Guida per l'utente

AWS PC

AWS PC: Guida per l’utente

Table of Contents

Cos'è il servizio AWS Parallel Computing?

AWS Parallel Computing Service (AWS PCS) è un servizio gestito che semplifica l'esecuzione e la scalabilità dei carichi di lavoro HPC (High Performance Computing) e la creazione di modelli scientifici e ingegneristici utilizzando Slurm. AWS Usa AWS PCS per creare cluster di elaborazione che integrano elaborazione, archiviazione, rete e AWS visualizzazione all'avanguardia. Esegui simulazioni o crea modelli scientifici e ingegneristici. Semplifica e semplifica le operazioni del cluster utilizzando funzionalità integrate di gestione e osservabilità. Consenti ai tuoi utenti di concentrarsi sulla ricerca e l'innovazione consentendo loro di eseguire applicazioni e lavori in un ambiente familiare.

Argomenti

Concetti in AWS PCS

Un cluster in AWS PCS ha 1 o più code, associate ad almeno 1 gruppo di nodi di calcolo. I lavori vengono inviati alle code ed eseguiti su EC2 istanze definite da gruppi di nodi di calcolo. È possibile utilizzare queste basi per implementare architetture HPC sofisticate.

Cluster

Un cluster è una risorsa per la gestione delle risorse e l'esecuzione dei carichi di lavoro. Un cluster è una risorsa AWS PCS che definisce un insieme di configurazione di elaborazione, rete, archiviazione, identità e pianificazione dei processi. È possibile creare un cluster specificando quale job scheduler si desidera utilizzare (attualmente Slurm), quale configurazione di scheduler si desidera, quale controller di servizio si desidera gestire il cluster e in quale VPC si desidera avviare le risorse del cluster. Lo scheduler accetta e pianifica i lavori e avvia anche i nodi di calcolo (istanze) che elaborano tali lavori. EC2

Gruppo di nodi di calcolo

Un gruppo di nodi di calcolo è una raccolta di nodi di elaborazione che AWS PCS utilizza per eseguire processi o fornire accesso interattivo a un cluster. Quando definisci un gruppo di nodi di calcolo, specifichi caratteristiche comuni come i tipi di EC2 istanze Amazon, il numero minimo e massimo di istanze, le sottoreti VPC di destinazione, Amazon Machine Image (AMI), l'opzione di

acquisto e la configurazione di avvio personalizzata. AWS PCS utilizza queste impostazioni per avviare, gestire e terminare in modo efficiente i nodi di calcolo in un gruppo di nodi di calcolo.

Queue

Quando si desidera eseguire un processo su un cluster specifico, lo si invia a una coda particolare (a volte chiamata anche partizione). Il processo rimane in coda finché AWS PCS non ne pianifica l'esecuzione su un gruppo di nodi di calcolo. Si associano uno o più gruppi di nodi di calcolo a ciascuna coda. È necessaria una coda per pianificare ed eseguire i lavori sulle risorse del gruppo di nodi di calcolo sottostanti utilizzando varie politiche di pianificazione offerte dal job scheduler. Gli utenti non inviano i lavori direttamente a un nodo di calcolo o a un gruppo di nodi di calcolo.

Amministratore di sistema

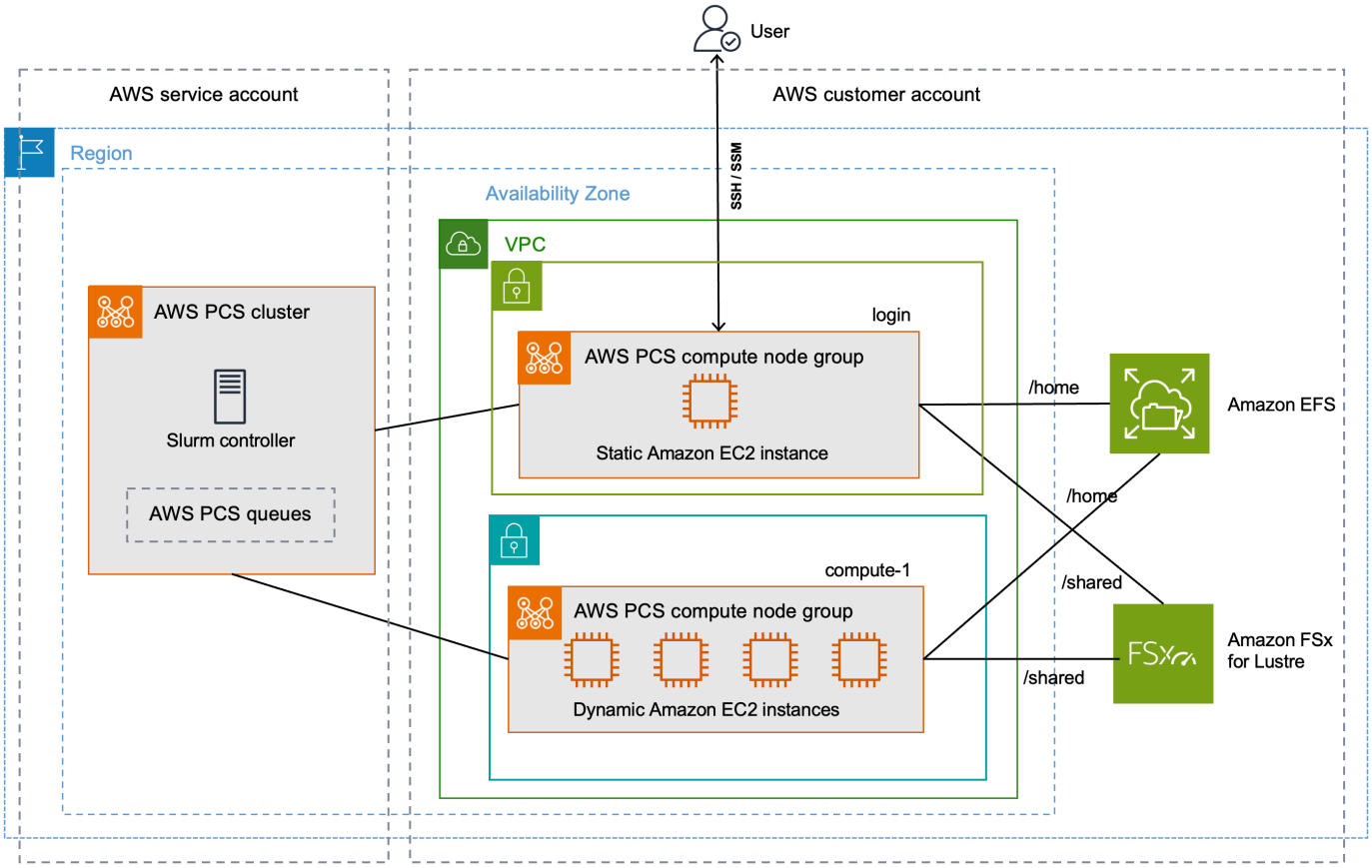
Un amministratore di sistema distribuisce, mantiene e gestisce un cluster. Possono accedere a AWS PCS tramite l' Console di gestione AWS API AWS PCS e l' AWS SDK. Hanno accesso a cluster specifici tramite SSH o AWS Systems Manager, dove possono eseguire attività amministrative, eseguire lavori, gestire dati ed eseguire altre attività basate su shell. Per ulteriori informazioni, consulta la documentazione di AWS Systems Manager.

Utente finale

Un utente finale non ha day-to-day la responsabilità di implementare o gestire un cluster. Utilizzano un'interfaccia terminale (come SSH) per accedere alle risorse del cluster, eseguire processi, gestire dati ed eseguire altre attività basate sulla shell.

Inizia a usare AWS Parallel Computing Service

Questo è un tutorial per creare un cluster semplice che puoi usare per provare AWS PCS. La figura seguente mostra il design del cluster.



Il tutorial sulla progettazione del cluster ha i seguenti componenti chiave:

- Un VPC e sottoreti che soddisfano AWS i requisiti di rete PCS.
- Un file system Amazon EFS, che verrà utilizzato come home directory condivisa.
- Un file system Amazon FSx for Lustre, che fornisce una directory condivisa ad alte prestazioni.
- Un cluster AWS PCS, che fornisce un controller Slurm.
- 2 gruppi di nodi di calcolo AWS PCS.
  - Il gruppo di `login` nodi, che fornisce un accesso interattivo basato su shell al sistema.
  - Il gruppo di `compute-1` nodi fornisce istanze con scalabilità elastica per eseguire i processi.
- 1 coda che invia i lavori alle istanze del gruppo di nodi. EC2 `compute-1`

Il cluster richiede AWS risorse aggiuntive, come gruppi di sicurezza, ruoli IAM e modelli di EC2 avvio, che non sono mostrati nel diagramma.

> **Note**
>
> Ti consigliamo di completare i passaggi della riga di comando descritti in questo argomento in una shell Bash. In alternativa, puoi apportare alcune modifiche alla tua shell per alcuni comandi di script, come i caratteri di continuazione della riga, e per il modo in cui le variabili vengono impostate e utilizzate. Inoltre, le regole di escape e di utilizzo delle virgolette per la shell (interprete di comandi) potrebbero essere diverse. Per ulteriori informazioni, consulta Virgolette e lettere con stringhe nella Guida per l' AWS CLIAWS Command Line Interfaceutente della versione 2.

Argomenti

Prerequisiti per iniziare a usare PCS AWS

Fate riferimento ai seguenti argomenti per preparare il vostro ambiente di sviluppo Account AWS e quello locale per AWS PCS.

Argomenti

Registrati AWS e crea un utente amministrativo

Completa le seguenti attività per configurare AWS Parallel Computing Service (AWS PCS).

Argomenti

Registrati per un Account AWS

Se non ne hai uno Account AWS, completa i seguenti passaggi per crearne uno.

Per iscriverti a un Account AWS

1. Apri la https://portal.aws.amazon.com/billing/registrazione.
2. Segui le istruzioni online.

   Nel corso della procedura di registrazione riceverai una telefonata o un messaggio di testo e ti verrà chiesto di inserire un codice di verifica attraverso la tastiera del telefono.

   Quando ti iscrivi a un Account AWS, Utente root dell'account AWSviene creato un. L'utente root dispone dell'accesso a tutte le risorse e tutti i Servizi AWS nell'account. Come best practice di sicurezza, assegna l'accesso amministrativo a un utente e utilizza solo l'utente root per eseguire attività che richiedono l'accesso di un utente root.

AWS ti invia un'email di conferma dopo il completamento della procedura di registrazione. In qualsiasi momento, puoi visualizzare l'attività corrente del tuo account e gestirlo accedendo a https://aws.amazon.com/e scegliendo Il mio account.

Crea un utente con accesso amministrativo

Dopo esserti registrato Account AWS, proteggi Utente root dell'account AWS AWS IAM Identity Center, abilita e crea un utente amministrativo in modo da non utilizzare l'utente root per le attività quotidiane.

Proteggi i tuoi Utente root dell'account AWS

1. Accedi Console di gestione AWScome proprietario dell'account scegliendo Utente root e inserendo il tuo indirizzo Account AWS email. Nella pagina successiva, inserisci la password.

   Per informazioni sull'accesso utilizzando un utente root, consulta la pagina Signing in as the root user della Guida per l'utente di Accedi ad AWS .
2. Abilita l'autenticazione a più fattori (MFA) per l'utente root.

   Per istruzioni, consulta Abilitare un dispositivo MFA virtuale per l'utente Account AWS root (console) nella Guida per l'utente IAM.

Crea un utente con accesso amministrativo

1. Abilita Centro identità IAM.

   Per istruzioni, consulta Abilitazione di AWS IAM Identity Center nella Guida per l'utente di AWS IAM Identity Center .
2. In IAM Identity Center, assegna l'accesso amministrativo a un utente.

   Per un tutorial sull'utilizzo di IAM Identity Center directory come fonte di identità, consulta Configurare l'accesso utente con l'impostazione predefinita IAM Identity Center directory nella Guida per l'AWS IAM Identity Center utente.

Accesso come utente amministratore

- Per accedere con l'utente IAM Identity Center, utilizza l'URL di accesso che è stato inviato al tuo indirizzo e-mail quando hai creato l'utente IAM Identity Center.

  Per informazioni sull'accesso utilizzando un utente IAM Identity Center, consulta AWS Accedere al portale di accesso nella Guida per l'Accedi ad AWS utente.

Assegna l'accesso a ulteriori utenti

1. In IAM Identity Center, crea un set di autorizzazioni conforme alla best practice dell'applicazione di autorizzazioni con il privilegio minimo.

   Segui le istruzioni riportate nella pagina Creazione di un set di autorizzazioni nella Guida per l'utente di AWS IAM Identity Center .
2. Assegna al gruppo prima gli utenti e poi l'accesso con autenticazione unica (Single Sign-On).

   Per istruzioni, consulta Aggiungere gruppi nella Guida per l'utente di AWS IAM Identity Center .

Installa il AWS CLI for AWS PCS

È necessario utilizzare la versione più recente di AWS CLI. Per informazioni, consulta Installare o aggiornare alla versione più recente di AWS CLI nella Guida per l'AWS Command Line Interface utente della versione 2.

È necessario configurare il AWS CLI. Per ulteriori informazioni, vedere Configurare il AWS CLI nella Guida per l'AWS Command Line Interface utente della versione 2.

Digitate il seguente comando al prompt dei comandi per verificarlo AWS CLI; dovrebbe visualizzare informazioni di aiuto.

```
aws pcs help
```

Autorizzazioni IAM richieste per AWS PCS

Il responsabile della sicurezza IAM che stai utilizzando deve disporre delle autorizzazioni per lavorare con i ruoli IAM AWS PCS, i ruoli collegati ai servizi AWS CloudFormation, un VPC e le risorse correlate. Per ulteriori informazioniServizio di Identity and Access Management per AWS Parallel Computing, consulta la sezione Creazione di un ruolo collegato ai servizi nella Guida per l'utente.AWS Identity and Access Management È necessario che tutti i passaggi di questa guida siano completati dallo stesso utente. Esegui il comando seguente per controllare l'utente corrente:

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

Utilizzo AWS CloudFormation con il tutorial AWS PCS

Il tutorial AWS PCS prevede molti passaggi e ha lo scopo di aiutarti a comprendere le parti di un cluster AWS PCS e le procedure necessarie per crearlo. Ti consigliamo di seguire i passaggi del tutorial almeno 1 volta. Dopo aver acquisito una buona conoscenza di ciò che si tratta, è possibile iniziare AWS CloudFormation a creare rapidamente il cluster di esempio con l'automazione.

CloudFormation è un AWS servizio che consente di creare e fornire implementazioni di AWS infrastrutture in modo prevedibile e ripetuto. È possibile utilizzare un CloudFormation modello per fornire automaticamente AWS le risorse per il cluster di esempio come una singola unità, denominata stack. È possibile eliminare lo stack quando lo si utilizza.

Per ulteriori informazioni, consulta Inizia con CloudFormation e AWS PCS.

Crea un VPC e sottoreti per PCS AWS

Puoi creare un VPC e delle sottoreti con un modello. CloudFormation Utilizza il seguente URL per scaricare il CloudFormation modello, quindi carica il modello nella CloudFormation console per creare un nuovo stack. CloudFormation Per ulteriori informazioni, consulta Uso della CloudFormation console nella Guida per l'AWS CloudFormation utente.

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

Con il modello aperto nella CloudFormation console, inserisci le seguenti opzioni. Puoi utilizzare i valori predefiniti forniti nel modello.

- In Fornisci un nome per lo stack:
  - In Nome dello stack, inserisci:

    ```
    hpc-networking
    ```
- In Parametri:
  - In VPC:
    - In CidrBlock, inserisci:

      ```
      10.3.0.0/16
      ```

- In Sottoreti A:
  - In CidrPublicSubnetA, inserisci:

```
10.3.0.0/20
```

  - In CidrPrivateSubnetA, inserisci:

```
10.3.128.0/20
```

- In Sottoreti B:
  - In CidrPublicSubnetB, inserisci:

```
10.3.16.0/20
```

  - In CidrPrivateSubnetB, inserisci:

```
10.3.144.0/20
```

- In Sottoreti C:
  - Per ProvisionSubnetsC, seleziona True
  - In CidrPublicSubnetC, inserisci:

```
10.3.32.0/20
```

  - In CidrPrivateSubnetC, inserisci:

```
10.3.160.0/20
```

- In Capacità:
  - Seleziona la casella Riconosco che AWS CloudFormation potrebbe creare risorse IAM.

Monitora lo stato dello CloudFormation stack. Quando raggiunge`CREATE_COMPLETE`, trova l'ID per il gruppo di sicurezza predefinito nel nuovo VPC. L'ID verrà utilizzato più avanti nel tutorial.

Trova il gruppo di sicurezza predefinito per il VPC del cluster

Per trovare l'ID per il gruppo di sicurezza predefinito nel nuovo VPC, segui questa procedura:

- Accedi alla console Amazon VPC.

- Nella dashboard VPC, seleziona Filtra per VPC.
  - Scegli il VPC con cui inizia il nome. `hpc-networking`
  - In Sicurezza, scegli Gruppi di sicurezza.
- Trova l'ID del gruppo di sicurezza per il gruppo denominato`default`. Ha la descrizione`default VPC security group`. L'ID verrà utilizzato successivamente per configurare i modelli di lancio di EC2.

Creare gruppi di sicurezza per AWS PCS

AWS PCS si affida a gruppi di sicurezza per gestire il traffico di rete in entrata e in uscita da un cluster e dai relativi gruppi di nodi di calcolo. Per informazioni dettagliate su questo argomento, vedere. Requisiti e considerazioni sui gruppi di sicurezza

In questo passaggio, utilizzerai un CloudFormation modello per creare due gruppi di sicurezza.

- Un gruppo di sicurezza del cluster, che consente le comunicazioni tra controller AWS PCS, nodi di elaborazione e nodi di accesso.
- Un gruppo di sicurezza SSH in entrata, che è possibile aggiungere facoltativamente ai nodi di accesso per supportare l'accesso SSH

Crea i gruppi di sicurezza per PCS AWS

È possibile utilizzare un CloudFormation modello per creare i gruppi di sicurezza. Utilizza il seguente URL per scaricare il CloudFormation modello, quindi carica il modello nella CloudFormation console per creare un nuovo CloudFormation stack. Per ulteriori informazioni, consulta Uso della CloudFormation console nella Guida per l'AWS CloudFormation utente.

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

Con il modello aperto nella AWS CloudFormation console, inserisci le seguenti opzioni. Tieni presente che alcune opzioni saranno precompilate nel modello: puoi semplicemente lasciarle come valori predefiniti.

- In Fornisci un nome per lo stack
  - In Nome dello stack, inserisci:

```
getstarted-sg
```

- In Parametri
  - In VpcId, scegli il VPC con cui inizia il nome. `hpc-networking`
  - (Facoltativo) In ClientIpCidr, inserisci un intervallo IP più restrittivo per il gruppo di sicurezza SSH in entrata. Ti consigliamo di limitarlo con il tuo IP/subnet (x.x.x.x/32 per il tuo IP o x.x.x.x/24 per l'intervallo. Sostituisci x.x.x.x con il tuo IP PUBBLICO. Puoi ottenere il tuo IP pubblico utilizzando strumenti come https://ifconfig.co/)

Monitora lo stato dello CloudFormation stack. Quando raggiunge `CREATE_COMPLETE` il gruppo di sicurezza, le risorse sono pronte.

Sono stati creati due gruppi di sicurezza, con i seguenti nomi:

- `cluster-getstarted-sg`— questo è il gruppo di sicurezza del cluster
- `inbound-ssh-getstarted-sg`— questo è un gruppo di sicurezza per consentire l'accesso SSH in entrata

Crea un cluster in AWS PCS

In AWS PCS, un cluster è una risorsa persistente per la gestione delle risorse e l'esecuzione dei carichi di lavoro. Si crea un cluster per uno scheduler specifico (AWS PCS attualmente supporta Slurm) in una sottorete di un VPC nuovo o esistente. Il cluster accetta e pianifica i lavori e avvia anche i nodi di calcolo (EC2 istanze) che elaborano tali lavori.

Creazione di un cluster

1. Apri la console AWS PCS e scegli Crea cluster.
2. Nella sezione Dettagli del cluster, inserisci i seguenti campi:
   - Nome del cluster: immettere `get-started`
   - Scheduler: seleziona la versione 25.05 di Slurm
   - Dimensioni del controller: seleziona Small
3. Nella sezione Rete, selezionate i valori per i seguenti campi:
   - VPC: scegli il VPC denominato `hpc-networking:Large-Scale-HPC`

- Subnet: seleziona la sottorete da cui inizia il nome `hpc-networking:PrivateSubnetA`
- Gruppi di sicurezza: selezionare il gruppo di sicurezza del cluster denominato `cluster-getstarted-sg`

4. Scegliere Crea cluster.

> **Note**
>
> Il campo Stato mostra Creazione durante il provisioning del cluster. La creazione del cluster può richiedere diversi minuti.

