin interfaz de consola, solo desde la línea de comandos.

Temas

Reinicie un nodo de cómputo mediante Slurm en PCS AWS

Utilice el comando reboot nativo de Slurm para resolver problemas de rendimiento, solucionar problemas de recursos o recuperarse de estados degradados sin perder la capacidad de la instancia EC2.

Requisitos previos

- Privilegios de administrador de Slurm (acceso de usuario root)
- Acceso a un nodo de inicio de sesión en el clúster de AWS PCS

Procedimiento

1. Conéctese a un nodo de inicio de sesión a través de la consola EC2.

a. En la consola de EC2, elija Instances (Instancias).

b. Seleccione su instancia de nodo de inicio de sesión.

c. Elija Conectar.

2. Identifique el nombre del nodo de procesamiento de destino mediante `sinfo` o `scontrol show node`.

```
sinfo
# or
scontrol show node
```

3. Ejecute el comando reboot mediante una de estas opciones:

> **Warning**
>
> No lo utilices `nextstate=DOWN` con el `scontrol reboot` comando. Este parámetro marca el nodo como en mal estado y activa el reemplazo de la instancia.

- Reinicio básico (espera a que el nodo quede inactivo):

```
scontrol reboot nodename
```

- Reinicio inmediato (drena el nodo y se reinicia cuando se completan los trabajos):

```
scontrol reboot ASAP nodename
```

- Reinicie con el motivo:

```
scontrol reboot ASAP reason="troubleshooting" nodename
```

- Reinicie con el estado de reanudación:

```
scontrol reboot ASAP nextstate=RESUME nodename
```

4. Supervise el progreso del reinicio mediante`scontrol show node`.

```
scontrol show node nodename
```

5. Compruebe que el nodo vuelva a funcionar una vez finalizado el reinicio.

Cancele un reinicio pendiente en el AWS PCS

Cancela un reinicio pendiente para evitar tiempos de inactividad innecesarios cuando el problema se haya resuelto o cuando ya no sea necesario reiniciar.

Requisitos previos

- Privilegios de administrador de Slurm
- El nodo debe tener un reinicio pendiente (que muestre el estado «reinicio iniciado»)
- Acceso al nodo de inicio de sesión para ejecutar el comando

Procedimiento

1. Conéctese al nodo de inicio de sesión.
2. Compruebe que el nodo tiene un reinicio pendiente mediante`scontrol show node`.

   ```
   scontrol show node nodename
   ```

   Busca «se ha producido un reinicio» en el estado del nodo.
3. Ejecuta el comando de cancelación.

   ```
   scontrol cancel_reboot nodename
   ```
4. Compruebe que la cancelación del reinicio y el estado del nodo vuelvan a la normalidad.

   ```
   scontrol show node nodename
   ```

Preguntas frecuentes sobre el reinicio de Slurm en PCS AWS

Encuentre respuestas a preguntas frecuentes sobre el uso del reinicio de Slurm en PCS. AWS

¿Qué es la compatibilidad con el reinicio de Slurm?

> Support para el comando Slurm `scontrol reboot` nativo. Utilice este comando para reiniciar los nodos de procesamiento sin reemplazar automáticamente las instancias, lo que preserva la capacidad de la instancia EC2 y reduce los costos operativos.

¿Quién puede usar los comandos de reinicio de Slurm?

Solo los usuarios de Slurm Admin (usuarios root) pueden ejecutar los comandos de reinicio. Los usuarios habituales que intenten utilizarlos `scontrol reboot` recibirán un error de permiso denegado por parte de Slurm sin que ello afecte al nodo.

¿Qué ocurre con los trabajos en ejecución durante un reinicio?

De forma predeterminada, los trabajos se completan normalmente antes de que se reinicie. Con la opción ASAP, el nodo se vacía para evitar nuevos trabajos y el reinicio se produce una vez finalizados los trabajos actuales. Los trabajos se pueden cancelar o volver a poner en cola para reiniciarse inmediatamente.

¿En qué se diferencia esto del reinicio de la consola EC2?

El reinicio mediante Slurm preserva la instancia EC2 y evita su sustitución, mientras que al reiniciar la consola EC2, el PCS reemplaza la instancia debido a que las comprobaciones de estado no se realizaron correctamente durante el proceso de reinicio.

¿Puedo configurar scripts de reinicio personalizados?

No, RebootProgram la configuración no se admite en la versión inicial. La función utiliza el comportamiento de reinicio estándar de Slurm sin compatibilidad con scripts personalizados.

¿Cuánto tarda un reinicio de Slurm?

El tiempo de reinicio varía según el tipo de instancia, los procesos de arranque del cliente, la configuración de la AMI y si los trabajos deben completarse primero. El proceso incluye esperar a que se completen los trabajos, el reinicio físico, las comprobaciones de estado y el registro del daemon slurmd.

¿Puedo ver un historial de reinicios?

Los eventos de reinicio se registran en los registros de Slurm (slurmctld y slurmd), que se pueden monitorear. CloudWatch El campo de motivo en el estado del nodo muestra el motivo del reinicio durante el proceso.

¿Qué pasa si un nodo se atasca durante el reinicio?

Si un nodo no completa el proceso de reinicio dentro de él ResumeTimeout, se marcará como INACTIVO. Compruebe si hay errores en los CloudWatch registros, compruebe la conectividad de la red y examine los registros slurmd. Póngase en contacto con AWS Support si los problemas persisten.

¿Puedo reiniciar varios nodos a la vez?

Sí, puede especificar varios nodos en el comando reboot:

```
scontrol reboot ASAP node1,node2,node3
```

¿Cómo puedo reiniciar un nodo sin esperar a que se completen las tareas?

Tienes dos opciones para que los nodos se reinicien inmediatamente cuando surjan problemas, como nodos problemáticos que afecten a tareas de varios nodos, una degradación significativa del rendimiento o un comportamiento inestable de la GPU:

- Cancelar y reiniciar: primero, cancele los trabajos afectados utilizando y, a continuación`scancel <job_id>`, inicie un reinicio inmediato utilizando. `scontrol reboot ASAP <nodename>` Los trabajos en ejecución finalizarán y deberán volver a enviarse una vez que el nodo se recupere.
- Drenar y volver a poner en cola (con menos impacto): comience por iniciar un vaciado y reinicie con`scontrol reboot ASAP <nodename>`, a continuación, vuelva a poner en cola los trabajos afectados utilizando. `scontrol requeue <job_id>` Esto vuelve a poner los trabajos en estado pendiente en lugar de cancelarlos.

¿Qué ocurre si especifico nextState=DOWN?

Si lo especifica`nextstate=DOWN`, el nodo se marcará como en mal estado tras el reinicio y se activará la sustitución de la instancia. Para evitar el reemplazo de la instancia, no especifiques el estado o el uso `nextstate=RESUME` del siguiente.

Recursos adicionales

- Para obtener información sobre los procedimientos básicos de reinicio, consulteReinicie un nodo de cómputo mediante Slurm en PCS AWS.
- Para solucionar problemas de reinicio, consulteSolución de problemas de reinicio de Slurm en PCS AWS.
- Para ver la documentación sobre el reinicio de Slurm, consulte la documentación de Slurm scontrol.

Solución de problemas de reinicio de Slurm en PCS AWS

Cuando tenga problemas con el reinicio del nodo, compruebe primero el estado del nodo mediante. `scontrol show node` *`nodename`* A continuación, examine CloudWatch los registros de Slurm (slurmctld y slurmd) y los registros del sistema para identificar posibles errores.

Para solucionar problemas básicos, compruebe la conectividad de la red, compruebe la configuración del grupo de seguridad y asegúrese de que todos los servicios necesarios estén funcionando tras el reinicio. Si los problemas persisten después de los pasos básicos de solución de problemas, ponte en contacto con AWS Support. Cuando te pongas en contacto con el servicio de asistencia, proporciona los extractos del registro pertinentes, la información sobre el estado del nodo y una cronología del intento de reinicio para acelerar el proceso de resolución.

Recursos adicionales

- Para monitorizar instancias de AWS PCS mediante Amazon CloudWatch, consulta Supervisión de instancias de AWS PCS CloudWatch.
- Para obtener información general sobre la solución de problemas, consulteSolución de problemas en AWS Parallel Computing Service.
- Para ver la documentación de Slurm, consulte la Guía de solución de problemas de Slurm.

Configuración de ajustes de Slurm personalizados en PCS AWS

Usa los ajustes de Slurm personalizados para configurar parámetros de Slurm adicionales en los recursos de clústeres, colas y grupos de nodos de cómputo. Esta versión añade compatibilidad con la configuración de Slurm en los recursos de Queue, lo que proporciona un control pormenorizado de los comportamientos específicos de las particiones.

Ventajas de la configuración personalizada de Slurm

La configuración personalizada de Slurm proporciona un control sofisticado sobre su entorno de HPC basado en PC AWS . Puede implementar una contabilidad detallada, aplicar controles de acceso y optimizar la ejecución de la carga de trabajo mediante quality-of-service configuraciones y políticas de prevención. Estas capacidades garantizan que los trabajos críticos reciban los recursos necesarios y, al mismo tiempo, mantienen una utilización eficiente del clúster. Ya sea que gestione cargas de trabajo aceleradas por la GPU, implemente una programación equitativa o controle los ciclos de vida de los trabajos, las configuraciones personalizadas ayudan a alinear su infraestructura de HPC con los requisitos operativos y los objetivos de investigación.

Configuración de ajustes personalizados

Los ajustes de Slurm personalizados se pueden configurar a través de la AWS consola, la CLI o SDKs durante la creación de los recursos, o se pueden modificar posteriormente mediante operaciones de actualización.

Consola de administración de AWS

Acceda a la configuración adicional del programador en la página de creación o edición para cualquier tipo de recurso (clúster, cola o grupo de nodos de cómputo).

Para añadir una nueva configuración

1. Selecciona Añadir nueva configuración.
2. Seleccione un nombre de parámetro en el menú desplegable (que incluye breves descripciones de los parámetros).
3. Proporcione el valor correspondiente.

Para anular una configuración personalizada

1. Selecciona Eliminar junto al parameter/value par correspondiente.
2. Cree o actualice el recurso.

AWS CLI

Para la administración programática de la configuración personalizada, utilice el `SlurmCustomSettings` campo en las operaciones de creación o actualización.

Example— Actualizar el Prolog parámetro en un clúster

```
aws pcs update-cluster --cluster-identifier my-cluster \
--slurm-configuration \
'SlurmCustomSettings=[{parameterName=Prolog,parameterValue="/path/to/prolog.sh"}]'
```

Example— Configurar una cola para que esté Default en un clúster

```
aws pcs update-queue \
    --cluster-identifier my-cluster \
```

```
    --queue-identifier my-queue \
    --slurm-configuration
 'SlurmCustomSettings=[{parameterName=Default,parameterValue=YES}]'
```

Example— Configuración personalizada Features en un grupo de nodos de cómputo

```
aws pcs update-compute-node-group \
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-cng-1 \
    --slurm-configuration \
    'SlurmCustomSettings=[{parameterName=Features,parameterValue="gpu,nvme"}]'
```

Validación y gestión de errores

AWS PCS implementa un proceso de validación de varios niveles para la configuración personalizada de Slurm. Durante las operaciones de creación y actualización, realizamos validaciones sincrónicas que incluyen:

- Comprobaciones a nivel de campo: validamos las configuraciones individuales para comprobar los tipos de datos, los valores permitidos y los requisitos de formato correctos. Por ejemplo, nos aseguramos de que los valores de tiempo estén en el formato Slurm correcto y los valores booleanos utilizan las representaciones booleanas de Slurm aceptadas por Slurm.
- Validaciones sensibles al contexto: algunos ajustes se comparan con el contexto de configuración más amplio. Por ejemplo, algunos parámetros solo son válidos cuando la contabilidad de Slurm está habilitada.
- Coherencia entre configuraciones: verificamos que las opciones que se excluyen mutuamente no estén configuradas juntas y que las configuraciones interdependientes estén configuradas correctamente.

Si la validación no se realiza correctamente, recibirás un mensaje `ValidationException` con un código de error específico (por ejemplo, InvalidInput), un mensaje de error claro en el que se describe el problema y una lista de los campos no válidos y sus respectivos detalles de error.

Si bien durante esta validación inicial se detectan muchos problemas, es posible que algunas interacciones complejas entre los ajustes solo se hagan evidentes al aplicar la configuración. En esos casos, la operación fallará con un mensaje de error informativo y se revertirá cualquier cambio parcial.

Limitaciones

AWS El PCS implementa un enfoque de lista de permitidos para proteger la seguridad del servicio y la estabilidad operativa. Los ajustes que puedan comprometer la seguridad de la cuenta de servicio o interferir con las capacidades del servicio gestionado están restringidos. Sin embargo, evaluamos continuamente las necesidades de los clientes y podemos añadir soporte para configuraciones adicionales en función de los comentarios de los clientes.

Temas

Configuración de Slurm personalizada para AWS clústeres de PCS

A nivel de clúster, se admiten las siguientes configuraciones personalizadas de Slurm:

- AccountingStorageEnforce

  > **Important**
  >
  > AWS PCS admite un subconjunto de las opciones para. `AccountingStorageEnforce` Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.

- AccountingStorageTRES
- AccountingStoreFlags
- DefMemPerCPU
- Epilog
- EnforcePartLimits
- FairShareDampeningFactor
- HealthCheckInterval
- HealthCheckNodeState
- HealthCheckProgram
- JobRequeue

- LaunchParameters
- Licenses
- MinJobAge

> **Note**
>
> AWS PCS admite un valor mínimo de 5 segundos para`MinJobAge`.

- OverTimeLimit
- PreemptExemptTime
- PreemptMode
- PreemptParameters
- PreemptType
- PriorityCalcPeriod
- PriorityDecayHalfLife
- PriorityFavorSmall
- PriorityFlags
- PriorityMaxAge
- PriorityUsageResetPeriod
- PriorityWeightAge
- PriorityWeightAssoc
- PriorityWeightFairshare
- PriorityWeightJobSize
- PriorityWeightPartition
- PriorityWeightQOS
- PriorityWeightTRES
- PrivateData
- Prolog
- PrologFlags
- PropagatePrioProcess
- PropagateResourceLimits
- PropagateResourceLimitsExcept

- RequeueExit
- RequeueExitHold
- SchedulerParameters
- SelectTypeParameters
- SrunPortRange
- TaskEpilog
- TaskPluginParam
- TaskProlog
- UnkillableStepProgram
- UnkillableStepTimeout

Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS

A nivel de grupo de nodos de cómputo, se admiten las siguientes configuraciones personalizadas de Slurm:

- CpuSpecList
- Features
- MemSpecLimit
- RealMemory
- Weight

Configuración de Slurm personalizada para colas de PCS AWS

A nivel de cola, se admiten las siguientes configuraciones personalizadas de Slurm:

- AllowAccounts
- AllowQoS
- Default
- DefaultTime
- DenyAccounts
- DenyQoS

- ExclusiveUser
- GraceTime
- MaxTime
- OverSubscribe
- OverTimeLimit
- PreemptMode
- PriorityJobFactor
- PriorityTier
- QOS
- TRESBillingWeights

Solución de problemas de la configuración de Slurm personalizada en PCS AWS

Si encuentra errores al crear o actualizar los recursos del AWS PCS con la configuración personalizada de Slurm, puede utilizar el registro para diagnosticar y resolver los problemas.

Solución de problemas de configuración personalizada de Slurm incompatibles

Problema: recibe un mensaje de error similar al siguiente al realizar operaciones de clúster, grupo de nodos de cómputo o cola:

```
{OPERATION} failed. The Slurm custom settings of the cluster might be incompatible.
 Check the settings and try again.
```

Este error puede producirse con las siguientes operaciones:

- CreateCluster
- CreateComputeNodeGroup
- UpdateComputeNodeGroup
- CreateQueue
- UpdateQueue

Solución: habilite el registro para comprender el problema específico y solucionar los problemas de configuración incompatibles.

Para solucionar problemas de configuración personalizada de Slurm incompatibles

1. Cree el clúster si aún no existe o asegúrese de que el clúster existente esté en un estado en el que se pueda habilitar el registro.
2. Habilita el registro en tu clúster. Para obtener instrucciones detalladas, consulte Registro y supervisión para AWS PCS.

   > **Note**
   >
   > El registro se puede habilitar una vez que se haya creado el clúster.

3. Revise los registros para identificar el problema específico de configuración de Slurm que está causando la incompatibilidad.
4. Corrija la configuración personalizada incompatible en función de la información del registro y vuelva a intentar la operación.

Para obtener información sobre los ajustes personalizados de Slurm compatibles, consulte:

- Configuración de Slurm personalizada para AWS clústeres de PCS
- Configuración de Slurm personalizada para grupos de nodos de cómputo de AWS PCS
- Configuración de Slurm personalizada para colas de PCS AWS

Amplíe la funcionalidad de Slurm en los PCS con los complementos de AWS SPANK

Utilice los complementos SPANK (arquitectura de complementos de Slurm para Node and Job Kontrol) para ampliar y modificar el comportamiento de Slurm durante el lanzamiento y la ejecución de tareas en clústeres de PCS. AWS Los complementos de SPANK proporcionan una interfaz genérica para interceptar y modificar las etapas de inicio de los trabajos.

Instale los complementos de SPANK en la AMI de su nodo de cómputo y configúrelos para personalizar el comportamiento del clúster de Slurm según sus requisitos de carga de trabajo. Para obtener más información sobre SPANK, consulta la documentación de SPANK en el sitio web de SchedMD.

Contenido

Instala los complementos de SPANK en PCS AWS

Siga la documentación del complemento para instalar los complementos de SPANK en su AMI.

Compila los complementos de SPANK para la versión específica de Slurm en tu clúster. El instalador de Slurm proporcionado por AWS PCS almacena Slurm en. `/opt/aws/pcs/scheduler/slurm-version` Al compilar el complemento, especifique la versión de Slurm.

El siguiente ejemplo muestra cómo especificar la versión de Slurm para algunos complementos:

```
export CFLAGS="-I/opt/aws/pcs/scheduler/slurm-version/include"
```

Si tiene varias versiones de Slurm en la AMI, compile el complemento para cada versión. Guarde los complementos compilados en carpetas versionadas.

El siguiente ejemplo muestra cómo especificar la carpeta de destino de algunos complementos:

```
export DESTDIR="your-preferred-versioned-path"
```

> **Important**
>
> Los complementos pueden requerir variables diferentes. Consulta la documentación oficial del complemento que estás instalando.

Configura los complementos de SPANK en AWS PCS

De forma predeterminada, almacene los archivos de configuración en`/etc/aws/pcs/scheduler/slurm-version/plugstack.conf.d/`.

Para almacenar su configuración de SPANK en una ubicación diferente, añada sus ubicaciones a un archivo de configuración en el directorio predeterminado.

El siguiente ejemplo muestra cómo incluir archivos de configuración de otros directorios:

```
# content of /etc/aws/pcs/scheduler/slurm-version/any-filename.conf
include path-to-your-configuration-folder/*.conf
include path-to-a-second-configuration-folder/*.conf
```

Guarde cada configuración en un archivo dedicado o en un archivo común. Puede utilizar varios archivos de configuración.

Los siguientes ejemplos muestran ejemplos de archivos de configuración:

```
# content of path-to-your-or-default-config-folder/filename-1.conf
required path-to-plugin-1 arguments
optional path-to-plugin-2 arguments
```

```
# content of path-to-your-or-default-config-folder/filename-2.conf
required path-to-plugin-3 arguments
```

Para obtener información adicional sobre cómo configurar sus complementos, consulte la documentación de configuración de SPANK en el sitio web de SchedMD.

Important

Configura los permisos de las carpetas para evitar cambios no autorizados en la configuración de tus complementos.

Note

AWS PCS no administra tus complementos de SPANK. Si recibes errores relacionados con los complementos, consulta los registros de errores de tus nodos de cómputo.

Note

Al cargar tu configuración de SPANK, Slurm registra incorrectamente un error similar al siguiente:

```
error: "Include" failed in file /etc/slurm/plugstack.conf line 3
```

Puede omitir este error. No afecta al funcionamiento de los complementos de SPANK.

Preguntas frecuentes sobre los complementos de SPANK en PCS AWS

En esta sección se abordan las preguntas más frecuentes sobre la instalación y configuración de los complementos de SPANK en clústeres de AWS PCS.

¿Necesito instalar los complementos de SPANK tanto en los nodos de inicio de sesión como en los nodos de cómputo?

Algunos complementos de SPANK no requieren instalación en todos los nodos, pero para una mayor compatibilidad, te recomendamos que instales todos los complementos de SPANK en todos los nodos.

¿Qué configuración adicional se necesita para el uso de producción de los complementos de SPANK?

Además de la instalación y la configuración básicas que se muestran en los ejemplos, las implementaciones de producción suelen requerir una configuración adicional. Los complementos basados en contenedores, como Pyxis, pueden requerir que establezca variables de entorno para Enroot, habilite la PMI (interfaz de administración de procesos) y configure los permisos para el tiempo de ejecución del contenedor. Consulte la documentación del complemento específico para conocer los requisitos detallados de implementación en producción.

¿Cómo puedo solucionar los problemas del plugin SPANK?

AWS PCS no administra los complementos de SPANK. Examine los registros de errores de sus nodos de cómputo para solucionar problemas.

Utilice los complementos de filtro CLI de Slurm para personalizar el envío de trabajos en PCS AWS

AWS PCS es compatible con los complementos de filtro CLI de Slurm para ejecutar scripts de Lua personalizados que validan y modifican los parámetros de envío de trabajos en los nodos de inicio de sesión y computación. Para obtener información detallada sobre los complementos de filtro CLI, consulte la documentación de la API del complemento cli_filter en el sitio web de SchedMD.

Requisitos

Los complementos de filtro CLI requieren la versión 24.11 o posterior de Slurm y un script de Lua implementado en todos los nodos de inicio de sesión y procesamiento.

> **Important**
>
> Para las versiones 24.11 y 25.05 de Slurm, los complementos de filtro CLI requieren la instalación de Slurm AWS mediante el instalador PCS Slurm (versión 24.11.6-2+ o 25.05.4-1+). Para obtener más información Paso 3: Instalar Slurm sobre la instalación de Slurm, consulte.

Limitaciones y consideraciones de seguridad

- Aplicación de la seguridad: cualquier usuario puede omitir fácilmente los complementos de filtro CLI y no deben usarse para políticas críticas para la seguridad. Los usuarios pueden deshabilitar los complementos de filtro CLI proporcionando una configuración personalizada que se `CLIFilterPlugins` deshabilita al enviar trabajos.
- Implementación de Lua únicamente: se admite la implementación de scripts de Lua. No se admite la implementación en C.

Temas

- Configurar los complementos de filtro CLI de Slurm en un AWS clúster de PCS
- Utilice Amazon S3 para implementar un script de complemento de filtro CLI en AWS PCS
- Traduzca un script del complemento Job Submit de Slurm para usar el complemento CLI Filter en PCS AWS
- Preguntas frecuentes sobre los complementos de filtro CLI de Slurm en PCS AWS
- Solución de problemas con el complemento de filtro CLI de Slurm en PCS AWS

Configurar los complementos de filtro CLI de Slurm en un AWS clúster de PCS

Configure los complementos de filtro CLI al crear un nuevo clúster de AWS PCS. Puede habilitar o deshabilitar los complementos de filtro CLI en los clústeres existentes mediante la API de actualización o la consola sin volver a crear el clúster.

Requisitos previos

Antes de configurar los complementos de filtro CLI, complete estas tareas:

- Escriba y pruebe un script de Lua que implemente la API del complemento CLI Filter
- Asigne un nombre exacto a su script de Lua `cli_filter.lua`
- Elija un método para implementar el script en todas las instancias del clúster (AMI, S3 o sistema de archivos)
- Compruebe que utiliza la versión 24.11 o posterior de Slurm

Habilitar los complementos de filtro CLI en un clúster nuevo

AWS PCS console

1. Abra la consola AWS PCS en https://console.aws.amazon.com/pcs/.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Elija Create cluster.
4. Seleccione una versión válida de Slurm (versión 24.11 o posterior).
5. En Configuración del programador, expanda Configuración adicional del programador.
6. Añada una nueva configuración personalizada de Slurm con el nombre del parámetro establecido en `CliFilterPlugins` y el valor del parámetro establecido en. `cli_filter/lua`
7. Complete la configuración del clúster restante y elija Crear clúster.

AWS PCS API

Proporciona la `slurmCustomSettings` configuración en tu llamada a la acción de la `CreateCluster` API. Defina el `parameterName` para `CliFilterPlugins` y

`parameterValue` el para`cli_filter/lua`. Para obtener más información, consulte la referencia CreateClusterde la API de AWS PCS.

En el siguiente ejemplo, se utiliza AWS CLI para llamar a la acción `CreateCluster` de la API. La configuración personalizada `CliFilterPlugins=cli_filter/lua` habilita los complementos de filtro CLI.

```
aws pcs create-cluster --cluster-name cluster-name \
--scheduler type=SLURM,version=24.11 \
--size SMALL \
--networking subnetIds=cluster-subnet-id,securityGroupIds=cluster-security-group-id
 \
--slurm-configuration \
'slurmCustomSettings=[{parameterName=CliFilterPlugins,parameterValue="cli_filter/
lua"}]'
```

Implementar scripts del complemento de filtro CLI

Para implementar scripts del complemento de filtro CLI en su clúster

1. Asegúrese de que todos los que AMIs se utilizan en los grupos de nodos de cómputo tengan Slurm instalado mediante el instalador AWS PCS Slurm.

   > **Note**
   >
   > Si usa la AMI de muestra de AWS PCS para todos los grupos de nodos de cómputo, omita este paso. Slurm ya está instalado.

2. Implemente su `cli_filter.lua` script `/etc/aws/pcs/scheduler/slurm-<version>/cli_filter.lua` en todas las instancias del clúster.

   Por ejemplo, para la versión 24.11 de Slurm:

   ```
   /etc/aws/pcs/scheduler/slurm-24.11/cli_filter.lua
   ```

3. Inicie todos los nodos de inicio de sesión y cálculo con el que esté preparado. AMIs
4. Pruebe el envío del trabajo para verificar que el complemento de filtro CLI se ejecute correctamente.