Crea storage condiviso per AWS PCS in Amazon Elastic File System

Amazon Elastic File System (Amazon EFS) è un AWS servizio che fornisce uno storage di file senza server e completamente elastico in modo da poter condividere i dati dei file senza fornire o gestire capacità e prestazioni di storage. Per ulteriori informazioni, consulta Cos'è Amazon Elastic File System? nella Amazon Elastic File System User Guide.

Il cluster dimostrativo AWS PCS utilizza un file system EFS per fornire una home directory condivisa tra i nodi del cluster. Crea un file system EFS nello stesso VPC del cluster.

Creazione del file system Amazon EFS

1. Vai alla console Amazon EFS.
2. Assicurati che sia impostato sullo stesso Regione AWS punto in cui proverai AWS PCS.
3. Scegliere Create file system (Crea file system).
4. Nella pagina Crea file system, imposta i seguenti parametri:
   - Per Nome immetti `getstarted-efs`.
   - In Virtual Private Cloud (VPC), scegli il VPC denominato `hpc-networking:Large-Scale-HPC`
   - Scegli Create (Crea) . Questo ti riporta alla pagina dei file system.
5. Prendi nota dell'ID del file system per il `getstarted-efs` file system. Queste informazioni serviranno in seguito.

Crea storage condiviso per AWS PCS in Amazon FSx for Lustre

Amazon FSx for Lustre semplifica ed economica l'avvio e l'esecuzione del popolare file system Lustre ad alte prestazioni. Usi Lustre per carichi di lavoro in cui la velocità è importante, come l'apprendimento automatico, l'elaborazione ad alte prestazioni (HPC), l'elaborazione video e la modellazione finanziaria. Per ulteriori informazioni, consulta Cos'è Amazon FSx for Lustre? nella Guida per l'utente di Amazon FSx for Lustre.

Il cluster dimostrativo AWS PCS può utilizzare un file system FSx for Lustre per fornire una directory condivisa ad alte prestazioni tra i nodi del cluster. Crea un file system FSx for Lustre nello stesso VPC del cluster.

Per creare il tuo file system FSx for Lustre

1. Vai alla FSx console Amazon.
2. Assicurati che la console sia impostata per l'utilizzo Regione AWS come il cluster.
3. Scegliere Create file system (Crea file system).
   - Per Seleziona il tipo di file system, scegli Amazon FSx for Lustre, quindi scegli Avanti.
4. Nella pagina Specificare i dettagli del file system, imposta i seguenti parametri:
   - In Dettagli del file system
     - Per Nome immetti `getstarted-fsx`.
     - Per il tipo di distribuzione e archiviazione, scegli Persistente, SSD
     - Per Throughput per unità di storage, scegli 125 MB/s/TiB
     - Per Capacità di archiviazione, immettere 1,2 TiB
     - Per Configurazione dei metadati, scegliete Automatico
     - Per Tipo di compressione dei dati, scegli LZ4
   - In Rete e sicurezza
     - Per Virtual Private Cloud (VPC), scegli il VPC denominato `hpc-networking:Large-Scale-HPC`
     - Per i gruppi di sicurezza VPC, lascia il nome al gruppo di sicurezza `default`
     - Per Subnet, scegli la sottorete con cui inizia il nome `hpc-networking:PrivateSubnetA`
   - Lasciate le altre opzioni impostate sui valori predefiniti.
   - Scegli Next (Successivo).

5. Nella pagina Rivedi e crea, scegli Crea file system. Verrà visualizzata di nuovo la pagina File system.
6. Vai alla pagina dei dettagli del file system FSx for Lustre che hai creato.
7. Prendi nota dell'ID del file system e del nome del montaggio. Queste informazioni serviranno in seguito.

> **Note**
>
> Il campo Stato mostra Creazione durante il provisioning del file system. La creazione del file system può richiedere diversi minuti. Attendi il completamento prima di procedere con il resto del tutorial.

Crea gruppi di nodi di calcolo in AWS PCS

Un gruppo di nodi di calcolo è una raccolta virtuale di nodi di calcolo (istanze EC2) che AWS PCS avvia e gestisce. Quando definisci un gruppo di nodi di calcolo, specifichi caratteristiche comuni come i tipi di istanze EC2, il numero minimo e massimo di istanze, le sottoreti VPC di destinazione, l'opzione di acquisto preferita e la configurazione di avvio personalizzata. AWS PCS avvia, gestisce e termina in modo efficiente i nodi di calcolo in un gruppo di nodi di calcolo, in base a queste impostazioni. Il cluster dimostrativo utilizza un gruppo di nodi di calcolo per fornire nodi di accesso per l'accesso degli utenti e un gruppo di nodi di calcolo separato per elaborare i lavori. I seguenti argomenti descrivono le procedure per configurare questi gruppi di nodi di calcolo nel cluster.

Argomenti

- Creare un profilo di istanza per AWS PCS
- Crea modelli di lancio per AWS PCS
- Crea un gruppo di nodi di calcolo per i nodi di accesso in AWS PCS
- Crea un gruppo di nodi di calcolo per eseguire lavori di elaborazione in PCS AWS

Creare un profilo di istanza per AWS PCS

I gruppi di nodi di calcolo richiedono un profilo di istanza al momento della creazione. Se utilizzi la Console di gestione AWS per creare un ruolo per Amazon EC2, la console crea automaticamente un

profilo dell'istanza e gli assegna lo stesso nome del ruolo. Per ulteriori informazioni, consulta Uso dei profili di istanza nella Guida per l'AWS Identity and Access Management utente.

Nella procedura seguente, usi per creare un ruolo per Amazon EC2, che crea anche il profilo di istanza per i tuoi gruppi di nodi di calcolo. Console di gestione AWS

Per creare il ruolo e il profilo dell'istanza

- Passare alla IAM console (Console IAM).
- In Gestione accessi scegli Policy.
  - Seleziona Create Policy (Crea policy).
  - In Specificare le autorizzazioni, per Policy editor, scegli JSON.
  - Sostituisci il contenuto dell'editor di testo con quanto segue:

    JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

  - Scegli Next (Successivo).
  - In Rivedi e crea, per Nome della politica, inserisci`AWSPCS-getstarted-policy`.
  - Scegli Crea policy.
- In Access management (Gestione accessi), scegli Roles (Ruoli).
- Scegli Crea ruolo.
- In Seleziona entità attendibile:
  - Per il tipo di entità affidabile, seleziona AWS servizio
  - In Caso d'uso, seleziona EC2.
    - Quindi, in Scegli un caso d'uso per il servizio specificato, scegli EC2.

    - Scegli Next (Successivo).
- In Aggiungi autorizzazioni:
    - In Politiche di autorizzazione, cerca AWSPCS-getstarted -policy.
    - Seleziona la casella accanto a AWSPCS-getstarted-policy per aggiungerla al ruolo.
    - In Politiche di autorizzazione, cerca Amazon SSMManaged InstanceCore.
    - Seleziona la casella accanto SSMManaged InstanceCore ad Amazon per aggiungerlo al ruolo.
    - Scegli Next (Successivo).
- In Nome, rivedi e crea:
    - In Dettagli del ruolo:
        - Per Nome ruolo, inserisci `AWSPCS-getstarted-role`.
    - Scegli Crea ruolo.

Crea modelli di lancio per AWS PCS

Quando crei un gruppo di nodi di calcolo, fornisci un modello di lancio EC2 che AWS PCS utilizza per configurare le istanze EC2 che lancia. Ciò include impostazioni come gruppi di sicurezza e script che vengono eseguiti all'avvio dell'istanza.

In questa fase, verrà utilizzato un CloudFormation modello per creare due modelli di lancio EC2. Un modello verrà utilizzato per creare nodi di accesso e l'altro verrà utilizzato per creare nodi di calcolo. La differenza fondamentale tra loro è che i nodi di accesso possono essere configurati per consentire l'accesso SSH in entrata.

Accedi al modello CloudFormation

Utilizza il seguente URL per scaricare il CloudFormation modello, quindi carica il modello nella CloudFormation console per creare un nuovo CloudFormation stack. Per ulteriori informazioni, consulta Uso della CloudFormation console nella Guida per l'AWS CloudFormation utente.

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

Usa il CloudFormation modello per creare modelli di lancio EC2

Utilizza la seguente procedura per completare il CloudFormation modello nella console CloudFormation

- In Fornisci un nome per lo stack:
  - In Nome dello stack, inserisci. `getstarted-lt`
- In Parametri:
  - In Sicurezza
    - Per VpcSecurityGroupId, seleziona il gruppo di sicurezza denominato `default` nel tuo VPC del cluster.
    - Per ClusterSecurityGroupId, seleziona il gruppo denominato `cluster-getstarted-sg`
    - Per SshSecurityGroupId, seleziona il gruppo denominato `inbound-ssh-getstarted-sg`
    - Per SshKeyName, seleziona la tua coppia di chiavi SSH preferita.
  - In File system
    - Per EfsFilesystemId, inserisci l'ID del file system dal file system EFS che hai creato in precedenza nel tutorial.
    - Ad FSxLustreFilesystemIdesempio, inserisci l'ID del file system del file system FSx for Lustre che hai creato in precedenza nel tutorial.
    - Per FSxLustreFilesystemMountName, inserisci il nome di montaggio corrispondente per il file system Lustre. FSx
- Scegliete Avanti, quindi scegliete nuovamente Avanti.
- Seleziona Invia.

Monitora lo stato dello CloudFormation stack. Quando raggiunge `CREATE_COMPLETE` il modello di lancio è pronto per essere utilizzato.

> **Note**
>
> Per vedere tutte le risorse create dal CloudFormation modello, apri la CloudFormation console. Scegli lo stack `getstarted-lt`, quindi la scheda Resources (Risorse).

Crea un gruppo di nodi di calcolo per i nodi di accesso in AWS PCS

Un gruppo di nodi di calcolo è una raccolta virtuale di nodi di calcolo (istanze EC2) che AWS PCS avvia e gestisce. Quando definisci un gruppo di nodi di calcolo, specifichi caratteristiche comuni come i tipi di istanze EC2, il numero minimo e massimo di istanze, le sottoreti VPC di destinazione, l'opzione di acquisto preferita e la configurazione di avvio personalizzata. AWS PCS avvia, gestisce

e termina in modo efficiente i nodi di calcolo in un gruppo di nodi di calcolo, in base a queste impostazioni.

In questo passaggio, lancerai un gruppo di nodi di calcolo statici che fornisce l'accesso interattivo al cluster. Puoi usare SSH o Amazon EC2 Systems Manager (SSM) per accedervi, quindi eseguire comandi shell e gestire i job Slurm.

Per creare il gruppo di nodi di calcolo

- Apri la console AWS PCS e vai a Clusters.
- Seleziona il cluster denominato `get-started`
- Vai ai gruppi di nodi di calcolo e scegli Crea.
- Nella sezione Configurazione del gruppo di nodi di calcolo, fornisci quanto segue:
  - Nome del gruppo di nodi di calcolo: immettere. `login`
- In Configurazione informatica, inserisci o seleziona questi valori:
  - Modello di lancio EC2: scegli il modello di lancio con il nome `login-getstarted-lt`
  - Profilo dell'istanza IAM: scegli il profilo di istanza denominato `AWSPCS-getstarted-role`
  - Sottoreti: seleziona la sottorete da cui inizia il nome. `hpc-networking:PublicSubnetA`
  - Istanze: seleziona. `c6i.xlarge`
  - Configurazione di scalabilità: per il numero minimo di istanze, immettere. 1 Per Numero massimo di istanze, immettete. 1
- In Impostazioni aggiuntive, specificate quanto segue:
  - ID AMI: seleziona un AMI che desideri utilizzare, con un nome nel seguente formato:

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

    Per ulteriori informazioni sull'esempio AMIs, vedereUtilizzo di Amazon Machine Images (AMIs) di esempio con AWS PCS.
- Scegli Crea gruppo di nodi di calcolo.

Il campo Stato mostra Creazione durante il provisioning del gruppo di nodi di calcolo. Puoi procedere al passaggio successivo del tutorial mentre è in corso.

Crea un gruppo di nodi di calcolo per eseguire lavori di elaborazione in PCS AWS

In questo passaggio, lancerai un gruppo di nodi di calcolo con scalabilità elastica per eseguire i lavori inviati al cluster.

Per creare il gruppo di nodi di calcolo

- Apri la console AWS PCS e vai a Clusters.
- Seleziona il cluster denominato `get-started`
- Passa ai gruppi di nodi di calcolo e scegli Crea.
- Nella sezione Configurazione del gruppo di nodi di calcolo, fornisci quanto segue:
  - Nome del gruppo di nodi di calcolo: immettere. `compute-1`
- In Configurazione informatica, inserisci o seleziona questi valori:
  - Modello di lancio EC2: scegli il modello di lancio con il nome `compute-getstarted-lt`
  - Profilo dell'istanza IAM: scegli il profilo di istanza denominato `AWSPCS-getstarted-role`
  - Sottoreti: seleziona la sottorete da cui inizia il nome. `hpc-networking:PrivateSubnetA`
  - Istanze: seleziona. `c6i.xlarge`
  - Configurazione di scalabilità: per il numero minimo di istanze, immettere. `0` Per Numero massimo di istanze, immettete. `4`
- In Impostazioni aggiuntive, specificate quanto segue:
  - ID AMI: seleziona un AMI che desideri utilizzare, con un nome nel seguente formato:

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

    Per ulteriori informazioni sull'esempio AMIs, vedereUtilizzo di Amazon Machine Images (AMIs) di esempio con AWS PCS.
- Scegli Crea gruppo di nodi di calcolo.

Il campo Stato mostra Creazione durante il provisioning del gruppo di nodi di calcolo.

> **Important**
>
> Attendi che il campo Stato mostri Attivo prima di procedere al passaggio successivo di questo tutorial.

Crea una coda per gestire i lavori in AWS PCS

Si invia un lavoro a una coda per eseguirlo. Il lavoro rimane in coda finché AWS PCS non ne pianifica l'esecuzione su un gruppo di nodi di calcolo. Ogni coda è associata a uno o più gruppi di nodi di calcolo, che forniscono le EC2 istanze necessarie per eseguire l'elaborazione.

In questo passaggio, creerai una coda che utilizza il gruppo di nodi di calcolo per elaborare i lavori.

Per creare una coda

- Apri la console AWS PCS.
- Seleziona il cluster denominato`get-started`.
- Passa ai gruppi di nodi di calcolo e assicurati che lo stato del `compute-1` gruppo sia Attivo.

> **Important**
>
> Lo stato del `compute-1` gruppo deve essere Attivo prima di procedere al passaggio successivo.

- Vai a Code e scegli Crea coda.
  - Nella sezione Configurazione della coda, fornisci i seguenti valori:
    - Nome della coda: immettete quanto segue: `demo`
    - Gruppi di nodi di calcolo: seleziona il gruppo di nodi di calcolo denominato. `compute-1`
- Scegliere Crea coda.

Il campo Stato mostra Creazione durante la creazione della coda.

> **Important**
>
> Attendi che il campo Stato mostri Attivo prima di procedere al passaggio successivo di questo tutorial.

Connect al cluster AWS PCS

Dopo che lo stato del gruppo di nodi di `login` calcolo diventa Attivo, puoi connetterti all' EC2 istanza che ha creato.

Per connettersi al nodo di accesso

- Apri la console AWS PCS e vai a Clusters.
- Seleziona il cluster denominato`get-started`.
- Scegli Gruppi di nodi Compute.
- Passa al gruppo di nodi di calcolo denominato. `login`
- Trova l'ID del gruppo di nodi Compute.
- In un'altra finestra o scheda del browser, apri la EC2 console Amazon.
  - Seleziona Instances (Istanze).
  - Cerca le EC2 istanze con il tag seguente. Sostituisci *`node-group-id`* con il valore dell'ID del gruppo di nodi Compute del passaggio precedente. Dovrebbe esserci 1 istanza.

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

  - Connect all' EC2 istanza. È possibile utilizzare Session Manager o SSH.

    Session Manager

    - Selezionare l'istanza.
    - Scegli Connetti.
    - In Connect to instance, seleziona Session Manager.
    - Scegli Connetti.
    - Scegli Connetti. Nel browser viene avviato un terminale interattivo.

    SSH

    - Selezionare l'istanza.
    - Scegli Connetti.
    - In Connect to instance, seleziona Client SSH.
    - Segui le istruzioni fornite dalla console.

> **Note**
>
> Il nome utente dell'istanza **`ec2-user`**non lo è`root`.

Esplora l'ambiente cluster in AWS PCS

Dopo aver effettuato l'accesso al cluster, puoi eseguire i comandi della shell. Ad esempio, puoi cambiare utente, lavorare con i dati su file system condivisi e interagire con Slurm.

Cambia utente

Se hai effettuato l'accesso al cluster utilizzando Session Manager, potresti essere connesso come`ssm-user`. Si tratta di un utente speciale creato per Session Manager. Passa all'utente predefinito su Amazon Linux 2 utilizzando il seguente comando. Non avrai bisogno di farlo se ti connetti tramite SSH.

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

Lavora con file system condivisi

È possibile confermare che il file system EFS e FSx per i file system Lustre sono disponibili con il comando. `df -h` L'output sul cluster dovrebbe essere simile al seguente:

```
[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
```

Il /home filesystem monta 127.0.0.1 e ha una capacità molto grande. Questo è il file system EFS creato in precedenza nel tutorial. Tutti i file scritti qui saranno disponibili /home in tutti i nodi del cluster.

Il `/shared` filesystem monta un IP privato e ha una capacità di 1,2 TB. Questo è il file system FSx for Lustre creato in precedenza nel tutorial. Tutti i file scritti qui saranno disponibili `/shared` in tutti i nodi del cluster.

Interagisci con Slurm

Argomenti

Elenca code e nodi

È possibile elencare le code e i nodi a cui sono associate. `sinfo` L'output del cluster dovrebbe essere simile al seguente:

```
[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 ~]$
```

Notate la partizione denominata. `demo` Il suo stato è up e ha un massimo di 4 nodi. È associato ai nodi del gruppo di `compute-1` nodi. Se modifichi il gruppo di nodi di calcolo e aumenti il numero massimo di istanze a 8, verrà letto il numero di nodi 8 e verrà letto l'elenco dei nodi. `compute-1-[1-8]` Se creassi un secondo gruppo di nodi di calcolo denominato `test` con 4 nodi e lo aggiungessi alla `demo` coda, tali nodi verranno visualizzati anche nell'elenco dei nodi.

Mostra offerte di lavoro

Puoi elencare tutti i lavori, in qualsiasi stato, sul sistema `consqueue`. L'output del cluster dovrebbe essere simile al seguente:

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

Prova a eseguire `squeue` di nuovo più tardi, quando hai un job Slurm in sospeso o in esecuzione.

Esegui un processo a nodo singolo in AWS PCS

Per eseguire un lavoro utilizzando Slurm, si prepara uno script di invio che specifica i requisiti del lavoro e lo si invia a una coda con il comando. `sbatch` In genere, questa operazione viene eseguita da una directory condivisa in modo che i nodi di accesso e di calcolo abbiano uno spazio comune per l'accesso ai file.

Connect al nodo di login del cluster ed esegui i seguenti comandi al prompt della shell.

- Diventa l'utente predefinito. Passa alla directory condivisa.

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

- Utilizzate i seguenti comandi per creare uno script di lavoro di esempio:

```
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
```

- Invia lo script di lavoro allo scheduler Slurm:

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

- Quando il lavoro viene inviato, restituirà un ID del lavoro come numero. Usa quell'ID per controllare lo stato del lavoro. Sostituisci *`job-id`* nel comando seguente con il numero restituito da`sbatch`.

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

  Example

```
squeue --job 1
```

Il `squeue` comando restituisce un output simile al seguente:

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

- Continuare a controllare lo stato del processo finché non raggiunge lo stato R (in esecuzione). Il lavoro è terminato quando `squeue` non restituisce nulla.
- Ispeziona il contenuto della `/shared` directory.

```
ls -alth /shared
```

L'output del comando è simile al seguente:

```
-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
```

I file `single.1.err` denominati `single.1.out` e scritti da uno dei nodi di calcolo del cluster. Poiché il processo è stato eseguito in una directory condivisa (`/shared`), sono disponibili anche nel nodo di accesso. Questo è il motivo per cui hai configurato un file system FSx for Lustre per questo cluster.

- Ispeziona il contenuto del `single.1.out` file.

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

L'output è simile a quello riportato di seguito:

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

Esegui un processo MPI multinodo con Slurm in PCS AWS

Queste istruzioni dimostrano l'utilizzo di Slurm per eseguire un processo MPI (Message Passing Interface) in PCS. AWS

Esegui i seguenti comandi al prompt della shell del tuo nodo di accesso.

- Diventa l'utente predefinito. Passa alla sua home directory.

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

- Crea codice sorgente nel linguaggio di programmazione 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
```

- Caricate il modulo OpenMPI.

```
module load openmpi
```

- Compila il programma C.

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

- Scrivi uno script per l'invio di lavori a 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
```

- Passa alla directory condivisa.

```
cd /shared
```

- Invia lo script del lavoro.

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

- Utilizzatelo `squeue` per monitorare il lavoro fino al termine.
- Controlla il contenuto di`multi.out`:

```
cat multi.out
```

L'output è simile a quello riportato di seguito. Nota che ogni rank ha il proprio indirizzo IP perché è stato eseguito su un nodo diverso.

```
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
```

Elimina le tue AWS risorse per AWS PCS

Dopo aver finito con i gruppi di cluster e nodi che hai creato per questo tutorial, dovresti eliminare le risorse che hai creato.

> **Important**
>
> Ti verranno addebitati i costi di fatturazione per tutte le risorse in esecuzione nel tuo Account AWS

Per eliminare le risorse AWS PCS che hai creato per questo tutorial

- Apri la console AWS PCS.
- Passa al cluster denominato get-started.
- Vai alla sezione Code.
- Seleziona la coda denominata demo.

- Scegli Elimina.

> **Important**
>
> Attendi che la coda sia stata eliminata prima di procedere.

- Vai alla sezione Compute node groups.
- Seleziona il gruppo di nodi di calcolo denominato compute-1.
- Scegli Elimina.
- Seleziona il gruppo di nodi di calcolo denominato login.
- Scegli Elimina.

> **Important**
>
> Attendi che entrambi i gruppi di nodi di calcolo siano stati eliminati prima di procedere.

- Nella pagina dei dettagli del cluster per iniziare, scegli Elimina.

> **Important**
>
> Attendi che il cluster sia stato eliminato prima di procedere con i passaggi successivi.

Per eliminare altre AWS risorse che hai creato per questo tutorial

- Apri la console IAM.
  - Scegli Ruoli.
  - Seleziona il ruolo denominato AWSPCS-getstarted-role, quindi scegli Elimina.
  - Dopo che il ruolo è stato eliminato, scegli Politiche.
  - Seleziona la politica denominata AWSPCS-getstarted-policy, quindi scegli Elimina.
- Apri la CloudFormation console.
  - Seleziona lo stack denominato getstarted-lt.
  - Scegli Elimina.

> **Important**
>
> Attendi che lo stack venga eliminato prima di procedere.

- Apri la Console di Amazon EFS.
  - Seleziona File system.
  - Seleziona il file system denominato getstarted-efs.
  - Scegli Elimina.

> **Important**
>
> Attendi l'eliminazione del file system prima di procedere.

- Apri la FSx console Amazon.
  - Seleziona File system.
  - Seleziona il file system denominato getstarted-fsx.
  - Scegli Elimina.

> **Important**
>
> Attendi l'eliminazione del file system prima di procedere.

- Apri la CloudFormation console.
  - Seleziona lo stack denominato getstarted-sg.
  - Scegli Elimina.
- Apri la CloudFormation console.
  - Seleziona lo stack denominato hpc-networking.
  - Scegli Delete (Elimina).

Inizia con CloudFormation e AWS PCS

Puoi usarlo AWS CloudFormation per creare un cluster AWS PCS. CloudFormation consente di creare e fornire implementazioni di AWS infrastrutture in modo prevedibile e ripetuto. È possibile CloudFormation utilizzare il provisioning automatico delle risorse di molti AWS servizi per creare applicazioni altamente affidabili, scalabili ed economiche Cloud AWS senza creare e configurare l'infrastruttura sottostante. AWS CloudFormation consente di utilizzare un file modello per creare ed eliminare una raccolta di risorse insieme come una singola unità, denominata stack. Per ulteriori informazioni su CloudFormation, consulta What is CloudFormation? nella Guida AWS CloudFormation per l'utente. Per ulteriori informazioni sui tipi di risorse AWS PCS in CloudFormation, vedere il riferimento ai tipi di risorse AWS PCS nella Guida per l'AWS CloudFormation utente.

Argomenti

Utilizzare CloudFormation per creare un cluster AWS PCS di esempio

La procedura seguente utilizza un CloudFormation modello Console di gestione AWS per creare un cluster AWS PCS di esempio. Per ulteriori informazioni su CloudFormation, consulta What is CloudFormation? nella Guida AWS CloudFormation per l'utente. Per ulteriori informazioni sui tipi di risorse AWS PCS in CloudFormation, vedere il riferimento ai tipi di risorse AWS PCS nella Guida per l'AWS CloudFormation utente.

Per creare il cluster di esempio

1. Scegli in cui Regione AWS creare il cluster (il link apre la CloudFormation console con il modello):

   - US East (N. Virginia) (Stati Uniti orientali (Virginia settentrionale)) (us-east-1)
   - US East (Ohio) (Stati Uniti orientali (Ohio)) (us-east-2)

- US West (Oregon) (Stati Uniti occidentali (Oregon)) (us-west-2)
- Asia Pacific (Singapore) (Asia Pacifico (Singapore)) (ap-southeast-1)
- Asia Pacific (Sydney) (Asia Pacifico (Sydney)) (ap-southeast-2)
- Asia Pacific (Tokyo) (Asia Pacifico (Tokyo)) (ap-northeast-1)
- Europa (Francoforte) (eu-central-1)
- Europa (Irlanda) (eu-west-1)
- Europa (Londra) (eu-west-2)
- Europa (Stoccolma) (eu-north-1)
- AWS GovCloud (Stati Uniti orientali) (-1) us-gov-east
- AWS GovCloud (Stati Uniti occidentali) (us-gov-west-1)

2. In Fornisci un nome per lo stack, inserisci un nome descrittivo. Questo è il nome del tuo CloudFormation stack. Il modello utilizza questo valore come nome per il cluster AWS PCS.
3. In Parametri:
   a. In SlurmVersion, scegli la versione di Slurm che desideri venga utilizzata dal tuo cluster.
   b. In NodeArchitecture, scegli x86 per distribuire un cluster che utilizza istanze compatibili con x86_64, oppure scegli Graviton per usare le istanze Arm64.
   c. Per KeyName, scegli una coppia di chiavi SSH per accedere ai nodi di accesso del cluster. Assicurati di avere il file PEM per la coppia di chiavi che scegli.
   d. Ad ClientIpCidresempio, inserisci un intervallo IP in formato CIDR per controllare l'accesso ai nodi di accesso.

      > ⚠ Warning
      >
      > Il valore predefinito di `0.0.0.0/0` consente l'accesso da tutti gli indirizzi IP.

   e. Lascia i valori per HpcRecipesS3Bucket e HpcRecipesBranchcome valori predefiniti.
4. In Capacità e trasformazioni:
   a. Seleziona la casella di controllo per confermare che CloudFormation verranno create risorse IAM.
   b. Seleziona la casella di controllo per confermare che CloudFormation verranno create risorse IAM con nomi personalizzati.

c. Seleziona la casella di controllo `CAPABILITY_AUTO_EXPAND` per confermare l'esistenza del nuovo stack. Per ulteriori informazioni, consulta CreateStack nella documentazione di riferimento dell'API AWS CloudFormation .

5. Seleziona Crea stack.
6. Monitora lo stato del tuo stack. Puoi connetterti al cluster dopo che lo stato dello stack è. `CREATE_COMPLETE`

Connect a un cluster AWS PCS creato con CloudFormation

Dopo aver creato un cluster AWS PCS da un CloudFormation modello, puoi utilizzare la console AWS PCS (in Console di gestione AWS) per amministrare il cluster. È inoltre possibile connettersi a 1 dei nodi di accesso del cluster per amministrare il cluster, eseguire processi e gestire i dati. Lo CloudFormation stack fornisce collegamenti che è possibile utilizzare per connettersi al cluster.

Per connetterti al tuo cluster

1. Apri la console CloudFormation
2. Scegli lo stack che hai creato.
3. Scegli la scheda Output dello stack.

   Lo stack fornisce i seguenti link:

   - PcsConsoleUrl— Scegliete questo collegamento per aprire la console AWS PCS con il cluster selezionato. Puoi usarlo per esplorare le configurazioni del cluster, del gruppo di nodi e della coda.
   - Ec2 ConsoleUrl: scegli questo link per aprire la console Amazon EC2, filtrata per mostrare le istanze gestite dal gruppo di nodi di accesso del cluster.

     Da questa vista, puoi selezionare un'istanza e scegliere Connect. L'istanza del cluster di esempio supporta SSH in ingresso e AWS Systems Manager connessioni in un browser Web. Per ulteriori informazioni, consulta Connect al cluster AWS PCS.

     Dopo esserti connesso a un'istanza di accesso, puoi seguire il tutorial all'indirizzo. Esplora l'ambiente cluster in AWS PCS

Pulisci un cluster AWS PCS in CloudFormation

Se in precedenza CloudFormation creavi il tuo cluster AWS PCS, puoi aprire la CloudFormation console ed eliminare lo stack per eliminare il cluster e tutte le risorse associate.

> **Important**
>
> Per il cluster di esempio, se nel cluster sono stati creati gruppi o code di nodi di calcolo aggiuntivi (oltre ai `compute-1` gruppi `login` e creati dal CloudFormation modello di esempio), è necessario utilizzare la console AWS PCS o AWS CLI eliminare tali risorse prima di eliminare lo stack. CloudFormation Per ulteriori informazioni, consulta Eliminazione di un cluster in AWS PCS.

Parti di un CloudFormation modello per AWS PCS

Un CloudFormation modello ha 1 o più sezioni, ognuna delle quali ha uno scopo specifico. CloudFormation definisce il formato, la sintassi e il linguaggio standard in un modello. Per ulteriori informazioni, consulta Lavorare con i CloudFormation modelli nella Guida per l'AWS CloudFormation utente.

CloudFormation i modelli sono altamente personalizzabili e pertanto i loro formati possono variare. Per comprendere le parti necessarie di un CloudFormation modello per creare un cluster AWS PCS, ti consigliamo di esaminare il modello di esempio che forniamo per creare un cluster di esempio. Questo argomento spiega brevemente le sezioni di quel modello di esempio.

> **Important**
>
> Gli esempi di codice in questo argomento non sono completi. La presenza di ellipsis (`[...]`) indica che esiste un codice aggiuntivo che non viene visualizzato. Per scaricare il modello completo in formato YAML CloudFormation , vedi. CloudFormation modelli per creare un cluster AWS PCS di esempio

Indice

- Parameters
- Mappature
- Resources
- Output

Header

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

`AWSTemplateFormatVersion`identifica la versione del formato del modello a cui il modello è conforme. Per ulteriori informazioni, consulta la sintassi della versione in formato CloudFormation modello nella Guida per l'AWS CloudFormation utente.

`Transform`specifica una macro che CloudFormation utilizza per elaborare il modello. Per ulteriori informazioni, consultate la sezione CloudFormation Template Transform nella Guida per l'AWS CloudFormation utente. La `AWS::Serverless-2016-10-31` trasformazione consente CloudFormation di elaborare un modello scritto nella sintassi AWS Serverless Application Model (AWS SAM). Per ulteriori informazioni, consulta `AWS::Serverless`transform nella Guida per l'AWS CloudFormation utente.

Metadati

```
### 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` sezione di un CloudFormation modello fornisce informazioni sul modello stesso. Il modello di esempio crea un cluster HPC (High Performance Computing) completo che utilizza AWS PCS. La sezione dei metadati del modello di esempio dichiara i parametri che controllano il modo in cui CloudFormation avvia (fornisce) lo stack corrispondente. Esistono parametri che controllano l'architettura choice (`NodeArchitecture`), la versione Slurm () e i controlli di accesso (and`SlurmVersion`). `KeyName` `ClientIpCidr`

Parameters

La `Parameters` sezione definisce i parametri personalizzati per il modello. CloudFormation utilizza queste definizioni di parametri per costruire e convalidare il modulo con cui interagisci quando avvii uno stack da questo modello.

```
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$)(?!.*/\.)(?!.*\\.\.)[a-zA-Z0-9-_\.]+$'
```

Mappature

La `Mappings` sezione definisce coppie chiave-valore che specificano i valori in base a determinate condizioni o dipendenze.

```
Mappings:

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

Resources

La `Resources` sezione dichiara le AWS risorse da fornire e configurare come parte dello stack.

```
Resources:

  [...]
```

Il modello fornisce l'infrastruttura del cluster di esempio a livelli. Inizia con `Networking` la configurazione VPC. Lo storage è fornito da due sistemi: `EfsStorage` per lo storage condiviso e `FSxLStorage` per lo storage ad alte prestazioni. Il cluster principale viene stabilito tramite`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 ]
```

Per le risorse di elaborazione, il modello crea due gruppi di nodi: `PCSNodeGroupLogin` per un singolo nodo di accesso e `PCSNodeGroupCompute` per un massimo di quattro nodi di elaborazione. Questi gruppi di nodi sono supportati da `PCSInstanceProfile` per le autorizzazioni e, ad `PCSLaunchTemplate` esempio, le configurazioni.

```
  # 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 pianificazione del lavoro viene gestita tramite. `PCSQueueCompute`

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

La selezione degli AMI avviene automaticamente tramite la funzione Pcs AMILookup Fn Lambda e le risorse correlate.

```
  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
```

Output

Il modello restituisce l'identificazione e la gestione del cluster URLs tramite`ClusterId`, `PcsConsoleUrl` e. `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 modelli per creare un cluster AWS PCS di esempio

Regione AWS nome	Regione AWS	Visualizza fonte	Stack di lancio
Stati Uniti orientali (Virginia settentrionale)	us-east-1	Scarica YAML	Launch Stack
Stati Uniti orientali (Ohio)	us-east-2	Scarica YAML	Launch Stack
Stati Uniti occidentali (Oregon)	us-west-2	Scarica YAML	Launch Stack
Asia Pacifico (Singapore)	ap-southeast-1	Scarica YAML	Launch Stack
Asia Pacifico (Sydney)	ap-southeast-2	Scarica YAML	Launch Stack
Asia Pacifico (Tokyo)	ap-northeast-1	Scarica YAML	Launch Stack

Regione AWS nome	Regione AWS	Visualizza fonte	Stack di lancio
Europa (Francoforte)	eu-central-1	Scarica YAML	Launch Stack
Europa (Irlanda)	eu-west-1	Scarica YAML	Launch Stack
Europe (London)	eu-west-2	Scarica YAML	Launch Stack
Europa (Stoccolma)	eu-north-1	Scarica YAML	Launch Stack
AWS GovCloud (Stati Uniti orientali)	us-gov-east-1	Scarica YAML	Launch Stack
AWS GovCloud (Stati Uniti occidentali)	us-gov-west-1	Scarica YAML	Launch Stack

AWS Cluster PCS

Un cluster AWS PCS è costituito dai seguenti componenti:

- Istanze gestite del software di pianificazione del sistema HPC, come il daemon di controllo Slurm (). `slurmctld`
- Componenti che si integrano con lo scheduler del sistema HPC per il provisioning e la gestione delle istanze Amazon EC2.
- Componenti che si integrano con lo scheduler del sistema HPC per trasmettere log e metriche ad Amazon. CloudWatch

Questi componenti vengono eseguiti in un account gestito da. AWS Collaborano per gestire le EC2 istanze Amazon nel tuo account cliente. AWS PCS fornisce interfacce di rete elastiche nella sottorete Amazon VPC per fornire connettività dal software di pianificazione alle istanze EC2 Amazon (ad esempio, per supportare la pianificazione di lavori in batch su di esse e consentire agli utenti di eseguire comandi di pianificazione per elencare e gestire tali lavori).

Argomenti

Creazione di un cluster in AWS PCS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive gli elementi da considerare quando si crea un cluster in AWS Parallel Computing Service (AWS PCS). Se è la prima volta che crei un cluster AWS PCS, ti consigliamo di seguirloInizia a usare AWS Parallel Computing Service. Il tutorial può aiutarti a creare un sistema HPC funzionante senza approfondire tutte le opzioni e le architetture di sistema disponibili possibili.

> **Note**
>
> Dopo aver creato un cluster, è possibile modificare molte impostazioni di configurazione senza ricostruire l'infrastruttura. Per ulteriori informazioni, consulta Aggiornamento di un cluster in AWS PCS.

> **Note**
>
> È possibile configurare impostazioni Slurm personalizzate per implementare politiche di pianificazione avanzate e gestione delle risorse. Per ulteriori informazioni, consulta Configurazione delle impostazioni Slurm personalizzate in PCS AWS.

Prerequisiti

- Un VPC e una sottorete esistenti che soddisfano i requisiti. AWS Rete PCS Prima di implementare un cluster da utilizzare in produzione, ti consigliamo di approfondire le nozioni relative ai requisiti del VPC e delle sottoreti. Per creare un VPC e una sottorete, vedere. Creazione di un VPC per il AWS cluster PCS
- Un preside IAM con autorizzazioni per creare e gestire AWS risorse PCS. Per ulteriori informazioni, consulta Servizio di Identity and Access Management per AWS Parallel Computing.

Crea un cluster AWS PCS

È possibile utilizzare Console di gestione AWS o AWS CLI per creare un cluster.

Console di gestione AWS

Come creare un cluster

1. Apri la console AWS PCS a https://console.aws.amazon.com/pcs/home#/clusters e scegli Crea cluster.
2. Nella sezione Configurazione del cluster, inserisci i seguenti campi:
   - Nome del cluster: un nome per il cluster. Il nome può contenere solo caratteri alfanumerici (con distinzione tra lettere maiuscole e minuscole) e trattini. Deve iniziare con un carattere

alfabetico e non può superare i 40 caratteri. Il nome deve essere univoco all'interno del Regione AWS e in Account AWS cui si sta creando il cluster.

- Scheduler: scegli uno scheduler e una versione. Per ulteriori informazioni, consulta Versioni Slurm in PCS AWS.
- Dimensioni del controller: scegli una dimensione per il controller. Ciò determina il numero di lavori e nodi di elaborazione simultanei che possono essere gestiti dal cluster AWS PCS. È possibile impostare la dimensione del controller solo al momento della creazione del cluster. Per ulteriori informazioni sul dimensionamento, vedereDimensione del cluster in AWS PCS.

3. Nella sezione Rete, selezionate i valori per i seguenti campi:

   - Tipo di rete: scegli il tipo di indirizzo IP per il tuo cluster. Il cluster può utilizzare uno IPv4 o entrambi IPv6, ma non entrambi. Il VPC e le sottoreti devono utilizzare lo stesso tipo di indirizzo di rete. Il blocco di indirizzi IP utilizzato per ogni sottorete deve avere almeno un indirizzo disponibile. AWS riserva alcuni indirizzi in ogni sottorete. Per ulteriori informazioni, consulta Blocchi CIDR della sottorete nella Guida per l'utente di Amazon VPC.
   - VPC: scegli un VPC esistente che soddisfi i requisiti PCS. AWS Per ulteriori informazioni, consulta AWS Requisiti e considerazioni su PCS, VPC e sottorete. Dopo aver creato il cluster, non puoi modificarne il VPC. Se non VPCs ne è elencato nessuno, devi prima crearne uno.
   - Subnet: vengono elencate tutte le sottoreti disponibili nel VPC selezionato. Scegli una sottorete che soddisfi i requisiti della sottorete PCS. AWS Per ulteriori informazioni, consulta AWS Requisiti e considerazioni su PCS, VPC e sottorete. Ti consigliamo di selezionare una sottorete privata per evitare di esporre gli endpoint dello scheduler alla rete Internet pubblica.
   - Gruppi di sicurezza: specificate i gruppi di sicurezza che desiderate che AWS PCS associ alle interfacce di rete create per il cluster. È necessario selezionare almeno un gruppo di sicurezza che consenta la comunicazione tra il cluster e i relativi nodi di elaborazione. È possibile selezionare Creazione rapida di un gruppo di sicurezza per fare in modo che AWS PCS ne crei uno con la configurazione necessaria nel VPC selezionato oppure selezionare un gruppo di sicurezza esistente. Per ulteriori informazioni, consulta Requisiti e considerazioni sui gruppi di sicurezza.

4. (Facoltativo) Nella sezione di configurazione della contabilità Slurm, è possibile abilitare la contabilità Slurm e impostare i parametri di contabilità. Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

5. (Facoltativo) Nella sezione di configurazione di Slurm, è possibile aggiungere coppie di nomi e valori dei parametri per configurare impostazioni Slurm aggiuntive. Per un elenco completo dei parametri supportati, vedere. Impostazioni Slurm personalizzate per AWS cluster PCS
6. (Facoltativo) In Tag, aggiungi qualsiasi tag al tuo cluster AWS PCS.
7. Scegli Crea cluster. Il campo Status mostra `Creating` mentre il AWS PCS crea il cluster. Questo processo può richiedere alcuni minuti.

> **Important**
>
> Può esserci solo 1 cluster in uno `Creating` stato Regione AWS per ogni stato Account AWS. AWS PCS restituisce un errore se c'è già un cluster in uno `Creating` stato quando si tenta di creare un cluster.

AWS CLI

Come creare un cluster

1. Crea un cluster con il comando seguente. Prima di eseguire il comando, apporta le modifiche seguenti:
   - Sostituiscilo *`region`* con l'ID in Regione AWS cui desideri creare il cluster, ad esempio`us-east-1`.
   - Sostituisci *`my-cluster`* con un nome da assegnare al cluster. Il nome può contenere solo caratteri alfanumerici (con distinzione tra lettere maiuscole e minuscole) e trattini. Deve iniziare con un carattere alfabetico e non può superare i 40 caratteri. Il nome deve essere univoco all'interno Regione AWS e nel Account AWS luogo in cui si sta creando il cluster.
   - *`25.05`*Sostituiscilo con qualsiasi versione supportata di Slurm.