Habilitar o deshabilitar los complementos de filtro CLI en los clústeres existentes

Puede habilitar o deshabilitar los complementos de filtro CLI en los clústeres existentes sin necesidad de reconstruir la infraestructura. Para obtener más información, consulte Actualización de un clúster en AWS PCS.

AWS PCS console

1. Abra la consola AWS PCS en https://console.aws.amazon.com/pcs/.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Seleccione el clúster que desee actualizar.
4. Elija Editar acción.
5. En la página Editar clúster, en Configuración adicional del programador:
   - Para habilitar los complementos de filtro CLI: agregue una nueva configuración personalizada de Slurm con el nombre del parámetro establecido en `CliFilterPlugins` y el valor del parámetro establecido en. `cli_filter/lua`
   - Para deshabilitar los complementos de filtro CLI: elimine la `CliFilterPlugins` configuración existente.
6. Seleccione Actualizar clúster para enviar los cambios.
7. Supervisa el estado del clúster, que aparece como «Actualizando» durante el proceso y «Activo» cuando se completa la actualización.

AWS PCS API

Use la acción de la `UpdateCluster` API para habilitar o deshabilitar los complementos de filtro CLI. Para obtener más información, consulte la referencia UpdateClusterde la API de AWS PCS.

Para habilitar los complementos de filtro CLI en un clúster existente:

```
aws pcs update-cluster --cluster-identifier my-cluster \
--slurm-configuration \
'slurmCustomSettings=[{parameterName=CliFilterPlugins,parameterValue="cli_filter/
lua"}]'
```

Para deshabilitar los complementos de filtro CLI en un clúster existente:

```
aws pcs update-cluster --cluster-identifier my-cluster \
```

```
--slurm-configuration \
'slurmCustomSettings=[]'
```

Resultados esperados

Después de completar la configuración:

- El clúster se crea con el complemento de filtro CLI activado
- Los envíos de trabajos activan su lógica de validación personalizada antes de llegar al controlador Slurm
- Los trabajos no conformes se rechazan con tus mensajes de error personalizados
- Los trabajos que cumplen con los requisitos se procesan normalmente a través del programador Slurm

Resolución de problemas

Falta el script del complemento de filtro CLI en ningún nodo

Síntomas: El envío del trabajo falla inmediatamente y se produce un error al cargar el plugin.

Causa probable: el script no se implementó en todas las instancias o la ruta o el nombre del archivo son incorrectos.

Solución: compruebe que el script existe en la ruta correcta en todos los nodos de inicio de sesión y procesamiento con el nombre de archivo exacto`cli_filter.lua`.

Configuración del complemento de filtro CLI no válida

Síntomas: la creación del clúster falla debido a un error de validación.

Causa probable: el `CliFilterPlugins` parámetro no está configurado para `cli_filter/lua` formatear.

Resolución: utilice el valor exacto del parámetro `cli_filter/lua` en`slurmCustomSettings`.

Utilice Amazon S3 para implementar un script de complemento de filtro CLI en AWS PCS

Utilice S3 para implementar el script del complemento de filtro CLI cuando desee actualizar la lógica de envío de trabajos en un clúster activo sin necesidad de volver a crearlo AMIs. Este enfoque descarga el script de S3 durante el lanzamiento de la instancia utilizando los datos del usuario.

Requisitos previos

Antes de implementar el script mediante S3, complete estas tareas:

- Cree un bucket de S3 con el script Lua del complemento de filtro CLI
- Configure el perfil de la instancia de IAM con acceso de lectura al bucket de S3
- Configure el punto de enlace S3 VPC Gateway para acceso directo sin Internet
- Prepare el script de datos de usuario para descargarlo de S3

Para implementar el script del complemento de filtro CLI mediante S3

1. Cargue el `cli_filter.lua` script en su bucket de S3.
2. Configure el perfil de su instancia de IAM con los permisos de lectura de S3 para el bucket.
3. Añada el código shell a los datos de usuario de la plantilla de lanzamiento para descargar el script:

   ```
   aws s3 cp s3://my-bucket/cli_filter.lua /etc/aws/pcs/scheduler/slurm-24.11/
   cli_filter.lua
   chmod 644 /etc/aws/pcs/scheduler/slurm-24.11/cli_filter.lua
   ```

4. Implemente grupos de nodos de cómputo con sus plantillas de lanzamiento actualizadas.
5. Pruebe el envío de los trabajos para verificar la funcionalidad del script.

Resultados esperados

Tras completar la implementación de S3:

- El script del complemento CLI Filter se descarga automáticamente en todas las instancias durante el lanzamiento

- Las actualizaciones de scripts en S3 se reflejan en las instancias recién lanzadas
- Las políticas de presentación de trabajos se aplican de forma coherente en todo el clúster

Resolución de problemas

Acceso denegado a S3

Síntomas: se produce un error en el lanzamiento de la instancia o no se ha descargado el script.

Causa probable: falta de permisos de IAM o punto final de VPC de S3.

Solución: compruebe que el perfil de la instancia de IAM tenga `s3:GetObject` permiso y que el punto de enlace de VPC de S3 esté configurado.

Traduzca un script del complemento Job Submit de Slurm para usar el complemento CLI Filter en PCS AWS

Traduzca el script Lua de su plugin Job Submit Plugin al CLI Filter Plugin cuando migre desde otros entornos de Slurm. El proceso de traducción implica actualizar los nombres de las funciones y los patrones de acceso a los campos para que funcionen con la API del complemento CLI Filter.

Requisitos previos

Antes de traducir el script, complete estas tareas:

- Revisa tu script Lua del plugin Job Submit existente
- Comprenda las diferencias entre el complemento Job Submit y el CLI Filter APIs
- Acceda a la documentación del complemento de filtro CLI de Slurm

Para traducir el script del complemento Job Submit al complemento de filtro CLI

1. Revisa las funciones de script del plugin Job Submit existentes (`slurm_job_submit`,`slurm_job_modify`).
2. Identifique las funciones equivalentes del complemento de filtro CLI:
   - `slurm_job_submit` se convertirá en `slurm_cli_pre_submit`
   - Agregar `slurm_cli_setup_defaults` para la configuración de parámetros predeterminada

- Añadir `slurm_cli_post_submit` para acciones posteriores al envío

3. Translate la lógica de validación de trabajos de `job_desc` los campos al acceso a la `options` matriz:

   - `job_desc.account` se convertirá en `options["account"]`
   - `job_desc.partition` se convertirá en `options["partition"]`
   - `job_desc.features` se convertirá en `options["constraint"]`
4. Actualice el registro de llamadas de `slurm.log_user()` a`slurm.log_error()`.
5. Pruebe su guion traducido en un clúster de desarrollo.
6. Implemente en su clúster de producción siguiendo el proceso de implementación estándar del complemento de filtro CLI.

Resultados esperados

Después de completar la traducción:

- El guion traducido proporciona una validación equivalente al envío de los trabajos
- Los usuarios ven mensajes de error y mensajes de error similares a los del plugin Job Submit original.
- Las políticas de envío de trabajos se mantienen durante la migración a AWS PCS

Resolución de problemas

Errores de traducción de guiones

Síntomas: Los envíos de trabajos fallan debido a errores de ejecución de Lua.

Causa probable: acceso incorrecto a los campos o llamadas a funciones en el guion traducido.

Solución: revise la documentación de la API del complemento CLI Filter y compare las asignaciones de campos entre las interfaces Job Submit y CLI Filter.

Preguntas frecuentes sobre los complementos de filtro CLI de Slurm en PCS AWS

Revise estas preguntas frecuentes sobre los complementos de filtro CLI.

¿Cuál es la diferencia entre el complemento CLI Filter y el complemento Job Submit?

El complemento CLI Filter se ejecuta en el lado del cliente en los nodos de inicio de sesión y procesamiento antes de que el envío del trabajo llegue al controlador, mientras que el complemento Job Submit se ejecuta en el lado del servidor en el controlador después del envío del trabajo. Los usuarios pueden omitir el complemento CLI Filter, pero no bloquea el controlador, mientras que Job Submit es seguro, pero puede afectar al rendimiento del clúster durante la ejecución.

¿ AWS PCS es compatible con el complemento Job Submit de Slurm?

No, el complemento Job Submit no es compatible con AWS PCS. En su lugar, utilice el complemento de filtro CLI para la validación y modificación del envío de trabajos.

¿Puedo usar el complemento de filtro CLI para reforzar la seguridad?

No, determinados usuarios pueden omitir el complemento de filtro CLI y no se debe confiar en él para aplicar medidas de seguridad. Úselo para mejorar la experiencia del usuario, establecer parámetros predeterminados y orientar políticas, en lugar de utilizar políticas críticas para la seguridad.

¿Por qué el script debe estar en todos los nodos de cómputo y no solo en los nodos de inicio de sesión?

Los comandos de Slurm, como estos, se `srun` pueden ejecutar dentro de los scripts de trabajo en los nodos de cómputo, lo que también desencadena la ejecución del complemento de filtro CLI. El script debe estar disponible en cualquier lugar donde se ejecuten los comandos de Slurm.

¿Puedo modificar el script del complemento de filtro CLI en un clúster activo?

Sí, si utiliza el enfoque de implementación de S3 o del sistema de archivos. Las instancias nuevas recibirán el script actualizado, pero las instancias existentes necesitan que el script se actualice manualmente o mediante el método de implementación que elija.

¿Puedo usar diferentes scripts del complemento de filtro CLI en diferentes grupos de nodos de cómputo?

Sí, pero no se recomienda. Puede proporcionar scripts con una lógica diferente a los diferentes grupos de nodos de procesamiento, pero es responsable de administrar las interdependencias y evitar la superposición de la lógica. La mayoría de los clientes proporcionan un conjunto de lógicas en todo el clúster.

¿Puedo usar el complemento de filtro CLI con la implementación de C en lugar de Lua?

No se admite la implementación en C. AWS PCS solo admite la implementación del script Lua. SchedMD recomienda a los clientes usar Lua en lugar de C para facilitar su uso al implementar los complementos de filtro CLI.

¿Puedo activar o desactivar el complemento de filtro CLI en un clúster existente?

Sí, puede habilitar o deshabilitar el complemento de filtro CLI en los clústeres existentes mediante la API de actualización sin volver a crear el clúster.

Solución de problemas con el complemento de filtro CLI de Slurm en PCS AWS

Utilice esta información de solución de problemas para resolver problemas comunes del complemento de filtro CLI.

El envío del trabajo falla inmediatamente y se produce un error al cargar el plugin

Síntomas: Los usuarios reciben mensajes de error sobre la falta o el error del complemento de filtro CLI al enviar trabajos.

Causas posibles:

- Falta el script del complemento de filtro CLI en uno o más nodos
- El nombre del archivo del script es incorrecto (debe ser exacto`cli_filter.lua`)
- El script se implementó en una ruta de directorio incorrecta
- El script tiene permisos de archivo incorrectos

Solución:

- Compruebe que el script existe `/etc/aws/pcs/scheduler/slurm-<version>/cli_filter.lua` en todos los nodos de inicio de sesión y de cómputo
- Compruebe que el nombre del archivo del script sea exacto `cli_filter.lua`
- Asegúrese de que el script tenga permisos de lectura (644 o similar)
- Pruebe la implementación del script en un único nodo de inicio de sesión antes de implementarlo en todo el clúster

La creación del clúster falla debido a un error de validación del complemento de filtro CLI

Síntomas: La creación del clúster falla y se produce un error sobre un `CliFilterPlugins` parámetro no válido.

Causas posibles:

- El formato del valor del parámetro es incorrecto en `slurmCustomSettings`
- Escriba el nombre o el valor del parámetro

Solución:

- Utilice el nombre exacto del parámetro: `CliFilterPlugins`
- Utilice el valor exacto del parámetro: `cli_filter/lua`
- Verifique la sintaxis de JSON en la `slurmCustomSettings` matriz

El script del complemento CLI Filter se ejecuta pero la validación del trabajo no funciona como se esperaba

Síntomas: Los trabajos se envían correctamente, pero la lógica de validación personalizada no se activa o produce resultados inesperados.

Causas posibles:

- Errores de sintaxis del script de Lua
- Patrones de acceso a los campos incorrectos (utilizando la sintaxis del complemento Job Submit en lugar del complemento de filtro CLI)
- Errores lógicos en las condiciones de validación

Solución:

- Revise el script de Lua para ver si hay errores de sintaxis
- Compruebe que el acceso al campo utilice el `options["field_name"]` formato en lugar de `job_desc.field_name`
- Agregue sentencias de registro para depurar el flujo de ejecución del script
- Pruebe primero la lógica del script con casos de validación simples

La implementación del script de S3 falla

Síntomas: Las instancias se inician pero el script del complemento de filtro CLI no se descarga de S3.

Causas posibles:

- El perfil de instancia de IAM carece de permisos de lectura de S3
- El punto final de VPC de S3 no está configurado
- La ruta del objeto o el depósito de S3 en los datos de usuario son incorrectos

Solución:

- Comprueba que el perfil de la instancia de IAM tenga `s3:GetObject` permiso para tu bucket
- Configure el punto final de la puerta de enlace de VPC S3 para el acceso directo
- Compruebe el nombre del bucket de S3 y la ruta del objeto en el script de datos de usuario
- Revise los registros de datos de usuario de la instancia para ver si hay errores de descarga en S3

Servicio de seguridad en la computación AWS paralela

La seguridad en la nube AWS es la máxima prioridad. Como AWS cliente, usted se beneficia de los centros de datos y las arquitecturas de red diseñados para cumplir con los requisitos de las organizaciones más sensibles a la seguridad.

La seguridad es una responsabilidad compartida entre usted AWS y usted. El modelo de responsabilidad compartida la describe como seguridad de la nube y seguridad en la nube:

- Seguridad de la nube: AWS es responsable de proteger la infraestructura que ejecuta AWS los servicios en la Nube de AWS. AWS también le proporciona servicios que puede utilizar de forma segura. Los auditores externos prueban y verifican periódicamente la eficacia de nuestra seguridad como parte de los AWS programas de de . Para obtener más información sobre los programas de conformidad que se aplican al Servicio de Computación AWS Paralela, consulte AWS el apartado AWS Servicios incluidos .
- Seguridad en la nube: su responsabilidad viene determinada por el AWS servicio que utilice. También es responsable de otros factores, incluida la confidencialidad de los datos, los requisitos de la empresa y la legislación y la normativa aplicables.

Esta documentación le ayuda a comprender cómo aplicar el modelo de responsabilidad compartida al utilizar AWS PCS. Los siguientes temas muestran cómo configurar el AWS PCS para cumplir sus objetivos de seguridad y conformidad. También aprenderá a utilizar otros AWS servicios que le ayudan a supervisar y proteger los recursos de su AWS PCS.

Temas

Protección de datos en AWS Parallel Computing Service

El modelo de se aplica a protección de datos en AWS Parallel Computing Service. Como se describe en este modelo, AWS es responsable de proteger la infraestructura global que ejecuta todos los Nube de AWS. Eres responsable de mantener el control sobre el contenido alojado en esta infraestructura. También eres responsable de las tareas de administración y configuración de seguridad para los Servicios de AWS que utiliza. Para obtener más información sobre la privacidad de los datos, consulte las Preguntas frecuentes sobre la privacidad de datos. Para obtener información sobre la protección de datos en Europa, consulte la publicación de blog sobre el Modelo de responsabilidad compartida de AWS y GDPR en el Blog de seguridad de AWS .

Con fines de protección de datos, le recomendamos que proteja Cuenta de AWS las credenciales y configure los usuarios individuales con AWS IAM Identity Center o AWS Identity and Access Management (IAM). De esta manera, solo se otorgan a cada usuario los permisos necesarios para cumplir sus obligaciones laborales. También recomendamos proteger sus datos de la siguiente manera:

- Utiliza la autenticación multifactor (MFA) en cada cuenta.
- Se utiliza SSL/TLS para comunicarse con AWS los recursos. Exigimos TLS 1.2 y recomendamos TLS 1.3.
- Configure la API y el registro de actividad de los usuarios con AWS CloudTrail. Para obtener información sobre el uso de CloudTrail senderos para capturar AWS actividades, consulte Cómo trabajar con CloudTrail senderos en la Guía del AWS CloudTrail usuario.
- Utilice soluciones de AWS cifrado, junto con todos los controles de seguridad predeterminados Servicios de AWS.
- Utiliza servicios de seguridad administrados avanzados, como Amazon Macie, que lo ayuden a detectar y proteger la información confidencial almacenada en Amazon S3.
- Si necesita módulos criptográficos validados por FIPS 140-3 para acceder a AWS través de una interfaz de línea de comandos o una API, utilice un punto final FIPS. Para obtener más información sobre los puntos de conexión de FIPS disponibles, consulte Estándar de procesamiento de la información federal (FIPS) 140-3.

Se recomienda encarecidamente no introducir nunca información confidencial o sensible, como por ejemplo, direcciones de correo electrónico de clientes, en etiquetas o campos de formato libre, tales como el campo Nombre. Esto incluye cuando trabaja con AWS PCS u otros dispositivos Servicios de AWS mediante la consola, la API o. AWS CLI AWS SDKs Cualquier dato que introduzca

en etiquetas o campos de formato libre utilizados para los nombres se pueden emplear para los registros de facturación o diagnóstico. Si proporciona una URL a un servidor externo, recomendamos encarecidamente que no incluya la información de las credenciales en la URL para validar la solicitud para ese servidor.

Cifrado en reposo

El cifrado está habilitado de forma predeterminada para los datos en reposo cuando se crea un clúster de AWS Parallel Computing Service (AWS PCS) con la Consola de administración de AWS API de AWS PCS o AWS SDKs. AWS CLI AWS El PCS utiliza AWS una clave KMS propia para cifrar los datos en reposo. Para obtener más información, consulte las claves y AWS claves del cliente en la Guía para AWS KMS desarrolladores. También puede utilizar una clave gestionada por el cliente. Para obtener más información, consulte Política de claves de KMS requerida para su uso con volúmenes de EBS cifrados en PCS AWS.

El secreto del clúster se almacena AWS Secrets Manager y cifra con la clave KMS gestionada por Secrets Manager. Para obtener más información, consulte Trabajar con secretos de clústeres en AWS PCS.

En un clúster de AWS PCS, los siguientes datos están en reposo:

- Estado del planificador: incluye datos sobre las tareas en ejecución y los nodos aprovisionados en el clúster. Estos son los datos en los que Slurm persiste según lo definido en su. `StateSaveLocation` `slurm.conf` Para obtener más información, consulte la descripción de la documentación StateSaveLocationde Slurm. AWS El PCS elimina los datos del trabajo una vez finalizado un trabajo.
- Secreto de autenticación del programador: AWS PCS lo usa para autenticar todas las comunicaciones del programador en el clúster.

Para obtener información sobre el estado del programador, el AWS PCS cifra automáticamente los datos y los metadatos antes de escribirlos en el sistema de archivos. El sistema de archivos cifrados utiliza el algoritmo de cifrado AES-256 estándar del sector para los datos en reposo.

Cifrado en tránsito

Sus conexiones a la API de AWS PCS utilizan el cifrado TLS con el proceso de firma Signature versión 4, independientemente de si utiliza AWS Command Line Interface ()AWS CLI o. AWS SDKs Para obtener más información, consulte Firmar las solicitudes de la AWS API en la Guía del AWS

Identity and Access Management usuario. AWS gestiona el control de acceso a través de la API con las políticas de IAM para las credenciales de seguridad que se utilizan para conectarse.

AWS PCS usa TLS para conectarse a otros AWS servicios.

Dentro de un clúster de Slurm, el programador se configura con el complemento de autenticación que proporciona la `auth/slurm` autenticación para todas las comunicaciones del programador. Slurm no proporciona cifrado a nivel de aplicación para sus comunicaciones, ya que todos los datos que fluyen entre las instancias del clúster permanecen locales en la VPC de EC2 y, por lo tanto, están sujetos al cifrado de VPC si esas instancias admiten el cifrado en tránsito. Para obtener más información, consulte Cifrado en tránsito en la Guía del usuario de Amazon Elastic Compute Cloud. La comunicación se cifra entre el controlador (aprovisionado en una cuenta de servicio) y los nodos del clúster de su cuenta.

Administración de claves

AWS PCS utiliza una clave AWS KMS propia para cifrar los datos. Para obtener más información, consulte Claves y AWS claves del cliente en la Guía para AWS KMS desarrolladores. También puede utilizar una clave gestionada por el cliente. Para obtener más información, consulte Política de claves de KMS requerida para su uso con volúmenes de EBS cifrados en PCS AWS.

El secreto del clúster se almacena AWS Secrets Manager y cifra con la clave KMS gestionada por Secrets Manager. Para obtener más información, consulte Trabajar con secretos de clústeres en AWS PCS.

Privacidad del tráfico entre redes

AWS Los recursos informáticos de PCS de un clúster residen en 1 VPC de la cuenta del cliente. Por lo tanto, todo el tráfico interno del servicio de AWS PCS dentro de un clúster permanece dentro de la AWS red y no viaja a través de Internet. La comunicación entre el usuario y los nodos AWS PCS puede viajar a través de Internet y recomendamos usar SSH o Systems Manager para conectarse a los nodos. Para obtener más información, consulte ¿Qué es? AWS Systems Manager en la Guía AWS Systems Manager del usuario.

También puede utilizar las siguientes ofertas para conectar su red local a AWS:

- AWS Site-to-Site VPN. Para obtener más información, consulte ¿Qué es AWS Site-to-Site VPN? en la Guía AWS Site-to-Site VPN del usuario.
- Un AWS Direct Connect. Para obtener más información, consulte ¿Qué es AWS Direct Connect? en la Guía AWS Direct Connect del usuario.

Se accede a la API de AWS PCS para realizar tareas administrativas para el servicio. Usted y sus usuarios acceden a los puertos de punto final de Slurm para interactuar directamente con el programador.

Cifrar el tráfico de la API

Para acceder a la API de AWS PCS, los clientes deben ser compatibles con Transport Layer Security (TLS) 1.2 o una versión posterior. Exigimos TLS 1.2 y recomendamos TLS 1.3. Los clientes también deben admitir conjuntos de cifrado con confidencialidad directa total (PFS) tales como Ephemeral Diffie-Hellman (DHE) o Elliptic Curve Diffie-Hellman Ephemeral (ECDHE). La mayoría de los sistemas modernos como Java 7 y posteriores son compatibles con estos modos. Además, las solicitudes deben estar firmadas mediante un ID de clave de acceso y una clave de acceso secreta que esté asociada a una entidad principal de IAM. También puede usar AWS Security Token Service (AWS STS) para generar credenciales de seguridad temporales para firmar las solicitudes.

Cifrado del tráfico de datos

El cifrado de los datos en tránsito se habilita desde las instancias EC2 compatibles que acceden al punto final del programador y entre ComputeNodeGroup instancias desde dentro del. Nube de AWS Para obtener más información, consulte Cifrado en tránsito.

Política de claves de KMS requerida para su uso con volúmenes de EBS cifrados en PCS AWS

AWS PCS utiliza funciones vinculadas a servicios para delegar permisos a otras personas. Servicios de AWS La función vinculada al servicio de AWS PCS está predefinida e incluye los permisos que AWS PCS necesita para llamar a otras Servicios de AWS personas en su nombre. Los permisos predefinidos también incluyen el acceso a sus claves gestionadas por el cliente Claves administradas por AWS , pero no a las que administra el cliente.

En este tema se describe cómo configurar la política de claves necesaria para lanzar instancias cuando se especifica una clave gestionada por el cliente para el cifrado de Amazon EBS.

> **Note**
>
> AWS El PCS no requiere una autorización adicional para usar la predeterminada Clave administrada de AWS a fin de proteger los volúmenes cifrados de su cuenta.

Contenido

Descripción general de

Puede usar lo siguiente AWS KMS keys para el cifrado de Amazon EBS cuando AWS PCS lance instancias:

- Clave administrada de AWS: una clave de cifrado de su cuenta que Amazon EBS crea, posee y administra. Esta es la clave de cifrado predeterminada en las cuentas nuevas. Amazon EBS la utiliza Clave administrada de AWS para el cifrado, a menos que especifique una clave gestionada por el cliente.
- Clave administrada por el cliente: una clave de cifrado personalizada que usted crea, posee y administra. Para obtener más información, consulte Crear una clave KMS en la Guía para AWS Key Management Service desarrolladores.

> **Note**
>
> La clave debe ser simétrica. Amazon EBS no admite claves asimétricas administradas por el cliente.

Las claves administradas por el cliente se configuran cuando se crean instantáneas cifradas o una plantilla de lanzamiento que especifica los volúmenes cifrados, o cuando se decide habilitar el cifrado de forma predeterminada.

Configuración de las políticas de claves

Sus claves de KMS deben tener una política de claves que permita a AWS PCS lanzar instancias con volúmenes de Amazon EBS cifrados con una clave administrada por el cliente.

Utilice los ejemplos de esta página para configurar una política de claves que permita a AWS PCS acceder a su clave gestionada por el cliente. Puede modificar la política de claves de la clave gestionada por el cliente al crear la clave o más adelante.

La política clave debe incluir las siguientes declaraciones:

- Una declaración que permite que la identidad de IAM especificada en el `Principal` elemento utilice directamente la clave gestionada por el cliente. Incluye permisos para realizar las `DescribeKey` operaciones AWS KMS `Encrypt`,`Decrypt`,`ReEncrypt*`,`GenerateDataKey*`, y en la clave.
- Una declaración que permite a la identidad de IAM especificada en el `Principal` elemento utilizar la `CreateGrant` operación para generar concesiones que deleguen un subconjunto de sus propios permisos en uno de los Servicios de AWS que estén integrados con AWS KMS o con otro principal. Esto les permite utilizar la clave para crear recursos cifrados en su nombre.

No modifique ninguna declaración existente en la política cuando añada las nuevas declaraciones de política a su política clave.