     > **Note**
     >
     > AWS PCS attualmente supporta Slurm 25.05 e 24.11.

   - Sostituiscilo *`SMALL`* con qualsiasi dimensione di cluster supportata. Ciò determina quanti processi e nodi di calcolo simultanei possono essere gestiti dal cluster AWS PCS. Può essere impostato solo al momento della creazione del cluster. Per ulteriori informazioni sul dimensionamento, vedereDimensione del cluster in AWS PCS.

- Sostituisci il valore di `subnetIds` con il tuo. Ti consigliamo di selezionare una sottorete privata per evitare di esporre gli endpoint dello scheduler alla rete Internet pubblica.
- Specificate `securityGroupIds` quello che desiderate che AWS PCS associ alle interfacce di rete che crea per il cluster. I gruppi di sicurezza devono trovarsi nello stesso VPC del cluster. È necessario selezionare almeno un gruppo di sicurezza che consenta la comunicazione tra il cluster e i relativi nodi di calcolo. Per ulteriori informazioni, consulta Requisiti e considerazioni sui gruppi di sicurezza.

```
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
```

- per utilizzarlo IPv6, aggiungilo `networkType=IPV6` alla `--networking` configurazione.

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

- Facoltativamente, è possibile aggiungere l'`--slurm-configration`opzione per personalizzare il comportamento di Slurm e specificare le opzioni di configurazione di Slurm. L'esempio seguente imposta il tempo di inattività della scala ridotta a 60 minuti (3600 secondi), abilita la contabilità Slurm e specifica le impostazioni come valore per. `slurm.conf` `slurmCustomSettings` Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

> **Note**
>
> La contabilità è supportata per Slurm 24.11 o versioni successive.

```
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. Il provisioning del cluster può richiedere diversi minuti. È possibile eseguire query sullo stato del cluster con il comando seguente. Non procedere alla creazione di code o gruppi di nodi di calcolo finché non viene visualizzato il campo di stato del cluster. `ACTIVE`

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

> **Important**
>
> Può esserci solo 1 cluster in uno `Creating` stato per ogni stato. Regione AWS Account AWS AWS PCS restituisce un errore se c'è già un cluster in uno `Creating` stato quando si tenta di creare un cluster.

Passaggi successivi consigliati per il cluster

- Aggiungi gruppi di nodi di calcolo.
- Aggiungi code.
- Attivare la registrazione nel log.

Aggiornamento di un cluster in AWS PCS

AWS PCS consente di aggiornare le configurazioni del cluster dopo la creazione tramite l' UpdateCluster API o la console. È possibile modificare le impostazioni del cluster senza ricostruire l'infrastruttura, il che riduce il sovraccarico operativo e minimizza le interruzioni.

Vantaggi degli aggiornamenti dei cluster

L'aggiornamento dei cluster AWS PCS consente di adattare l'infrastruttura HPC ai nuovi requisiti senza interruzioni del servizio. Le modifiche alla configurazione richiedono pochi minuti anziché l'ora o più necessaria per ricostruire i cluster. Questa funzionalità è importante per gli ambienti di produzione che richiedono tempi di inattività minimi e per i team che devono modificare le impostazioni dei cluster al variare dei modelli di carico di lavoro.

Modifiche alla configurazione supportate

È possibile modificare tre categorie principali di impostazioni:

- Configurazione della contabilità: abilita o disabilita la contabilità gestita e configura le impostazioni di conservazione.
- Comportamento ridotto: modifica il `scaleDownIdleTime` parametro, che controlla per quanto tempo le istanze dinamiche rimangono inattive prima che AWS PCS le interrompa automaticamente.
- Impostazioni personalizzate Slurm: modifica tutte le impostazioni Slurm supportate che si applicano a livello di cluster, tra cui Prolog, Epilog e. SelectTypeParameters

Limitazioni

Non è possibile modificare determinate configurazioni dopo la creazione del cluster. Ciò include:

- Configurazioni dei gruppi di sicurezza
- Selezione della sottorete VPC
- Dimensione del cluster
- Versione Slurm
- Nome cluster

Queste impostazioni sono fondamentali per l'architettura del cluster e richiedono la creazione di un nuovo cluster per modificarle.

Prerequisiti per gli aggiornamenti del cluster

Prima di aggiornare un cluster, assicurati che siano soddisfatte le seguenti condizioni:

- Il cluster deve essere in `ACTIVEUPDATE_FAILED`, o deve essere `SUSPENDED` stato
- Tutte le risorse associate (Queues, Compute Node Groups) devono essere in stato `ACTIVE`
- È necessario disporre delle autorizzazioni IAM appropriate per l'operazione UpdateCluster
- Non possono essere in corso altre operazioni di aggiornamento

Processo di aggiornamento e impatto sul lavoro

Durante un'operazione di aggiornamento, i nodi di elaborazione continuano a eseguire i job esistenti anche quando il controller del cluster diventa irraggiungibile per un breve periodo. Tuttavia, il sistema non può accettare nuove candidature di lavoro o prendere decisioni di pianificazione durante questo periodo.

È possibile monitorare gli aggiornamenti del cluster tramite le interfacce della console e dell'API. Il cluster passerà attraverso i seguenti stati durante un aggiornamento:

- `UPDATING`- Aggiornamento in corso
- `ACTIVE`- Aggiornamento completato con successo
- `UPDATE_FAILED`- L'aggiornamento ha rilevato un errore

Fatturazione durante gli aggiornamenti

Le tariffe orarie standard per il cluster AWS PCS continuano durante le operazioni di aggiornamento. Quando aggiorni un cluster per disabilitare la contabilità, la fatturazione per la funzionalità di contabilità si interrompe non appena il cluster entra nello stato. `UPDATING` Quando si abilita la contabilità, la fatturazione non inizia finché il cluster non completa correttamente l'aggiornamento e torna allo stato. `ACTIVE`

Argomenti

- Aggiornare un cluster AWS PCS
- Domande frequenti sull'aggiornamento dei cluster in AWS PCS
- Risoluzione dei problemi degli aggiornamenti del cluster AWS PCS

Aggiornare un cluster AWS PCS

Usa questi passaggi per modificare le impostazioni dello scheduler, la configurazione contabile e le impostazioni personalizzate di Slurm sul tuo cluster. Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per AWS cluster PCS.

Prerequisiti

- Il cluster deve essere in`ACTIVE`, `UPDATE_FAILED` o deve essere in stato `SUSPENDED`
- Tutte le risorse associate (Queues, Compute Node Groups) devono essere in stato `ACTIVE`

- Non possono essere in corso altre operazioni di aggiornamento

Procedura

Console di gestione AWS

1. Apri la console AWS PCS all'indirizzo https://console.aws.amazon.com/pcs/.
2. Nel pannello di navigazione scegliere Cluster.
3. Seleziona il cluster da aggiornare.
4. Scegli Modifica.
5. Nella pagina Modifica cluster, modifica le impostazioni desiderate:
   - Nella configurazione Scheduler, aggiorna il tempo di inattività di Scale-down per controllare per quanto tempo le istanze dinamiche rimangono inattive prima della chiusura automatica.
   - Modificate le impostazioni dei parametri Prolog, Epilog e Select-type secondo necessità.
   - Abilita, disabilita o configura il tempo di conservazione per la contabilità gestita.
   - In Impostazioni aggiuntive dello scheduler, aggiungi, modifica o rimuovi le impostazioni personalizzate di Slurm. Per ulteriori informazioni sui parametri supportati, vedere. Impostazioni Slurm personalizzate per AWS cluster PCS

   > **Note**
   >
   > I campi che non possono essere modificati vengono visualizzati in sola lettura e mostrano i valori correnti.
6. Scegli Aggiorna per inviare le modifiche.
7. Monitora lo stato del cluster, che viene visualizzato come «Aggiornamento» durante il processo. Lo stato cambia quando l'aggiornamento viene completato correttamente.

AWS CLI

1. Aprire un terminale o un prompt dei comandi.
2. Verifica lo stato del cluster utilizzando il seguente comando:

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

3. Invia una richiesta di aggiornamento utilizzando uno dei seguenti esempi:

   - Per abilitare la contabilità gestita:

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

   - Per aggiornare un'impostazione di Slurm Prolog:

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

   - Per aggiornare il tempo di inattività con scale-down:

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

4. Monitora l'avanzamento dell'aggiornamento controllando lo stato del cluster:

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

Dopo una richiesta di aggiornamento riuscita, il comando restituisce l'oggetto Cluster con tutte le modifiche. Lo stato del cluster cambia da `UPDATING` a `ACTIVE` quando è completo.

Domande frequenti sull'aggiornamento dei cluster in AWS PCS

Ottieni risposte alle domande più comuni sull'aggiornamento delle configurazioni dei cluster in AWS PCS.

Quali impostazioni posso modificare?

È possibile modificare la configurazione della contabilità (abilitare/disabilitare la contabilità gestita), il comportamento di ridimensionamento (parametro scaleDownIdle Time) e qualsiasi impostazione personalizzata Slurm supportata che si applica a livello di cluster. Non è possibile modificare i gruppi di sicurezza, le sottoreti VPC, la dimensione del cluster, la versione Slurm o il nome del cluster.

Posso mettere in coda più aggiornamenti?

No. È necessario attendere che il cluster ritorni allo `ACTIVE` stato prima di inviare un altro aggiornamento. Anche tutte le risorse associate (Queues, Compute Node Groups) devono essere in stato. `ACTIVE`

Posso annullare un'operazione di aggiornamento del cluster?

No, non è possibile annullare un'operazione di aggiornamento del cluster in corso.

Posso inviare lavori mentre il mio cluster è in fase di aggiornamento?

Ti consigliamo di evitare di inviare lavori durante gli aggiornamenti del cluster. Il controller Slurm potrebbe non essere disponibile durante il processo di aggiornamento.

I miei lavori continueranno a funzionare durante gli aggiornamenti del cluster?

Sì, i lavori in esecuzione continuano a essere eseguiti sui nodi di calcolo anche quando il controller del cluster diventa irraggiungibile per un breve periodo durante il processo di aggiornamento. Tuttavia, lo stato del processo potrebbe non aggiornarsi fino a quando il controller non sarà nuovamente disponibile.

In che modo viene influenzata la fatturazione durante gli aggiornamenti?

Le tariffe orarie standard continuano durante le operazioni di aggiornamento. Quando si disabilita la contabilità, la fatturazione si interrompe quando il cluster entra in stato. `UPDATING` Quando si abilita la contabilità, la fatturazione inizia quando il cluster torna correttamente allo stato. `ACTIVE`

Risoluzione dei problemi degli aggiornamenti del cluster AWS PCS

Questo argomento consente di identificare e risolvere i problemi più comuni che possono verificarsi durante l'aggiornamento delle configurazioni del cluster.

L'aggiornamento non riesce a causa di un errore di configurazione contabile

Cause comuni

Il cluster entra `UPDATE_FAILED` nello stato e il messaggio di errore indica un problema di configurazione dell'account. Ciò si verifica in genere quando la configurazione dell'accounting è incompatibile con la versione corrente di Slurm o contiene impostazioni non valide.

Risoluzione

Controlla le impostazioni di contabilità per verificarne la compatibilità con la versione Slurm del tuo cluster e invia una richiesta di aggiornamento corretta con parametri di configurazione validi.

L'aggiornamento non riesce a causa di un errore nelle impostazioni personalizzate

Cause comuni

Il cluster entra `UPDATE_FAILED` nello stato e il messaggio di errore indica un problema di impostazioni personalizzate di Slurm. Ciò si verifica quando si forniscono valori dei parametri Slurm non validi o combinazioni di parametri non supportate.

Risoluzione

Convalida le impostazioni personalizzate di Slurm rispetto ai parametri supportati e invia una richiesta di aggiornamento corretta con valori e combinazioni di parametri validi.

Impossibile inviare una richiesta di aggiornamento

Cause comuni

Il pulsante di aggiornamento è disabilitato nella console o l'API restituisce un errore di 400 livelli. Ciò si verifica quando il cluster non si trova in uno stato appropriato, le risorse associate non sono attive o si verificano errori di convalida nella configurazione.

Risoluzione

Attendi che il cluster e tutte le risorse associate raggiungano `ACTIVE` lo stato, quindi esamina la configurazione per individuare eventuali errori di convalida prima di inviare nuovamente la richiesta di aggiornamento.

Errori di convalida

Cause comuni

Il comando viene restituito immediatamente con un errore HTTP di 400 livelli e un messaggio descrittivo. Ciò si verifica a causa dello stato del cluster, dello stato della risorsa o dei parametri di configurazione non validi.

Risoluzione

Risolvi l'errore di convalida specifico menzionato nella risposta e riprova l'operazione di aggiornamento.

Eliminazione di un cluster in AWS PCS

Questo argomento fornisce una panoramica su come eliminare un cluster AWS PCS.

Considerazioni sull'eliminazione di un AWS cluster PCS

- Tutte le code associate al cluster devono essere eliminate prima che il cluster possa essere eliminato. Per ulteriori informazioni, consulta Eliminazione di una coda in PCS AWS.
- Tutti i gruppi di nodi di calcolo associati al cluster devono essere eliminati prima che il cluster possa essere eliminato. Per ulteriori informazioni, consulta Eliminazione di un gruppo di nodi di calcolo in PCS AWS.

Eliminare il cluster

È possibile utilizzare Console di gestione AWS o AWS CLI per eliminare un cluster.

Console di gestione AWS

Come eliminare un cluster

1. Aprire la console AWS PCS.
2. Seleziona il cluster da eliminare.
3. Scegli Elimina.
4. Viene visualizzato il campo Stato del cluster`Deleting`. Per il completamento possono essere necessari alcuni minuti.

AWS CLI

Come eliminare un cluster

1. Utilizzate il seguente comando per eliminare un cluster, con queste sostituzioni:
   - Sostituisci *`region-code`* con Regione AWS il cluster in cui si trova.

- Sostituiscilo *my-cluster* con il nome o l'ID del tuo cluster.

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

2. L'eliminazione del cluster può richiedere diversi minuti. Puoi controllare lo stato del tuo cluster con il seguente comando.

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

Dimensione del cluster in AWS PCS

AWS PCS fornisce cluster ad alta disponibilità e sicuri, automatizzando al contempo attività chiave come l'applicazione di patch, il provisioning dei nodi e gli aggiornamenti.

Quando si crea un cluster, si seleziona una dimensione in base a due fattori:

- Il numero di nodi di elaborazione che gestirà
- Il numero di lavori attivi e in coda che si prevede di eseguire nel cluster

> **Important**
>
> Non è possibile modificare la dimensione del cluster dopo averlo creato. Se è necessario modificare le dimensioni, è necessario creare un nuovo cluster.

Dimensione del cluster Slurm	Numero di istanze gestite	Numero di lavori attivi e in coda
Small	Fino a 32	Fino a 256
Media	Fino a 512	Fino a 8192
Large	Fino al 2048	Fino a 16384

Esempi

- Se il tuo cluster avrà fino a 24 istanze gestite ed eseguirà fino a 100 job, scegli Small.
- Se il cluster avrà fino a 24 istanze gestite e gestirà fino a 1000 job, scegli Medium.
- Se il cluster avrà fino a 1000 istanze gestite e gestirà fino a 100 job, scegli Large.
- Se il tuo cluster avrà fino a 1000 istanze gestite e gestirà fino a 10.000 job, scegli Large.

Utilizzo dei segreti del cluster in AWS PCS

Come parte della creazione di un cluster, AWS PCS crea un cluster secret necessario per connettersi al job scheduler del cluster. È inoltre possibile creare gruppi di nodi di calcolo AWS PCS, che definiscono set di istanze da avviare in risposta a eventi di scalabilità. AWS PCS configura le istanze lanciate da tali gruppi di nodi di calcolo con il cluster secret in modo che possano connettersi al job scheduler. In alcuni casi potresti voler configurare i client Slurm manualmente. Gli esempi includono la creazione di un nodo di accesso persistente o la configurazione di un gestore del flusso di lavoro con funzionalità di gestione dei lavori.

AWS PCS memorizza il segreto del cluster come segreto gestito con il prefisso inseritopcs!. Gestione dei segreti AWS Il costo del segreto è incluso nel costo per l'utilizzo di AWS PCS. È possibile modificare i segreti del cluster Gestione dei segreti AWS per mantenere la conformità alla sicurezza e correggere potenziali compromessi in materia di sicurezza.

Argomenti

- Usa per trovare Gestione dei segreti AWS il segreto del cluster
- Usa AWS PCS per trovare il segreto del cluster
- Ottieni il segreto del cluster Slurm
- Segreti dei cluster rotanti in AWS PCS

Usa per trovare Gestione dei segreti AWS il segreto del cluster

Console di gestione AWS

1. Vai alla console Secrets Manager.
2. Scegli Segreti, quindi cerca il `pcs!` prefisso.

> **Note**
>
> Un segreto del cluster AWS PCS ha un nome nel formato in `pcs!slurm-secret-`*`cluster-id`* cui *`cluster-id`* è l'ID del cluster AWS PCS.

AWS CLI

Ogni segreto del cluster AWS PCS è inoltre etichettato con`aws:pcs:`*`cluster-id`*. È possibile ottenere l'ID segreto di un cluster con il comando seguente. Effettua queste sostituzioni prima di eseguire il comando:

- Sostituisci *`region`* con il Regione AWS per creare il cluster, ad esempio. `us-east-1`
- Sostituisci *`cluster-id`* con l'ID del cluster AWS PCS per trovare il segreto del cluster.

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

Usa AWS PCS per trovare il segreto del cluster

È possibile utilizzare il AWS CLI per trovare l'ARN di un segreto del cluster AWS PCS. Immettete il comando che segue, effettuando le seguenti sostituzioni:

- Sostituisci *`region`* con il Regione AWS per creare il tuo cluster, ad esempio. `us-east-1`
- Sostituiscilo *`my-cluster`* con il nome o l'identificatore del cluster.

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

L'output di esempio seguente proviene dal `get-cluster` comando. Potete usare `secretArn` e `secretVersion` insieme per ottenere il segreto.

```
{
    "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"
            }
        ]
    }
}
```

Ottieni il segreto del cluster Slurm

È possibile utilizzare Secrets Manager per ottenere la versione corrente con codifica base64 di un segreto del cluster Slurm. L'esempio seguente utilizza il. AWS CLI Effettua le seguenti sostituzioni prima di eseguire il comando.

- Sostituisci *`region`* con il Regione AWS per creare il cluster, ad esempio. `us-east-1`
- Sostituisci *`secret-arn`* con quello `secretArn` proveniente da un cluster AWS PCS.

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

Per informazioni su come utilizzare il segreto del cluster Slurm, vedere. Utilizzo di istanze autonome come nodi di accesso AWS PCS

Permissions

Si utilizza un principale IAM per ottenere il segreto del cluster Slurm. Il preside IAM deve avere il permesso di leggere il segreto. Per ulteriori informazioni, consulta i termini e i concetti relativi ai ruoli nella Guida AWS Identity and Access Management per l'utente.

La seguente policy IAM di esempio consente l'accesso a un cluster secret di esempio.

```
{
    "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"
        }
    ]
}
```

Segreti dei cluster rotanti in AWS PCS

Usa Gestione dei segreti AWS Managed Rotation per ruotare i segreti del cluster in AWS PCS. La rotazione segreta regolare è una best practice di sicurezza per mantenere un elevato livello di sicurezza negli ambienti HPC. Questa funzionalità consente di soddisfare gli standard di conformità del settore, tra cui HIPAA e FedRAMP, che impongono una rotazione regolare delle credenziali.

Il segreto del cluster ha un duplice scopo: autenticare i nodi di calcolo che si uniscono al cluster e fungere da chiave JWT per l'autenticazione dell'API REST di Slurm. Quando viene ruotato, entrambi gli aspetti vengono influenzati contemporaneamente.

Come funziona la rotazione segreta dei cluster

Preparati manualmente per mantenere la stabilità del cluster durante la rotazione segreta:

1. Preparazione: ridimensiona tutti i gruppi di nodi di elaborazione fino a una capacità pari a 0 e assicurati che nessun processo sia in esecuzione
2. Rotazione: avvia la rotazione tramite la console o l'API di Secrets Manager
3. Monitoraggio: monitora i progressi attraverso gli eventi CloudTrail
4. Ripristino: ridimensiona i gruppi di nodi di elaborazione fino alla capacità desiderata

Durante la rotazione, il cluster rimane invariato e `ACTIVE` la fatturazione prosegue normalmente. Il processo richiede in genere alcuni minuti.

Requisiti e limitazioni

Prima di modificare i segreti dei cluster, completa i seguenti requisiti:

- Il cluster deve essere nel nostro `ACTIVE` stato `UPDATE_FAILED`
- Il ruolo IAM deve avere `secretsmanager:RotateSecret` l'autorizzazione
- Tutti i gruppi di nodi di calcolo devono essere scalati fino a una capacità pari a 0
- Interrompi tutti i lavori prima della rotazione

Restrizioni:

- Preparazione manuale richiesta per ogni rotazione
- I token JWT esistenti non sono più validi e richiedono una riemissione
- I nodi di accesso BYO richiedono un aggiornamento segreto manuale dopo la rotazione

Argomenti

- Ruota un cluster segreto in AWS PCS
- Domande frequenti sulla rotazione segreta dei cluster in AWS PCS

- Risoluzione dei problemi di rotazione segreta del cluster in AWS PCS

Ruota un cluster segreto in AWS PCS

Ruota il segreto del cluster per soddisfare i requisiti di sicurezza e risolvere potenziali compromessi. Questo processo richiede l'attivazione della modalità di manutenzione del cluster.

Prerequisiti

- Ruolo IAM con `secretsmanager:RotateSecret` autorizzazione
- Cluster nel `ACTIVE` nostro `UPDATE_FAILED` stato

Procedura

1. Notifica agli utenti del cluster la prossima finestra di manutenzione.
2. Metti il cluster in modalità di manutenzione scalando tutti i gruppi di nodi di calcolo a 0 capacità.
    a. Usa l' UpdateComputeNodeGroup API per impostare entrambi minInstanceCount e su 0 maxInstanceCount per tutti i gruppi di nodi di calcolo.
    b. Attendi che tutti i nodi si fermino.
    c. Facoltativo: svuota le code dello scheduler con i comandi Slurm prima di terminare la capacità per una corretta gestione dei lavori.
3. Avvia la rotazione tramite Secrets Manager.
    - Metodo della console:
        - Vai a Secrets Manager, seleziona il segreto del cluster e scegli Ruota segreto.
    - Metodo API:
        - Usa l'`rotate-secret`API Secrets Manager.
4. Monitora l'avanzamento della rotazione.
    a. Tieni traccia dei progressi attraverso CloudTrail gli eventi.
    b. Verifica `lastRotatedDate` tramite la console Secrets Manager o l'`secretsmanager:describeSecret`API.
    c. Attendi `RotationSucceeded` il `RotationFailed` CloudTrail nostro evento.
5. Dopo una rotazione riuscita, ripristina la capacità del cluster.

a. Utilizza l' UpdateComputeNodeGroup API per ripristinare i gruppi di nodi alla min/max capacità desiderata.
b. Per i nodi di accesso AWS gestiti da PCS: non è richiesta alcuna azione aggiuntiva.
c. Per i nodi di accesso BYO:

    i. Connect ai nodi di accesso.
    ii. Aggiorna `/etc/slurm/slurm.key` con il nuovo segreto di Secrets Manager.
    iii. Riavvia Slurm Auth e Cred Kiosk Daemon (sackd).

Domande frequenti sulla rotazione segreta dei cluster in AWS PCS

Trova le risposte alle domande più comuni sulla rotazione segreta dei cluster in AWS PCS.

Cos'è un cluster secret?

Un cluster secret è una credenziale sicura che consente comunicazioni sicure tra il controller Slurm e i nodi di calcolo AWS PCS. Serve anche come chiave JSON Web Token (JWT) per l'autenticazione dell'API REST di Slurm.

Qual è la differenza tra il segreto del cluster e la chiave JWT?

In AWS PCS, il segreto del cluster e la chiave JWT sono la stessa risorsa che serve a scopi diversi. Il segreto del cluster autentica le comunicazioni interne di Slurm, mentre la chiave JWT firma i token per l'autenticazione dell'API REST. Quando viene ruotato, entrambi gli aspetti vengono influenzati contemporaneamente.

Quanto dura la rotazione?

Il processo di rotazione richiede in genere alcuni minuti. Il cluster rimane nello stato ATTIVO e la fatturazione continua normalmente durante la rotazione.

Posso programmare le rotazioni automatiche?

È possibile abilitare la rotazione pianificata in Secrets Manager. Tuttavia, la versione iniziale richiede una preparazione manuale (ridimensionamento dei gruppi di nodi a 0) prima di ogni rotazione.

I miei token JWT esistenti funzioneranno ancora dopo la rotazione?

No, i token JWT esistenti non sono più validi dopo la rotazione. Emetti nuovi token per i client API REST.

Dove posso trovare il segreto del mio cluster?

È possibile trovare il segreto del cluster nella console Secrets Manager o tramite la console AWS PCS. Per istruzioni dettagliate, consulta Usa per trovare Gestione dei segreti AWS il segreto del cluster eUsa AWS PCS per trovare il segreto del cluster.

Perché la rotazione richiede il ridimensionamento dei gruppi di nodi a 0?

La rotazione non richiede l'esecuzione di istanze per garantire la stabilità del cluster durante il processo di aggiornamento segreto. In questo modo si evitano conflitti di autenticazione tra vecchi e nuovi segreti.

Quali requisiti di conformità supporta questa funzionalità?

Questa funzionalità consente a AWS PCS di soddisfare gli standard di conformità del settore, tra cui HIPAA e FedRAMP, che impongono la rotazione regolare delle credenziali come parte dei controlli di sicurezza.

Risoluzione dei problemi di rotazione segreta del cluster in AWS PCS

La rotazione segreta del cluster fallisce se l'ambiente non è preparato correttamente. La causa più comune sono le istanze attive nel cluster. Per evitare errori:

1. Imposta tutti i gruppi di nodi sulla capacità 0.
2. Attendi che i nodi si fermino.
3. Verifica che il cluster non sia in questi stati: `CREATE_FAILED`,`DELETE_FAILED`,`RESUMING`,`SUSPENDING`, o`SUSPENDED`.

Se la rotazione fallisce:

- Viene visualizzato un RotationFailed CloudTrail evento
- Il segreto del cluster rimane invariato
- Controlla l' RotationFailed evento CloudTrail per i dettagli
- Completa tutte le fasi di preparazione per una rotazione corretta

AWS Gruppi di nodi di calcolo PCS

Un gruppo di nodi di calcolo AWS PCS è una raccolta logica di nodi (EC2 istanze Amazon). Questi nodi possono essere utilizzati per eseguire processi di elaborazione e per fornire un accesso interattivo basato su shell a un sistema HPC. Un gruppo di nodi di calcolo è costituito da regole per la creazione di nodi, tra cui quali tipi di EC2 istanze Amazon utilizzare, quante istanze eseguire, se utilizzare istanze Spot o istanze On-demand, quali sottoreti e gruppi di sicurezza utilizzare e come configurare ogni istanza all'avvio. Quando tali regole vengono aggiornate, AWS PCS aggiorna le risorse associate al gruppo di nodi di calcolo in modo che corrispondano.

Argomenti

Creazione di un gruppo di nodi di calcolo in AWS PCS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive gli elementi da considerare quando si crea un gruppo di nodi di calcolo in AWS Parallel Computing Service (AWS PCS). Se è la prima volta che crei un gruppo di nodi di calcolo in AWS PCS, ti consigliamo di seguire il tutorial in. Inizia a usare AWS Parallel Computing Service Il tutorial può aiutarti a creare un sistema HPC funzionante senza approfondire tutte le opzioni disponibili e le architetture di sistema possibili.

> **Note**
>
> È possibile configurare impostazioni Slurm personalizzate sui gruppi di nodi di calcolo per controllare l'utilizzo delle risorse e i comportamenti a livello di nodo. Per ulteriori informazioni, consulta Configurazione delle impostazioni Slurm personalizzate in PCS AWS.

> **Important**
>
> AWS PCS attualmente richiede un kernel con IPv4 supporto per la comunicazione tra nodi locali, anche quando si utilizza PCS in una rete di sola rete. AWS IPv6 Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Prerequisiti

- Quote di servizio sufficienti per avviare il numero desiderato di istanze EC2 nel tuo. Regione AWS Puoi utilizzarle Console di gestione AWSper controllare e richiedere aumenti delle quote di servizio.
- Un VPC e una o più sottoreti esistenti che soddisfano i requisiti di rete AWS PCS. Si consiglia di comprendere a fondo questi requisiti prima di implementare un cluster per l'uso in produzione. Per ulteriori informazioni, consulta AWS Requisiti e considerazioni su PCS, VPC e sottorete. Puoi anche usare un CloudFormation modello per creare un VPC e delle sottoreti. AWS fornisce una ricetta HPC per il modello. CloudFormation Per ulteriori informazioni, vedere aws-hpc-recipeson. GitHub
- Un profilo di istanza IAM con autorizzazioni per richiamare l'azione dell'`RegisterComputeNodeGroupInstance`API AWS PCS e accedere a qualsiasi altra AWS risorsa richiesta per le istanze del gruppo di nodi. Per ulteriori informazioni, consulta Profili di istanza IAM per AWS Parallel Computing Service.
- Un modello di avvio per le istanze del gruppo di nodi. Per ulteriori informazioni, consulta Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS.
- Per creare un gruppo di nodi di calcolo che utilizzi istanze Spot di Amazon EC2, devi avere AWSServiceRoleForEC2il ruolo collegato al servizio Spot nel tuo. Account AWS Per ulteriori informazioni, consulta Ruolo Spot di Amazon EC2 per PCS AWS.

Crea un gruppo di nodi di calcolo in PCS AWS

È possibile creare un gruppo di nodi di calcolo utilizzando Console di gestione AWS o il. AWS CLI

Console di gestione AWS

Per creare il gruppo di nodi di calcolo utilizzando la console

1. Apri la console AWS PCS.

2. Seleziona il cluster in cui desideri creare un gruppo di nodi di calcolo. Passa ai gruppi di nodi di calcolo e scegli Crea.
3. Nella sezione Configurazione del gruppo di nodi di calcolo, fornisci un nome per il tuo gruppo di nodi. Il nome può contenere solo caratteri alfanumerici e trattini con distinzione tra maiuscole e minuscole. Deve iniziare con un carattere alfabetico e non può superare i 25 caratteri. Il nome deve essere univoco all'interno del cluster.
4. In Computing configuration, inserisci o seleziona questi valori:

   a. Modello di lancio EC2: seleziona un modello di avvio personalizzato da utilizzare per questo gruppo di nodi. I modelli di avvio possono essere utilizzati per personalizzare le impostazioni di rete come sottorete e gruppi di sicurezza, configurazione di monitoraggio e archiviazione a livello di istanza. Se non hai preparato un modello di lancio, scopri come Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS crearne uno.

   > **Important**
   >
   > AWS PCS crea un modello di avvio gestito per ogni gruppo di nodi di calcolo. Questi sono `pcs-identifier-do-not-delete` denominati. Non selezionarli quando crei o aggiorni un gruppo di nodi di calcolo, altrimenti il gruppo di nodi non funzionerà correttamente.

   b. Versione del modello di lancio EC2: devi selezionare una versione del modello di lancio personalizzato. Se modifichi la versione in un secondo momento, devi aggiornare il gruppo di nodi di calcolo per rilevare le modifiche nel modello di lancio. Per ulteriori informazioni, consulta Aggiornamento di un gruppo di nodi di calcolo AWS PCS.

   c. ID AMI: se il modello di lancio non include un ID AMI o se desideri sovrascrivere il valore nel modello di lancio, fornisci qui un ID AMI. Nota che l'AMI utilizzata per il gruppo di nodi deve essere compatibile con AWS PCS. Puoi anche selezionare un AMI di esempio fornito da AWS. Per ulteriori informazioni su questo argomento, vedereAmazon Machine Images (AMIs) per AWS PCS.

   d. Profilo di istanza IAM: scegli un profilo di istanza per il gruppo di nodi. Un profilo di istanza concede all'istanza le autorizzazioni per accedere a AWS risorse e servizi in modo sicuro. Se non ne hai uno pronto, puoi selezionare Crea un profilo di base per fare in modo che AWS PCS ne crei uno per te con la politica minima, oppure consulta. Profili di istanza IAM per AWS Parallel Computing Service

e. Sottoreti: scegli una o più sottoreti nel VPC in cui è distribuito il cluster PCS. AWS Se si selezionano più sottoreti, le comunicazioni EFA non saranno disponibili tra i nodi e la comunicazione tra nodi in sottoreti diverse potrebbe avere una latenza maggiore. Assicurati che le sottoreti che specifichi qui corrispondano a quelle definite nel modello di lancio EC2.

f. Istanze: scegli uno o più tipi di istanze per soddisfare le richieste di scalabilità nel gruppo di nodi. Tutti i tipi di istanza devono avere la stessa architettura del processore (x86_64 o arm64) e lo stesso numero di v. CPUs Se le istanze lo sono GPUs, tutti i tipi di istanza devono avere lo stesso numero di. GPUs

g. Configurazione di scalabilità: specifica il numero minimo e massimo di istanze per il gruppo di nodi. È possibile definire una configurazione statica, in cui è in esecuzione un numero fisso di nodi, o una configurazione dinamica, in cui è possibile eseguire fino al numero massimo di nodi. Per una configurazione statica, imposta minimo e massimo sullo stesso numero, maggiore di zero. Per una configurazione dinamica, imposta il numero minimo di istanze su zero e il numero massimo di istanze su un numero maggiore di zero. AWS PCS non supporta gruppi di nodi di calcolo con un mix di istanze statiche e dinamiche.

5. (Facoltativo) In Impostazioni aggiuntive, specifica quanto segue:

   a. Opzione di acquisto: seleziona Istanze On-Demand, Istanze Spot o un Capacity Block esistente. Scegli anche On-Demand se prevedi di utilizzare una On-Demand Capacity Reservation (ODCR). Per ulteriori informazioni, consulta Utilizzo ODCRs con AWS PCS. Scegli Capacity Block per utilizzare una prenotazione esistente di Amazon EC2 Capacity Blocks for ML. Per ulteriori informazioni, consulta Utilizzo dei blocchi di capacità di Amazon EC2 per ML con PCS AWS.

   b. Strategia di allocazione: se hai selezionato l'opzione di acquisto Spot, puoi specificare come vengono scelti i pool di capacità Spot al momento del lancio delle istanze nel gruppo di nodi. Per ulteriori informazioni, consulta Strategie di allocazione per istanze Spot nella Guida per l'utente di Amazon Elastic Compute Cloud. Questa opzione non ha effetto se hai selezionato l'opzione di acquisto On-demand.

6. (Facoltativo) Nella sezione delle impostazioni Slurm personalizzate, è possibile aggiungere coppie di nomi e valori dei parametri per configurare impostazioni Slurm aggiuntive. Per un elenco completo dei parametri supportati, vedere. Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS

7. (Facoltativo) In Tag, aggiungi qualsiasi tag al gruppo di nodi di calcolo.

8. Scegli Crea gruppo di nodi di calcolo. Il campo Status viene visualizzato `Creating` mentre AWS PCS esegue il provisioning del gruppo di nodi. Questo processo può richiedere diversi minuti.

Fase successiva consigliata

- Aggiungi il tuo gruppo di nodi a una coda in AWS PCS per consentirgli di elaborare i lavori.

AWS CLI

Per creare il tuo gruppo di nodi di calcolo utilizzando AWS CLI

Crea la tua coda con il comando che segue. Prima di eseguire il comando, apporta le modifiche seguenti:

1. Sostituisci *`region`* con l'ID di in Regione AWS cui creare il cluster, ad esempio`us-east-1`.
2. Sostituiscilo *`my-cluster`* con il nome o con il nome `clusterId` del cluster.
3. Sostituiscilo *`my-node-group`* con il nome del tuo gruppo di nodi di calcolo. Il nome può contenere solo caratteri alfanumerici (con distinzione tra lettere maiuscole e minuscole) e trattini. Deve iniziare con un carattere alfabetico e non può essere più lungo di 25 caratteri. Il nome deve essere univoco all'interno del cluster.
4. Sostituisci *`subnet-ExampleID1`* con una o più sottoreti IDs dal tuo VPC del cluster.
5. *`lt-ExampleID1`*Sostituiscilo con l'ID del tuo modello di lancio personalizzato. Se non ne hai uno già pronto, scopri Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS come crearne uno.

   > **Important**
   >
   > AWS PCS crea un modello di avvio gestito per ogni gruppo di nodi di calcolo. Questi sono `pcs-`*`identifier`*`-do-not-delete` denominati. Non selezionarli quando crei o aggiorni un gruppo di nodi di calcolo, altrimenti il gruppo di nodi non funzionerà correttamente.

6. *`launch-template-version`*Sostituiscilo con una versione specifica del modello di lancio. AWS PCS associa il gruppo di nodi a quella versione specifica del modello di lancio.

7. Sostituisci *arn:InstanceProfile* con l'ARN del tuo profilo di istanza IAM. Se non ne hai uno pronto, consulta la sezione Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS per maggiori informazioni.
8. Sostituisci *min-instances* e *max-instances* con valori interi. È possibile definire una configurazione statica, in cui è in esecuzione un numero fisso di nodi, o una configurazione dinamica, in cui è possibile eseguire fino al numero massimo di nodi. Per una configurazione statica, imposta minimo e massimo sullo stesso numero, maggiore di zero. Per una configurazione dinamica, imposta il numero minimo di istanze su zero e il numero massimo di istanze su un numero maggiore di zero. AWS PCS non supporta gruppi di nodi di calcolo con un mix di istanze statiche e dinamiche.
9. Sostituisci *t3.large* con un altro tipo di istanza. È possibile aggiungere altri tipi di istanza specificando un elenco di `instanceType` impostazioni. Ad esempio, *--instance-configs instanceType=c6i.16xlarge instanceType=c6a.16xlarge*. Tutti i tipi di istanza devono avere la stessa architettura del processore (x86_64 o arm64) e lo stesso numero di v. CPUs Se le istanze lo sono GPUs, tutti i tipi di istanza devono avere lo stesso numero di. 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— Creazione di un gruppo di nodi di calcolo con impostazioni Slurm personalizzate