Para obtener más información, consulte lo siguiente:

- create-key en la Referencia de comandos AWS CLI
- put-key-policy en la Referencia de comandos de la AWS CLI
- Busque el ID de clave y el ARN clave en la AWS Key Management Service Guía para desarrolladores
- Funciones vinculadas al servicio para PCS AWS
- Cifrado de Amazon EBS en la Guía del usuario de Amazon EBS
- AWS Key Management Service en la Guía para desarrolladores de AWS Key Management Service .

Ejemplo 1: secciones de la política de claves que permiten el acceso a la clave administrada por el cliente

Añada las siguientes declaraciones de política a la política clave de la clave gestionada por el cliente. Sustituya el ARN de ejemplo por el ARN de su función vinculada al servicio. `AWSServiceRoleForPCS` Esta política de ejemplo otorga a la función vinculada al servicio de AWS PCS permisos para usar la clave `AWSServiceRoleForPCS` administrada por el cliente.

```
{
   "Sid": "Allow service-linked role use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/pcs.amazonaws.com/
AWSServiceRoleForPCS"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
   "Sid": "Allow attachment of persistent resources",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::account-id:role/aws-service-role/pcs.amazonaws.com/
AWSServiceRoleForPCS"
       ]
   },
   "Action": [
       "kms:CreateGrant"
   ],
   "Resource": "*",
   "Condition": {
       "Bool": {
           "kms:GrantIsForAWSResource": true
       }
    }
}
```

Ejemplo 2: secciones de la política de claves que permiten el acceso entre cuentas a la clave administrada por el cliente

Si crea una clave gestionada por el cliente en una cuenta diferente a la de su clúster de AWS PCS, debe utilizar una concesión en combinación con la política de claves para permitir el acceso a la clave entre cuentas.

Para conceder el acceso a la clave

1. Añada las siguientes declaraciones de política a la política clave de la clave gestionada por el cliente. Sustituya el ARN de ejemplo por el ARN de la otra cuenta. *111122223333*Sustitúyalo por el ID de cuenta real en el Cuenta de AWS que desea crear el clúster de AWS PCS. Esto le permite otorgar a un usuario o rol de IAM en la cuenta especificada permiso para crear una concesión para la clave mediante el siguiente comando CLI. De forma predeterminada, los usuarios no tienen acceso a la clave.

```
{.
   "Sid": "Allow external account 111122223333 use of the customer managed key",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::111122223333:root"
       ]
   },
   "Action": [
       "kms:Encrypt",
       "kms:Decrypt",
       "kms:ReEncrypt*",
       "kms:GenerateDataKey*",
       "kms:DescribeKey"
   ],
   "Resource": "*"
}
```

```
{
   "Sid": "Allow attachment of persistent resources in external
 account 111122223333",
   "Effect": "Allow",
   "Principal": {
       "AWS": [
           "arn:aws:iam::111122223333:root"
```

```
        ]
    },
    "Action": [
        "kms:CreateGrant"
    ],
    "Resource": "*"
}
```

2. Desde la cuenta en la que desee crear el clúster de AWS PCS, cree una concesión que delegue los permisos pertinentes a la función vinculada al servicio de AWS PCS. El valor de `grantee-principal` es el ARN del rol vinculado al servicio. El valor de `key-id` es el ARN de la clave.

   El siguiente ejemplo de comando CLI create-grant otorga *111122223333* permisos al rol vinculado al servicio mencionado `AWSServiceRoleForPCS` en la cuenta para usar la clave administrada por el cliente en la cuenta. *444455556666*

   ```
   aws kms create-grant \
     --region us-west-2 \
     --key-id arn:aws:kms:us-
   west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d \
     --grantee-principal arn:aws:iam::111122223333:role/aws-service-role/
   pcs.amazonaws.com/AWSServiceRoleForPCS \
     --operations "Encrypt" "Decrypt" "ReEncryptFrom" "ReEncryptTo" "GenerateDataKey"
    "GenerateDataKeyWithoutPlaintext" "DescribeKey" "CreateGrant"
   ```

   > **Note**
   >
   > El usuario que realiza la solicitud debe tener permisos para usar la acción. `kms:CreateGrant`

   El siguiente ejemplo de política de IAM permite que una identidad de IAM (usuario o rol) en la cuenta *111122223333* cree una concesión para la cuenta clave gestionada por el cliente. *444455556666*

   JSON

   ```
   {
     "Version":"2012-10-17",
     "Statement": [
   ```

```
    {
      "Sid": "AllowCreationOfGrantForTheKMSKeyinExternalAccount444455556666",
      "Effect": "Allow",
      "Action": "kms:CreateGrant",
      "Resource": "arn:aws:kms:us-
west-2:444455556666:key/1a2b3c4d-5e6f-1a2b-3c4d-5e6f1a2b3c4d"
    }
  ]
}
```

Para obtener más información acerca de cómo crear una concesión para una clave KMS en una Cuenta de AWS diferente, consulte Concesiones en AWS KMS en laAWS Key Management Service Guía para desarrolladores.

> **Important**
>
> El nombre del rol vinculado al servicio especificado como principal del beneficiario debe ser el nombre de un rol existente. Tras crear la concesión, para asegurarse de que permite a AWS PCS utilizar la clave de KMS especificada, no elimine ni vuelva a crear la función vinculada al servicio.

Edite las políticas clave en la consola AWS KMS

En los ejemplos que aparecen en las secciones anteriores, solo se explica cómo agregar instrucciones a una política de claves, que es una de las múltiples formas de cambiar una de dichas políticas. La forma más sencilla de cambiar una política clave consiste en utilizar la vista predeterminada de la AWS KMS consola para las políticas clave y convertir una identidad de IAM (usuario o rol) en uno de los usuarios clave de la política clave correspondiente. Para obtener más información, consulte Uso de la vista Consola de administración de AWS predeterminada en la Guía para AWS Key Management Service desarrolladores.

> **Warning**
>
> Las declaraciones de política de visualización predeterminadas de la consola incluyen permisos para realizar AWS KMS `Revoke` operaciones en la clave gestionada por el cliente. Si revocas una concesión que daba Cuenta de AWS acceso a una clave gestionada por el

cliente en tu cuenta, los usuarios de esa cuenta Cuenta de AWS pierden el acceso a los datos cifrados y a la clave.

Acceda AWS Parallel Computing Service mediante un punto final de interfaz ()AWS PrivateLink

Puede usarlo AWS PrivateLink para crear una conexión privada entre su VPC y AWS Parallel Computing Service ()AWS PCS. Puede acceder AWS PCS como si estuviera en su VPC, sin el uso de una puerta de enlace a Internet, un dispositivo NAT, una conexión VPN o Direct Connect una conexión. Las instancias de la VPC no necesitan direcciones IP públicas para acceder a AWS PCS.

Esta conexión privada se establece mediante la creación de un punto de conexión de interfaz alimentado por AWS PrivateLink. Creamos una interfaz de red de punto de conexión en cada subred habilitada para el punto de conexión de interfaz. Se trata de interfaces de red administradas por el solicitante que sirven como punto de entrada para el tráfico destinado a AWS PCS.

Para obtener más información, consulte Acceso Servicios de AWS directo AWS PrivateLink en la AWS PrivateLink Guía.

Consideraciones sobre AWS PCS

Antes de configurar un punto de enlace de interfaz para AWS PCS, consulte Acceder a un servicio de AWS mediante un punto de enlace de VPC de interfaz en la AWS PrivateLink Guía.

AWS PCS admite realizar llamadas a todas sus acciones de API a través del punto final de la interfaz.

Si tu VPC no tiene acceso directo a Internet, debes configurar un punto de conexión de VPC para permitir que las instancias del grupo de nodos de cómputo invoquen la acción de la API. AWS PCS `RegisterComputeNodeGroupInstance`

Crea un punto final de interfaz para AWS PCS

Puede crear un punto final de interfaz para AWS PCS usar la consola de Amazon VPC o AWS Command Line Interface ()AWS CLI. Para obtener más información, consulte Creación de un punto de conexión de interfaz en la Guía de AWS PrivateLink .

Cree un punto final de interfaz para AWS PCS usar el siguiente nombre de servicio:

```
com.amazonaws.region.pcs
```

*region*Sustitúyalo por el ID del punto final en el que se va Región de AWS a crear el punto final, por ejemplo`us-east-1`.

Si habilita DNS privado para el punto de conexión de interfaz, puede realizar solicitudes a la API para AWS PCS usando su nombre de DNS predeterminado para la región. Por ejemplo, `pcs.us-east-1.amazonaws.com`.

Creación de una política de puntos de conexión para el punto de conexión de interfaz

Una política de punto de conexión es un recurso de IAM que puede adjuntar al punto de conexión de su interfaz. La política de puntos finales predeterminada permite el acceso total a AWS PCS través del punto final de la interfaz. Para controlar el acceso permitido AWS PCS desde su VPC, adjunte una política de punto final personalizada al punto final de la interfaz.

Una política de punto de conexión especifica la siguiente información:

- Las entidades principales que pueden llevar a cabo acciones (Cuentas de AWS, usuarios de IAM y roles de IAM).
- Las acciones que se pueden realizar.
- El recurso en el que se pueden realizar las acciones.

Para obtener más información, consulte Control del acceso a los servicios con políticas de punto de conexión en la Guía del usuario de AWS PrivateLink .

Ejemplo: política de puntos finales de VPC para acciones AWS PCS

El siguiente es un ejemplo de una política de un punto de conexión personalizado. Al adjuntar esta política al punto final de la interfaz, se concede acceso a las AWS PCS acciones enumeradas a todos los principales del clúster con las especificadas. *cluster-id region*Sustitúyalo por el ID Región de AWS del clúster, por ejemplo`us-east-1`. *account-id*Sustitúyalo por el Cuenta de AWS número del clúster.

```
{
    "Statement": [
            {
```

```
                "Action": [
                "pcs:CreateCluster",
                "pcs:ListClusters",
                "pcs:DeleteCluster",
                "pcs:GetCluster",
                ],
                "Effect": "Allow",
                "Principal": "*",
                "Resource": [
                    "arn:aws:pcs:region:account-id:cluster/cluster-id*"
                ]
            }
        ]
}
```

Servicio de Gestión de Identidad y Acceso para Computación AWS Paralela

AWS Identity and Access Management (IAM) es una herramienta Servicio de AWS que ayuda al administrador a controlar de forma segura el acceso a los AWS recursos. Los administradores de IAM controlan quién puede autenticarse (iniciar sesión) y quién puede autorizarse (tener permisos) para usar AWS los recursos del PCS. El IAM es un Servicio de AWS servicio que puede utilizar sin coste adicional.

Temas

- Público
- Autenticación con identidades
- Administración del acceso con políticas
- Cómo funciona AWS Parallel Computing Service con IAM
- Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS
- AWS políticas administradas para AWS Parallel Computing Service
- Funciones vinculadas al servicio para PCS AWS
- Función Amazon EC2 Spot para PCS AWS
- Permisos mínimos para PCS AWS
- Perfiles de instancia de IAM para AWS Parallel Computing Service
- Solución de problemas de identidad y acceso a AWS Parallel Computing Service

Público

La forma de usar AWS Identity and Access Management (IAM) varía según la función que desempeñes:

- Usuario del servicio: solicite permisos al administrador si no puede acceder a las características (consulte Solución de problemas de identidad y acceso a AWS Parallel Computing Service).
- Administrador del servicio: determine el acceso de los usuarios y envíe las solicitudes de permiso (consulte Cómo funciona AWS Parallel Computing Service con IAM).
- Administrador de IAM: escribe las políticas para administrar el acceso (consulte Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS).

Autenticación con identidades

La autenticación es la forma en que inicias sesión AWS con tus credenciales de identidad. Debe autenticarse como usuario de Usuario raíz de la cuenta de AWS IAM o asumir una función de IAM.

Puede iniciar sesión como una identidad federada con las credenciales de una fuente de identidad, como AWS IAM Identity Center (IAM Identity Center), la autenticación de inicio de sesión único o las credenciales. Google/Facebook Para obtener más información sobre el inicio de sesión, consulte Cómo iniciar sesión en la Cuenta de AWS en la Guía del usuario de AWS Sign-In .

Para el acceso programático, AWS proporciona un SDK y una CLI para firmar criptográficamente las solicitudes. Para obtener más información, consulte AWS Signature Version 4 para solicitudes de API en la Guía del usuario de IAM.

Cuenta de AWS usuario root

Al crear un Cuenta de AWS, se comienza con una identidad de inicio de sesión denominada usuario Cuenta de AWS raíz que tiene acceso completo a todos Servicios de AWS los recursos. Se recomiendaencarecidamente que no utilice el usuario raíz para las tareas diarias. Para ver las tareas que requieren credenciales de usuario raíz, consulte Tareas que requieren credenciales de usuario raíz en la Guía del usuario de IAM.

Identidad federada

Como práctica recomendada, exija a los usuarios humanos que utilicen la federación con un proveedor de identidades para acceder Servicios de AWS mediante credenciales temporales.

Una identidad federada es un usuario del directorio empresarial, del proveedor de identidades web o al Directory Service que se accede Servicios de AWS mediante credenciales de una fuente de identidad. Las identidades federadas asumen roles que proporcionan credenciales temporales.

Para una administración de acceso centralizada, se recomienda AWS IAM Identity Center. Para obtener más información, consulte ¿Qué es el Centro de identidades de IAM? en la Guía del usuario de AWS IAM Identity Center .

Usuarios y grupos de IAM

Un usuario de IAM es una identidad con permisos específicos para una sola persona o aplicación. Recomendamos el uso de credenciales temporales en lugar de usuarios de IAM con credenciales de larga duración. Para obtener más información, consulte Exigir a los usuarios humanos que utilicen la federación con un proveedor de identidad para acceder AWS mediante credenciales temporales en la Guía del usuario de IAM.

Un grupo de IAM especifica un conjunto de usuarios de IAM y facilita la administración de los permisos para grupos grandes de usuarios. Para obtener más información, consulte Casos de uso para usuarios de IAM en la Guía del usuario de IAM.

Roles de IAM

Un Rol de IAM es una identidad con permisos específicos que proporciona credenciales temporales. Puede asumir un rol cambiando de un rol de usuario a uno de IAM (consola) o llamando a una AWS CLI operación de AWS API. Para obtener más información, consulte Métodos para asumir un rol en la Guía del usuario de IAM.

Los roles de IAM son útiles para el acceso de usuario federado, los permisos de usuario de IAM temporales, el acceso entre cuentas, el acceso entre servicios y las aplicaciones que se ejecutan en Amazon EC2. Para obtener más información, consulte Acceso a recursos entre cuentas en IAM en la Guía del usuario de IAM.

Administración del acceso con políticas

AWS Para controlar el acceso, puede crear políticas y adjuntarlas a AWS identidades o recursos. Una política define los permisos cuando están asociados a una identidad o un recurso. AWS evalúa estas políticas cuando un director hace una solicitud. La mayoría de las políticas se almacenan AWS como documentos JSON. Para obtener más información sobre los documentos de políticas de JSON, consulte Información general de políticas de JSON en la Guía del usuario de IAM.

Mediante las políticas, los administradores especifican quién tiene acceso a qué, definiendo qué entidad principal puede realizar acciones sobre qué recursos y en qué condiciones.

De forma predeterminada, los usuarios y los roles no tienen permisos. Un administrador de IAM crea políticas de IAM y las agrega a roles, que los usuarios pueden asumir posteriormente. Las políticas de IAM definen permisos independientemente del método que se utilice para realizar la operación.

Políticas basadas en identidades

Las políticas basadas en identidad son documentos de política de permisos JSON que asocia a una identidad (usuario, grupo o rol). Estas políticas controlan qué acciones pueden realizar las identidades, en qué recursos y en qué condiciones. Para obtener más información sobre cómo crear una política basada en la identidad, consulte Definición de permisos de IAM personalizados con políticas administradas por el cliente en la Guía del usuario de IAM.

Las políticas basadas en identidad pueden ser políticas insertadas (incrustadas directamente en una sola identidad) o políticas administradas (políticas independientes asociadas a varias identidades). Para obtener información sobre cómo elegir entre políticas administradas e insertadas, consulte Selección entre políticas administradas y políticas insertadas en la Guía del usuario de IAM.

Políticas basadas en recursos

Las políticas basadas en recursos son documentos de políticas JSON que se asocian a un recurso. Los ejemplos incluyen las Políticas de confianza de roles de IAM y las Políticas de bucket de Amazon S3. En los servicios que admiten políticas basadas en recursos, los administradores de servicios pueden utilizarlos para controlar el acceso a un recurso específico. Debe especificar una entidad principal en una política basada en recursos.

Las políticas basadas en recursos son políticas insertadas que se encuentran en ese servicio. No puedes usar políticas AWS gestionadas de IAM en una política basada en recursos.

Otros tipos de políticas

AWS admite tipos de políticas adicionales que pueden establecer los permisos máximos que conceden los tipos de políticas más comunes:

- Límites de permisos: establecen los permisos máximos que una política basada en identidad puede conceder a una entidad de IAM. Para obtener más información, consulte Límites de permisos para las entidades de IAM en la Guía del usuario de IAM.

- Políticas de control de servicios (SCPs): especifican los permisos máximos para una organización o unidad organizativa en AWS Organizations. Para obtener más información, consulte Políticas de control de servicios en la Guía del usuario de AWS Organizations .
- Políticas de control de recursos (RCPs): establece los permisos máximos disponibles para los recursos de tus cuentas. Para obtener más información, consulte Políticas de control de recursos (RCPs) en la Guía del AWS Organizations usuario.
- Políticas de sesión: políticas avanzadas que se pasan como parámetro cuando se crea una sesión temporal para un rol o un usuario federado. Para obtener más información, consulte Políticas de sesión en la Guía del usuario de IAM.

Varios tipos de políticas

Cuando se aplican varios tipos de políticas a una solicitud, los permisos resultantes son más complicados de entender. Para saber cómo se AWS determina si se debe permitir una solicitud cuando se trata de varios tipos de políticas, consulte la lógica de evaluación de políticas en la Guía del usuario de IAM.

Cómo funciona AWS Parallel Computing Service con IAM

Antes de utilizar IAM para gestionar el acceso al AWS PCS, conozca las funciones de IAM disponibles para su uso con el PCS. AWS

Funciones de IAM que puede utilizar con AWS Parallel Computing Service

Característica de IAM	AWS Soporte para PCS
Políticas basadas en identidades	Sí
Políticas basadas en recursos	No
Acciones de políticas	Sí
Recursos de políticas	Sí
Claves de condición de política (específicas del servicio)	Sí
ACLs	No

Característica de IAM	AWS Soporte para PCS
ABAC (etiquetas en políticas)	Sí
Credenciales temporales	Sí
Permisos de entidades principales	Sí
Roles de servicio	No
Roles vinculados al servicio	Sí

Para obtener una visión general de cómo funcionan los AWS PCS y otros AWS servicios con la mayoría de las funciones de IAM, consulte AWS los servicios que funcionan con IAM en la Guía del usuario de IAM.

Políticas de PCS basadas en la identidad AWS

Compatibilidad con las políticas basadas en identidad: sí

Las políticas basadas en identidad son documentos de políticas de permisos JSON que puede asociar a una identidad, como un usuario de IAM, un grupo de usuarios o un rol. Estas políticas controlan qué acciones pueden realizar los usuarios y los roles, en qué recursos y en qué condiciones. Para obtener más información sobre cómo crear una política basada en la identidad, consulte Definición de permisos de IAM personalizados con políticas administradas por el cliente en la Guía del usuario de IAM.

Con las políticas basadas en identidades de IAM, puede especificar las acciones y los recursos permitidos o denegados, así como las condiciones en las que se permiten o deniegan las acciones. Para obtener más información sobre los elementos que puede utilizar en una política de JSON, consulte Referencia de los elementos de la política de JSON de IAM en la Guía del usuario de IAM.

Ejemplos de políticas basadas en la identidad para PCS AWS

Para ver ejemplos de políticas de AWS PCS basadas en la identidad, consulte. Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS

Políticas basadas en recursos dentro de PCS AWS

Admite políticas basadas en recursos: no

Las políticas basadas en recursos son documentos de política JSON que se asocian a un recurso. Los ejemplos de políticas basadas en recursos son las políticas de confianza de roles de IAM y las políticas de bucket de Amazon S3. En los servicios que admiten políticas basadas en recursos, los administradores de servicios pueden utilizarlos para controlar el acceso a un recurso específico. Para el recurso al que se asocia la política, la política define qué acciones puede realizar una entidad principal especificada en ese recurso y en qué condiciones. Debe especificar una entidad principal en una política basada en recursos. Los principales pueden incluir cuentas, usuarios, roles, usuarios federados o. Servicios de AWS

Para habilitar el acceso entre cuentas, puede especificar toda una cuenta o entidades de IAM de otra cuenta como la entidad principal de una política en función de recursos. Para obtener más información, consulte Acceso a recursos entre cuentas en IAM en la Guía del usuario de IAM.

Acciones políticas para PCS AWS

Compatibilidad con las acciones de políticas: sí

Los administradores pueden usar las políticas de AWS JSON para especificar quién tiene acceso a qué. Es decir, qué entidad principal puede realizar acciones en qué recursos y en qué condiciones.

El elemento `Action` de una política JSON describe las acciones que puede utilizar para conceder o denegar el acceso en una política. Incluya acciones en una política para conceder permisos y así llevar a cabo la operación asociada.

Para ver una lista de las acciones de AWS PCS, consulte las acciones definidas por AWS Parallel Computing Service en la referencia de autorización del servicio.

Las acciones políticas del AWS PCS utilizan el siguiente prefijo antes de la acción:

```
pcs
```

Para especificar varias acciones en una única instrucción, sepárelas con comas.

```
"Action": [
      "pcs:action1",
      "pcs:action2"
         ]
```

Recursos de políticas para PCS AWS

Compatibilidad con los recursos de políticas: sí

Los administradores pueden usar las políticas de AWS JSON para especificar quién tiene acceso a qué. Es decir, qué entidad principal puede realizar acciones en qué recursos y en qué condiciones.

El elemento `Resource` de la política JSON especifica el objeto u objetos a los que se aplica la acción. Como práctica recomendada, especifique un recurso utilizando el Nombre de recurso de Amazon (ARN). En el caso de las acciones que no admiten permisos por recurso, utilice un carácter comodín (*) para indicar que la instrucción se aplica a todos los recursos.

```
"Resource": "*"
```

Para ver una lista de los tipos de recursos de AWS PCS y sus tipos ARNs, consulte los recursos definidos por AWS Parallel Computing Service en la referencia de autorización de servicios. Para saber con qué acciones puede especificar el ARN de cada recurso, consulte Acciones definidas por AWS Parallel Computing Service.

Para ver ejemplos de políticas de AWS PCS basadas en la identidad, consulte. Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS

Claves de condición de la política para PCS AWS

Compatibilidad con claves de condición de políticas específicas del servicio: sí

Los administradores pueden usar las políticas de AWS JSON para especificar quién tiene acceso a qué. Es decir, qué entidad principal puede realizar acciones en qué recursos y en qué condiciones.

El elemento `Condition` especifica cuándo se ejecutan las instrucciones en función de criterios definidos. Puede crear expresiones condicionales que utilizan operadores de condición, tales como igual o menor que, para que la condición de la política coincida con los valores de la solicitud. Para ver todas las claves de condición AWS globales, consulte las claves de contexto de condición AWS globales en la Guía del usuario de IAM.

Para ver una lista de las claves de condición del AWS PCS, consulte las claves de condición del servicio de computación AWS paralela en la Referencia de autorización del servicio. Para saber con qué acciones y recursos puede utilizar una clave de condición, consulte Acciones definidas por AWS Parallel Computing Service.

Para ver ejemplos de políticas de AWS PCS basadas en la identidad, consulte. Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS

ACLs en PCS AWS

Soportes ACLs: No

Las listas de control de acceso (ACLs) controlan qué directores (miembros de la cuenta, usuarios o roles) tienen permisos para acceder a un recurso. ACLs son similares a las políticas basadas en recursos, aunque no utilizan el formato de documento de políticas JSON.

ABAC con PCS AWS

Admite ABAC (etiquetas en las políticas): sí

El control de acceso basado en atributos (ABAC) es una estrategia de autorización que define permisos en función de atributos denominados etiquetas. Puede adjuntar etiquetas a las entidades y AWS los recursos de IAM y, a continuación, diseñar políticas de ABAC para permitir las operaciones cuando la etiqueta del principal coincida con la etiqueta del recurso.

Para controlar el acceso en función de etiquetas, debe proporcionar información de las etiquetas en el elemento de condición de una política utilizando las claves de condición `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` o `aws:TagKeys`.

Si un servicio admite las tres claves de condición para cada tipo de recurso, el valor es Sí para el servicio. Si un servicio admite las tres claves de condición solo para algunos tipos de recursos, el valor es Parcial.

Para obtener más información sobre ABAC, consulte Definición de permisos con la autorización de ABAC en la Guía del usuario de IAM. Para ver un tutorial con los pasos para configurar ABAC, consulte Uso del control de acceso basado en atributos (ABAC) en la Guía del usuario de IAM.

Uso de credenciales temporales con PCS AWS

Compatibilidad con credenciales temporales: sí

Las credenciales temporales proporcionan acceso a AWS los recursos a corto plazo y se crean automáticamente cuando se utiliza la federación o se cambia de rol. AWS recomienda generar credenciales temporales de forma dinámica en lugar de utilizar claves de acceso a largo plazo. Para obtener más información, consulte Credenciales de seguridad temporales en IAM y Servicios de AWS que funcionan con IAM en la Guía del usuario de IAM.

Permisos principales entre servicios para PCS AWS

Admite sesiones de acceso directo (FAS): sí

Las sesiones de acceso directo (FAS) utilizan los permisos del operador principal que realiza la llamada Servicio de AWS, junto con los que solicitan, Servicio de AWS para realizar solicitudes a los servicios descendentes. Para obtener información sobre las políticas a la hora de realizar solicitudes de FAS, consulte Sesiones de acceso directo.

Funciones de servicio para PCS AWS

Compatible con roles de servicio: No

Un rol de servicio es un rol de IAM que asume un servicio para realizar acciones en su nombre. Un administrador de IAM puede crear, modificar y eliminar un rol de servicio desde IAM. Para obtener más información, consulte Crear un rol para delegar permisos a un Servicio de AWS en la Guía del usuario de IAM.