```
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"}]'
```

Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS.

Esistono diverse impostazioni di configurazione opzionali che è possibile aggiungere al comando. `create-compute-node-group`

- Puoi specificare `--amiId` se il tuo modello di avvio personalizzato non include un riferimento a un AMI o se desideri sovrascrivere quel valore. Nota che l'AMI utilizzata per il gruppo di nodi deve essere compatibile con AWS PCS. Puoi anche selezionare un AMI di esempio fornito da AWS. Per ulteriori informazioni su questo argomento, vedereAmazon Machine Images (AMIs) per AWS PCS.
- Utilizzalo `--purchase-option` per scegliere il modo in cui AWS PCS acquista le istanze EC2 per il tuo gruppo di nodi di calcolo. On-Demand è l'impostazione predefinita.
  - `ONDEMAND`— Utilizza istanze On-Demand. Scegli questa opzione anche se prevedi di utilizzare una prenotazione di capacità su richiesta (ODCR). Per ulteriori informazioni, consulta Utilizzo ODCRs con AWS PCS.
  - `SPOT`— Utilizza le istanze Spot. Se scegli le istanze Spot, puoi anche utilizzarle `--allocation-strategy` per definire in che modo AWS PCS sceglie i pool di capacità Spot quando avvia le istanze nel gruppo di nodi. Per ulteriori informazioni, consulta Strategie di allocazione per istanze Spot nella Guida per l'utente di Amazon Elastic Compute Cloud.
  - `CAPACITY_BLOCK`— Utilizza una prenotazione esistente di Amazon EC2 Capacity Blocks for ML. Per ulteriori informazioni, consulta Utilizzo dei blocchi di capacità di Amazon EC2 per ML con PCS AWS.
- È possibile fornire opzioni di Slurm configurazione per i nodi del gruppo di nodi utilizzando`--slurm-configuration`. È possibile impostare il peso (priorità di pianificazione) e la memoria reale. I nodi con pesi inferiori hanno una priorità più alta e le unità sono arbitrarie. Per ulteriori informazioni, consulta Weight nella Slurm documentazione. La memoria reale imposta la dimensione (in GB) della memoria reale sui nodi del gruppo di nodi. È pensata per essere utilizzata insieme all'`CR_CPU_Memory`opzione per il cluster in AWS PCS nella Slurm configurazione. Per ulteriori informazioni, consulta RealMemory nella documentazione Slurm.

> **Important**
>
> La creazione del gruppo di nodi di calcolo può richiedere diversi minuti.

Puoi interrogare lo stato del tuo gruppo di nodi con il seguente comando. Non sarai in grado di associare il gruppo di nodi a una coda finché non ne raggiungerà `ACTIVE` lo stato.

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

Aggiornamento di un gruppo di nodi di calcolo AWS PCS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive cosa prendere in considerazione quando si aggiorna un gruppo di nodi di calcolo AWS PCS. Per informazioni sulle impostazioni personalizzate di Slurm, consulta. Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS

Opzioni per l'aggiornamento di un gruppo di nodi di calcolo AWS PCS

L'aggiornamento di un gruppo di nodi di calcolo AWS PCS consente di modificare le proprietà delle istanze lanciate da AWS PCS, nonché le regole per il lancio di tali istanze. Ad esempio, puoi sostituire l'AMI per le istanze del gruppo di nodi con un'altra in cui è installato un software diverso. In alternativa, è possibile aggiornare i gruppi di sicurezza per modificare la connettività di rete in entrata o in uscita. Puoi anche modificare la configurazione di scalabilità e l'opzione di acquisto preferita.

Le seguenti impostazioni del gruppo di nodi non possono essere modificate dopo la creazione:

- Name
- Istanze

Considerazioni sull'aggiornamento di un gruppo di nodi di calcolo AWS PCS

I gruppi di nodi di calcolo definiscono le istanze EC2 utilizzate per elaborare lavori, fornire l'accesso interattivo alla shell e altre attività. Sono spesso associati a una o più AWS code PCS. Quando aggiorni il gruppo di nodi di calcolo per modificarne il comportamento (o quello dei nodi), considera quanto segue:

- Le modifiche alle proprietà del gruppo di nodi di calcolo diventano effettive quando lo stato del gruppo di nodi di calcolo passa da Aggiornamento ad Attivo. Le nuove istanze vengono avviate con le proprietà aggiornate.

- Gli aggiornamenti che non influiscono sulla configurazione di nodi specifici non influiscono sui nodi in esecuzione. Ad esempio, l'aggiunta di una sottorete e la modifica della strategia di allocazione.
- Se si aggiorna il modello di avvio per un gruppo di nodi di calcolo, è necessario aggiornare il gruppo di nodi di calcolo per utilizzare la nuova versione.
- Per aggiungere o rimuovere un gruppo di sicurezza dai nodi di un gruppo di nodi di calcolo, modifica il relativo modello di avvio e aggiorna il gruppo di nodi di calcolo. Le nuove istanze vengono lanciate con il set aggiornato di gruppi di sicurezza.
- Se modifichi direttamente un gruppo di sicurezza utilizzato da un gruppo di nodi di calcolo, ciò ha effetto immediato sulle istanze in esecuzione e future.
- Se aggiungi o rimuovi le autorizzazioni dal profilo dell'istanza IAM utilizzato da un gruppo di nodi di calcolo, ha effetto immediato sulle istanze in esecuzione e future.
- Per modificare l'AMI utilizzata dalle istanze di un gruppo di nodi di calcolo, aggiorna il gruppo di nodi di calcolo (o il relativo modello di avvio) per utilizzare la nuova AMI e attendi che AWS PCS sostituisca le istanze.
- AWS PCS sostituisce le istanze esistenti nel gruppo di nodi dopo un'operazione di aggiornamento del gruppo di nodi. Se ci sono lavori in esecuzione su un nodo, tali processi possono essere completati prima che AWS PCS sostituisca il nodo. I processi utente interattivi (ad esempio sulle istanze del nodo di accesso) vengono terminati. Lo stato del gruppo di nodi torna a `Active` quando AWS PCS contrassegna le istanze per la sostituzione, ma la sostituzione effettiva avviene quando le istanze sono inattive.
- Se riduci il numero massimo di istanze consentite in un gruppo di nodi di calcolo, AWS PCS rimuove i nodi da Slurm per raggiungere il nuovo numero massimo. AWS PCS interrompe l'esecuzione delle istanze associate ai nodi Slurm rimossi. I job in esecuzione sui nodi rimossi falliscono e ritornano nelle rispettive code.
- AWS PCS crea un modello di avvio gestito per ogni gruppo di nodi di calcolo. Sono denominati `pcs-`*`identifier`*`-do-not-delete`. Non selezionarli quando crei o aggiorni un gruppo di nodi di calcolo, altrimenti il gruppo di nodi non funzionerà correttamente.
- Se aggiorni un gruppo di nodi di calcolo per utilizzare Spot come opzione di acquisto, devi avere il ruolo collegato al servizio AWSServiceRoleForEC2Spot nel tuo account. Per ulteriori informazioni, consulta Ruolo Spot di Amazon EC2 per PCS AWS.

Per aggiornare un gruppo di nodi di calcolo AWS PCS

Puoi aggiornare un gruppo di nodi utilizzando la Console di gestione AWS o la CLI AWS.

Console di gestione AWS

Per aggiornare un gruppo di nodi di calcolo

1. Apri la console AWS PCS all'indirizzo `https://console.aws.amazon.com/pcs/home#/clusters`
2. Seleziona il cluster in cui desideri aggiornare un gruppo di nodi di calcolo.
3. Passa ai gruppi di nodi di calcolo, vai al gruppo di nodi che desideri aggiornare, quindi seleziona Modifica.
4. Nelle sezioni Configurazione informatica, Impostazioni aggiuntive e Impostazioni di Slurmpersonalizzazione, aggiorna tutti i valori tranne:

   - Istanze: non è possibile modificare le istanze in un gruppo di nodi di calcolo.

   Per ulteriori informazioni sulle impostazioni personalizzate di Slurm, consulta. Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS
5. Scegliere Aggiorna. Il campo Stato mostrerà Aggiornamento durante l'applicazione delle modifiche.

> **Important**
>
> Gli aggiornamenti dei gruppi di nodi di calcolo possono richiedere diversi minuti.

AWS CLI

Per aggiornare un gruppo di nodi di calcolo

1. Aggiorna il tuo gruppo di nodi di calcolo con il comando che segue. Prima di eseguire il comando, apporta le modifiche seguenti:

   a. Sostituisci *`region-code`* con la regione AWS in cui desideri creare il cluster.

   b. Sostituiscilo *`my-node-group`* con il nome o con `computeNodeGroupId` il gruppo di nodi di calcolo.

   c. *`my-cluster`*Sostituiscilo con il nome o con il nome `clusterId` del tuo cluster.

```
aws pcs update-compute-node-group --region region-code \
```

```
    --cluster-identifier my-cluster \
    --compute-node-group-identifier my-node-group
```

Example— Aggiornamento di un gruppo di nodi di calcolo con impostazioni Slurm personalizzate

```
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"}]'
```

Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS.

2. Aggiorna tutti i parametri del gruppo di nodi ad eccezione di. `--instance-configs` Ad esempio, per impostare un nuovo ID AMI, il `--amiId my-custom-ami-id` comando pass where *my-custom-ami-id* viene sostituito dall'AMI scelto.

> **Important**
>
> L'aggiornamento del gruppo di nodi di calcolo può richiedere diversi minuti.

Puoi interrogare lo stato del tuo gruppo di nodi con il seguente comando.

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

Eliminazione di un gruppo di nodi di calcolo in PCS AWS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive gli aspetti da considerare quando si elimina un gruppo di nodi di calcolo in PCS. AWS

Considerazioni sull'eliminazione di un gruppo di nodi di calcolo

I gruppi di nodi di calcolo definiscono le istanze EC2 utilizzate per elaborare lavori, fornire l'accesso interattivo alla shell e altre attività. Sono spesso associati a una o più AWS code PCS. Prima di eliminare un gruppo di nodi di calcolo, considerate quanto segue:

- Tutte le istanze EC2 lanciate dal gruppo di nodi di calcolo verranno terminate. Ciò annullerà i lavori in esecuzione su queste istanze e interromperà l'esecuzione dei processi interattivi.
- È necessario dissociare il gruppo di nodi di calcolo da tutte le code prima di poterlo eliminare. Per ulteriori informazioni, consulta Aggiornamento di una coda AWS PCS.

Eliminare il gruppo di nodi di calcolo

È possibile utilizzare Console di gestione AWS o AWS CLI per eliminare un gruppo di nodi di calcolo.

Console di gestione AWS

Per eliminare un gruppo di nodi di calcolo

1. Aprire la console AWS PCS.
2. Seleziona il cluster del gruppo di nodi di calcolo.
3. Passa ai gruppi di nodi di calcolo e seleziona il gruppo di nodi di calcolo da eliminare.
4. Scegli Elimina.
5. Viene visualizzato il campo Status. `Deleting` Per il completamento possono essere necessari alcuni minuti.

> **Note**
>
> È possibile utilizzare i comandi nativi dello scheduler per confermare l'eliminazione del gruppo di nodi di calcolo. Ad esempio, usa `sinfo` o `squeue` per Slurm.

AWS CLI

Per eliminare un gruppo di nodi di calcolo

- Usa il comando seguente per eliminare un gruppo di nodi di calcolo, con queste sostituzioni:

- Sostituisci *region-code* con quello in cui si trova Regione AWS il cluster.
- Sostituisci *my-node-group* con il nome o l'ID del tuo gruppo di nodi di calcolo.
- Sostituiscilo *my-cluster* con il nome o l'ID del tuo cluster.

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

L'eliminazione del gruppo di nodi di calcolo può richiedere diversi minuti.

> **Note**
>
> È possibile utilizzare i comandi nativi dello scheduler per confermare l'eliminazione del gruppo di nodi di calcolo. Ad esempio, usa `sinfo` o `squeue` per Slurm.

Ottieni i dettagli del gruppo di nodi di calcolo in AWS PCS

Puoi utilizzare Console di gestione AWS o AWS CLI per ottenere dettagli su un gruppo di nodi di calcolo, come l'ID del gruppo di nodi di calcolo, Amazon Resource Name (ARN) e l'ID Amazon Machine Image (AMI). Questi dettagli sono spesso valori obbligatori per le azioni e le configurazioni dell'API AWS PCS.

Console di gestione AWS

Per ottenere i dettagli del gruppo di nodi di calcolo

1. Apri la console AWS PCS.
2. Seleziona il cluster .
3. Scegli i gruppi di nodi Compute.
4. Scegli un gruppo di nodi di calcolo dal riquadro elenco.

AWS CLI

Per ottenere i dettagli del gruppo di nodi di calcolo

1. Utilizza l'azione ListClustersAPI per trovare il nome o l'ID del cluster.

```
aws pcs list-clusters
```

Output di esempio:

```
{
    "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. Utilizza l'azione ListComputeNodeGroupsAPI per elencare i gruppi di nodi di calcolo in un cluster.

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

Esempio di chiamata:

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

Output di esempio:

```
{
    "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. Utilizza l'azione GetComputeNodeGroupAPI per ottenere dettagli aggiuntivi per un gruppo di nodi di calcolo.

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

Esempio di chiamata:

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

Output di esempio:

```
{
    "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"
            }
        ]
    }
}
```

Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS

Ogni gruppo di nodi di calcolo AWS PCS può avviare istanze EC2 con configurazioni condivise. Puoi utilizzare i tag EC2 per trovare istanze in un gruppo di nodi di calcolo in o con. Console di gestione AWS AWS CLI

Console di gestione AWS

Per trovare le istanze del tuo gruppo di nodi di calcolo

1. Apri la console AWS PCS.
2. Seleziona il cluster .
3. Scegli i gruppi di nodi Compute.
4. Trova l'ID per il gruppo di nodi di accesso che hai creato.
5. Vai alla console EC2 e scegli Istanze.
6. Cerca le istanze con il tag seguente. Sostituiscilo *`node-group-id`* con l'ID (non il nome) del tuo gruppo di nodi di calcolo.

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

7. (Facoltativo) Puoi modificare il valore dello stato dell'istanza nel campo di ricerca per trovare le istanze che sono in fase di configurazione o che sono state terminate di recente.
8. Trova l'ID e l'indirizzo IP dell'istanza per ogni istanza nell'elenco delle istanze con tag.

AWS CLI

Per trovare le istanze del tuo gruppo di nodi, usa i comandi che seguono. Prima di eseguire i comandi, apporta le seguenti sostituzioni:

- Sostituisci *region-code* con il Regione AWS del tuo cluster. Ad esempio: `us-east-1`
- Sostituisci *node-group-id* con l'ID (non il nome) del tuo gruppo di nodi di calcolo. Per trovare l'ID di un gruppo di nodi di calcolo, vedi. Ottieni i dettagli del gruppo di nodi di calcolo in AWS PCS
- Sostituisci `running` con altri stati di istanza come `pending` o `terminated` per trovare istanze EC2 in altri stati.

```
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}
```

Il comando restituisce un risultato simile al seguente. Il valore di `PublicIP` è `null` se l'istanza si trova in una sottorete privata.

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

```
]
```

> **Note**
>
> Se prevedi `describe-instances` di restituire un numero elevato di istanze, devi utilizzare le opzioni per più pagine. Per ulteriori informazioni, consulta DescribeInstancesAmazon Elastic Compute Cloud API Reference.

Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS

In Amazon EC2, un modello di lancio può memorizzare una serie di preferenze in modo da non doverle specificare singolarmente all'avvio delle istanze. AWS PCS incorpora modelli di lancio come modo flessibile per configurare i gruppi di nodi di calcolo. Quando crei un gruppo di nodi, fornisci un modello di lancio. AWS PCS ne crea un modello di lancio derivato che include trasformazioni per garantire che funzioni con il servizio.

Capire quali sono le opzioni e le considerazioni da prendere in considerazione quando si scrive un modello di lancio personalizzato può aiutarvi a scriverne uno da utilizzare con AWS PCS. Per ulteriori informazioni sui modelli di lancio, consulta Launching an Instance from a Launch an instance from a launch template nella Amazon EC2 User Guide.

Argomenti

Panoramica dei modelli di lancio nei PC AWS

Sono disponibili oltre 30 parametri che puoi includere in un modello di lancio di EC2, che controllano molti aspetti della configurazione delle istanze. La maggior parte sono completamente compatibili con AWS PCS, ma ci sono alcune eccezioni.

I seguenti parametri del modello EC2 Launch verranno ignorati da AWS PCS poiché queste proprietà devono essere gestite direttamente dal servizio:

- Attributi del tipo di type/Specify istanza (`InstanceRequirements`): AWS PCS non supporta la selezione delle istanze basata sugli attributi.
- Tipo di istanza (`InstanceType`): specifica i tipi di istanza quando crei un gruppo di nodi.
- Profilo di details/IAM istanza avanzato (`IamInstanceProfile`): viene fornito quando si crea o si aggiorna il gruppo di nodi.

- Terminazione details/Disable API avanzata (`DisableApiTermination`): il AWS PCS deve controllare il ciclo di vita delle istanze del gruppo di nodi che avvia.
- Advanced details/Disable API stop (`DisableApiStop`): il AWS PCS deve controllare il ciclo di vita delle istanze del gruppo di nodi che avvia.
- Advanced details/Stop — Hibernate behavior (`HibernationOptions`) — AWS PCS non supporta l'ibernazione delle istanze.
- Advanced details/Elastic GPU (`ElasticGpuSpecifications`) — Amazon Elastic Graphics ha raggiunto la fine del ciclo di vita l'8 gennaio 2024.
- Advanced details/Elastic inference (`ElasticInferenceAccelerators`): Amazon Elastic Inference non è più disponibile per i nuovi clienti.
- AAdvanced details/Specify CPU options/Threadsper core (`ThreadsPerCore`): AWS PCS imposta il numero di thread per core su 1.

Questi parametri hanno requisiti speciali che supportano la compatibilità con AWS PCS:

- Dati utente (`UserData`): devono essere codificati in più parti. Per informazioni, consulta Utilizzo dei dati utente di Amazon EC2 per PCS AWS.
- Immagini dell'applicazione e del sistema operativo (`ImageId`): puoi includerle. Tuttavia, se specifichi un ID AMI quando crei o aggiorni il gruppo di nodi, questo sovrascriverà il valore nel modello di avvio. L'AMI che fornisci deve essere compatibile con AWS PCS. Per ulteriori informazioni, consulta "Amazon Machine Images (AMIs) per AWS PCS.
- Network settings/Firewall (security groups) (**`SecurityGroups`**): non è possibile impostare un elenco di nomi di gruppi di sicurezza in un modello di avvio AWS PCS. È possibile impostare un elenco di gruppi di sicurezza IDs (`SecurityGroupIds`), a meno che non si definiscano interfacce di rete nel modello di avvio. Quindi, è necessario specificare il gruppo di sicurezza IDs per ogni interfaccia. Per ulteriori informazioni, consulta Gruppi di sicurezza in AWS PCS.
- Configurazione settings/Advanced della rete di rete (`NetworkInterfaces`): se utilizzi istanze EC2 con una singola scheda di rete e non richiedi alcuna configurazione di rete specializzata, AWS PCS può configurare il networking delle istanze per te. Per configurare più schede di rete o abilitare Elastic Fabric Adapter sulle tue istanze, usa. `NetworkInterfaces` Ogni interfaccia di rete deve avere un elenco di gruppi di sicurezza IDs in`Groups`. Per ulteriori informazioni, consulta Interfacce di rete multiple in AWS PCS.
- Dettagli avanzati/Prenotazione della capacità (`CapacityReservationSpecification`): può essere impostato, ma non può fare riferimento a uno specifico `CapacityReservationId` quando si lavora con AWS PCS. Tuttavia, è possibile fare riferimento a un gruppo di prenotazione

di capacità, laddove tale gruppo contenga una o più prenotazioni di capacità. Per ulteriori informazioni, consulta Prenotazioni di capacità in AWS PCS.

Creare un modello di avvio di base

È possibile creare un modello di lancio utilizzando Console di gestione AWS o il AWS CLI.

Console di gestione AWS

Per creare un modello di avvio

1. Apri la EC2console Amazon e seleziona Launch templates.
2. Scegli Crea modello di avvio.
3. In Nome e descrizione del modello Launch, inserisci un nome univoco e distintivo per il nome del modello Launch
4. In Key pair (login) in Key pair name, seleziona la coppia di chiavi SSH che verrà utilizzata per accedere alle EC2 istanze gestite da AWS PCS. Questo passaggio è facoltativo, ma è consigliato.
5. In Impostazioni di rete, quindi Firewall (gruppi di sicurezza), scegli i gruppi di sicurezza da collegare all'interfaccia di rete. Tutti i gruppi di sicurezza nel modello di avvio devono provenire dal AWS VPC del cluster PCS. Come minimo, scegli:
   - Un gruppo di sicurezza che consente la comunicazione con il cluster AWS PCS
   - Un gruppo di sicurezza che consente la comunicazione tra EC2 istanze lanciate da AWS PCS
   - (Facoltativo) Un gruppo di sicurezza che consente l'accesso SSH in entrata a istanze interattive
   - (Facoltativo) Un gruppo di sicurezza che consente ai nodi di elaborazione di effettuare connessioni in uscita a Internet
   - (Facoltativo) Gruppi di sicurezza che consentono l'accesso a risorse di rete come file system condivisi o un server di database.
6. Il tuo nuovo ID del modello di lancio sarà accessibile nella EC2 console Amazon alla voce Launch templates. L'ID del modello di lancio avrà il modulo`lt-0123456789abcdef01`.

Fase successiva consigliata

- Usa il nuovo modello di lancio per creare o aggiornare un gruppo di nodi di calcolo AWS PCS.

AWS CLI

Per creare un modello di avvio

Crea il tuo modello di lancio con il comando che segue.

- Prima di eseguire il comando, apporta le modifiche seguenti:
  a. Sostituiscilo *region-code* con quello Regione AWS in cui stai lavorando con AWS PCS
  b. Sostituiscilo *my-launch-template-name* con un nome per il tuo modello. Deve essere univoco per Account AWS e Regione AWS che stai utilizzando.
  c. Sostituisci *my-ssh-key-name* con il nome della tua chiave SSH preferita.
  d. Sostituisci *sg-ExampleID1* e *sg-ExampleID2* con un gruppo di sicurezza IDs che consente la comunicazione tra le EC2 istanze e lo scheduler e la comunicazione tra le istanze. EC2 Se disponi di un solo gruppo di sicurezza che abilita tutto questo traffico, puoi rimuovere `sg-ExampleID2` anche la virgola che lo precede. Puoi anche aggiungere altri gruppi IDs di sicurezza. Tutti i gruppi di sicurezza inclusi nel modello di avvio devono provenire dal AWS VPC del cluster 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"]}'
```

AWS CLI Verrà emesso un testo simile al seguente. L'ID del modello di avvio si trova in`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"
    }
}
```

Fase successiva consigliata

- Usa il nuovo modello di lancio per creare o aggiornare un gruppo di nodi di calcolo AWS PCS.

Utilizzo dei dati utente di Amazon EC2 per PCS AWS

Puoi fornire i dati utente EC2 nel modello di lancio che `cloud-init` viene eseguito all'avvio delle istanze. I blocchi di dati utente con il tipo di contenuto `cloud-config` vengono eseguiti prima che l'istanza si registri con l'API AWS PCS, mentre i blocchi di dati utente con il tipo di contenuto vengono `text/x-shellscript` eseguiti dopo il completamento della registrazione, ma prima dell'avvio del demone Slurm. Per ulteriori informazioni sui tipi di contenuto, consultare la documentazione di cloud-init.

i nostri dati utente possono eseguire scenari di configurazione comuni, tra cui, a titolo esemplificativo ma non esaustivo, i seguenti:

- Inclusi utenti o gruppi
- Installazione di pacchetti
- Creazione di partizioni e file system
- Montaggio di file system di rete

I dati utente nei modelli di avvio devono essere in formato di archivio multiparte MIME. Questo perché i dati utente vengono uniti ad altri dati utente AWS PCS necessari per configurare i nodi nel gruppo di nodi. È possibile unire più blocchi di dati utente in un unico blocco, detto file MIME in più parti.

Un file MIME in più parti è composto dai seguenti elementi:

- Il tipo di contenuto e la dichiarazione di delimitazione della parte: `Content-Type: multipart/mixed; boundary="==BOUNDARY=="`
- La dichiarazione della versione MIME: `MIME-Version: 1.0`
- Uno o più blocchi di dati utente che contengono i seguenti componenti:
  - Il limite di apertura che segnala l'inizio di un blocco di dati utente:`--==BOUNDARY==`. È necessario mantenere vuota la linea prima di questo limite.

  - La dichiarazione del tipo di contenuto per il blocco: `Content-Type: text/cloud-config; charset="us-ascii"` o`Content-Type: text/x-shellscript; charset="us-ascii"`. È necessario lasciare vuota la riga dopo la dichiarazione del tipo di contenuto.
  - Il contenuto dei dati utente, ad esempio un elenco di comandi o `cloud-config` direttive di shell.
- Il limite di chiusura che segnala la fine del file multiparte MIME:. `--==BOUNDARY==--` È necessario mantenere vuota la linea prima del limite di chiusura.

> **Note**
>
> Se aggiungi dati utente a un modello di lancio nella console Amazon EC2, puoi incollarli come testo normale. In alternativa, puoi caricarli da un file. Se utilizzi AWS CLI o un AWS SDK, devi prima codificare in base64 i dati utente e inviare quella stringa come valore del `UserData` parametro quando chiami CreateLaunchTemplate, come mostrato in questo file JSON.

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

Esempi

- Esempio: installa il software da un repository di pacchetti
- Esempio: esegui script da un bucket S3
- Esempio: imposta le variabili di ambiente globali
- Utilizzo di file system di rete con AWS PCS
- Esempio: utilizzare un file system EFS come home directory condivisa

Esempio: installazione del software per AWS PCS da un archivio di pacchetti

Fornisci questo script come valore di `"userData"` nel tuo modello di lancio. Per ulteriori informazioni, consulta Utilizzo dei dati utente di Amazon EC2 per PCS AWS.

Questo script utilizza cloud-config per installare pacchetti software su istanze di gruppi di nodi al momento del lancio. Per ulteriori informazioni, consulta i formati dei dati utente nella documentazione di cloud-init. Questo esempio installa and. `curl` `llvm`

> **Note**
>
> Le istanze devono essere in grado di connettersi agli archivi di pacchetti configurati.