> **Warning**
>
> Cambiar los permisos de un rol de servicio podría interrumpir la funcionalidad del AWS PCS. Edite las funciones de servicio solo cuando el AWS PCS le dé instrucciones para hacerlo.

Funciones vinculadas al servicio para PCS AWS

Compatible con roles vinculados al servicio: sí

Un rol vinculado a un servicio es un tipo de rol de servicio que está vinculado a un. Servicio de AWS El servicio puede asumir el rol para realizar una acción en su nombre. Los roles vinculados al servicio aparecen en usted Cuenta de AWS y son propiedad del servicio. Un administrador de IAM puede ver, pero no editar, los permisos de los roles vinculados a servicios.

Para obtener más información sobre la creación o administración de funciones vinculadas al servicio de AWS PCS, consulte. Funciones vinculadas al servicio para PCS AWS

Ejemplos de políticas basadas en la identidad para Parallel Computing Service AWS

De forma predeterminada, los usuarios y los roles no tienen permiso para crear o modificar los recursos de AWS PCS. Un administrador de IAM puede crear políticas de IAM para conceder permisos a los usuarios para realizar acciones en los recursos que necesitan.

Para obtener información acerca de cómo crear una política basada en identidades de IAM mediante el uso de estos documentos de políticas JSON de ejemplo, consulte Creación de políticas de IAM (consola) en la Guía del usuario de IAM.

Para obtener más información sobre las acciones y los tipos de recursos definidos por el AWS PCS, incluido el ARNs formato de cada uno de los tipos de recursos, consulte las claves de condición, recursos y acciones del servicio de computación AWS paralela en la Referencia de autorización de servicios.

Temas

- Prácticas recomendadas sobre las políticas
- Uso de la consola PCS AWS
- Cómo permitir a los usuarios consultar sus propios permisos

Prácticas recomendadas sobre las políticas

Las políticas basadas en la identidad determinan si alguien puede crear, acceder o eliminar los recursos de AWS PCS de su cuenta. Estas acciones pueden generar costos adicionales para su Cuenta de AWS. Siga estas directrices y recomendaciones al crear o editar políticas basadas en identidades:

- Comience con las políticas AWS administradas y avance hacia los permisos con privilegios mínimos: para empezar a conceder permisos a sus usuarios y cargas de trabajo, utilice las políticas AWS administradas que otorgan permisos para muchos casos de uso comunes. Están disponibles en su. Cuenta de AWS Le recomendamos que reduzca aún más los permisos definiendo políticas administradas por el AWS cliente que sean específicas para sus casos de uso. Con el fin de obtener más información, consulte las políticas administradas por AWS o las políticas administradas por AWS para funciones de tarea en la Guía de usuario de IAM.
- Aplique permisos de privilegio mínimo: cuando establezca permisos con políticas de IAM, conceda solo los permisos necesarios para realizar una tarea. Para ello, debe definir las acciones que se pueden llevar a cabo en determinados recursos en condiciones específicas, también conocidos como permisos de privilegios mínimos. Con el fin de obtener más información sobre el uso de IAM para aplicar permisos, consulte Políticas y permisos en IAM en la Guía del usuario de IAM.
- Utilice condiciones en las políticas de IAM para restringir aún más el acceso: puede agregar una condición a sus políticas para limitar el acceso a las acciones y los recursos. Por ejemplo, puede escribir una condición de políticas para especificar que todas las solicitudes deben

enviarse utilizando SSL. También puedes usar condiciones para conceder el acceso a las acciones del servicio si se utilizan a través de una acción específica Servicio de AWS, por ejemplo CloudFormation. Para obtener más información, consulte Elementos de la política de JSON de IAM: Condición en la Guía del usuario de IAM.

- Utiliza el analizador de acceso de IAM para validar las políticas de IAM con el fin de garantizar la seguridad y funcionalidad de los permisos: el analizador de acceso de IAM valida políticas nuevas y existentes para que respeten el lenguaje (JSON) de las políticas de IAM y las prácticas recomendadas de IAM. El analizador de acceso de IAM proporciona más de 100 verificaciones de políticas y recomendaciones procesables para ayudar a crear políticas seguras y funcionales. Para más información, consulte Validación de políticas con el Analizador de acceso de IAM en la Guía del usuario de IAM.
- Requerir autenticación multifactor (MFA): si tiene un escenario que requiere usuarios de IAM o un usuario raíz en Cuenta de AWS su cuenta, active la MFA para mayor seguridad. Para exigir la MFA cuando se invoquen las operaciones de la API, añada condiciones de MFA a sus políticas. Para más información, consulte Acceso seguro a la API con MFA en la Guía del usuario de IAM.

Para obtener más información sobre las prácticas recomendadas de IAM, consulte Prácticas recomendadas de seguridad en IAM en la Guía del usuario de IAM.

Uso de la consola PCS AWS

Para acceder a la consola del Servicio de Computación AWS Paralela, debe tener un conjunto mínimo de permisos. Estos permisos deben permitirle enumerar y ver detalles sobre los recursos de AWS PCS de su propiedad Cuenta de AWS. Si crea una política basada en identidades que sea más restrictiva que el mínimo de permisos necesarios, la consola no funcionará del modo esperado para las entidades (usuarios o roles) que tengan esa política.

No es necesario que concedas permisos mínimos de consola a los usuarios que solo realicen llamadas a la API AWS CLI o a la AWS API. En su lugar, permita el acceso únicamente a las acciones que coincidan con la operación de API que intentan realizar.

Para obtener más información sobre los permisos mínimos necesarios para usar la consola AWS PCS, consultePermisos mínimos para PCS AWS.

Cómo permitir a los usuarios consultar sus propios permisos

En este ejemplo, se muestra cómo podría crear una política que permita a los usuarios de IAM ver las políticas administradas e insertadas que se asocian a la identidad de sus usuarios. Esta política

incluye permisos para completar esta acción en la consola o mediante programación mediante la API AWS CLI o AWS .

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ViewOwnUserInfo",
            "Effect": "Allow",
            "Action": [
                "iam:GetUserPolicy",
                "iam:ListGroupsForUser",
                "iam:ListAttachedUserPolicies",
                "iam:ListUserPolicies",
                "iam:GetUser"
            ],
            "Resource": ["arn:aws:iam::*:user/${aws:username}"]
        },
        {
            "Sid": "NavigateInConsole",
            "Effect": "Allow",
            "Action": [
                "iam:GetGroupPolicy",
                "iam:GetPolicyVersion",
                "iam:GetPolicy",
                "iam:ListAttachedGroupPolicies",
                "iam:ListGroupPolicies",
                "iam:ListPolicyVersions",
                "iam:ListPolicies",
                "iam:ListUsers"
            ],
            "Resource": "*"
        }
    ]
}
```

AWS políticas administradas para AWS Parallel Computing Service

Una política AWS administrada es una política independiente creada y administrada por AWS. AWS Las políticas administradas están diseñadas para proporcionar permisos para muchos casos de uso comunes, de modo que pueda empezar a asignar permisos a usuarios, grupos y funciones.

Ten en cuenta que es posible que las políticas AWS administradas no otorguen permisos con privilegios mínimos para tus casos de uso específicos, ya que están disponibles para que los usen todos los AWS clientes. Se recomienda definir políticas administradas por el cliente específicas para sus casos de uso a fin de reducir aún más los permisos.

No puedes cambiar los permisos definidos en AWS las políticas administradas. Si AWS actualiza los permisos definidos en una política AWS administrada, la actualización afecta a todas las identidades principales (usuarios, grupos y roles) a las que está asociada la política. AWS es más probable que actualice una política AWS administrada cuando Servicio de AWS se lance una nueva o cuando estén disponibles nuevas operaciones de API para los servicios existentes.

Para obtener más información, consulte Políticas administradas por AWS en la Guía del usuario de IAM.

AWS política gestionada: AWSPCSCompute NodePolicy

Puede adjuntarla AWSPCSCompute NodePolicy a sus entidades de IAM. Puede adjuntar esta política a una función de IAM de nodo de cómputo de AWS PCS que especifique para permitir que los nodos que utilizan esa función se conecten a un clúster de AWS PCS.

AWS PCS asocia esta política a un rol de grupo de nodos de cómputo cuando se usa la consola para crear un grupo de nodos de cómputo.

Detalles de los permisos

Esta política incluye los siguientes permisos.

- `pcs:RegisterComputeNodeGroupInstance`— Permita que un nodo de cómputo de AWS PCS (instancia EC2) se registre en un clúster de AWS PCS.

Para ver los permisos de esta política, consulte AWSPCSComputeNodePolicy en la Referencia de la política administrada de AWS .

AWS política gestionada: AWSPCSService RolePolicy

No puede adjuntarse AWSPCSService RolePolicy a sus entidades de IAM. Esta política está asociada a una función vinculada al servicio que permite a AWS PCS realizar acciones en su nombre. Para obtener más información, consulte Funciones vinculadas al servicio para PCS AWS.

Detalles de los permisos

Esta política incluye los siguientes permisos.

- `ec2`— Permite a AWS PCS crear y gestionar los recursos de Amazon EC2.
- `iam`— Permite a AWS PCS crear una función vinculada a un servicio para la flota de Amazon EC2 y transferirla a Amazon EC2.
- `cloudwatch`— Permite a AWS PCS publicar métricas de servicio en Amazon CloudWatch.
- `secretsmanager`— Permite a AWS PCS gestionar los secretos de los recursos del clúster de AWS PCS.

Para ver los permisos de esta política, consulte AWSPCSServiceRolePolicy en la Referencia de la política administrada de AWS .

AWS PCS actualiza las políticas AWS gestionadas

Consulte los detalles sobre las actualizaciones de las políticas AWS administradas para AWS PCS desde que este servicio comenzó a rastrear estos cambios. Para recibir alertas automáticas sobre los cambios en esta página, suscríbase a la fuente RSS de la página del historial de documentos de AWS PCS.

Cambio	Descripción	Fecha
AWSPCSServiceRolePolicy: actualización de una política actual	AWS PCS agregó nuevos permisos para admitir los bloques de capacidad y lograr una capacidad de cómputo predecible. Se agregó un `ec2:DescribeCapacityReserva`	11 de septiembre de 2025

Cambio	Descripción	Fecha
	`tions` permiso para que el AWS PCS pueda detectar y usar las reservas de bloques de capacidad para grupos de nodos de cómputo.	
AWSPCSComputeNodePolicy: política nueva	AWS PCS agregó una nueva política para conceder permiso a los nodos de cómputo de AWS PCS para conectarse a los clústeres de AWS PCS. AWS PCS asocia esta política a una función de IAM al crear un grupo de nodos de procesamiento en la consola de AWS PCS.	23 de junio de 2025
Se actualizó el JSON en este documento	Se corrigió el JSON de este documento para incluirlo `"arn:aws:ec2:*:*:spot-instances-request/*"` .	5 de septiembre de 2024
AWS PCS comenzó a rastrear los cambios	AWS PCS comenzó a realizar un seguimiento de los cambios en sus políticas AWS gestionadas.	28 de agosto de 2024

Funciones vinculadas al servicio para PCS AWS

AWS Parallel Computing Service utiliza funciones vinculadas a servicios AWS Identity and Access Management (IAM). Un rol vinculado a un servicio es un tipo único de rol de IAM que está vinculado directamente a PCS. AWS Los roles vinculados al servicio están predefinidos por AWS PCS e incluyen todos los permisos que el servicio requiere para llamar a otros AWS servicios en su nombre.

Un rol vinculado a un servicio facilita la configuración del AWS PCS, ya que no es necesario añadir manualmente los permisos necesarios. AWS PCS define los permisos de sus funciones vinculadas al servicio y, a menos que se defina lo contrario, solo AWS PCS puede asumir sus funciones. Los permisos definidos incluyen las políticas de confianza y de permisos, y que la política de permisos no se pueda asociar a ninguna otra entidad de IAM.

Solo puede eliminar un rol vinculado a servicios después de eliminar sus recursos relacionados. Esto protege sus recursos de AWS PCS porque no puede eliminar accidentalmente el permiso de acceso a los recursos.

Para obtener información sobre otros servicios que admiten funciones vinculadas a servicios, consulte AWS los servicios que funcionan con IAM y busque los servicios con la palabra Sí en la columna Funciones vinculadas a servicios. Elija una opción Sí con un enlace para ver la documentación acerca del rol vinculado al servicio en cuestión.

Permisos de roles vinculados al servicio para PCS AWS

AWS PCS usa el rol vinculado al servicio denominado AWSServiceRoleForPCS: otorga permiso a AWS PCS para administrar los recursos de Amazon EC2.

El rol vinculado al servicio de AWSService RoleFor PCS confía en los siguientes servicios para asumir el rol:

- `pcs.amazonaws.com`

La política de permisos de roles denominada AWSPCSServiceRolePolicypermite a AWS PCS completar acciones en recursos específicos.

Debe configurar los permisos para permitir a sus usuarios, grupos o funciones, crear, editar o eliminar la descripción de un rol vinculado al servicio. Para obtener más información, consulte Permisos de roles vinculados a servicios en la Guía del usuario de IAM.

Crear un rol vinculado a un servicio para PCS AWS

No es necesario crear manualmente un rol vinculado a un servicio. AWS PCS le crea un rol vinculado a un servicio cuando crea un clúster.

Edición de un rol vinculado a un servicio para PCS AWS

AWS PCS no le permite editar el rol vinculado al servicio de AWSService RoleFor PCS. Después de crear un rol vinculado al servicio, no podrá cambiar el nombre del rol, ya que varias entidades

podrían hacer referencia al rol. Sin embargo, sí puede editar la descripción del rol con IAM. Para obtener más información, consulte Editar un rol vinculado a servicios en la Guía del usuario de IAM.

Eliminar un rol vinculado a un servicio para PCS AWS

Si ya no necesita usar una característica o servicio que requieran un rol vinculado a un servicio, le recomendamos que elimine dicho rol. Así no tendrá una entidad no utilizada que no se supervise ni mantenga de forma activa. Sin embargo, debe limpiar los recursos de su rol vinculado al servicio antes de eliminarlo manualmente.

> **Note**
>
> Si el servicio AWS PCS utiliza la función al intentar eliminar los recursos, es posible que la eliminación no se realice correctamente. En tal caso, espere unos minutos e intente de nuevo la operación.

Para eliminar los recursos del AWS PCS utilizados por el AWSService RoleFor PCS

Debe eliminar todos los clústeres para eliminar la función vinculada al servicio de AWSService RoleFor PCS. Para obtener más información, consulte Eliminar un clúster.

Para eliminar manualmente el rol vinculado a servicios mediante IAM

Utilice la consola de IAM AWS CLI, la o la AWS API para eliminar la función vinculada al servicio de AWSService RoleFor PCS. Para obtener más información, consulte Eliminación de un rol vinculado a servicios en la Guía del usuario de IAM.

Regiones compatibles para AWS las funciones vinculadas al servicio de PCS

AWS PCS admite el uso de funciones vinculadas al servicio en todas las regiones en las que el servicio está disponible. Para obtener más información, consulte Puntos de conexión y Regiones de AWS.

Función Amazon EC2 Spot para PCS AWS

Si desea crear un grupo de nodos de cómputo de AWS PCS que utilice Spot como opción de compra, también debe tener en su seno la función vinculada al servicio de AWSServiceRoleForEC2Spot. Cuenta de AWS Puede usar el siguiente AWS CLI comando para crear el rol. Para obtener más información, consulte Crear un rol vinculado a un servicio y Crear un

rol para delegar permisos a un AWS servicio en la Guía del AWS Identity and Access Management usuario.

```
aws iam create-service-linked-role --aws-service-name spot.amazonaws.com
```

> **Note**
>
> Aparece el siguiente error si Cuenta de AWS ya tiene un rol de `AWSServiceRoleForEC2Spot` IAM.
>
> ```
> An error occurred (InvalidInput) when calling the CreateServiceLinkedRole
> operation: Service role name AWSServiceRoleForEC2Spot has been taken in this
> account, please try a different suffix.
> ```

Permisos mínimos para PCS AWS

En esta sección se describen los permisos de IAM mínimos necesarios para que una identidad de IAM (usuario, grupo o rol) utilice el servicio.

Contenido

Permisos mínimos para usar las acciones de la API

Acción de la API	Permisos mínimos	Permisos adicionales para la consola
CreateCluster	`ec2:CreateNetworkInterface,`	

Acción de la API	Permisos mínimos	Permisos adicionales para la consola
	`ec2:DescribeVpcs, ec2:DescribeSubnets, ec2:DescribeSecurityGroups, ec2:GetSecurityGroupsForVpc, iam:CreateServiceLinkedRole, secretsmanager:CreateSecret, secretsmanager:TagResource, secretsmanager:RotateSecret, pcs:CreateCluster`	
ListClusters	`pcs:ListClusters`	
GetCluster	`pcs:GetCluster`	`ec2:DescribeSubnets`
DeleteCluster	`pcs:DeleteCluster`	

Acción de la API	Permisos mínimos	Permisos adicionales para la consola
CreateComputeNodeGroup	`ec2:DescribeVpcs, ec2:DescribeSubnets, ec2:DescribeSecurityGroups, ec2:DescribeLaunchTemplates, ec2:DescribeLaunchTemplateVersions, ec2:DescribeInstanceTypes, ec2:DescribeInstanceTypeOfferings, ec2:RunInstances, ec2:CreateFleet, ec2:CreateTags, iam:PassRole, iam:GetInstanceProfile, pcs:CreateComputeNodeGroup`	`iam:ListInstanceProfiles, ec2:DescribeImages, pcs:GetCluster`
ListComputerNodeGroups	`pcs:ListComputeNodeGroups`	`pcs:GetCluster`
GetComputeNodeGroup	`pcs:GetComputeNodeGroup`	`ec2:DescribeSubnets`

Acción de la API	Permisos mínimos	Permisos adicionales para la consola
UpdateComputeNodeGroup	`ec2:DescribeVpcs, ec2:DescribeSubnets, ec2:DescribeSecurityGroups, ec2:DescribeLaunchTemplates, ec2:DescribeLaunchTemplateVersions, ec2:DescribeInstanceTypes, ec2:DescribeInstanceTypeOfferings, ec2:RunInstances, ec2:CreateFleet, ec2:CreateTags, iam:PassRole, iam:GetInstanceProfile, pcs:UpdateComputeNodeGroup`	`pcs:GetComputeNodeGroup, iam:ListInstanceProfiles, ec2:DescribeImages, pcs:GetCluster`
DeleteComputeNodeGroup	`pcs:DeleteComputeNodeGroup`	
CreateQueue	`pcs:CreateQueue`	`pcs:ListComputeNodeGroups, pcs:GetCluster`
ListQueues	`pcs:ListQueues`	`pcs:GetCluster`
GetQueue	`pcs:GetQueue`	

Acción de la API	Permisos mínimos	Permisos adicionales para la consola
UpdateQueue	`pcs:UpdateQueue`	`pcs:ListComputeNodeGroups, pcs:GetQueue`
DeleteQueue	`pcs:DeleteQueue`	

Permisos mínimos para usar etiquetas

Se requieren los siguientes permisos para usar etiquetas con sus recursos en AWS PCS.

```
pcs:ListTagsForResource,
pcs:TagResource,
pcs:UntagResource
```

Permisos mínimos para admitir registros

AWS PCS envía los datos de registro a Amazon CloudWatch Logs (CloudWatch Logs). Debe asegurarse de que su identidad tiene los permisos mínimos para usar CloudWatch Logs. Para obtener más información, consulte Descripción general de la gestión de los permisos de acceso a sus recursos de CloudWatch Logs en la Guía del usuario de Amazon CloudWatch Logs.

Para obtener información sobre los permisos necesarios para que un servicio envíe CloudWatch registros a Logs, consulte Habilitar el registro desde AWS los servicios en la Guía del usuario de Amazon CloudWatch Logs.

Permisos mínimos para usar los bloques de capacidad

Amazon EC2 Capacity Blocks for ML es una opción de compra de Amazon EC2 que le permite pagar por adelantado la reserva de instancias de computación acelerada basadas en GPU dentro de un intervalo de fechas y horas específico para soportar cargas de trabajo de corta duración. Para obtener más información, consulte Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS.

Puede elegir usar bloques de capacidad al crear o actualizar un grupo de nodos de cómputo. La identidad de IAM que utilice para crear o actualizar el grupo de nodos de cómputo debe tener el siguiente permiso:

```
ec2:DescribeCapacityReservations
```

Permisos mínimos para un administrador de servicios

La siguiente política de IAM especifica los permisos mínimos necesarios para que una identidad de IAM (usuario, grupo o rol) configure y administre el servicio AWS PCS.

> **Note**
>
> Los usuarios que no configuran ni administran el servicio no necesitan estos permisos. Los usuarios que solo ejecutan trabajos utilizan un shell seguro (SSH) para conectarse al clúster. AWS Identity and Access Management (IAM) no gestiona la autenticación ni la autorización de SSH.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "PCSAccess",
      "Effect": "Allow",
      "Action": [
        "pcs:*"
      ],
      "Resource": "*"
    },
    {
      "Sid": "EC2Access",
      "Effect": "Allow",
      "Action": [
        "ec2:CreateNetworkInterface",
        "ec2:DescribeImages",
        "ec2:GetSecurityGroupsForVpc",
        "ec2:DescribeSubnets",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeVpcs",
        "ec2:DescribeLaunchTemplates",
```

```
        "ec2:DescribeLaunchTemplateVersions",
        "ec2:DescribeInstanceTypes",
        "ec2:DescribeInstanceTypeOfferings",
        "ec2:RunInstances",
        "ec2:CreateFleet",
        "ec2:CreateTags",
        "ec2:DescribeCapacityReservations"
      ],
      "Resource": "*"
    },
    {
      "Sid": "IamInstanceProfile",
      "Effect": "Allow",
      "Action": [
        "iam:GetInstanceProfile"
      ],
      "Resource": "*"
    },
    {
      "Sid": "IamPassRole",
      "Effect": "Allow",
      "Action": [
        "iam:PassRole"
      ],
      "Resource": [
        "arn:aws:iam::*:role/*/AWSPCS*",
        "arn:aws:iam::*:role/AWSPCS*",
        "arn:aws:iam::*:role/aws-pcs/*",
        "arn:aws:iam::*:role/*/aws-pcs/*"
      ],
      "Condition": {
        "StringEquals": {
           "iam:PassedToService": [
             "ec2.amazonaws.com"
           ]
        }
      }
    },
    {
      "Sid": "SLRAccess",
      "Effect": "Allow",
      "Action": [
        "iam:CreateServiceLinkedRole"
      ],
```

```
      "Resource": [
        "arn:aws:iam::*:role/aws-service-role/pcs.amazonaws.com/AWSServiceRoleFor*",
        "arn:aws:iam::*:role/aws-service-role/spot.amazonaws.com/AWSServiceRoleFor*"
      ],
      "Condition": {
        "StringLike": {
          "iam:AWSServiceName": [
            "pcs.amazonaws.com",
            "spot.amazonaws.com"
          ]
        }
      }
    },
    {
      "Sid": "AccessKMSKey",
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt",
        "kms:Encrypt",
        "kms:GenerateDataKey",
        "kms:CreateGrant",
        "kms:DescribeKey"
      ],
      "Resource": "*"
    },
    {
      "Sid": "SecretManagementAccess",
      "Effect": "Allow",
      "Action": [
        "secretsmanager:CreateSecret",
        "secretsmanager:TagResource",
        "secretsmanager:UpdateSecret",
        "secretsmanager:RotateSecret"
      ],
      "Resource": "*"
    },
    {
       "Sid": "ServiceLogsDelivery",
       "Effect": "Allow",
       "Action": [
         "pcs:AllowVendedLogDeliveryForResource",
         "logs:PutDeliverySource",
         "logs:PutDeliveryDestination",
         "logs:CreateDelivery"
```

```
        ],
        "Resource": "*"
    }
  ]
}
```

Perfiles de instancia de IAM para AWS Parallel Computing Service

Las aplicaciones que se ejecutan en una instancia EC2 deben incluir AWS credenciales en todas las solicitudes de AWS API que realicen. Le recomendamos que utilice un rol de IAM para administrar las credenciales temporales en la instancia EC2. Para ello, puede definir un perfil de instancia y adjuntarlo a sus instancias. Para obtener más información, consulte las funciones de IAM para Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud.

> **Note**
>
> Cuando se utiliza Consola de administración de AWS para crear un rol de IAM para Amazon EC2, la consola crea un perfil de instancia automáticamente y le asigna el mismo nombre que el rol de IAM. Si utiliza las acciones de la AWS CLI AWS API o un AWS SDK para crear el rol de IAM, crea el perfil de instancia como una acción independiente. Para obtener más información, consulte Perfiles de instancia en la Guía del usuario de Amazon Elastic Compute Cloud.

Debe especificar el nombre de recurso de Amazon (ARN) de un perfil de instancia al crear un grupo de nodos de cómputo. Puede elegir diferentes perfiles de instancia para algunos o todos los grupos de nodos de cómputo.

Requisitos

Función de IAM del perfil de instancia

El rol de IAM asociado al perfil de la instancia debe estar `/aws-pcs/` en su ruta o su nombre debe empezar por. `AWSPCS`

Ejemplo de rol de IAM ARNs

- `arn:aws:iam::*:role/AWSPCS-example-role-1`
- `arn:aws:iam::*:role/aws-pcs/example-role-2`

Permisos

La función de IAM asociada al perfil de instancia de AWS PCS debe incluir la siguiente política.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

Políticas adicionales

Considere la posibilidad de añadir políticas administradas al perfil de la instancia. Por ejemplo:

- AmazonS3 ReadOnlyAccess proporciona acceso de solo lectura a todos los buckets de S3.
- Amazon SSMManaged InstanceCore habilita las funciones principales del servicio AWS Systems Manager, como el acceso remoto directamente desde la consola de administración de Amazon.
- CloudWatchAgentServerPolicycontiene los permisos necesarios para su uso AmazonCloudWatchAgent en los servidores.

También puede incluir sus propias políticas de IAM que respalden su caso de uso específico.

Cree un perfil de instancia para PCS AWS

AWS PCS console

Seleccione Crear un perfil básico al crear un grupo de nodos de procesamiento para que AWS PCS cree uno por usted con la política mínima requerida.

Amazon EC2 console

Puede crear un perfil de instancia directamente desde la consola Amazon EC2. Para obtener más información, consulte Uso de perfiles de instancia en la Guía del AWS Identity and Access Management usuario.

> **Important**
>
> Asegúrese de utilizar el prefijo necesario `AWSPCS` en el nombre del rol de IAM.

AWS CLI

Configuración del perfil de instancia básico mediante AWS CLI

> **Note**
>
> Sustitúyalo *`example-role`* en los siguientes ejemplos por el nombre de su función de IAM.

1. Cree el rol de IAM `/aws-pcs/` como atributo de ruta o un nombre que comience por. `AWSPCS`

   a. Copia y pega el siguiente contenido en un nuevo archivo de texto denominado`trust_policy.json`.

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
             "Principal": {
                "Service": [
                    "ec2.amazonaws.com"
                ]
            },
            "Action": [
                "sts:AssumeRole"
            ]
```

```
        }
    ]
}
```

b. Utilice uno de los siguientes comandos para crear el rol de IAM.

```
aws iam create-role --path /aws-pcs/ --role-name example-role --assume-role-
policy-document file://trust_policy.json
```

o

```
aws iam create-role --role-name AWSPCS-example-role --assume-role-policy-
document file://trust_policy.json
```

2. Adjunte permisos.

a. Copia y pega el siguiente contenido en un nuevo archivo de texto denominado`policy_document.json`.

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

b. Adjunte el documento de política al rol. Este comando adjunta la política como una política en línea.

```
aws iam put-role-policy \
    --role-name example-role \
    --policy-name pcsRegisterInstancePolicy \
    --policy-document file://policy_document.json
```

3. Cree un perfil de instancia. ***example-profile***Sustitúyalo por el nombre de tu perfil de instancia.

```
aws iam create-instance-profile --instance-profile-name example-profile
```

4. Asocia la función de IAM al perfil de la instancia.

```
aws iam add-role-to-instance-profile \
   --instance-profile-name example-profile \
   --role-name example-role
```

Busque los perfiles de instancia utilizados con PCS AWS

1. Si no conoce los nombres exactos de las funciones de IAM para AWS PCS, utilice el siguiente AWS CLI comando para enumerar las funciones de IAM que cumplen los requisitos de nombre de AWS PCS.

```
aws iam list-roles --query "Roles[?starts_with(RoleName, 'AWSPCS') ||
 contains(Path, '/aws-pcs/')].[RoleName]" --output text
```

2. Utilice el siguiente AWS CLI comando para enumerar los perfiles de instancia asociados a un rol de IAM específico. *role-name*Sustitúyalo por el nombre de un rol de IAM que cumpla con los requisitos de nombre del AWS PCS.

```
aws iam list-instance-profiles-for-role --role-name role-name
```

Solución de problemas de identidad y acceso a AWS Parallel Computing Service

Utilice la siguiente información como ayuda para diagnosticar y solucionar los problemas habituales que pueden surgir al trabajar con AWS PCS e IAM.

Temas

- No estoy autorizado a realizar ninguna acción en PCS AWS
- No estoy autorizado a realizar tareas como: PassRole
- Quiero permitir que personas ajenas a mí accedan Cuenta de AWS a mis recursos de AWS PCS

No estoy autorizado a realizar ninguna acción en PCS AWS

Si recibe un error que indica que no tiene autorización para realizar una acción, las políticas se deben actualizar para permitirle realizar la acción.

En el siguiente ejemplo, el error se produce cuando el usuario de IAM `mateojackson` intenta utilizar la consola para consultar los detalles acerca de un recurso ficticio *`my-example-widget`*, pero no tiene los permisos ficticios `pcs:`*`GetWidget`*.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 pcs:GetWidget on resource: my-example-widget
```

En este caso, la política del usuario `mateojackson` debe actualizarse para permitir el acceso al recurso *`my-example-widget`* mediante la acción `pcs:`*`GetWidget`*.

Si necesita ayuda, póngase en contacto con su AWS administrador. El administrador es la persona que le proporcionó las credenciales de inicio de sesión.

No estoy autorizado a realizar tareas como: PassRole

Si recibe un mensaje de error que indica que no está autorizado a realizar la `iam:PassRole` acción, sus políticas deben actualizarse para que pueda transferir una función a AWS PCS.

Algunas Servicios de AWS permiten transferir una función existente a ese servicio en lugar de crear una nueva función de servicio o una función vinculada al servicio. Para ello, debe tener permisos para transferir la función al servicio.

El siguiente ejemplo de error se produce cuando un usuario de IAM denominado `marymajor` intenta utilizar la consola para realizar una acción en AWS PCS. Sin embargo, la acción requiere que el servicio cuente con permisos que otorguen un rol de servicio. Mary no tiene permisos para transferir la función al servicio.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

En este caso, las políticas de Mary se deben actualizar para permitirle realizar la acción `iam:PassRole`.

Si necesita ayuda, póngase en contacto con su AWS administrador. El administrador es la persona que le proporcionó las credenciales de inicio de sesión.

Quiero permitir que personas ajenas a mí accedan Cuenta de AWS a mis recursos de AWS PCS

Se puede crear un rol que los usuarios de otras cuentas o las personas externas a la organización puedan utilizar para acceder a sus recursos. Se puede especificar una persona de confianza para que asuma el rol. En el caso de los servicios que admiten políticas basadas en recursos o listas de control de acceso (ACLs), puede utilizar esas políticas para permitir que las personas accedan a sus recursos.

Para obtener más información, consulte lo siguiente:

- Para saber si AWS PCS admite estas funciones, consulte. Cómo funciona AWS Parallel Computing Service con IAM
- Para obtener información sobre cómo proporcionar acceso a los recursos de su Cuentas de AWS propiedad, consulte Proporcionar acceso a un usuario de IAM en otro usuario de su propiedad Cuenta de AWS en la Guía del usuario de IAM.
- Para obtener información sobre cómo proporcionar acceso a tus recursos a terceros Cuentas de AWS, consulta Cómo proporcionar acceso a recursos que Cuentas de AWS son propiedad de terceros en la Guía del usuario de IAM.
- Para obtener información sobre cómo proporcionar acceso mediante una federación de identidades, consulte Proporcionar acceso a usuarios autenticados externamente (identidad federada) en la Guía del usuario de IAM.
- Para conocer sobre la diferencia entre las políticas basadas en roles y en recursos para el acceso entre cuentas, consulte Acceso a recursos entre cuentas en IAM en la Guía del usuario de IAM.

Validación de la conformidad del servicio de computación AWS paralela

Para saber si uno Servicio de AWS está dentro del ámbito de aplicación de programas de cumplimiento específicos, consulte Servicios de AWS Alcance por programa de cumplimiento Servicios de AWS de cumplimiento y elija el programa de cumplimiento que le interese. Para obtener información general, consulte Programas de AWS cumplimiento > Programas AWS .

Puede descargar informes de auditoría de terceros utilizando AWS Artifact. Para obtener más información, consulte Descarga de informes en AWS Artifact .

Su responsabilidad de cumplimiento al Servicios de AWS utilizarlos viene determinada por la confidencialidad de sus datos, los objetivos de cumplimiento de su empresa y las leyes y reglamentos aplicables. Para obtener más información sobre su responsabilidad de conformidad al utilizarlos Servicios de AWS, consulte AWS la documentación de seguridad.

Servicio de resiliencia en la computación AWS paralela

La infraestructura AWS global se basa en zonas Regiones de AWS de disponibilidad. Regiones de AWS proporcionan varias zonas de disponibilidad aisladas y separadas físicamente, que están conectadas mediante redes de baja latencia, alto rendimiento y alta redundancia. Con las zonas de disponibilidad, puede diseñar y utilizar aplicaciones y bases de datos que realizan una conmutación por error automática entre las zonas sin interrupciones. Las zonas de disponibilidad tienen una mayor disponibilidad, tolerancia a errores y escalabilidad que las infraestructuras tradicionales de uno o varios centros de datos.

Para obtener más información sobre las zonas de disponibilidad Regiones de AWS y las zonas de disponibilidad, consulte Infraestructura global.AWS

Servicio de seguridad de infraestructura en computación AWS paralela

Como servicio gestionado, AWS Parallel Computing Service está protegido por la seguridad de la red AWS global. Para obtener información sobre los servicios AWS de seguridad y cómo se AWS protege la infraestructura, consulte Seguridad AWS en la nube. Para diseñar su AWS entorno utilizando las mejores prácticas de seguridad de la infraestructura, consulte Protección de infraestructuras en un marco de buena AWS arquitectura basado en el pilar de la seguridad.

Las llamadas a la API AWS publicadas se utilizan para acceder al AWS PCS a través de la red. Los clientes deben admitir lo siguiente:

- Seguridad de la capa de transporte (TLS). Exigimos TLS 1.2 y recomendamos TLS 1.3.
- Conjuntos de cifrado con confidencialidad directa total (PFS) como DHE (Ephemeral Diffie-Hellman) o ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). La mayoría de los sistemas modernos como Java 7 y posteriores son compatibles con estos modos.

Cuando AWS PCS crea un clúster, el servicio lanza el controlador Slurm en una cuenta propiedad del servicio, independiente de los nodos de procesamiento de su cuenta. Para conectar la

comunicación entre el controlador y los nodos de cómputo, AWS PCS crea una interfaz de red elástica (ENI) multicuenta en la VPC. El controlador Slurm utiliza el ENI para gestionar y comunicarse con los distintos nodos de cómputo Cuentas de AWS, manteniendo la seguridad y el aislamiento de los recursos y, al mismo tiempo, facilitando la eficiencia de la HPC y de las operaciones. AI/ML

Análisis y gestión de vulnerabilidades en AWS Parallel Computing Service

La configuración y los controles de TI son una responsabilidad compartida entre usted AWS y usted. Para obtener más información, consulte el modelo de responsabilidad AWS compartida. AWS gestiona las tareas de seguridad básicas de la infraestructura subyacente de la cuenta de servicio, como la aplicación de parches al sistema operativo en las instancias del controlador, la configuración del firewall y la recuperación ante desastres de la AWS infraestructura. Estos procedimientos han sido revisados y certificados por los terceros pertinentes. Para obtener más información, consulte las Prácticas recomendadas sobre seguridad, identidad y conformidad.

> **Note**
>
> Los controladores Slurm no estarán disponibles mientras los actualizamos. Los trabajos en ejecución no se ven afectados. Los trabajos enviados cuando el controlador del clúster no está disponible se retienen hasta que el controlador esté disponible.

Usted es responsable de la seguridad de la infraestructura subyacente de su Cuenta de AWS:

- Mantenga su código, incluidas las actualizaciones y los parches de seguridad.
- Aplica parches y actualiza el sistema operativo en Amazon Machine Image (AMI) para tus grupos de nodos de cómputo y actualiza tus grupos de nodos de cómputo para usar la AMI actualizada.
- Actualice el programador para mantenerlo dentro de las versiones compatibles. Actualice la AMI de sus grupos de nodos de cómputo y actualice su grupo de nodos de cómputo para usar la AMI actualizada.
- Autentica y cifra la comunicación entre los clientes de los usuarios y los nodos a los que se conectan.

Para obtener más información sobre la actualización de la AMI para sus grupos de nodos de procesamiento, consulteAmazon Machine Images (AMIs) para AWS PCS.

Prevención de la sustitución confusa entre servicios

El problema del suplente confuso es una cuestión de seguridad en la que una entidad que no tiene permiso para realizar una acción puede obligar a una entidad con más privilegios a realizar la acción. En AWS, la suplantación de identidad entre servicios puede provocar el confuso problema de un diputado. La suplantación entre servicios puede producirse cuando un servicio (el servicio que lleva a cabo las llamadas) llama a otro servicio (el servicio al que se llama). El servicio que lleva a cabo las llamadas se puedes manipular para utilizar sus permisos a fin de actuar en función de los recursos de otro cliente de una manera en la que no debe tener permiso para acceder. Para evitarlo, AWS proporciona herramientas que lo ayudan a proteger sus datos para todos los servicios con entidades principales de servicio a las que se les ha dado acceso a los recursos de su cuenta.

Se recomienda utilizar las claves de contexto de condición `aws:SourceAccount`global `aws:SourceArn`y las claves de contexto en las políticas de recursos para limitar los permisos que AWS Parallel Computing Service (AWS PCS) concede a otro servicio al recurso. Utiliza `aws:SourceArn` si desea que solo se asocie un recurso al acceso entre servicios. Utiliza `aws:SourceAccount` si quiere permitir que cualquier recurso de esa cuenta se asocie al uso entre servicios.

La forma más eficaz de protegerse contra el problema de la sustitución confusa es utilizar la clave de contexto de condición global de `aws:SourceArn` con el ARN completo del recurso. Si no conoce el ARN completo del recurso o si está especificando varios recursos, utilice la clave de condición de contexto global `aws:SourceArn` con caracteres comodines (*) para las partes desconocidas del ARN. Por ejemplo, `arn:aws:`*`servicename`*`:*:`*`123456789012`*`:*`.

Si el valor de `aws:SourceArn` no contiene el ID de cuenta, como un ARN de bucket de Amazon S3, debe utilizar ambas claves de contexto de condición global para limitar los permisos.

El valor de `aws:SourceArn` debe ser un ARN de clúster.

El siguiente ejemplo muestra cómo se pueden utilizar las claves de contexto de condición `aws:SourceAccount` global `aws:SourceArn` y las claves de contexto del AWS PCS para evitar el confuso problema de los diputados.

```
{
"Version": "2012-10-17",
  "Statement": {
"Sid": "ConfusedDeputyPreventionExamplePolicy",
    "Effect": "Allow",
```

```
    "Principal": {
      "Service": "pcs.amazonaws.com"
    },
    "Action": "sts:AssumeRole",
    "Condition": {
      "ArnLike": {
        "aws:SourceArn": [
          "arn:aws:pcs:us-east-1:123456789012:cluster/*"
        ]
      },
      "StringEquals": {
        "aws:SourceAccount": "123456789012"
      }
    }
  }
}
```

Función de IAM para instancias de Amazon EC2 aprovisionadas como parte de un grupo de nodos de cómputo

AWS PCS organiza automáticamente la capacidad de Amazon EC2 para cada uno de los grupos de nodos de procesamiento configurados de un clúster. Al crear un grupo de nodos de cómputo, los usuarios deben proporcionar un perfil de instancia de IAM a través del campo. `iamInstanceProfileArn` El perfil de instancia especifica los permisos asociados a las instancias de EC2 aprovisionadas. AWS PCS acepta cualquier rol que tenga `AWSPCS` como prefijo de nombre de rol o `/aws-pcs/` como parte de la ruta del rol. El `iam:PassRole` permiso es obligatorio para la identidad de IAM (usuario o rol) que crea o actualiza un grupo de nodos de procesamiento. Cuando un usuario llama a las acciones `CreateComputeNodeGroup` o a la `UpdateComputeNodeGroup` API, AWS PCS comprueba si el usuario tiene permiso para realizar la `iam:PassRole` acción.

La siguiente política de ejemplo otorga permisos para transferir únicamente los roles de IAM cuyo nombre comience por `AWSPCS`.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:PassRole",
            "Resource": "arn:aws:iam::123456789012:role/AWSPCS*",
            "Condition": {
```

```
                "StringEquals": {
                    "iam:PassedToService": [
                        "ec2.amazonaws.com"
                    ]
                }
            }
        }
    ]
}
```

Prácticas recomendadas de seguridad para AWS Parallel Computing Service

En esta sección se describen las mejores prácticas de seguridad específicas de AWS Parallel Computing Service (AWS PCS). Para obtener más información sobre las prácticas recomendadas de seguridad AWS, consulte Prácticas recomendadas en materia de seguridad, identidad y conformidad.

Seguridad relacionada con la AMI

- No utilice la muestra de AWS PCS AMIs para las cargas de trabajo de producción. Las muestras no AMIs son compatibles y solo están destinadas a ser probadas.
- Actualice periódicamente el sistema operativo y el software de la AMI para sus grupos de nodos de cómputo a fin de mitigar las vulnerabilidades.
- Utilice únicamente paquetes de AWS PCS oficiales autenticados y descargados de AWS fuentes oficiales.
- Actualice periódicamente los paquetes de AWS PCS en la AMI para los grupos de nodos de cómputo y actualice los nodos de cómputo para usar la AMI actualizada. Considere la posibilidad de automatizar este proceso para minimizar las vulnerabilidades.

Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Seguridad de Slurm Workload Manager

- Implemente controles de acceso y restricciones de red para proteger los nodos de control y cómputo de Slurm. Permita que solo los usuarios y sistemas de confianza envíen trabajos y accedan a los comandos de administración de Slurm.

- Utilice las funciones de seguridad integradas de Slurm, como la autenticación de Slurm, para garantizar que las solicitudes de trabajo y las comunicaciones estén autenticadas.
- Actualice las versiones de Slurm para mantener un funcionamiento fluido y la compatibilidad con clústeres.

  > **Important**
  >
  > Cualquier clúster que utilice una versión de Slurm que haya llegado al final de su vida útil (EOSL) se detendrá inmediatamente. Utilice el enlace que aparece en la parte superior de las páginas de la guía del usuario para suscribirse a la fuente RSS de documentación del AWS PCS y recibir notificaciones cuando una versión de Slurm se acerque a la EOSL.

  Para obtener más información, consulte Versiones de Slurm en PCS AWS.
- Cambie periódicamente los secretos de los clústeres para mantener el cumplimiento de las normas de seguridad y solucionar posibles problemas de seguridad. Esto es obligatorio para cumplir con la HIPAA y el FedRAMP.

  Para obtener más información, consulte Secretos de clústeres rotativos en AWS PCS.

Supervisión y registro

- Use Amazon CloudWatch Logs AWS CloudTrail para monitorear y registrar las acciones en sus clústeres y Cuenta de AWS. Utilice los datos para la solución de problemas y la auditoría.

Seguridad de la red

- Implemente sus clústeres de AWS PCS en una VPC independiente para aislar su entorno de HPC del resto del tráfico de red.
- Utilice grupos de seguridad y listas de control de acceso a la red (ACLs) para controlar el tráfico entrante y saliente a las instancias y subredes de AWS PCS.
- Utilice AWS PrivateLink nuestros puntos de enlace de VPC para mantener el tráfico de red entre los clústeres y otros AWS servicios de la red. AWS Para obtener más información, consulte Acceda AWS Parallel Computing Service mediante un punto final de interfaz ()AWS PrivateLink.

Registro y supervisión para AWS PCS

La supervisión es una parte importante del mantenimiento de la fiabilidad, la disponibilidad y el rendimiento de AWS PCS y del resto de los recursos de AWS. AWS proporciona las siguientes herramientas de supervisión para vigilar el AWS PCS, informar cuando algo va mal y tomar medidas automáticas cuando sea necesario:

- Amazon CloudWatch monitorea tus AWS recursos y las aplicaciones en las que AWS ejecutas en tiempo real. Puede recopilar métricas y realizar un seguimiento de las métricas, crear paneles personalizados y definir alarmas que le advierten o que toman medidas cuando una métrica determinada alcanza el umbral que se especifique. Por ejemplo, puede CloudWatch hacer un seguimiento del uso de la CPU u otras métricas de sus instancias de Amazon EC2 y lanzar automáticamente nuevas instancias cuando sea necesario. Para obtener más información, consulta la Guía del CloudWatch usuario de Amazon.
- Amazon CloudWatch Logs le permite supervisar, almacenar y acceder a sus archivos de registro desde instancias de Amazon EC2 y otras fuentes. CloudTrail CloudWatch Los registros pueden monitorear la información de los archivos de registro y notificarle cuando se alcancen ciertos umbrales. También se pueden archivar los datos del registro en un almacenamiento de larga duración. Para obtener más información, consulta la Guía del usuario CloudWatch de Amazon Logs.
- AWS CloudTrailcaptura las llamadas a la API y los eventos relacionados realizados por su AWS cuenta o en su nombre y entrega los archivos de registro a un bucket de Amazon S3 que especifique. Puede identificar qué usuarios y cuentas llamaron AWS, la dirección IP de origen desde la que se realizaron las llamadas y cuándo se produjeron las llamadas. Para obtener más información, consulte la Guía del usuario de AWS CloudTrail.

Registros de finalización de trabajos en AWS PCS

Los registros de finalización de trabajos le proporcionan detalles clave sobre los trabajos del Servicio de Computación AWS Paralela (AWS PCS) una vez finalizados, sin coste adicional. Puede utilizar otros AWS servicios para acceder a sus datos de registro y procesarlos, como Amazon CloudWatch Logs, Amazon Simple Storage Service (Amazon S3) y Amazon Data Firehose AWS ; PCS registra metadatos sobre sus trabajos, como los siguientes.

- ID y nombre del trabajo
- Información de usuario y grupo

- Estado del trabajo (por ejemplo`COMPLETED`,`FAILED`,`CANCELLED`)
- Partición utilizada
- Límites de tiempo
- Horarios de inicio, finalización, envío y disponibilidad
- Lista y recuento de nodos
- Recuento de procesadores
- Directorio de trabajo
- Uso de recursos (CPU, memoria)
- Códigos de salida
- Detalles de los nodos (nombres, instancias IDs, tipos de instancias)

Contenido

Requisitos previos

El principal de IAM que administra el clúster de AWS PCS debe permitir la `pcs:AllowVendedLogDeliveryForResource` acción.

El siguiente ejemplo de política de IAM concede los permisos necesarios.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [
      {
         "Sid": "PcsAllowVendedLogsDelivery",
```

```
            "Effect": "Allow",
            "Action": ["pcs:AllowVendedLogDeliveryForResource"],
            "Resource": [
               "arn:aws:pcs:*::cluster/*"
            ]
         }
      ]
}
```

Configure los registros de finalización de trabajos

Puede configurar los registros de finalización de tareas para su clúster de AWS PCS con la tecla Consola de administración de AWS o AWS CLI.

Consola de administración de AWS

Para configurar los registros de finalización de tareas con la consola

1. Abra la consola AWS PCS.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Elija el clúster en el que desee añadir los registros de finalización de tareas.
4. En la página de detalles del clúster, seleccione la pestaña Registros.
5. En Registros de finalización de trabajos, selecciona Añadir para añadir hasta 3 destinos de entrega de CloudWatch registros de entre Logs, Amazon S3 y Firehose.
6. Selecciona Actualizar entregas de registros.

AWS CLI

Para configurar los registros de finalización de trabajos con el AWS CLI

1. Cree un destino de entrega de registros:

```
aws logs put-delivery-destination --region region \
  --name pcs-logs-destination \
  --delivery-destination-configuration \
  destinationResourceArn=resource-arn
```

Reemplace:

- *region*— El Región de AWS lugar donde desea crear el destino, por ejemplo `us-east-1`
- *pcs-logs-destination*— Un nombre para el destino
- *resource-arn*— El nombre de recurso de Amazon (ARN) de un grupo de CloudWatch registros de Logs, un bucket de S3 o una transmisión de entrega de Firehose.

Para obtener más información, consulta PutDeliveryDestinationla referencia de la API CloudWatch de Amazon Logs.

2. Configure el clúster de PCS como fuente de entrega de registros:

```
aws logs put-delivery-source --region region \
  --name cluster-logs-source-name \
  --resource-arn cluster-arn \
  --log-type PCS_JOBCOMP_LOGS
```

Reemplace:

- *region*— El Región de AWS de su clúster, como `us-east-1`
- *cluster-logs-source-name*— Un nombre para la fuente
- *cluster-arn*— el ARN de su AWS clúster PCS

Para obtener más información, consulta PutDeliverySourcela referencia de la API CloudWatch de Amazon Logs.

3. Conecta la fuente de entrega al destino de entrega:

```
aws logs create-delivery --region region \
  --delivery-source-name cluster-logs-source \
  --delivery-destination-arn destination-arn
```

Reemplace:

- *region*— Los Región de AWS, como `us-east-1`
- *cluster-logs-source*— El nombre de su fuente de entrega
- *destination-arn*— El ARN de su destino de entrega

Para obtener más información, consulta CreateDeliveryla referencia de la API CloudWatch de Amazon Logs.

¿Cómo encontrar los registros de finalización de trabajos

Puede configurar los destinos de los CloudWatch registros en Logs y Amazon S3. AWS PCS utiliza los siguientes nombres de rutas y archivos estructurados.

CloudWatch Registros

AWS El PCS utiliza el siguiente formato de nombre para el flujo de CloudWatch registros:

```
AWSLogs/PCS/cluster-id/jobcomp.log
```

Por ejemplo: `AWSLogs/PCS/pcs_abc123de45/jobcomp.log`

Amazon S3

AWS PCS utiliza el siguiente formato de nombre para la ruta S3:

```
AWSLogs/account-id/PCS/region/cluster-id/jobcomp/year/month/day/hour/
```

Por ejemplo: `AWSLogs/111122223333/PCS/us-east-1/pcs_abc123de45/jobcomp/2025/06/19/11/`

AWS PCS utiliza el siguiente formato de nombre para los archivos de registro:

```
PCS_jobcomp_year-month-day-hour_cluster-id_random-id.log.gz
```

Por ejemplo: `PCS_jobcomp_2025-06-19-11_pcs_abc123de45_04be080b.log.gz`

Campos del registro de finalización de trabajos

AWS PCS escribe los datos del registro de finalización del trabajo como objetos JSON. El contenedor JSON `jobcomp` contiene los detalles del trabajo. En la siguiente tabla se describen los campos del `jobcomp` contenedor. Algunos campos solo están presentes en circunstancias específicas, como en trabajos de matriz o trabajos heterogéneos.