```
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==--
```

Esempio: eseguire script aggiuntivi per AWS PCS da un bucket S3

Fornisci questo script come valore di `"userData"` nel tuo modello di lancio. Per ulteriori informazioni, consulta Utilizzo dei dati utente di Amazon EC2 per PCS AWS.

Il seguente script di dati utente utilizza cloud-config per importare uno script da un bucket S3 ed eseguirlo su istanze di gruppi di nodi all'avvio. Per ulteriori informazioni, consulta i formati dei dati utente nella documentazione di cloud-init.

Sostituisci i seguenti valori con i tuoi dati:

- *`amzn-s3-demo-bucket`*— Il nome di un bucket S3 da cui il tuo account può leggere.

- *object-key*— La chiave oggetto S3 dello script da importare. Ciò include il nome dello script e la sua posizione nella struttura delle cartelle del bucket. Ad esempio, `scripts/script.sh`. Per ulteriori informazioni, consulta Organizzare gli oggetti nella console Amazon S3 utilizzando le cartelle nella Guida per l'utente di Amazon Simple Storage Service.
- *shell*— La shell Linux da usare per eseguire lo script, ad esempio`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==--
```

Il profilo di istanza IAM per il gruppo di nodi deve avere accesso al bucket. La seguente policy IAM è un esempio del bucket nello script di dati utente riportato sopra.

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/*"
            ]
        }
    ]
}
```

Esempio: imposta le variabili di ambiente globali per AWS PCS

Fornisci questo script come valore di `"userData"` nel tuo modello di lancio. Per ulteriori informazioni, consulta Utilizzo dei dati utente di Amazon EC2 per PCS AWS.

L'esempio seguente utilizza `/etc/profile.d` per impostare variabili globali su istanze di gruppi di nodi.

```
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==--
```

Esempio: utilizzare un file system EFS come home directory condivisa per AWS PCS

Fornisci questo script come valore di `"userData"` nel tuo modello di lancio. Per ulteriori informazioni, consulta Utilizzo dei dati utente di Amazon EC2 per PCS AWS.

Questo esempio estende l'esempio EFS mount in Utilizzo di file system di rete con AWS PCS per implementare una home directory condivisa. Il contenuto di /home viene sottoposto a backup prima del montaggio del file system EFS. I contenuti vengono quindi rapidamente copiati nella memoria condivisa dopo il completamento del montaggio.

Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`/mount-point-directory`*— Il percorso su un'istanza in cui si desidera montare il file system EFS.
- *`filesystem-id`*— L'ID del file system per il file system 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==--
```

Esempio: attivazione di SSH senza password

È possibile basarsi sull'esempio della home directory condivisa per implementare connessioni SSH tra istanze del cluster utilizzando chiavi SSH. Per ogni utente che utilizza il file system home condiviso, esegui uno script simile al seguente:

```
#!/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**
>
> Le istanze devono utilizzare un gruppo di sicurezza che consenta connessioni SSH tra i nodi del cluster.

Prenotazioni di capacità in AWS PCS

Puoi prenotare la EC2 capacità di Amazon in una zona di disponibilità specifica e per una durata specifica utilizzando On-Demand Capacity Reservations o Amazon EC2 Capacity Blocks for ML per assicurarti di avere la capacità di elaborazione necessaria quando ne hai bisogno.

Le prenotazioni di capacità su richiesta (ODCRs) ti consentono di riservare la capacità di calcolo per le tue EC2 istanze Amazon in una zona di disponibilità specifica per qualsiasi durata. Puoi creare e cancellare prenotazioni in qualsiasi momento, senza impegni a lungo termine o pagamenti anticipati. ODCRs sono ideali quando hai bisogno di prenotazioni di capacità flessibili che puoi modificare al variare delle tue esigenze. Per ulteriori informazioni, consulta la sezione Prenotazioni di capacità on demand nella Guida per l'utente di Amazon Elastic Compute Cloud.

Amazon EC2 Capacity Blocks for ML ti consente di prenotare istanze di elaborazione accelerata basate su GPU per utilizzi futuri, fino a 8 settimane in anticipo. Puoi prenotare blocchi da 1 a 64 istanze per durate da 1 giorno a 6 mesi. I Capacity Blocks sono ideali per carichi di lavoro di machine learning che richiedono un accesso garantito alla capacità della GPU in momenti specifici. Per ulteriori informazioni, consulta Capacity Blocks for ML nella Amazon Elastic Compute Cloud User Guide.

Argomenti

Utilizzo ODCRs con AWS PCS

Puoi scegliere in che modo AWS PCS utilizza le tue istanze riservate. Se crei un ODCR aperto, tutte le istanze corrispondenti avviate da AWS PCS o da altri processi nel tuo account vengono conteggiate nella prenotazione. Con un ODCR mirato, solo le istanze avviate con lo specifico ID di prenotazione vengono conteggiate ai fini della prenotazione. Per i carichi di lavoro urgenti, i target ODCRs sono più comuni.

Puoi configurare un gruppo di nodi di calcolo AWS PCS per utilizzare un ODCR mirato aggiungendolo a un modello di avvio. Ecco i passaggi per farlo:

1. Crea una prenotazione di capacità su richiesta (ODCR) mirata utilizzando la guida per l'utente Create a Capacity Reservation di Amazon EC2.

2. Associa l'ODCR a un modello di lancio. Ci sono due modi per farlo:

    a. Associazione ODCR diretta: fai riferimento all'ID ODCR direttamente nel modello di lancio. Questo approccio offre un controllo rigoroso della capacità e non supporta il backfilling delle istanze (se il gruppo di nodi di calcolo richiede più istanze di quelle disponibili nell'ODCR, non verrà avviata alcuna istanza aggiuntiva).

    b. Associazione del gruppo di prenotazione della capacità: aggiungi l'ODCR a un gruppo di prenotazione della capacità e fai riferimento al gruppo nel modello di lancio. Questo approccio supporta il backfilling delle istanze, consentendo a AWS PCS di avviare istanze On-Demand aggiuntive se la capacità di prenotazione viene superata.

3. Crea o aggiorna un gruppo di nodi di calcolo AWS PCS per utilizzare il modello di avvio. Per ulteriori informazioni, consulta la Guida per l'utente di AWS PCS Compute Node Groups.

    - Imposta il gruppo `purchaseOption` di nodi di calcolo su. `ONDEMAND`

Esempio: prenota e utilizza istanze hpc6a.48xlarge con un ODCR mirato

Questo comando di esempio crea un ODCR mirato per 32 istanze hpc6a.48xlarge. Per avviare le istanze riservate in un gruppo di posizionamento, aggiungetele al comando. `--placement-group-arn` È possibile definire una data di fine con `--end-date` e`--end-date-type`, in caso contrario, la prenotazione continuerà fino a quando non verrà terminata 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
```

Il risultato di questo comando sarà un ARN per il nuovo ODCR. L'ID ODCR può essere recuperato dall'ARN `"arn:aws:ec2:us-east-2:123456789012:capacity-reservation/ODCR-ID"` o utilizzando Amazon EC2. DescribeCapacityReservations

Associazione ODCR diretta: aggiungi l'ID ODCR al modello di lancio. Ecco un esempio di modello di lancio che fa riferimento all'ID ODCR.

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

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

Associazione del gruppo di prenotazione della capacità: crea un gruppo di prenotazione della capacità e aggiungi il gruppo al modello di lancio. Il comando seguente crea un gruppo di prenotazione di capacità denominatoEXAMPLE-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"]}]}'
```

Il comando seguente aggiunge l'ODCR al gruppo di prenotazione della capacità.

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

Dopo aver creato e aggiunto l'ODCR a un gruppo di prenotazione della capacità, ora può essere collegato a un gruppo di nodi di calcolo AWS PCS aggiungendolo a un modello di avvio. Ecco un esempio di modello di lancio che fa riferimento al gruppo Capacity Reservation.

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

Infine, crea o aggiorna un gruppo di nodi di calcolo AWS PCS per utilizzare le istanze hpc6a.48xlarge e usa il modello di avvio che fa riferimento all'ODCR. Per un gruppo di nodi statico, imposta il numero minimo e massimo di istanze in base alla dimensione della prenotazione (32). Per un gruppo di nodi dinamico, imposta il numero minimo di istanze su 0 e il massimo sulla dimensione dell'istanza desiderata.

Questo esempio è una semplice implementazione di un singolo ODCR che viene fornito per un gruppo di nodi di calcolo. Tuttavia, AWS PCS supporta molti altri design. Ad esempio, è possibile suddividere un gruppo ODCR o Capacity Reservation di grandi dimensioni tra più gruppi di nodi di elaborazione. In alternativa, puoi utilizzare ODCRs quello che un altro account AWS ha creato e condiviso con il tuo.

Per ulteriori informazioni, consulta On-Demand Capacity Reservations e Capacity Blocks for ML nella Amazon Elastic Compute Cloud User Guide.

Utilizzo dei blocchi di capacità di Amazon EC2 per ML con PCS AWS

Amazon EC2 Capacity Blocks for ML è un'opzione di acquisto di Amazon EC2 che consente di pagare in anticipo per prenotare istanze di elaborazione accelerata basate su GPU entro un intervallo di data e ora specifico per supportare carichi di lavoro di breve durata. Le istanze eseguite all'interno di un Capacity Block vengono automaticamente posizionate vicine tra loro all'interno di Amazon UltraClusters EC2, per reti a bassa latenza, su scala petabit e non bloccanti. Per ulteriori informazioni, consulta Capacity Blocks for ML nella Amazon Elastic Compute Cloud User Guide.

Puoi utilizzare un modello di lancio per fare in modo che i AWS PCS utilizzino un Capacity Block quando avvia istanze per un gruppo di nodi di calcolo.

> **Note**
>
> AWS PCS ha introdotto il supporto per Capacity Blocks a partire dalla versione 24.05 di Slurm.

Limitazioni

- AWS PCS supporta solo Capacity Blocks con famiglie di istanze P5en, P5e, P5 e P4d.
- È possibile associare un gruppo di nodi di calcolo solo a 1 blocco di capacità alla volta.
- Non è possibile associare un gruppo di nodi di calcolo a un gruppo di prenotazione della capacità che combina più blocchi di capacità.
- I blocchi di capacità devono trovarsi in uno `active` stato `scheduled` o per poter essere utilizzati con AWS PCS. Non puoi utilizzare i Capacity Blocks in altri stati, ad esempio`payment-failed`. Per ulteriori informazioni, consulta View Capacity Blocks nella Amazon Elastic Compute Cloud User Guide.

Scadenza del blocco di capacità

I Capacity Block sono limitati a un intervallo di data e ora specifico. Quando un Capacity Block scade:

- Il gruppo di nodi di calcolo associato a quel Capacity Block continua a esistere e rimane associato alle stesse code.
- Tutte le istanze del gruppo di nodi di calcolo vengono terminate e i lavori attivi potrebbero non riuscire, in base alle impostazioni di Slurm.
- AWS PCS non può avviare nuove istanze nel gruppo di nodi di calcolo.
- Tutti i lavori in coda o appena inviati rimangono in sospeso fino a quando un altro gruppo di nodi di calcolo non viene collegato alla coda o non si aggiorna il gruppo di nodi di calcolo per utilizzare un nuovo modello di avvio che specifica un nuovo blocco di capacità.

Configurare un gruppo di nodi di calcolo AWS PCS per utilizzare un Capacity Block

Per associare un Capacity Block a un gruppo di nodi di calcolo

1. Crea un modello di EC2 lancio Amazon per AWS PCS che specifichi il tuo Capacity Block. Per ulteriori informazioni sulla creazione di un modello di lancio per AWS PCS, consultaUtilizzo dei modelli di lancio di Amazon EC2 con PCS AWS.

   Il modello di lancio deve includere:

   - Il valore `MarketType` di `InstanceMarketOptions` deve essere impostato su`capacity-block`.
   - A `CapacityReservationSpecification` con un valore valido `CapacityReservationId`
   - Una versione valida `InstanceType` che corrisponde al tipo di istanza del Capacity Block acquistato.
2. Crea un gruppo di nodi di calcolo che utilizza il modello di avvio. Per ulteriori informazioni, consulta Creazione di un gruppo di nodi di calcolo in AWS PCS. Puoi anche aggiornare un gruppo di nodi di calcolo esistente per utilizzare il modello di avvio. Per ulteriori informazioni, consulta Aggiornamento di un gruppo di nodi di calcolo AWS PCS.

   Quando crei o aggiorni il gruppo di nodi di calcolo:

- L'identità IAM che usi per creare o aggiornare il gruppo di nodi di calcolo deve avere la seguente autorizzazione:

  ```
  ec2:DescribeCapacityReservations
  ```

  Per ulteriori informazioni, consulta Autorizzazioni minime per AWS PCS.
- Il Capacity Block deve trovarsi in uno `active` stato `scheduled` o.
- Imposta il gruppo `purchaseOption` di nodi di calcolo su`CAPACITY_BLOCK`.
- La dimensione `maxInstanceCount` del gruppo di nodi di calcolo non deve superare la dimensione del Capacity Block.
- La zona di disponibilità del gruppo di nodi di calcolo deve corrispondere a 1 delle zone di disponibilità della sottorete del gruppo di nodi di calcolo.

> **Important**
>
> Non è possibile modificare il tipo di istanza di un gruppo di nodi di calcolo quando lo si aggiorna. Puoi utilizzare solo un Capacity Block con lo stesso tipo di istanza del gruppo di nodi di calcolo. Se desideri utilizzare un Capacity Block con un tipo di istanza diverso, devi creare un nuovo gruppo di nodi di calcolo.

Domande frequenti sull'utilizzo di Capacity Blocks con AWS PCS

Ho appena pagato un Capacity Block e ho subito provato a usarlo con AWS PCS, ma la creazione del gruppo di nodi di calcolo non è riuscita. Che cos'è successo?

Il tuo Capacity Block potrebbe non essere in uno stato `scheduled` o`active`. Riprova dopo che il Capacity Block è `scheduled` o`active`.

Sto usando un Capacity Block in AWS PCS e ho acquistato un'estensione prima della scadenza. Come posso continuare a usarlo in AWS PCS?

Non devi fare nulla per continuare a utilizzare il Capacity Block in AWS PCS. La data di fine del Capacity Block viene aggiornata dopo che il pagamento dell'estensione è andato a buon fine. Finché il Capacity Block non scade, il gruppo di nodi di calcolo continua a funzionare. Se il pagamento dell'estensione fallisce, il Capacity Block rimane `active` e il gruppo di nodi di calcolo funziona fino alla scadenza del Capacity Block alla data di fine originale.

Cosa succede ai miei lavori in coda e in esecuzione se il mio Capacity Block scade?

I lavori in coda che non sono iniziati prima della scadenza del Capacity Block rimangono in sospeso finché non si collega un altro gruppo di nodi di elaborazione alla coda o si aggiorna il gruppo di nodi di calcolo con un nuovo Capacity Block. Puoi comunque inviare lavori alla coda. Le impostazioni di Slurm influiscono sui lavori attivi. Per impostazione predefinita, i lavori attivi vengono automaticamente rimessi in coda, ma potrebbero presentare errori o fallire.

Il mio Capacity Block è scaduto. Devo fare qualcosa?

Non devi fare niente. Puoi controllare lo stato delle tue prenotazioni di capacità EC2 sulla console Amazon EC2. Quando un Capacity Block scade, il gruppo di nodi di calcolo associato a quel Capacity Block continua a esistere e a gestire le stesse code. Il gruppo di nodi di calcolo non dispone di istanze per eseguire i job. Puoi eliminare il gruppo di nodi di calcolo o dissociarlo dalle code per impedire agli utenti di inviare lavori che non verranno eseguiti.

Voglio usare un nuovo Capacity Block con il mio gruppo di nodi di calcolo AWS PCS. Cosa devo fare?

Ti consigliamo di creare un nuovo gruppo di nodi di calcolo per utilizzare il nuovo Capacity Block. Per ulteriori informazioni, consulta Configurare un gruppo di nodi di calcolo AWS PCS per utilizzare un Capacity Block.

Come posso condividere 1 Capacity Block tra cluster e servizi?

È possibile suddividere un Capacity Block tra più cluster e servizi. Ad esempio, per dividere un Capacity Block con 64 `p5.48xlarge` istanze con 20 nodi su PCS-Cluster-1, 16 nodi su PCS-Cluster-2 e i nodi rimanenti per altri servizi, impostate entrambi `minInstanceCount` e su 20 per PCS-Cluster-1 e 16 per PCS-Cluster-2. `maxInstanceCount`

Posso usare più di 1 Capacity Block o una capacità combinata con 1 gruppo di nodi di calcolo?

No. È possibile associare solo 1 blocco di capacità a un singolo gruppo di nodi di elaborazione. AWS PCS non supporta gruppi di prenotazione della capacità che combinano più blocchi di capacità.

Come faccio a sapere quando iniziano o scadono i miei Capacity Block?

Indipendentemente dal AWS PCS, Amazon EC2 invia un `Capacity Block Reservation Delivered` evento EventBridge quando inizia una prenotazione Capacity Block e un `Capacity Block Reservation Expiration Warning` evento 40 minuti prima della scadenza della prenotazione Capacity Block. Per ulteriori informazioni, consulta Monitora i blocchi di capacità utilizzati EventBridge nella Guida per l'utente di Amazon Elastic Compute Cloud.

In che modo Slurm monitora lo stato del mio Capacity Block?

Puoi correre `sinfo` per capire come AWS PCS utilizza il Capacity Block. Nell'output di esempio seguente, una coda è associata a un gruppo di nodi di calcolo che esegue 4 istanze da un `active` Capacity Block. I nodi sono nello stato `idle` Slurm (disponibili per l'uso e non ancora assegnati a nessun lavoro).

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

Se invece i nodi sono in `maint` stato, puoi correre `scontrol show res` per vedere i dettagli sulla prenotazione Slurm che controlla questo stato. Nell'output di esempio seguente, il Capacity Block ha una data di inizio futura. `scheduled`

```
$ 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
```

Come posso sapere se gli errori che ricevo durante l'avvio della capacità sono dovuti al fatto che il mio Capacity Block è condiviso?

Controlla le prenotazioni di capacità nella console Amazon EC2 per scoprire quante istanze del Capacity Block vengono fornite attivamente. Controlla i tag di ogni istanza per scoprire quale servizio o cluster la utilizza. Ad esempio, tutte le istanze di AWS PCS hanno tag AWS PCS come quelli `aws:pcs:cluster-id = pcs_l0mizqyk5o | aws:pcs:compute-node-group-id = pcs_ic7onkmfqk` che indicano a quali cluster e gruppi di nodi di calcolo appartiene l'istanza. È quindi possibile verificare se il Capacity Block ha la capacità massima.

`scontrol show nodes`Per verificare se un nodo Capacity Block in un cluster AWS PCS si sta attivando`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]
```

Quando più gruppi di nodi di elaborazione sono collegati alla stessa coda, come posso forzare l'esecuzione di un processo su istanze supportate da Capacity Block?