Campos del registro de finalización de trabajos

Name	Ejemplo de valor	Obligatorio	Notas
`job_id`	`11`	yes	Siempre presentes y con valor
`user`	`"root"`	yes	Siempre presente con valor
`user_id`	`0`	yes	Siempre presente con valor
`group`	`"root"`	yes	Siempre presente con valor
`group_id`	`0`	yes	Siempre presente con valor
`name`	`"wrap"`	yes	Siempre presente con valor
`job_state`	`"COMPLETED"`	yes	Siempre presente con valor
`partition`	`"Hydra-MpiQueue-abcdef01-7"`	yes	Siempre presente con valor
`time_limit`	`"UNLIMITED"`	yes	Siempre presente, pero podría estarlo `"UNLIMITED"`
`start_time`	`"2025-06-19T10:58:57"`	yes	Siempre presente, pero podría estarlo `"Unknown"`
`end_time`	`"2025-06-19T10:58:57"`	yes	Siempre presente, pero podría estarlo `"Unknown"`
`node_list`	`"Hydra-MpiNG-abcdef01-2345-1"`	yes	Siempre presente con valor
`node_cnt`	`1`	yes	Siempre presente con valor
`proc_cnt`	`1`	yes	Siempre presente con valor

Name	Ejemplo de valor	Obligatorio	Notas
work_dir	"/root"	yes	Siempre presente, pero podría estarlo "Unknown"
reservation_name	"weekly_maintenance"	yes	Siempre presente, pero puede ser una cadena vacía ""
tres.cpu	1	yes	Siempre presente con valor
tres.mem.val	600	yes	Siempre presente con valor
tres.mem.unit	"M"	yes	Puede ser "M" o "bb"
tres.node	1	yes	Siempre presente con valor
tres.billing	1	yes	Siempre presente con valor
account	"finance"	yes	Siempre presente, pero puede ser una cadena vacía ""
qos	"normal"	yes	Siempre presente, pero puede ser una cadena vacía ""
wc_key	"project_1"	yes	Siempre presente, pero puede ser una cadena vacía ""
cluster	"unknown"	yes	Siempre presente, pero podría estarlo "unknown"
submit_time	"2025-06-19T10:55:46"	yes	Siempre presente, pero podría estarlo "Unknown"

Name	Ejemplo de valor	Obligatorio	Notas
`eligible_time`	`"2025-06-19T10:55:46"`	yes	Siempre presente, pero podría estarlo `"Unknown"`
`array_job_id`	`12`	no	Solo está presente si el trabajo es un trabajo de matriz
`array_task_id`	`1`	no	Solo está presente si el trabajo es un trabajo de matriz
`het_job_id`	`10`	no	Solo está presente si el trabajo es heterogéneo
`het_job_offset`	`0`	no	Solo está presente si el trabajo es heterogéneo
`derived_exit_code_status`	`0`	yes	Siempre presente con valor
`derived_exit_code_signal`	`0`	yes	Siempre presente con valor
`exit_code_status`	`0`	yes	Siempre presente con valor
`exit_code_signal`	`0`	yes	Siempre presente con valor
`node_details[0].name`	`"Hydra-MpiNG-abcdef01-2345-1"`	no	Siempre presente, pero `node_details` podría estarlo `"[]"`

Name	Ejemplo de valor	Obligatorio	Notas
`node_details[0].instance_id`	`"i-0abcdef01234567a"`	no	Siempre presente, pero `node_details` podría estarlo `"[]"`
`node_details[0].instance_type`	`"t4g.micro"`	no	Siempre presente, pero `node_details` podría estarlo `"[]"`

Ejemplos de registros de finalización de trabajos

En los siguientes ejemplos se muestran los registros de finalización de tareas para varios tipos y estados de tareas:


```
{ "jobcomp": { "job_id": 1, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-
abcdef01-7", "time_limit": "UNLIMITED", "start_time": "2025-06-19T16:32:57",
 "end_time": "2025-06-19T16:33:03", "node_list": "Hydra-MpiNG-abcdef01-2345-
[1-2]", "node_cnt": 2, "proc_cnt": 2, "work_dir": "/usr/bin", "reservation_name":
 "", "tres": { "cpu": 2, "mem": { "val": 1944, "unit": "M" }, "node": 2,
 "billing": 2 }, "account": "", "qos": "", "wc_key": "", "cluster": "unknown",
 "submit_time": "2025-06-19T16:29:40", "eligible_time": "2025-06-19T16:29:41",
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc123def45678", "instance_type": "t4g.micro" }, { "name":
 "Hydra-MpiNG-abcdef01-2345-2", "instance_id": "i-0def456abc78901", "instance_type":
 "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 2, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-
abcdef01-7", "time_limit": "UNLIMITED", "start_time": "2025-06-19T16:33:13",
 "end_time": "2025-06-19T16:33:14", "node_list": "Hydra-MpiNG-abcdef01-2345-
[1-2]", "node_cnt": 2, "proc_cnt": 2, "work_dir": "/usr/bin", "reservation_name":
 "", "tres": { "cpu": 2, "mem": { "val": 1944, "unit": "M" }, "node": 2,
 "billing": 2 }, "account": "", "qos": "", "wc_key": "", "cluster": "unknown",
 "submit_time": "2025-06-19T16:33:13", "eligible_time": "2025-06-19T16:33:13",
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc123def45678", "instance_type": "t4g.micro" }, { "name":
```

```
 "Hydra-MpiNG-abcdef01-2345-2", "instance_id": "i-0def456abc78901", "instance_type":
 "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 3, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T22:58:57", "end_time":
 "2025-06-19T22:58:57", "node_list": "Hydra-MpiNG-abcdef01-2345-1", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 972, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T22:55:46",
 "eligible_time": "2025-06-19T22:55:46", "derived_exit_code_status": 0,
 "derived_exit_code_signal": 0, "exit_code_status": 0, "exit_code_signal":
 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1", "instance_id":
 "i-0abc234def56789", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 4, "user": "root", "user_id": 0, "group": "root",
 "group_id": 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-
MpiQueue-abcdef01-7", "time_limit": "525600", "start_time": "2025-06-19T23:04:27",
 "end_time": "2025-06-19T23:04:27", "node_list": "Hydra-MpiNG-abcdef01-2345-
[1-2]", "node_cnt": 2, "proc_cnt": 2, "work_dir": "/root", "reservation_name":
 "", "tres": { "cpu": 2, "mem": { "val": 1944, "unit": "M" }, "node": 2,
 "billing": 2 }, "account": "", "qos": "", "wc_key": "", "cluster": "unknown",
 "submit_time": "2025-06-19T23:01:38", "eligible_time": "2025-06-19T23:01:38",
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc234def56789", "instance_type": "t4g.micro" }, { "name":
 "Hydra-MpiNG-abcdef01-2345-2", "instance_id": "i-0def345abc67890", "instance_type":
 "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 5, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "FAILED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:09:00", "end_time":
 "2025-06-19T23:09:00", "node_list": "(null)", "node_cnt": 0, "proc_cnt": 0,
 "work_dir": "/root", "reservation_name": "", "tres": { "cpu": 1, "mem": { "val":
 1, "unit": "G" }, "node": 1, "billing": 1 }, "account": "", "qos": "", "wc_key":
 "", "cluster": "unknown", "submit_time": "2025-06-19T23:09:00", "eligible_time":
 "2025-06-19T23:09:00", "derived_exit_code_status": 0, "derived_exit_code_signal": 0,
 "exit_code_status": 0, "exit_code_signal": 1, "node_details": [] } }
{ "jobcomp": { "job_id": 6, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "CANCELLED", "partition": "Hydra-MpiQueue-
abcdef01-7", "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:09:36",
 "end_time": "2025-06-19T23:09:36", "node_list": "(null)", "node_cnt": 0, "proc_cnt":
 0, "work_dir": "/root", "reservation_name": "", "tres": { "cpu": 1, "mem":
 { "val": 400, "unit": "M" }, "node": 1, "billing": 1 }, "account": "", "qos":
 "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:09:35",
 "eligible_time": "2025-06-19T23:09:36", "het_job_id": 6, "het_job_offset": 0,
```

```
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status": 0,
 "exit_code_signal": 1, "node_details": [] } }
{ "jobcomp": { "job_id": 7, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "CANCELLED", "partition": "Hydra-MpiQueue-
abcdef01-7", "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:10:03",
 "end_time": "2025-06-19T23:10:03", "node_list": "(null)", "node_cnt": 0, "proc_cnt":
 0, "work_dir": "/root", "reservation_name": "", "tres": { "cpu": 1, "mem":
 { "val": 400, "unit": "M" }, "node": 1, "billing": 1 }, "account": "", "qos":
 "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:10:03",
 "eligible_time": "2025-06-19T23:10:03", "het_job_id": 7, "het_job_offset": 0,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status": 0,
 "exit_code_signal": 1, "node_details": [] } }
{ "jobcomp": { "job_id": 8, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:11:24", "end_time":
 "2025-06-19T23:11:24", "node_list": "Hydra-MpiNG-abcdef01-2345-1", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 400, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:11:23",
 "eligible_time": "2025-06-19T23:11:23", "het_job_id": 8, "het_job_offset": 0,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc234def56789", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 9, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:11:24", "end_time":
 "2025-06-19T23:11:24", "node_list": "Hydra-MpiNG-abcdef01-2345-2", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 400, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:11:23",
 "eligible_time": "2025-06-19T23:11:23", "het_job_id": 8, "het_job_offset": 1,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-2",
 "instance_id": "i-0def345abc67890", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 10, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:12:24", "end_time":
 "2025-06-19T23:12:24", "node_list":"Hydra-MpiNG-abcdef01-2345-1", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 400, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:12:14",
 "eligible_time": "2025-06-19T23:12:14", "het_job_id": 10, "het_job_offset": 0,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
```

```
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc234def56789", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 11, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:12:24", "end_time":
 "2025-06-19T23:12:24", "node_list":"Hydra-MpiNG-abcdef01-2345-2", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 600, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:12:14",
 "eligible_time": "2025-06-19T23:12:14", "het_job_id": 10, "het_job_offset": 1,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-2",
 "instance_id": "i-0def345abc67890", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 13, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:47:57", "end_time":
 "2025-06-19T23:47:58", "node_list":"Hydra-MpiNG-abcdef01-2345-1", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 972, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:43:56",
 "eligible_time": "2025-06-19T23:43:56" , "array_job_id": 12, "array_task_id": 1,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc345def67890", "instance_type": "t4g.micro" } ] } }
{ "jobcomp": { "job_id": 12, "user": "root", "user_id": 0, "group": "root", "group_id":
 0, "name": "wrap", "job_state": "COMPLETED", "partition": "Hydra-MpiQueue-abcdef01-7",
 "time_limit": "UNLIMITED", "start_time": "2025-06-19T23:47:58", "end_time":
 "2025-06-19T23:47:58", "node_list":"Hydra-MpiNG-abcdef01-2345-1", "node_cnt":
 1, "proc_cnt": 1, "work_dir": "/root", "reservation_name": "", "tres": { "cpu":
 1, "mem": { "val": 972, "unit": "M" }, "node": 1, "billing": 1 }, "account": "",
 "qos": "", "wc_key": "", "cluster": "unknown", "submit_time": "2025-06-19T23:43:56",
 "eligible_time": "2025-06-19T23:43:56" , "array_job_id": 12, "array_task_id": 2,
 "derived_exit_code_status": 0, "derived_exit_code_signal": 0, "exit_code_status":
 0, "exit_code_signal": 0, "node_details": [ { "name": "Hydra-MpiNG-abcdef01-2345-1",
 "instance_id": "i-0abc345def67890", "instance_type": "t4g.micro" } ] } }
```


El planificador inicia sesión en AWS PCS

Puede configurar AWS PCS para que envíe datos de registro detallados desde el programador de clústeres a Amazon CloudWatch Logs, Amazon Simple Storage Service (Amazon S3) y Amazon Data Firehose. Esto puede ayudar con la supervisión y la solución de problemas.

Contenido

Requisitos previos

El director de IAM que administra el clúster de AWS PCS debe permitir la `pcs:AllowVendedLogDeliveryForResource` acción.

El siguiente ejemplo de política de IAM concede los permisos necesarios.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [
      {
         "Sid": "PcsAllowVendedLogsDelivery",
         "Effect": "Allow",
         "Action": ["pcs:AllowVendedLogDeliveryForResource"],
         "Resource": [
            "arn:aws:pcs:*::cluster/*"
         ]
      }
   ]
}
```

Configura los registros del programador

Puede configurar los registros del planificador para su clúster de AWS PCS con la Consola de administración de AWS tecla o. AWS CLI

Consola de administración de AWS

Para configurar los registros del programador con la consola

1. Abra la consola AWS PCS.
2. En el panel de navegación, seleccione Clusters (Clústeres).
3. Elija el clúster en el que desee añadir los registros del programador.
4. En la página de detalles del clúster, selecciona la pestaña Registros.
5. En Scheduler Logs, selecciona Añadir para añadir hasta 3 destinos de entrega de CloudWatch registros de entre Logs, Amazon S3 y Firehose.
6. Selecciona Actualizar entregas de registros.

AWS CLI

Para configurar los registros del programador con AWS CLI

1. Cree un destino de entrega de registros:

```
aws logs put-delivery-destination --region region \
  --name pcs-logs-destination \
  --delivery-destination-configuration \
  destinationResourceArn=resource-arn
```

Reemplace:

- *region*— El Región de AWS lugar en el que desea crear el destino, por ejemplo `us-east-1`
- *pcs-logs-destination*— Un nombre para el destino
- *resource-arn*— El nombre de recurso de Amazon (ARN) de un grupo de CloudWatch registros de Logs, un bucket de S3 o una transmisión de entrega de Firehose.

Para obtener más información, consulta PutDeliveryDestinationla referencia de la API CloudWatch de Amazon Logs.

2. Configure el clúster de PCS como fuente de entrega de registros:

```
aws logs put-delivery-source --region region \
  --name cluster-logs-source-name \
```

```
  --resource-arn cluster-arn \
  --log-type PCS_SCHEDULER_LOGS
```

Reemplace:

- *region*— El Región de AWS de su clúster, como `us-east-1`
- *cluster-logs-source-name*— Un nombre para la fuente
- *cluster-arn*— el ARN de su AWS clúster PCS

Para obtener más información, consulta PutDeliverySourcela referencia de la API CloudWatch de Amazon Logs.

3. Conecta la fuente de entrega al destino de entrega:

```
aws logs create-delivery --region region \
  --delivery-source-name cluster-logs-source \
  --delivery-destination-arn destination-arn
```

Reemplace:

- *region*— Los Región de AWS, como `us-east-1`
- *cluster-logs-source*— El nombre de su fuente de entrega
- *destination-arn*— El ARN de su destino de entrega

Para obtener más información, consulta CreateDeliveryla referencia de la API CloudWatch de Amazon Logs.

Rutas y nombres de las transmisiones de registros del programador

La ruta y el nombre de los registros del programador de AWS PCS dependen del tipo de destino.

- CloudWatch Registros
  - Un flujo CloudWatch de registros sigue esta convención de nomenclatura.

```
AWSLogs/PCS/${cluster_id}/${log_name}_${scheduler_major_version}.log
```

Example

```
AWSLogs/PCS/abcdef0123/slurmctld_24.05.log
```

- Bucket de S3
  - La ruta de salida de un bucket de S3 sigue esta convención de nomenclatura:

```
AWSLogs/${account-id}/PCS/${region}/${cluster_id}/${log_name}/
${scheduler_major_version}/yyyy/MM/dd/HH/
```

Example

```
AWSLogs/111111111111/PCS/us-east-2/abcdef0123/slurmctld/24.05/2024/09/01/00.
```

  - El nombre de un objeto de S3 sigue esta convención:

```
PCS_${log_name}_${scheduler_major_version}_#{expr date 'event_timestamp', format:
 "yyyy-MM-dd-HH"}_${cluster_id}_${hash}.log
```

Example

```
PCS_slurmctld_24.05_2024-09-01-00_abcdef0123_0123abcdef.log
```

Ejemplo de registro del programador

AWS Los registros del programador de PCS están estructurados. Incluyen campos como el identificador del clúster, el tipo de programador y las versiones principales y de parche, además del mensaje de registro emitido por el proceso del controlador Slurm. A continuación se muestra un ejemplo.

```
{
    "resource_id": "s3431v9rx2",
    "resource_type": "PCS_CLUSTER",
    "event_timestamp": 1721230979,
    "log_level": "info",
    "log_name": "slurmctld",
    "scheduler_type": "slurm",
    "scheduler_major_version": "25.05",
```

```
    "scheduler_patch_version": "3",
    "node_type": "controller_primary",
    "message": "[2024-07-17T15:42:58.614+00:00] Running as primary controller\n"
}
```

Servicio de monitorización de computación AWS paralela con Amazon CloudWatch

Amazon CloudWatch supervisa el estado y el rendimiento del clúster de AWS Parallel Computing Service (AWS PCS) mediante la recopilación de métricas del clúster a intervalos. Estas métricas se conservan, lo que le permite acceder a los datos históricos y obtener información sobre el rendimiento de su clúster a lo largo del tiempo.

CloudWatch también le permite monitorear las instancias EC2 lanzadas por AWS PCS para cumplir con sus requisitos de escalado. Si bien puede inspeccionar los registros de las instancias en ejecución, CloudWatch las métricas y los datos de registro normalmente se eliminan una vez que se cierran las instancias. Sin embargo, puede configurar el CloudWatch agente en las instancias mediante una plantilla de lanzamiento de EC2 para conservar las métricas y los registros incluso después de la finalización de la instancia, lo que permite la supervisión y el análisis a largo plazo.

Explore los temas de esta sección para obtener más información sobre la supervisión del uso CloudWatch de AWS PCS.

Temas

- Monitorear las métricas del AWS PCS mediante CloudWatch
- Supervisión de instancias de AWS PCS mediante Amazon CloudWatch

Monitorear las métricas del AWS PCS mediante CloudWatch

Puede supervisar el estado del clúster de AWS PCS con Amazon CloudWatch, que recopila datos de su clúster y los convierte en métricas prácticamente en tiempo real. Estas estadísticas se conservan durante un período de 15 meses, para que pueda acceder a la información histórica y obtener una mejor perspectiva del rendimiento de su clúster. Las métricas del clúster se envían CloudWatch en períodos de 1 minuto. Para obtener más información CloudWatch, consulta ¿Qué es Amazon CloudWatch? en la Guía del CloudWatch usuario de Amazon.

AWS PCS publica las siguientes métricas en el espacio de nombres AWS/PCS en. CloudWatch Tienen una sola dimensión,. `ClusterId`

Name	Description (Descripción)	Unidades
ActualCapacity	IdleCapacity + UtilizedCapacity	Recuento
CapacityUtilization	UtilizedCapacity / ActualCapacity	Recuento
DesiredCapacity	ActualCapacity + PendingCapacity	Recuento
IdleCapacity	Recuento de instancias que se están ejecutando pero que no están asignadas a trabajos	Recuento
UtilizedCapacity	Recuento de instancias que se están ejecutando y asignadas a trabajos	Recuento

Supervisión de instancias de AWS PCS mediante Amazon CloudWatch

AWS PCS lanza las instancias de Amazon EC2 según sea necesario para cumplir los requisitos de escalado definidos en los grupos de nodos de cómputo de PCS. Puedes monitorizar estas instancias mientras están en ejecución con Amazon CloudWatch. Puede inspeccionar los registros de las instancias en ejecución iniciando sesión en ellas y utilizando herramientas de línea de comandos interactivas. Sin embargo, de forma predeterminada, los datos de CloudWatch las métricas solo se conservan durante un período limitado una vez que se cierra una instancia y, por lo general, los registros de la instancia se eliminan junto con los volúmenes de EBS que respaldan la instancia. Para conservar las métricas o los datos de registro de las instancias lanzadas por PCS una vez finalizadas, puede configurar el CloudWatch agente de las instancias con una plantilla de lanzamiento de EC2. En este tema se proporciona información general sobre la supervisión de las instancias en ejecución y se proporcionan ejemplos de cómo configurar las métricas y los registros de las instancias persistentes.

Supervisión de instancias en ejecución

Búsqueda de instancias de AWS PCS

Para monitorear las instancias lanzadas por PCS, busque las instancias en ejecución asociadas a un clúster o grupo de nodos de cómputo. A continuación, en la consola EC2 de una instancia determinada, inspeccione las secciones de estado y alarmas y de supervisión. Si el acceso de inicio de sesión está configurado para esas instancias, puede conectarse a ellas e inspeccionar los distintos archivos de registro de las instancias. Para obtener más información sobre cómo identificar qué instancias administra PCS, consulteBúsqueda de instancias de grupos de nodos de cómputo en AWS PCS.

Habilitar métricas detalladas

De forma predeterminada, las métricas de las instancias se recopilan en intervalos de 5 minutos. Para recopilar métricas en intervalos de un minuto, habilita la CloudWatch supervisión detallada en la plantilla de lanzamiento de tu grupo de nodos de cómputo. Para obtener más información, consulte Active la CloudWatch supervisión detallada.

Configurar las métricas y los registros de las instancias persistentes

Puedes conservar las métricas y los registros de tus instancias instalando y configurando el CloudWatch agente de Amazon en ellas. Consta de tres pasos principales:

1. Cree una configuración de CloudWatch agente.
2. Guarde la configuración en un lugar donde las instancias de PCS puedan recuperarla.
3. Escriba una plantilla de lanzamiento de EC2 que instale el software del CloudWatch agente, busque la configuración e inicie el CloudWatch agente mediante la configuración.

Para obtener más información, consulte Recopilar métricas, registros y seguimientos con el CloudWatch agente en la Guía del CloudWatch usuario de Amazon yUso de plantillas de lanzamiento de Amazon EC2 con PCS AWS.

Cree una configuración CloudWatch de agente

Antes de implementar el CloudWatch agente en las instancias, debe generar un archivo de configuración JSON que especifique las métricas, los registros y los seguimientos que desea recopilar. Los archivos de configuración se pueden crear mediante un asistente o manualmente,

mediante un editor de texto. El archivo de configuración se creará manualmente para esta demostración.

En un equipo en el que tenga instalada la AWS CLI, cree un archivo de CloudWatch configuración denominado config.json con el contenido siguiente. También puede usar la siguiente URL para descargar una copia del archivo.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/cloudwatch/assets/config.json
```

Notas

- Las rutas de registro del archivo de ejemplo son para Amazon Linux 2. Si sus instancias utilizarán un sistema operativo base diferente, cambie las rutas según corresponda.
- Para capturar otros registros, añada entradas adicionales en`collect_list`.
- Los valores incluidos `{brackets}` son variables modeladas. Para ver la lista completa de variables admitidas, consulte Crear o editar manualmente el archivo de configuración del CloudWatch agente en la Guía del CloudWatch usuario de Amazon.
- Puede optar por omitir `logs` `metrics` o no recopilar estos tipos de información.

```
{
    "agent": {
        "metrics_collection_interval": 60
    },
    "logs": {
        "logs_collected": {
            "files": {
                "collect_list": [
                    {
                        "file_path": "/var/log/cloud-init.log",
                        "log_group_class": "STANDARD",
                        "log_group_name": "/PCSLogs/instances",
                        "log_stream_name": "{instance_id}.cloud-init.log",
                        "retention_in_days": 30
                    },
                    {
                        "file_path": "/var/log/cloud-init-output.log",
                        "log_group_class": "STANDARD",
                        "log_stream_name": "{instance_id}.cloud-init-output.log",
                        "log_group_name": "/PCSLogs/instances",
                        "retention_in_days": 30
```

```
                },
                {
                    "file_path": "/var/log/amazon/pcs/bootstrap.log",
                    "log_group_class": "STANDARD",
                    "log_stream_name": "{instance_id}.bootstrap.log",
                    "log_group_name": "/PCSLogs/instances",
                    "retention_in_days": 30
                },
                {
                    "file_path": "/var/log/slurmd.log",
                    "log_group_class": "STANDARD",
                    "log_stream_name": "{instance_id}.slurmd.log",
                    "log_group_name": "/PCSLogs/instances",
                    "retention_in_days": 30
                },
                {
                    "file_path": "/var/log/messages",
                    "log_group_class": "STANDARD",
                    "log_stream_name": "{instance_id}.messages",
                    "log_group_name": "/PCSLogs/instances",
                    "retention_in_days": 30
                },
                {
                    "file_path": "/var/log/secure",
                    "log_group_class": "STANDARD",
                    "log_stream_name": "{instance_id}.secure",
                    "log_group_name": "/PCSLogs/instances",
                    "retention_in_days": 30
                }
            ]
        }
    }
},
"metrics": {
    "aggregation_dimensions": [
        [
            "InstanceId"
        ]
    ],
    "append_dimensions": {
        "AutoScalingGroupName": "${aws:AutoScalingGroupName}",
        "ImageId": "${aws:ImageId}",
        "InstanceId": "${aws:InstanceId}",
        "InstanceType": "${aws:InstanceType}"
```

```
        },
        "metrics_collected": {
            "cpu": {
                "measurement": [
                    "cpu_usage_idle",
                    "cpu_usage_iowait",
                    "cpu_usage_user",
                    "cpu_usage_system"
                ],
                "metrics_collection_interval": 60,
                "resources": [
                    "*"
                ],
                "totalcpu": false
            },
            "disk": {
                "measurement": [
                    "used_percent",
                    "inodes_free"
                ],
                "metrics_collection_interval": 60,
                "resources": [
                    "*"
                ]
            },
            "diskio": {
                "measurement": [
                    "io_time"
                ],
                "metrics_collection_interval": 60,
                "resources": [
                    "*"
                ]
            },
            "mem": {
                "measurement": [
                    "mem_used_percent"
                ],
                "metrics_collection_interval": 60
            },
            "swap": {
                "measurement": [
                    "swap_used_percent"
                ],
```

```
                "metrics_collection_interval": 60
            }
        }
    }
}
```

Este archivo indica al CloudWatch agente que supervise varios archivos que pueden ser útiles para diagnosticar errores, como el arranque, la autenticación y el inicio de sesión, y otros ámbitos de solución de problemas. Entre ellos se incluyen:

- `/var/log/cloud-init.log`— Resultado de la fase inicial de configuración de la instancia
- `/var/log/cloud-init-output.log`— Resultado de los comandos que se ejecutan durante la configuración de la instancia
- `/var/log/amazon/pcs/bootstrap.log`— Resultado de las operaciones específicas de PCS que se ejecutan durante la configuración de la instancia
- `/var/log/slurmd.log`— Resultado del daemon slurmd del administrador de cargas de trabajo de Slurm
- `/var/log/messages`— Mensajes del sistema desde el núcleo, los servicios del sistema y las aplicaciones
- `/var/log/secure`— Registros relacionados con los intentos de autenticación, como SSH, sudo y otros eventos de seguridad

Los archivos de registro se envían a un grupo de CloudWatch registros denominado. `/PCSLogs/instances` Los flujos de registro son una combinación del ID de instancia y el nombre base del archivo de registro. El grupo de registros tiene un tiempo de retención de 30 días.

Además, el archivo indica al CloudWatch agente que recopile varias métricas comunes y las agregue por ID de instancia.

Guarde la configuración

El archivo de configuración del CloudWatch agente debe almacenarse en un lugar donde las instancias del nodo de cómputo de PCS puedan acceder a él. Existen dos formas comunes de hacerlo. Puede cargarlo en un bucket de Amazon S3 al que las instancias de su grupo de nodos de cómputo tendrán acceso a través de su perfil de instancia. También puede almacenarlo como un parámetro de SSM en el almacén de parámetros de Amazon Systems Manager.

Cárguelo a un bucket de S3

Para almacenar el archivo en S3, utilice los siguientes comandos de la AWS CLI. Antes de ejecutar el comando, realice los siguientes reemplazos:

- *amzn-s3-demo-bucket*Sustitúyalo por el nombre de su propio bucket de S3

En primer lugar (esto es opcional si tiene un depósito existente), cree un depósito para almacenar sus archivos de configuración.

```
aws s3 mb s3://amzn-s3-demo-bucket
```

A continuación, sube el archivo al depósito.

```
aws s3 cp ./config.json s3://amzn-s3-demo-bucket/
```

Almacénelo como un parámetro SSM

Para almacenar el archivo como un parámetro SSM, utilice el siguiente comando. Antes de ejecutar el comando, realice las siguientes sustituciones:

- *region-code*Sustitúyalo por la región de AWS en la que trabaja con AWS PCS.
- (Opcional) *AmazonCloudWatch-PCS* Sustitúyalo por su propio nombre para el parámetro. Ten en cuenta que si cambias el prefijo del nombre de, `AmazonCloudWatch-` tendrás que añadir específicamente el acceso de lectura al parámetro SSM en el perfil de instancia de tu grupo de nodos.

```
aws ssm put-parameter \
   --region region-code \
   --name "AmazonCloudWatch-PCS" \
   --type String \
   --value file://config.json
```

Escriba una plantilla de lanzamiento de EC2

Los detalles específicos de la plantilla de lanzamiento dependen de si el archivo de configuración está almacenado en S3 o SSM.

Utilice una configuración almacenada en S3

Este script instala el CloudWatch agente, importa un archivo de configuración de un bucket de S3 e inicia el CloudWatch agente con él. Sustituya los siguientes valores de este script por sus propios detalles:

- *amzn-s3-demo-bucket*— El nombre de un bucket de S3 desde el que puede leer su cuenta
- */config.json*— Ruta relativa a la raíz del bucket de S3 donde se almacena la configuración

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
- amazon-cloudwatch-agent

runcmd:
- aws s3 cp s3://amzn-s3-demo-bucket/config.json /etc/s3-cw-config.json
- /opt/aws/amazon-cloudwatch-agent/bin/amazon-cloudwatch-agent-ctl -a fetch-config -m
 ec2 -s -c file://etc/s3-cw-config.json

--==MYBOUNDARY==--
```

El perfil de instancia de IAM del grupo de nodos debe tener acceso al bucket. A continuación, se muestra un ejemplo de política de IAM para el depósito en el script de datos de usuario anterior.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource": [
```

```
                "arn:aws:s3:::amzn-s3-demo-bucket",
                "arn:aws:s3:::amzn-s3-demo-bucket/*"
            ]
        }
    ]
}
```

Tenga en cuenta también que las instancias deben permitir el tráfico saliente al S3 y CloudWatch a los puntos finales. Esto se puede lograr mediante grupos de seguridad o puntos finales de VPC, según la arquitectura del clúster.

Utilice una configuración almacenada en SSM

Este script instala el CloudWatch agente, importa un archivo de configuración desde un parámetro de SSM e inicia el CloudWatch agente con él. Sustituya los siguientes valores de este script por sus propios detalles:

- (Opcional) *AmazonCloudWatch-PCS* Sustitúyalo por su propio nombre para el parámetro.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
- amazon-cloudwatch-agent

runcmd:
- /opt/aws/amazon-cloudwatch-agent/bin/amazon-cloudwatch-agent-ctl -a fetch-config -m
 ec2 -s -c ssm:AmazonCloudWatch-PCS

--==MYBOUNDARY==--
```

La política de instancias de IAM para el grupo de nodos debe tener la CloudWatchAgentServerPolicyinformación adjunta.

Si el nombre de su parámetro no comienza por, `AmazonCloudWatch-` tendrá que añadir específicamente el acceso de lectura al parámetro SSM en el perfil de instancia de su grupo de

nodos. A continuación, se muestra un ejemplo de política de IAM que ilustra esto para el prefijo. *DOC-EXAMPLE-PREFIX*

JSON

```
{
  "Version":"2012-10-17",
  "Statement" : [
    {
      "Sid" : "CustomCwSsmMParamReadOnly",
      "Effect" : "Allow",
      "Action" : [
        "ssm:GetParameter"
      ],
      "Resource" : "arn:aws:ssm:*:*:parameter/DOC-EXAMPLE-PREFIX*"
    }
  ]
}
```

Tenga en cuenta también que las instancias deben permitir el tráfico saliente hacia el SSM y los puntos finales. CloudWatch Esto se puede lograr mediante grupos de seguridad o puntos finales de VPC, según la arquitectura del clúster.

Registro de llamadas a la API de AWS Parallel Computing Service mediante AWS CloudTrail

AWS PCS está integrado con AWS CloudTrail un servicio que proporciona un registro de las acciones realizadas por un usuario, un rol o un AWS servicio en AWS PCS. CloudTrail captura todas las llamadas a la API de AWS PCS como eventos. Las llamadas capturadas incluyen las llamadas desde la consola del AWS PCS y las llamadas en código a las operaciones de la API del AWS PCS. Si crea una ruta, puede habilitar la entrega continua de CloudTrail eventos a un bucket de Amazon S3, incluidos los eventos para AWS PCS. Si no configura una ruta, podrá ver los eventos más recientes en la CloudTrail consola, en el historial de eventos. Con la información recopilada por CloudTrail, puede determinar la solicitud que se realizó a AWS PCS, la dirección IP desde la que se realizó la solicitud, quién la hizo, cuándo se realizó y detalles adicionales.

Para obtener más información CloudTrail, consulte la Guía AWS CloudTrail del usuario.

AWS Información sobre el PCS en CloudTrail

CloudTrail está habilitada Cuenta de AWS cuando crea la cuenta. Cuando se produce una actividad en AWS PCS, esa actividad se registra en un CloudTrail evento junto con otros eventos de AWS servicio en el historial de eventos. Puede ver, buscar y descargar eventos recientes en su Cuenta de AWS. Para obtener más información, consulte Visualización de eventos con el historial de CloudTrail eventos.

Para tener un registro continuo de tus eventos Cuenta de AWS, incluidos los eventos de AWS PCS, crea una ruta. Un rastro permite CloudTrail entregar archivos de registro a un bucket de Amazon S3. De forma predeterminada, cuando se crea un registro de seguimiento en la consola, el registro de seguimiento se aplica a todas las Regiones de AWS. La ruta registra los eventos de todas las regiones de la AWS partición y envía los archivos de registro al bucket de Amazon S3 que especifique. Además, puede configurar otros AWS servicios para analizar más a fondo los datos de eventos recopilados en los CloudTrail registros y actuar en función de ellos. Para más información, consulte los siguientes temas:

- Introducción a la creación de registros de seguimiento
- CloudTrail servicios e integraciones compatibles
- Configuración de las notificaciones de Amazon SNS para CloudTrail
- Recibir archivos de CloudTrail registro de varias regiones y recibir archivos de CloudTrail registro de varias cuentas

Todas las acciones del AWS PCS se registran CloudTrail y se documentan en la referencia de la API del Servicio de Computación AWS Paralela. Por ejemplo, las llamadas a las `CreateComputeNodeGroup` `DeleteCluster` acciones y las llamadas generan entradas en los archivos de CloudTrail registro. `UpdateQueue`

Cada entrada de registro o evento contiene información sobre quién generó la solicitud. La información de identidad del usuario le ayuda a determinar lo siguiente:

- Si la solicitud se realizó con credenciales de usuario root o AWS Identity and Access Management (IAM).
- Si la solicitud se realizó con credenciales de seguridad temporales de un rol o fue un usuario federado.
- Si la solicitud la realizó otro AWS servicio.

Para obtener más información, consulte el elemento userIdentity de CloudTrail .

Descripción de las entradas de los archivos de CloudTrail registro del AWS PCS

Un registro es una configuración que permite la entrega de eventos como archivos de registro a un bucket de S3 que usted especifique. CloudTrail Los archivos de registro contienen una o más entradas de registro. Un evento representa una solicitud única de cualquier fuente e incluye información sobre la acción solicitada, la fecha y la hora de la acción, los parámetros de la solicitud, etc. CloudTrail Los archivos de registro no son un registro ordenado de las llamadas a la API pública, por lo que no aparecen en ningún orden específico.

En el siguiente ejemplo, se muestra una entrada de CloudTrail registro para una `CreateQueue` acción.


```
{
    "eventVersion": "1.09",
    "userIdentity": {
        "type": "AssumedRole",
        "principalId": "AIDACKCEVSQ6C2EXAMPLE:admin",
        "arn": "arn:aws:sts::012345678910:assumed-role/Admin/admin",
        "accountId": "012345678910",
        "accessKeyId": "ASIAY36PTPIEXAMPLE",
        "sessionContext": {
            "sessionIssuer": {
                "type": "Role",
                "principalId": "AROAY36PTPIEEXAMPLE",
                "arn": "arn:aws:iam::012345678910:role/Admin",
                "accountId": "012345678910",
                "userName": "Admin"
            },
            "attributes": {
                "creationDate": "2024-07-16T17:05:51Z",
                "mfaAuthenticated": "false"
            }
        }
    },
    "eventTime": "2024-07-16T17:13:09Z",
    "eventSource": "pcs.amazonaws.com",
    "eventName": "CreateQueue",
    "awsRegion": "us-east-1",
    "sourceIPAddress": "127.0.0.1",
```

```
    "userAgent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36
 (KHTML, like Gecko) Chrome/126.0.0.0 Safari/537.36",
    "requestParameters": {
        "clientToken": "c13b7baf-2894-42e8-acec-example",
        "clusterIdentifier": "abcdef0123",
        "computeNodeGroupConfigurations": [
            {
                "computeNodeGroupId": "abcdef0123"
            }
        ],
        "queueName": "all"
    },
    "responseElements": {
        "queue": {
            "arn": "arn:aws:pcs:us-east-1:609783872011:cluster/abcdef0123/queue/
abcdef0123",
            "clusterId": "abcdef0123",
            "computeNodeGroupConfigurations": [
                {
                    "computeNodeGroupId": "abcdef0123"
                }
            ],
            "createdAt": "2024-07-16T17:13:09.276069393Z",
            "id": "abcdef0123",
            "modifiedAt": "2024-07-16T17:13:09.276069393Z",
            "name": "all",
            "status": "CREATING"
        }
    },
    "requestID": "a9df46d7-3f6d-43a0-9e3f-example",
    "eventID": "7ab18f88-0040-47f5-8388-example",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "012345678910",
    "eventCategory": "Management",
    "tlsDetails": {
        "tlsVersion": "TLSv1.3",
        "cipherSuite": "TLS_AES_128_GCM_SHA256",
        "clientProvidedHostHeader": "pcs.us-east-1.amazonaws.com"
    },
    "sessionCredentialFromConsole": "true"
}
```

Puntos finales y cuotas de servicio para PCS AWS

En las siguientes secciones se describen los puntos finales y las cuotas de servicio de AWS Parallel Computing Service (AWS PCS). Las cuotas de servicio, anteriormente denominadas límites, son la cantidad máxima de recursos u operaciones de servicio para usted Cuenta de AWS.

Cuenta de AWS Tiene cuotas predeterminadas para cada AWS servicio. A menos que se indique lo contrario, cada cuota es específica de la región de . Puede solicitar el aumento de algunas cuotas, pero otras no se pueden aumentar.

Para obtener más información, consulte AWS service quotas en la Referencia general de AWS .

Contenido

Puntos de conexión de servicio

Nombre de la región	Región	Punto de conexión	Protocolo
Este de EE. UU. (Ohio)	us-east-2	pcs.us-east-2.amazonaws.com pcs-fips.us-east-2.amazonaws.com pcs-fips.us-east-2.api.aws pcs.us-east-2.api.aws	HTTPS
Este de EE. UU. (Norte de Virginia)	us-east-1	pcs.us-east-1.amazonaws.com	HTTPS

Nombre de la región	Región	Punto de conexión	Protocolo
		pcs-fips.us-east-1.amazonaws.com pcs-fips.us-east-1.api.aws pcs.us-east-1.api.aws	
Oeste de EE. UU. (Oregón)	us-west-2	pcs.us-west-2.amazonaws.com pcs-fips.us-west-2.amazonaws.com pcs-fips.us-west-2.api.aws pcs.us-west-2.api.aws	HTTPS
Asia-Pacífico (Singapur)	ap-southeast-1	pcs.ap-southeast-1.amazonaws.com pcs.ap-southeast-1.api.aws	HTTPS
Asia-Pacífico (Sídney)	ap-southeast-2	pcs.ap-southeast-2.amazonaws.com pcs.ap-southeast-2.api.aws	HTTPS
Asia-Pacífico (Tokio)	ap-northeast-1	pcs.ap-northeast-1.amazonaws.com pcs.ap-northeast-1.api.aws	HTTPS

Nombre de la región	Región	Punto de conexión	Protocolo
Europa (Fráncfort)	eu-central-1	pcs.eu-central-1.amazonaws.com pcs.eu-central-1.api.aws	HTTPS
Europa (Irlanda)	eu-west-1	pcs.eu-west-1.amazonaws.com pcs.eu-west-1.api.aws	HTTPS
Europa (Londres)	eu-west-2	pcs.eu-west-2.amazonaws.com pcs.eu-west-2.api.aws	HTTPS
Europa (Estocolmo)	eu-north-1	pcs.eu-north-1.amazonaws.com pcs.eu-north-1.api.aws	HTTPS
AWS GovCloud (Este de EE. UU.)	us-gov-east-1	piezas. us-gov-east-1.amazonaws.com consejos para PC. us-gov-east-1.amazonaws.com consejos para PC. us-gov-east-1.api.aws piezas. us-gov-east-1.api.aws	HTTPS

Nombre de la región	Región	Punto de conexión	Protocolo
AWS GovCloud (EE. UU.-Oeste)	us-gov-west-1	piezas. us-gov-west-1.amazonaws.com consejos para PC. us-gov-west-1.amazonaws.com consejos para PC. us-gov-west-1.api.aws piezas. us-gov-west-1.api.aws	HTTPS

Cuotas de servicio

Nombre	Predeterminado	Ajustable	Descripción
Clústeres	5	Sí	El número máximo de clústeres por. Región de AWS

> **Note**
>
> Los valores predeterminados son las cuotas iniciales establecidas por AWS. Estos valores predeterminados son independientes de los valores reales de la cuota aplicada y de las cuotas de servicio máximas posibles. Para obtener más información, consulte Terminología de Service Quotas en la Guía del usuario de Service Quotas.

Estas cuotas de servicio se enumeran en AWS Parallel Computing Service (PCS) en el Consola de administración de AWS. Para solicitar un aumento de cuota para los valores que se muestran como ajustables, consulte Solicitar un aumento de cuota en la Guía del usuario de Service Quotas.

> **Important**
>
> Recuerde comprobar la Región de AWS configuración actual en Consola de administración de AWS.

Cuotas internas

Las cuotas siguientes son ajustables.

Nombre	Predeterminado	Ajustable	Descripción
Creación simultánea de clústeres	1	No	El número máximo de clústeres en el estado `Creating` por Región de AWS.
Calcule los grupos de nodos por clúster	10	No	El número máximo de grupos de nodos de cómputo por clúster.
Colas por clúster	10	No	El número máximo de colas por clúster.

Cuotas relevantes para otros servicios AWS

AWS PCS utiliza otros AWS servicios. Sus cuotas de servicio para esos servicios afectan al uso que hace de AWS PCS.

Cuotas de servicio de Amazon EC2 que afectan a PCS AWS

- Solicitudes de instancia de spot
- Ejecución de instancias P bajo demanda
- Plantillas de inicialización
- Versiones de plantillas de lanzamiento
- Solicitudes de API de Amazon EC2

Para obtener más información, consulte las cuotas de servicio de Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud.

Solución de problemas en AWS Parallel Computing Service

Los siguientes temas proporcionan orientación para solucionar algunos problemas que pueden surgir en el AWS PCS.

Una instancia EC2 del AWS PCS se cierra y se reemplaza tras el reinicio

Descripción general del problema

Tras reiniciar una instancia EC2 de un grupo de nodos de procesamiento, AWS PCS finaliza y reemplaza automáticamente la instancia.

¿Por qué sucede esto

AWS PCS no admite el rearranque de instancias. Si se reinicia una instancia EC2, AWS PCS considera que la instancia está en mal estado y la reemplaza. Si AWS PCS termina y reemplaza las instancias de forma continua, puede deberse a que algo las reinicia después del lanzamiento. Algunos ejemplos incluyen el reinicio automático de la instancia EC2 (por ejemplo, un reinicio automático después de aplicar los parches), la automatización externa a la instancia EC2 (como una aplicación de administración de redes), otro AWS servicio (por ejemplo) o el reinicio manual AWS Systems Manager realizado por una persona.

Solución

Puede comprobar sus `slurmd` registros `slurmctld` o registros para comprobar si la instancia se ha reiniciado. Para obtener más información, consulte El planificador inicia sesión en AWS PCS y Supervisión de instancias de AWS PCS mediante Amazon CloudWatch. El siguiente ejemplo de entrada de `slurmctld` registro indica que la instancia se reinició:

Example

```
[2024-09-12T06:42:50.393+00:00] validate_node_specs: Node Login-1 unexpectedly rebooted
 boot_time=1726123354 last response=1726123285
```

Se está reiniciando debido a un parche

A menudo es necesario reiniciar el equipo después de aplicar los parches. No aplique los parches directamente a una instancia EC2 que forme parte de un grupo de nodos de cómputo de AWS PCS. Si debe aplicar parches a las instancias EC2, debe aplicar los parches a una Amazon Machine Image (AMI) actualizada y actualizar los grupos de nodos de cómputo para usar la AMI actualizada. Las nuevas instancias EC2 que AWS PCS lance para esos grupos de nodos de cómputo utilizarán la AMI actualizada (parcheada). Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Solucione los problemas de arranque y registro de los nodos de cómputo en PCS AWS

Si los nodos de cómputo no se inician o se registran correctamente en el clúster de AWS PCS, es posible que se presenten los siguientes síntomas:

- Los trabajos no comienzan
- No puedes conectarte a instancias en AWS Systems Manager
- Las instancias se cierran inesperadamente
- Las instancias se sustituyen continuamente

Estos errores pueden deberse a problemas durante el lanzamiento de la instancia EC2 o durante el proceso de arranque del nodo de cómputo de AWS PCS. En este tema se describen los procedimientos que le ayudarán a solucionar problemas durante el proceso de arranque del nodo AWS PCS. Para obtener más información sobre cómo solucionar problemas de lanzamiento de instancias EC2, consulte Solución de problemas de lanzamiento de instancias de Amazon EC2 en la Guía del usuario de Amazon Elastic Compute Cloud.

Los errores de Bootstrap se producen cuando una instancia EC2 se lanza correctamente, pero se produce un error durante el proceso de unión al clúster de PCS. AWS El proceso de arranque incluye dos fases principales:

- Registro de nodos: la instancia EC2 invoca la acción de la API de RegisterComputeNodeGroupInstance AWS PCS para registrarse en el servicio AWS PCS. Se pueden producir errores debido a los siguientes problemas:
  - Permisos
    - Perfil de instancia incorrecto
  - Red
    - No se puede conectar a los puntos finales de AWS PCS
    - Punto final de PCS mal configurado AWS
    - Instancia en una subred pública sin IP pública
    - Instancia de varias NIC en una subred pública
  - Secreto del clúster
    - El secreto del clúster se ha eliminado o marcado para su eliminación
- Integración con Slurm: la instancia se ejecuta `slurmd` y se une al clúster de Slurm. Se pueden producir errores debido a los siguientes problemas:
  - Permisos
    - Configuración del grupo de seguridad
    - Slurmctld no puede hacer ping al nodo de cómputo
  - Configuración de AMI personalizada
    - Faltan los controladores NVIDIA
    - ResumeTimeout alcanzado

Cómo funciona Slurm en PCS AWS

Podría ayudarlo a comparar la forma estándar en que Slurm funciona con la forma en que Slurm funciona en PCS. AWS

Procesamiento de trabajos estándar de Slurm

Los siguientes pasos se producen en el procesamiento de trabajos estándar de Slurm:

1. Al enviar un trabajo, lo `slurmctld` valida y lo pone en cola.
2. Cuando los recursos estén disponibles, `slurmctld` asigna los nodos existentes.
3. `slurmd`los daemons ejecutan tareas en los nodos asignados.

Procesamiento de tareas de Slurm en PCS AWS

Los siguientes pasos se producen en el procesamiento de trabajos de AWS PCS:

1. Al enviar un trabajo, lo `slurmctld` valida y lo pone en cola.
2. Cuando se necesita capacidad adicional, AWS PCS utiliza la plantilla de lanzamiento del grupo de nodos de cómputo para lanzar nuevas instancias de EC2.
3. Las nuevas instancias se incorporan al clúster:
   a. Las instancias se registran en AWS PCS.
   b. Las instancias se unen al clúster de Slurm.
4. Cuando los recursos están listos, `slurmctld` asigna los nodos (incluidos los que se han iniciado recientemente).
5. `slurmd`los daemons ejecutan tareas en los nodos asignados.

Recupera los registros de instancias

El primer paso para solucionar los problemas de arranque de los nodos de cómputo es recuperar los registros de las instancias. Puede usar uno de los métodos siguientes:

AWS CLI

Recupera la salida de la consola desde el nodo de cómputo mediante el siguiente comando:

```
aws ec2 get-console-output --region us-east-1 --instance-id i-1234567890abcdef0 --
output text
```

*us-east-1*Sustitúyala por tu AWS región y *i-1234567890abcdef0* por tu ID de instancia.

AWS Systems Manager

Si puedes conectarte a la instancia mediante Systems Manager, puedes ver el archivo de registro de arranque directamente:

1. Conéctese a la instancia mediante Systems Manager. Para obtener más información, consulte Iniciar una sesión en la Guía del usuario de Systems Manager.
2. Consulte el archivo de registro de arranque:

```
sudo cat /var/log/amazon/pcs/bootstrap.log
```

> **Note**
>
> Si se produce un problema durante la fase de inicialización, es posible que tengas que esperar unos 20 minutos antes de poder conectarte a la instancia. Los servicios Systems Manager y SSH se inician solo después de que se complete la inicialización o cuando la ejecución del bootstrap agote el tiempo de espera en caso de fallo.

Recupera VPC/Subnet/Security grupos de un ID de instancia

Para solucionar problemas con los nodos de procesamiento, es posible que tengas que recuperar información sobre la VPC, la subred y los grupos de seguridad asociados a tus instancias. Si no conoces tu instancia IDs, consulta. Búsqueda de instancias de grupos de nodos de cómputo en AWS PCS

Consola de administración de AWS

Para obtener grupos de VPC, subred y seguridad

1. Abra la consola de Amazon EC2.
2. Elija Instances.
3. En la tabla de instancias, elija el ID de la instancia.
4. Busca el ID de VPC y el ID de subred en el resumen de la instancia que se muestra.
5. En el resumen de la instancia, selecciona la pestaña Seguridad.
6. Busca los grupos de seguridad en la pestaña Seguridad.

AWS CLI

Usa el siguiente comando para recuperar la información de la VPC, la subred y el grupo de seguridad de la instancia:

```
aws ec2 describe-instances --instance-ids i-1234567890abcdef0 --query
 'Reservations[*].Instances[*].
{InstanceId:InstanceId,VpcId:VpcId,SubnetId:SubnetId,SecurityGroups:SecurityGroups[*].GroupI
 --output table
```

Problemas de registro de nodos

El registro de nodos es la primera acción que ejecuta un nodo de cómputo durante el arranque. El nodo llama al punto final de la API de AWS PCS para registrarse en el AWS PCS. Los errores de registro suelen mostrar mensajes de error similares a los siguientes:

```
<13>Nov  5 08:10:27 user-data: Recipe: aws-pcs-environment::node_registration
<13>Nov  5 08:10:27 user-data:   * ruby_block[Register NodeGroup Instance] action
 run[2024-11-05T08:10:27+00:00] INFO: Processing ruby_block[Register NodeGroup
 Instance] action run (aws-pcs-environment::node_registration line 19)
<13>Nov  5 08:15:46 user-data:
<13>Nov  5 08:15:46 user-data:
<13>Nov  5 08:15:46 user-data:
 ================================================================================
<13>Nov  5 08:15:46 user-data:     Error executing action `run` on resource
 'ruby_block[Register NodeGroup Instance]'
<13>Nov  5 08:15:46 user-data:
 ================================================================================
<13>Nov  5 08:15:46 user-data:
<13>Nov  5 08:15:46 user-data:     EOFError
```

Perfil de instancia incorrecto

Si la instancia no se puede registrar, compruebe que el perfil de instancia asociado al nodo de cómputo tenga el `pcs:RegisterComputeNodeGroupInstance` permiso.

Para obtener más información sobre cómo crear un perfil de instancia válido, consultaCrear un perfil de instancia para AWS PCS.

No se puede conectar a los puntos finales de AWS PCS

Si sus nodos de cómputo están en una subred privada, asegúrese de haber configurado puntos de enlace de VPC AWS para PCS o de que su subred tenga una ruta a una puerta de enlace NAT para el acceso a Internet. Para obtener más información, consulte los siguientes temas:

- Acceda a un AWS servicio mediante un punto final de VPC de interfaz en la guía Amazon Virtual Private Cloud AWS PrivateLink.
- Puntos finales y cuotas de servicio para PCS AWS.
- Conecta tu VPC a otras redes en la Guía del usuario de Amazon Virtual Private Cloud
- AWS Redes PCS

Punto final de PCS mal configurado AWS

Si aparece un mensaje de error similar al siguiente, compruebe la política asociada a su punto final de AWS VPC:

```
com.amazon.coral.security.AccessDeniedException: User: arn:aws:sts::xxx:assumed-
role/rolename/i-instanceid is not authorized to perform:
 pcs:RegisterComputeNodeGroupInstance on resource: arn:aws:pcs:us-west-2:xxx:cluster/
cluster-id as either the resource does not exist, some policy explicitly denies access,
 or no policy grants access
```

Para obtener más información sobre cómo configurar los puntos finales de la interfaz de VPC para AWS PCS, consulte. Acceda AWS Parallel Computing Service mediante un punto final de interfaz ()AWS PrivateLink

Instancia en una subred pública sin IP pública

Si la subred no tiene habilitada la asignación automática de IP pública y la configuración de la ruta usa una puerta de enlace a Internet, las instancias no se pueden comunicar con la API de AWS PCS.

Las instancias de una subred con una puerta de enlace a Internet deben tener una dirección IP pública. Para resolver este problema, elige una de las siguientes opciones:

- Agregue un punto de enlace de VPC para AWS PCS a la VPC de su clúster. Esto permite que las instancias se comuniquen con el AWS PCS sin necesidad de que una dirección IP pública pase por la puerta de enlace de Internet.
- Utilice una subred privada con una puerta de enlace NAT, de modo que no sea necesaria una dirección IP pública.
- Habilite la asignación automática de direcciones IP públicas a través de su subred o plantilla de lanzamiento para que las instancias puedan contactar con la API a través de la puerta de enlace de Internet. Ten en cuenta que esta opción no es válida para instancias de interfaz de varias redes.

Instancia de varias NIC en una subred pública

Debe usar una subred privada si usa un tipo de instancia que tiene varias interfaces de red (). NICs

AWS Las direcciones IP públicas solo se pueden asignar a instancias lanzadas con una única interfaz de red. Para obtener más información sobre las direcciones IP, consulte Asignar una IPv4

dirección pública durante el lanzamiento de una instancia en la Guía del usuario de Amazon EC2 para instancias de Linux.

Los tipos de instancias con varias NIC requieren una puerta de enlace NAT o un proxy interno en la subred para acceder al punto final del AWS PCS. Como alternativa, puede añadir un punto de enlace de VPC para AWS PCS a la VPC de su clúster.

El secreto del clúster se ha eliminado o marcado para su eliminación

Si el secreto compartido de Slurm en AWS Secrets Manager se ha eliminado o marcado para su eliminación, los nodos de procesamiento no se registrarán y el clúster se verá afectado.

AWS PCS crea automáticamente un secreto compartido de Slurm en AWS Secrets Manager (con el formato de nombre:`pcs!slurm-secret-<cluster-id>`) al crear un clúster. Este secreto es necesario para garantizar la seguridad de las comunicaciones en el clúster. Para obtener más información, consulte Trabajar con secretos de clústeres en AWS PCS.

Si este secreto se elimina o se marca para eliminarlo, los nodos nuevos no podrán unirse al clúster y es posible que el controlador u otros demonios del clúster (como `slurmd` y`slurmdbd`) no puedan volver a unirse al clúster si se reinician.

Para resolver este problema, puedes restaurar el secreto eliminado si aún se encuentra dentro del período de recuperación. Para obtener instrucciones detalladas, consulte Restaurar un secreto de AWS Secrets Manager.

Si el período de recuperación caduca, el secreto no se puede restaurar ni el clúster de AWS PCS afectado. Debe crear un clúster nuevo con la misma configuración. AWS PCS crea automáticamente un nuevo secreto de programador.

Problemas de unión al clúster de Slurm

Tras el registro correcto del nodo, el nodo de cómputo intenta unirse al clúster de Slurm. El `slurmd` daemon del nodo contacta con el controlador Slurm para registrarse en el clúster. Los errores de unión a Slurm suelen mostrar mensajes de error similares a los siguientes:

```
<13>Nov  5 17:20:29 user-data: [2024-11-05T17:20:28+00:00] FATAL:
 Mixlib::ShellOut::ShellCommandFailed: service[slurmd] (aws-pcs-slurm::finalize_slurm
 line 18) had an error: Mixlib::ShellOut::ShellCommandFailed: Expected process to exit
 with [0], but received '1'
```

```
<13>Nov  5 17:20:29 user-data: ---- Begin output of ["/usr/bin/systemctl", "--system",
 "start", "slurmd"] ----
<13>Nov  5 17:20:29 user-data: STDOUT:
<13>Nov  5 17:20:29 user-data: STDERR: Job for slurmd.service failed because the
 control process exited with error code. See "systemctl status slurmd.service" and
 "journalctl -xe" for details.
<13>Nov  5 17:20:29 user-data: ---- End output of ["/usr/bin/systemctl", "--system",
 "start", "slurmd"] ----
```

Configuración del grupo de seguridad

Compruebe que sus grupos de seguridad estén configurados correctamente para permitir la comunicación entre los nodos de procesamiento y el controlador Slurm. Los grupos de seguridad deben permitir el siguiente tráfico:

- Puerto 6817 para `slurmd` comunicarse con `slurmctld`
- Puerto 6818 para hacer ping `slurmctld` `slurmd`

Para obtener más información sobre los requisitos de los grupos de seguridad, consulte los temas siguientes:

- Crear grupos de seguridad para AWS PCS
- Cree plantillas de lanzamiento para AWS PCS
- Requisitos y consideraciones sobre los grupos de seguridad

> **Important**
>
> El grupo de seguridad de clúster que asoció al clúster durante la creación del clúster también debe configurarse en los grupos de seguridad del grupo de nodos de procesamiento para permitir que los nodos de procesamiento se comuniquen con el controlador.

Faltan los controladores NVIDIA

Si la instancia se inicia correctamente, pero los trabajos no se inician y ves mensajes de error similares a los siguientes en los registros de la instancia, es posible que te falten los controladores de NVIDIA:

```
<13>Dec  2 13:52:00 user-data: [2024-12-02T13:52:00.094+00:00] - /opt/aws/pcs/bin/
pcs_bootstrap_config_always.sh: INFO: nvidia-smi not found!
...
<13>Dec  2 13:54:10 user-data: Job for slurmd.service failed because the control
 process exited with error code. See "systemctl status slurmd.service" and "journalctl
 -xe" for details.
<13>Dec  2 13:54:12 user-data: [2024-12-02T13:54:12.718+00:00] - /opt/aws/pcs/bin/
pcs_bootstrap_finalize.sh: INFO: systemctl could not start slurmd!
```

Si te conectas a la instancia y compruebas el estado del `slurmd` daemon, es posible que aparezca un error similar al siguiente:

```
$ systemctl status slurmd
...
fatal: can't stat gres.conf file /dev/nvidia0: No such file or directory
```

Para resolver este problema, instale los controladores NVIDIA en la AMI personalizada. Para obtener más información, consulte Paso 4: (opcional) Instalar controladores, bibliotecas y software de aplicación adicionales.

ResumeTimeout alcanzado

Si un nodo de procesamiento y su instancia EC2 se cierran porque el nodo está en mal estado, es posible que el AWS PCS no admita la AMI o que haya problemas de red. La instancia EC2 se ejecuta durante aproximadamente 30 minutos hasta que se llega a la de Slurm y ResumeTimeout se marca el nodo como. `DOWN`

Si la instancia no se inicia correctamente y no está registrada en AWS PCS (no `RegisterComputeNodeGroupInstance` se requiere la instancia EC2), compruebe los registros de la instancia para ver si hay mensajes de error similares a los siguientes:

```
/opt/aws/pcs/bin/pcs_bootstrap_init.sh: No such file or directory
```

Este error indica que el software de arranque del AWS PCS no forma parte de la AMI. Para resolver este problema, asegúrese de que la AMI personalizada incluya el software de arranque de AWS PCS. Para obtener más información, consulte Imágenes personalizadas de Amazon Machine (AMIs) para AWS PCS.

Slurmctld no puede hacer ping al nodo de cómputo

Si la instancia ejecuta correctamente el procedimiento de arranque y está registrada en AWS PCS, pero `slurmctld` no puede verla ni enviarle trabajos, la instancia se configura después de un tiempo y, `DOWN` después, se cierra.

Esto puede deberse a una mala configuración de los grupos de seguridad. Por ejemplo, si el puerto 6817 está habilitado `slurmd` para permitir la comunicación con él`slurmctld`, pero falta el puerto 6818 para permitir `slurmctld` el ping. `slurmd`

Compruebe que sus grupos de seguridad incluyen todas las reglas obligatorias, tal como se indica en. Requisitos y consideraciones sobre los grupos de seguridad

Historial de documentos de la Guía del usuario de AWS PCS

La siguiente tabla describe los cambios importantes en la documentación del AWS PCS.

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
10 de marzo de 2026	Agente de PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.3.2-1. Se ha corregido un problema que afectaba al arranque de los nodos de cómputo de RHEL 8.10 y Rocky Linux 8.10. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y AWS Versiones del agente PCS.	N/A
11 de febrero de 2026	AWS PCS lanzado en Asia Pacífico (Mumbai) y Europa (París)	AWS PCS ya está disponible en Asia Pacífico (Bombay) (ap-south-1) y Europa (París) (eu-west-3). CloudFormation Las plantillas están disponibles para empezar en Asia Pacífico (Bombay) Región de AWS y Europa (París). Región de AWS Para obtener más	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
		información, consulte Se utiliza CloudFormation para crear un clúster de AWS PCS de muestra y CloudFormation plantillas para crear un clúster de AWS PCS de muestra.	
18 de noviembre de 2025	Nueva función: API REST de Slurm	La API REST de Slurm ahora es compatible con Slurm 25.05 o versiones posteriores. Para obtener más información, consulte API REST de Slurm en PCS AWS.	SDK DE AWS: 18 DE NOVIEMBRE DE 2021

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
10 de noviembre de 2025	Nueva función: compatibilidad con el complemento de filtro CLI de Slurm	AWS PCS ahora admite los complementos de filtro CLI de Slurm para ejecutar scripts de Lua personalizados que validan y modifican los parámetros de envío de trabajos antes de que lleguen al controlador de Slurm. Utilice filtros CLI para aplicar políticas personalizadas, establecer parámetros predeterminados y proporcionar orientación al usuario durante el envío de trabajos. Esta función requiere la versión 25.05 o posterior de Slurm. Para obtener más información, consulte Utilice los complementos de filtro CLI de Slurm para personalizar el envío de trabajos en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
7 de noviembre de 2025	Agente de PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.3.1-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y AWS Versiones del agente PCS.	N/A
3 de noviembre de 2025	Se actualizaron el agente de PCS y los instaladores de Slurm	Se actualizó el tema de AMI para el agente AWS PCS 1.3.0-1 y los instaladores de Slurm 24.11.6-2, 24.05.8-2 y 23.11.10-4. Lista actualizada de sistemas operativos compatibles. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y AWS Versiones del agente PCS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
23 de octubre de 2025	Contenido actualizado: pcs-multi-cluster-login - configure.sh	Se corrigieron algunos errores en el script de configuración del nodo de inicio de sesión multiclús ter. Para obtener más información, consulte AWS Código de script de configuración del nodo de inicio de sesión multiclús ter PCS.	N/A
21 de octubre de 2025	Nueva función: rotación secreta de clústeres	AWS El PCS ahora admite la rotación de secretos de los clústeres para mejorar la seguridad . Para obtener más información, consulte Secretos de clústeres rotativos en AWS PCS. Se han actualizado los permisos mínimos de administrador para permitir la rotación de secretos de los clústeres . Para obtener más información, consulte Permisos mínimos para PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
17 de octubre de 2025	Tema nuevo: script de configuración de nodos de inicio de sesión multiclúster	Se agregó un tema nuevo que proporciona un script para configurar un nodo de inicio de sesión independiente para conectarse a varios clústeres de AWS PCS. El script automatiza la configuración de varios demonios de Slurm y crea scripts de `sackd` activación para la interacción entre clústeres. Para obtener más información, consulte Conexión de un nodo de inicio de sesión independiente a varios clústeres en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
16 de octubre de 2025	Actualizado para Slurm 25.05	Se actualizó la guía de usuario para la compatibilidad con Slurm 25.05. Slurm 25.05 es ahora la versión por defecto. Para obtener más información, consulte los siguientes temas: • Versiones de Slurm en PCS AWS • Instaladores de software para crear PCS personalizados AMIs AWS • Notas de publicación para un ejemplo de AWS PCS AMIs	N/A
16 de octubre de 2025	Agente de PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.2.2-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y AWS Versiones del agente PCS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
2 de octubre de 2025	Nuevas funciones: reinicio del nodo de Slurm, actualizaciones del clúster y configuración personalizada de Slurm	AWS El PCS añade compatibilidad con varias funciones nuevas: • Reinicio del nodo Slurm: utilice el `scontrol reboot` comando nativo de Slurm para reiniciar los nodos de cómputo sin reemplazar la instancia . Para obtener más información, consulte Reiniciar los nodos de cómputo con Slurm en PCS AWS. • Actualizaciones de clústeres: modifique las configuraciones de los clústeres después de la creación sin necesidad de recompilarlos. Para obtener más informaci ón, consulte Actualiza ción de un clúster en AWS PCS. • Configuración personali zada de Slurm: configure los parámetro s avanzados de Slurm en los recursos del clúster, la cola y el grupo de nodos de	01/10/2025

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
		cómputo. Para obtener más información, consulte Configuración de ajustes de Slurm personalizados en PCS AWS.	
23 de septiembre de 2025	Nuevo tema de solución de problemas: problemas de arranque de nodos de cómputo	Se agregó una guía de solución de problemas para diagnosticar y resolver los problemas de arranque de los nodos de cómputo. Para obtener más información, consulte Solucione los problemas de arranque y registro de los nodos de cómputo en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
17 de septiembre de 2025	Nueva función: bloques de capacidad para aprendizaje automático	AWS PCS ahora es compatible con los bloques de capacidad de Amazon EC2 para ML, que le permiten reservar instancias de computación acelerada basadas en GPU para sus clústeres. Para obtener más información, consulte Uso de bloques de capacidad de Amazon EC2 para aprendizaje automático con PCS AWS. Los permisos mínimos para admitir los bloques de capacidad ahora forman parte de los permisos mínimos para un administrador de servicios. Para obtener más información, consulte Permisos mínimos para PCS AWS.	17-09-2025

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
11 de septiembre de 2025	Actualización de la política gestionada por AWS	AWS PCS lo actualizó AWSPCSService RolePolicy para admitir los bloques de capacidad . Para obtener más información, consulte AWS políticas administradas para AWS Parallel Computing Service.	N/A
14 de agosto de 2025	Documentación actualizada del perfil de instancia	Se mejoró la documentación del perfil de instancia con instrucciones CLI completas para crear roles de IAM y perfiles de instancia. Se agregaron step-by-step procedimientos para configurar los perfiles de instancia mediante el PCS AWS CLI y se mejoró la guía para encontrar los perfiles de instancia que se utilizan con el AWS PCS. Para obtener más información, consulte Perfiles de instancia de IAM para AWS Parallel Computing Service.	14 de agosto de 2025

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
1 de agosto de 2025	Tema nuevo: los complementos de SPANK	Se ha añadido documentación sobre los complementos SPANK (Slurm Plug-in Architecture for Node and job Kontrol) que puede utilizar para ampliar y modificar el comportamiento de Slurm durante el lanzamiento y la ejecución de tareas en clústeres de PCS. AWS Para obtener más información, consulte Amplíe la funcionalidad de Slurm en los PCS con los complementos de AWS SPANK .	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
1 de agosto de 2025	IPv6 soporte de redes	Se agregó soporte para IPv6 redes al crear clústeres de AWS PCS. Ahora puede elegir IPv6 el tipo de red para su clúster, con las actualizaciones correspondientes a los requisitos de la VPC, la configuración de la subred, la configuración del grupo de seguridad y los procedimientos de creación del clúster. Para obtener más información, consulte AWS Requisitos y consideraciones sobre la VPC y la subred y Creación de un clúster en AWS PCS.	2025-08-01

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
3 de julio de 2025	AWS PCS lanzado en Europa (Londres)	AWS PCS ya está disponible en Europa (Londres) (eu-west-2). CloudFormation Hay plantillas disponibles para empezar en Europa (Londres). Región de AWS Para obtener más información, consulte Se utiliza CloudFormation para crear un clúster de AWS PCS de muestra y CloudFormation plantillas para crear un clúster de AWS PCS de muestra.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
1 de julio de 2025	Instrucciones de consola actualizadas	Ahora puede hacer que AWS PCS cree un perfil de instancia básico y un grupo de seguridad para usted al crear un clúster y un grupo de nodos de cómputo en la consola. Para obtener más información, consulte lo siguiente: • Creación de un clúster en AWS PCS • Creación de un grupo de nodos de cómputo en AWS PCS • Perfiles de instancia de IAM para AWS Parallel Computing Service	N/A
23 de junio de 2025	Nueva política gestionada: AWSPCSComputeNodePolicy	Se agregó una nueva política administrada que otorga permiso a los nodos de cómputo de AWS PCS para conectarse a los clústeres de AWS PCS. Para obtener más información, consulte AWS política gestionada: AWSPCSComputeNodePolicy .	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
19 de junio de 2025	Tema nuevo: registros de finalización de trabajos	Utilice los registros de finalización de tareas para registrar los detalles sobre las tareas una vez finalizadas, sin coste adicional. Para obtener más informaci ón, consulte Registros de finalización de trabajos en AWS PCS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
18 de junio de 2025	AWS PCS lanzado en AWS GovCloud (US)	AWS El PCS ahora está disponible en AWS GovCloud (EE. UU. Este) (us-gov-east-1) y AWS GovCloud (EE. UU. Oeste) (us-gov-west-1). CloudFormation hay plantillas disponibles para empezar en. AWS GovCloud (US) Regions Para obtener más información, consulte Se utiliza CloudFormation para crear un clúster de AWS PCS de muestra y CloudFormation plantillas para crear un clúster de AWS PCS de muestra. Para obtener más información sobre los puntos finales del servicio AWS PCS en AWS GovCloud (US) Regions, consultePuntos finales y cuotas de servicio para PCS AWS. Para obtener más información sobre las diferencias entre AWS GovCloud (US) Regions, consulte AWS PCS in	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
		AWS GovCloud (US) en la Guía del AWS GovCloud (US) usuario.	
18 de junio de 2025	Agente PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.2.1-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A
15 de mayo de 2025	Nueva función: contabilidad	La contabilidad de Slurm ahora es compatible con Slurm 24.11 o versiones posteriores. Para obtener más información, consulte Contabilidad de Slurm en PCS AWS.	SDK DE AWS: 15 DE MAYO DE 2020

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
15 de mayo de 2025	Actualizado para Slurm 24.11	Se actualizó la guía del usuario para la compatibilidad con Slurm 24.11.5. Para obtener más información, consulte los siguientes temas: • Versiones de Slurm en PCS AWS • Instaladores de software para crear PCS personalizados AMIs AWS • Notas de publicación para un ejemplo de AWS PCS AMIs	N/A
5 de mayo de 2025	Preguntas frecuentes sobre las versiones actualizadas de Slurm	Se actualizaron las preguntas frecuentes (FAQ) de las versiones de Slurm sobre las versiones de Slurm que se acercan o superan el final de su vida útil (EOL). Para obtener más información, consulte Preguntas frecuentes sobre las versiones de Slurm en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
17 de abril de 2025	Tema nuevo: cómo obtener los detalles de un grupo de nodos de cómputo	Aprenda a obtener detalles de un grupo de nodos de cómputo de AWS PCS, como su ID, ARN e ID de AMI. Para obtener más información, consulte Obtenga detalles del grupo de nodos de cómputo en AWS PCS.	N/A
2 de abril de 2025	Instalador de Slurm actualizado	Se actualizó el tema de AMI para el instalador Slurm 24.05.7-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A
28 de marzo de 2025	Se agregaron cuotas para el número máximo de colas y grupos de nodos de cómputo	Se agregaron cuotas internas no ajustables para el número máximo de grupos de nodos de cómputo por clúster y el número máximo de colas por clúster. Para obtener más información, consulte Cuotas internas.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
14 de marzo de 2025	Se ha modificado una clave de propiedad en la plantilla CloudFormation	`Id` ahora es `TemplateId` para la `CustomLaunchTemplate` propiedad de la CloudFormation plantilla . Para obtener más información, consulta Recursos en Partes de una CloudFormation plantilla para AWS PCS.	N/A
13 de marzo de 2025	Se agregó información sobre la versión del agente AWS PCS y Slurm	Se agregó un tema nuevo que describe los cambios para cada versión del agente AWS PCS. Para obtener más información, consulte AWS Versiones del agente PCS. Se agregó más información al tema sobre las versiones de Slurm, en el que se describen las fechas de soporte importantes y las notas de lanzamiento detalladas sobre el soporte de AWS PCS para Slurm. Para obtener más información, consulte Versiones de Slurm en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
7 de marzo de 2025	Agente de PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.2.0-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A
3 de febrero de 2025	Se agregó un tema sobre el uso con PCS AWS CloudFormation AWS	Se agregó un tema a la guía del usuario que proporciona un ejemplo de cómo usarlo CloudFormation con AWS PCS. En este tema se proporciona un procedimiento para utilizar una CloudFormation plantilla de ejemplo para crear el clúster de AWS PCS de muestra y se describen brevemente las secciones de esa plantilla. Para obtener más información, consulte Comience con un CloudFormation AWS PCS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
18 de diciembre de 2024	Actualizado para Slurm 24.05	Se actualizó la guía del usuario para la compatibilidad con Slurm 24.05. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y Notas de publicación para un ejemplo de AWS PCS AMIs.	N/A
18 de diciembre de 2024	Ejemplo de versiones actualizadas de NVIDIA para Slurm 23.11 AMIs	Se actualizaron las versiones CUDA y del controlador NVIDIA en la muestra de Slurm 23.11. AMIs Para obtener más información, consulte Notas de publicación para un ejemplo de AWS PCS AMIs.	N/A
17 de diciembre de 2024	Instalador de Slurm actualizado	Se actualizó el tema de AMI para el instalador de Slurm 23.11.10-3. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
13 de diciembre de 2024	Agente PCS actualizado	Se actualizó el tema de AMI para el agente AWS PCS 1.1.1-1. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A
6 de diciembre de 2024	Se actualizaron el agente PCS y el instalador de Slurm	Se actualizó el tema de AMI para el agente AWS PCS 1.1.0-1 y el instalador de Slurm 23.11.10-2. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS.	N/A
6 de diciembre de 2024	Se agregó un tema sobre la compatibilidad con el sistema operativo	Para obtener más información, consulte Sistemas operativos compatibles en AWS PCS.	N/A
8 de noviembre de 2024	Guía de usuario reorganizada	Reorganizamos la guía del usuario para llevar los temas al nivel más alto, trasladamos algunos temas a sus propias páginas y agrupamos temas similares.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
7 de noviembre de 2024	Temas de AMI actualizados	Se actualizó el tema de AMI para Slurm 23.11.10 y libjwt 17.0. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y Paso 3: Instalar Slurm. Se han simplificado y corregido las notas de la versión de. AMIs Para obtener más información, consulte Notas de publicación para un ejemplo de AWS PCS AMIs.	N/A
7 de noviembre de 2024	Se agregó un tema nuevo sobre el uso de volúmenes de EBS cifrados con AWS PCS	Se agregó un tema que describe la política de claves de KMS requerida para los volúmenes de EBS cifrados en AWS PCS. Para obtener más información, consulte Política de claves de KMS requerida para su uso con volúmenes de EBS cifrados en PCS AWS.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
18 de octubre de 2024	AWS Lanzamiento del agente PCS 1.0.1-1	Se actualizó la documentación relacionada con la AMI para hacer referencia a la versión AWS 1.0.1-1 del agente PCS. Para obtener más información, consulte Instaladores de software para crear PCS personalizados AMIs AWS y Paso 2: Instalar el agente AWS PCS.	N/A
10 de octubre de 2024	Se agregó un capítulo de solución de problemas	Se agregó un capítulo de solución de problemas con un tema sobre el reemplazo automático de las instancias EC2 tras un reinicio. Para obtener más información, consulte Solución de problemas en AWS Parallel Computing Service.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
23 de septiembre de 2024	Se actualizaron los permisos mínimos para usar las acciones de la API y para un administrador de servicios	Ahora se requiere el `ec2:DescribeInstanceTypeOfferings` permiso para las acciones de la `UpdateComputeNodeGroup` API `CreateComputeNodeGroup` y las de la API. Para obtener más información, consulte Permisos mínimos para PCS AWS.	N/A
5 de septiembre de 2024	Se actualizó el ejemplo de política de IAM para los permisos mínimos para un administrador de servicios	Para obtener más información, consulte Permisos mínimos para un administrador de servicios.	N/A
5 de septiembre de 2024	Se agregó un permiso faltante al JSON en la página de políticas administradas	Se trataba únicamente de una corrección de la documentación. La política gestionada propiamente dicha no se modificó. Para obtener más información, consulte AWS políticas administradas para AWS Parallel Computing Service.	N/A

Date	Cambio	Actualizaciones de la documentación	Versiones de API actualizadas
28 de agosto de 2024	Se agregó la página de políticas administradas	Para obtener más información, consulte AWS políticas administradas para AWS Parallel Computing Service.	N/A
28 de agosto de 2024	AWS Versión de PCS	Versión inicial de la guía del usuario del AWS PCS.	AWS SDK: 28 de agosto de 2020

AWS Glosario

Para obtener la AWS terminología más reciente, consulte el AWS glosario de la Glosario de AWS Referencia.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.