Puoi utilizzare le funzionalità e i vincoli di Slurm per bloccare un lavoro su un determinato set di nodi. Ti consigliamo di non impostare i pesi Slurm per ogni gruppo di nodi di calcolo perché funziona solo con nodi che non si trovano nello stato. `maint`

Parametri utili del modello di lancio

Questa sezione descrive alcuni parametri del modello di lancio che possono essere ampiamente utili con AWS PCS.

Attiva il monitoraggio dettagliato CloudWatch

Puoi abilitare la raccolta di CloudWatch metriche a intervalli più brevi utilizzando un parametro del modello di avvio.

Console di gestione AWS

Nelle pagine della console per la creazione o la modifica dei modelli di avvio, questa opzione si trova nella sezione Dettagli avanzati. Imposta CloudWatch il monitoraggio dettagliato su Abilita.

YAML

```
Monitoring:
    Enabled: True
```

JSON

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

Per ulteriori informazioni, consulta Attivare o disattivare il monitoraggio dettagliato per le istanze nella Amazon Elastic Compute Cloud User Guide for Linux Instances.

Instance Metadata Service versione 2 (IMDS v2)

L'utilizzo di IMDS v2 con le istanze EC2 offre significativi miglioramenti della sicurezza e aiuta a mitigare i potenziali rischi associati all'accesso ai metadati delle istanze negli ambienti. AWS

Console di gestione AWS

Nelle pagine della console per la creazione o la modifica dei modelli di avvio, questa opzione si trova nella sezione Dettagli avanzati. Imposta Metadati accessibili su Enabled, la versione Metadata solo su V2 (token richiesto) e il limite dell'hop di risposta dei metadati su 4.

YAML

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

JSON

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

.

AWS Code PCS

Una coda AWS PCS è un'astrazione leggera rispetto all'implementazione nativa di una coda di lavoro da parte dello scheduler. Nel caso di Slurm, una coda AWS PCS è equivalente a una partizione Slurm.

Gli utenti inviano i lavori a una coda in cui risiedono fino a quando non è possibile programmarne l'esecuzione sui nodi forniti da uno o più gruppi di nodi di elaborazione. Un cluster AWS PCS può avere più code di lavoro. Ad esempio, puoi creare una coda che utilizza Amazon EC2 On-demand Instances per lavori ad alta priorità e un'altra coda che utilizza Amazon EC2 Spot Instances per lavori a bassa priorità.

Argomenti

Creazione di una coda in AWS PCS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive gli elementi da considerare quando si crea una coda in AWS PCS.

> **Note**
>
> È possibile configurare impostazioni Slurm personalizzate sulle code per implementare politiche di pianificazione e gestione delle risorse specifiche delle partizioni. Per ulteriori informazioni, consulta Configurazione delle impostazioni Slurm personalizzate in PCS AWS.

Prerequisiti

- Un cluster AWS PCS: le code possono essere create solo in associazione con un cluster PCS specifico. AWS
- Uno o più gruppi di nodi di calcolo AWS PCS: una coda deve essere associata ad almeno un gruppo di nodi di calcolo AWS PCS.

Per creare una coda in PCS AWS

È possibile creare una coda utilizzando Console di gestione AWS o il. AWS CLI

Console di gestione AWS

Per creare una coda utilizzando la console

1. Aprire la console AWS PCS.
2. Seleziona il cluster per la coda. Passa a Queues e scegli Crea coda.
3. Nella sezione Configurazione della coda, fornisci i seguenti valori:
   a. Nome della coda: un nome per la coda. Il nome può contenere solo caratteri alfanumerici (con distinzione tra lettere maiuscole e minuscole) e trattini. Deve iniziare con un carattere alfabetico e non può superare i 25 caratteri. Il nome deve essere univoco all'interno del cluster.
   b. Gruppi di nodi di calcolo: seleziona uno o più gruppi di nodi di calcolo per servire questa coda. Un gruppo di nodi di calcolo può essere associato a più di una coda.
4. (Facoltativo) Nella sezione Impostazioni aggiuntive dello scheduler, puoi aggiungere coppie di nomi e valori dei parametri per configurare impostazioni Slurm aggiuntive. Per un elenco completo dei parametri supportati, vedere. Impostazioni Slurm personalizzate per le code PCS AWS
5. (Facoltativo) In Tag, aggiungi qualsiasi tag alla coda AWS PCS
6. Scegliere Crea coda. Il campo Stato mostrerà Creazione mentre AWS PCS crea la coda. La creazione della coda può richiedere diversi minuti.

Passaggio successivo consigliato

- Invia un lavoro alla tua nuova coda.

AWS CLI

Per creare una coda utilizzando AWS CLI

Usa il seguente comando per creare la tua coda. Effettua le seguenti sostituzioni:

1. Sostituisci *region-code* con la AWS regione del cluster. Ad esempio, `us-east-1`.

2. Sostituisci *my-queue* con il nome della coda. Il nome può contenere solo caratteri alfanumerici (con distinzione tra lettere maiuscole e minuscole) e trattini. Deve iniziare con un carattere alfabetico e non può superare i 25 caratteri. Il nome deve essere univoco all'interno del cluster.
3. Sostituiscilo *my-cluster* con il nome o l'ID del cluster.
4. Sostituiscilo *compute-node-group-id* con l'ID del gruppo di nodi di calcolo per servire la coda. Ad esempio, `pcs_abcdef12345`.

> **Note**
>
> Quando crei una coda, devi fornire l'ID del gruppo di nodi di calcolo e non il suo nome.

```
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— Creazione di una coda con impostazioni Slurm personalizzate

```
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}]'
```

Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per le code PCS AWS.

La creazione della coda può richiedere diversi minuti. È possibile interrogare lo stato della coda con il seguente comando. Non potrai inviare lavori alla coda finché non verrà raggiunto lo stato corrispondente. `ACTIVE`

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

Fase successiva consigliata

- Invia un lavoro alla tua nuova coda

Aggiornamento di una coda AWS PCS

Questo argomento fornisce una panoramica delle opzioni disponibili e descrive gli elementi da considerare quando si aggiorna una coda AWS PCS. Per informazioni sulle impostazioni personalizzate di Slurm, vedere. Impostazioni Slurm personalizzate per le code PCS AWS

Considerazioni sull'aggiornamento di una coda PCS AWS

Gli aggiornamenti delle code non influiranno sui lavori in esecuzione, ma il cluster potrebbe non essere in grado di accettare nuovi lavori durante l'aggiornamento della coda.

Per aggiornare una coda AWS PCS

È possibile utilizzare Console di gestione AWS o AWS CLI per aggiornare una coda.

Console di gestione AWS

Per aggiornare una coda

1. Aprire la console AWS PCS all'indirizzo `https://console.aws.amazon.com/pcs/home#/clusters`
2. Seleziona il cluster in cui desideri aggiornare una coda.
3. Vai a Code, vai alla coda che desideri aggiornare, quindi seleziona Modifica.
4. Nella sezione di configurazione della coda, aggiorna uno dei seguenti valori:
   - Gruppi di nodi: aggiungi o rimuovi i gruppi di nodi di calcolo dall'associazione alla coda.
   - Impostazioni aggiuntive dello scheduler: aggiungi, modifica o rimuovi le impostazioni Slurm personalizzate per la coda. Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per le code PCS AWS.
   - Tag: aggiungi o rimuovi tag per la coda.
5. Scegliere Aggiorna. Il campo Stato mostrerà Aggiornamento durante l'applicazione delle modifiche.

> **Important**
>
> Gli aggiornamenti delle code possono richiedere diversi minuti.

AWS CLI

Per aggiornare una coda

1. Aggiorna la coda con il comando che segue. Prima di eseguire il comando, apporta le modifiche seguenti:

   a. Sostituiscila *region-code* con Regione AWS quella in cui vuoi creare il cluster.

   b. Sostituiscilo *my-queue* con il nome o con `computeNodeGroupId` la tua coda.

   c. *my-cluster*Sostituiscilo con il nome o con il nome `clusterId` del tuo cluster.

   d. Per modificare le associazioni dei gruppi di nodi di calcolo, fornisci un elenco aggiornato per`--compute-node-group-configurations`.

      - Ad esempio, per aggiungere un secondo gruppo di nodi di calcolo: `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— Aggiornamento di una coda con impostazioni Slurm personalizzate

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

Per ulteriori informazioni, consulta Impostazioni Slurm personalizzate per le code PCS AWS.

2. L'aggiornamento della coda può richiedere diversi minuti. È possibile interrogare lo stato della coda con il seguente comando. Non potrai inviare lavori alla coda finché non verrà raggiunto lo stato corrispondente. `ACTIVE`

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

Passaggi successivi consigliati

- Invia un lavoro alla tua coda aggiornata.

Eliminazione di una coda in PCS AWS

Questo argomento fornisce una panoramica su come eliminare una coda in PCS. AWS

Considerazioni sull'eliminazione di una coda

- Se ci sono lavori in esecuzione nella coda, questi verranno terminati dallo scheduler quando la coda viene eliminata. I lavori in sospeso in coda verranno annullati. Valuta la possibilità di attendere che i lavori in coda finiscano o di stop/cancel eseguirli manualmente utilizzando i comandi nativi dello scheduler (come per Slurm). `scancel`

Eliminare la coda

È possibile utilizzare Console di gestione AWS o AWS CLI per eliminare una coda.

Console di gestione AWS

Come eliminare una coda

1. Aprire la console AWS PCS.
2. Seleziona il cluster della coda.
3. Vai a Code e seleziona la coda da eliminare.
4. Scegli Elimina.

5. Viene visualizzato il campo Stato. `Deleting` Per il completamento possono essere necessari alcuni minuti.

> **Note**
>
> È possibile utilizzare i comandi nativi dello scheduler per confermare l'eliminazione della coda. Ad esempio, usa `sinfo` o `squeue` per Slurm.

AWS CLI

Come eliminare una coda

- Utilizzate il seguente comando per eliminare una coda, con queste sostituzioni:
  - Sostituisci *`region-code`* con quello in cui si trova Regione AWS il cluster.
  - Sostituisci *`my-queue`* con il nome o l'ID della coda.
  - Sostituiscilo *`my-cluster`* con il nome o l'ID del cluster.

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

L'eliminazione della coda può richiedere diversi minuti.

> **Note**
>
> È possibile utilizzare i comandi nativi dello scheduler per confermare l'eliminazione della coda. Ad esempio, usa `sinfo` o `squeue` per Slurm.

AWS Nodi di accesso PCS

Un cluster AWS PCS di solito necessita di almeno 1 nodo di accesso per supportare l'accesso interattivo e la gestione dei lavori. Un modo per farlo è utilizzare un gruppo di nodi di calcolo AWS PCS statico configurato per la funzionalità del nodo di accesso. Puoi anche configurare un'istanza EC2 autonoma che funga da nodo di accesso.

Argomenti

Utilizzo di un gruppo di nodi di calcolo AWS PCS per fornire nodi di accesso

Questo argomento fornisce una panoramica delle opzioni di configurazione suggerite e descrive cosa prendere in considerazione quando si utilizza un gruppo di nodi di calcolo AWS PCS per fornire un accesso persistente e interattivo al cluster.

Creazione di un gruppo di nodi di calcolo AWS PCS per i nodi di accesso

Dal punto di vista operativo, questo non è molto diverso dalla creazione di un normale gruppo di nodi di calcolo. Tuttavia, ci sono alcune scelte di configurazione chiave:

- Imposta una configurazione di scalabilità statica di almeno un'istanza EC2 nel gruppo di nodi di calcolo.
- Scegli l'opzione di acquisto su richiesta per evitare che le tue istanze vengano recuperate.
- Scegli un nome informativo per il gruppo di nodi di calcolo, ad esempio login.
- Se desideri che le istanze del nodo di accesso siano accessibili al di fuori del tuo VPC, prendi in considerazione l'utilizzo di una sottorete pubblica.
- Se intendi consentire l'accesso SSH, il modello di avvio dovrà disporre di un gruppo di sicurezza che esponga la porta SSH agli indirizzi IP che hai scelto.
- Il profilo dell'istanza IAM dovrebbe avere solo le autorizzazioni AWS che desideri siano concesse ai tuoi utenti finali. Per informazioni dettagliate, vedi Profili di istanza IAM per AWS Parallel Computing Service.

- Prendi in considerazione la possibilità di consentire ad AWS Systems Manager Session Manager di gestire le tue istanze di accesso.
- Prendi in considerazione la possibilità di limitare l'accesso alle credenziali AWS dell'istanza ai soli utenti amministrativi
- Seleziona tipi di istanze meno costosi rispetto ai normali gruppi di nodi di calcolo, poiché i nodi di accesso funzioneranno continuamente.
- Utilizza la stessa AMI (o una derivata) degli altri gruppi di nodi di calcolo per garantire che su tutte le istanze sia installato lo stesso software. Per ulteriori informazioni sulla personalizzazione, consulta AMIs Amazon Machine Images (AMIs) per AWS PCS
- Configura gli stessi supporti del file system di rete (Amazon EFS, Amazon FSx for Lustre, ecc.) sui nodi di accesso e sulle istanze di calcolo. Per ulteriori informazioni, consulta Utilizzo di file system di rete con AWS PCS.

Accedi ai tuoi nodi di accesso

Una volta che il tuo nuovo gruppo di nodi di calcolo raggiunge lo stato ATTIVO, puoi trovare le istanze EC2 che ha creato e accedere ad esse. Per ulteriori informazioni, consulta Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS.

Aggiornamento di un gruppo di nodi di calcolo AWS PCS per i nodi di accesso

È possibile aggiornare un gruppo di nodi di accesso utilizzando UpdateComputeNodeGroup. Come parte del processo di aggiornamento del gruppo di nodi, le istanze in esecuzione verranno sostituite. Tieni presente che ciò interromperà tutte le sessioni o i processi utente attivi sull'istanza. I job Slurm in esecuzione o in coda non subiranno alcuna modifica. Per ulteriori informazioni, consulta Aggiornamento di un gruppo di nodi di calcolo AWS PCS.

Puoi anche modificare il modello di avvio utilizzato dal tuo gruppo di nodi di calcolo. È necessario utilizzare UpdateComputeNodeGroup per applicare il modello di avvio aggiornato al gruppo di nodi di calcolo. Le nuove istanze EC2 lanciate nel gruppo di nodi di calcolo utilizzano il modello di avvio aggiornato. Per ulteriori informazioni, consulta Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS.

Eliminazione di un gruppo di nodi di calcolo AWS PCS per i nodi di accesso

È possibile aggiornare un gruppo di nodi di accesso utilizzando il meccanismo di eliminazione del gruppo di nodi di calcolo in PCS. AWS Le istanze in esecuzione verranno terminate come parte dell'eliminazione del gruppo di nodi. Tieni presente che ciò interromperà tutte le sessioni o i processi utente attivi sull'istanza. I job Slurm in esecuzione o in coda non subiranno alcuna modifica. Per ulteriori informazioni, consulta Eliminazione di un gruppo di nodi di calcolo in PCS AWS.

Utilizzo di istanze autonome come nodi di accesso AWS PCS

Puoi configurare istanze EC2 indipendenti per interagire con lo scheduler Slurm di un cluster AWS PCS. Ciò è utile per creare nodi di accesso, workstation o host dedicati alla gestione del flusso di lavoro che funzionano con i cluster AWS PCS ma operano al di fuori della gestione PCS. AWS A tale scopo, ogni istanza autonoma deve:

1. Avere installata una versione del software Slurm compatibile.
2. Essere in grado di connettersi all'endpoint Slurmctld del cluster AWS PCS.
3. Configurate correttamente Slurm Auth e Cred Kiosk Daemon () con l'endpoint e il segreto del cluster PCS. `sackd` AWS Per ulteriori informazioni, vedete sackd nella documentazione di Slurm.

Questo tutorial ti aiuta a configurare un'istanza indipendente che si connette a un cluster PCS. AWS

Indice

Passaggio 1: recuperare l'indirizzo e il segreto per il cluster AWS PCS di destinazione

Recupera i dettagli sul cluster AWS PCS di destinazione utilizzando AWS CLI il comando che segue. Prima di eseguire il comando, apporta le modifiche seguenti:

- Sostituisci *region-code* con il Regione AWS punto in cui è in esecuzione il cluster di destinazione.
- Sostituisci *cluster-ident* con il nome o l'identificatore del cluster di destinazione

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

Il comando restituirà un output simile a questo esempio.

```
{
    "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"
```

```
            }
        ]
    }
}
```

In questo esempio, l'endpoint del controller Slurm del cluster ha un indirizzo IP di `10.3.149.220` ed è in esecuzione sulla porta. `6817` `secretArn`Verrà utilizzato nei passaggi successivi per recuperare il segreto del cluster. L'indirizzo IP e la porta verranno utilizzati nei passaggi successivi per configurare il `sackd` servizio.

Fase 2: Avvio di un'istanza EC2

Per avviare un'istanza EC2

1. Aprire la console di Amazon EC2.
2. Nel pannello di navigazione, scegli Instances (Istanze) e quindi scegli Launch instances (Avvia istanze) per aprire la nuova procedura guidata di avvio dell'istanza.
3. (Facoltativo) Nella sezione Nome e tag, fornisci un nome per l'istanza, ad esempio. `PCS-LoginNode` Il nome viene assegnato all'istanza come tag di risorsa (`Name=PCS-LoginNode`).
4. Nella sezione Immagini dell'applicazione e del sistema operativo, seleziona un AMI per uno dei sistemi operativi supportati da AWS PCS. Per ulteriori informazioni, consulta Sistemi operativi supportati.
5. Nella sezione Tipo di istanza, seleziona un tipo di istanza supportato. Per ulteriori informazioni, consulta Tipi di istanze supportati.
6. Nella sezione Coppia di chiavi, seleziona la coppia di chiavi SSH da usare per l'istanza.
7. Nella sezione Impostazioni di rete:
    - Scegli Modifica.
        i. Seleziona il VPC del tuo cluster AWS PCS.
        ii. Per Firewall (gruppi di sicurezza), scegli Seleziona un gruppo di sicurezza esistente.
            A. Seleziona un gruppo di sicurezza che consenta il traffico tra l'istanza e il controller Slurm del cluster AWS PCS di destinazione. Per ulteriori informazioni, consulta Requisiti e considerazioni sui gruppi di sicurezza.
            B. (Facoltativo) Seleziona un gruppo di sicurezza che consenta l'accesso SSH in entrata all'istanza.

8. Nella sezione Archiviazione, configura i volumi di archiviazione in base alle esigenze. Assicurati di configurare uno spazio sufficiente per installare applicazioni e librerie adatte al tuo caso d'uso.
9. In Avanzato, scegli un ruolo IAM che consenta l'accesso al segreto del cluster. Per ulteriori informazioni, consulta Ottieni il segreto del cluster Slurm.
10. Nel riquadro Riepilogo, scegli Launch instance.

Passaggio 3: installa Slurm sull'istanza

Quando l'istanza è stata lanciata e diventa attiva, connettiti ad essa utilizzando il tuo meccanismo preferito. Utilizza il programma di installazione Slurm fornito da AWS per installare Slurm sull'istanza. Per ulteriori informazioni, consulta Programma di installazione Slurm.

Scarica il programma di installazione di Slurm, decomprimilo e usa lo script per installare Slurm. `installer.sh` Per ulteriori informazioni, consulta Fase 3 — Installare Slurm.

Fase 4 — Recuperare e archiviare il segreto del cluster

Queste istruzioni richiedono il. AWS CLI Per ulteriori informazioni, vedere Installazione o aggiornamento alla versione più recente di AWS CLI nella Guida per l'AWS Command Line Interface utente della versione 2.

Memorizza il segreto del cluster con i seguenti comandi.

- Crea la directory di configurazione per Slurm.

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

- Recupera, decodifica e archivia il segreto del cluster. Prima di eseguire questo comando, *region-code* sostituiscilo con la regione in cui è in esecuzione il cluster di destinazione e sostituiscilo *secret-arn* con il valore `secretArn` recuperato nel passaggio 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**
>
> In un ambiente multiutente, qualsiasi utente con accesso all'istanza potrebbe essere in grado di recuperare il segreto del cluster se può accedere al servizio di metadati dell'istanza (IMDS). Questo, a sua volta, potrebbe consentire loro di impersonare altri utenti. Prendi in considerazione la possibilità di limitare l'accesso a IMDS solo agli utenti root o amministrativi. In alternativa, prendi in considerazione l'utilizzo di un meccanismo diverso che non si basi sul profilo dell'istanza per recuperare e configurare il segreto.

- Imposta proprietà e autorizzazioni sul file chiave Slurm.

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

> **Note**
>
> La chiave Slurm deve essere di proprietà dell'utente e del gruppo con cui viene eseguito il servizio. `sackd`

Fase 5 — Configurare la connessione al cluster PCS AWS

Per stabilire una connessione al cluster AWS PCS, `sackd` avvialo come servizio di sistema seguendo questi passaggi.

> **Note**
>
> Se utilizzi Slurm 25.05 o versioni successive, puoi utilizzare uno script per configurare il nodo di accesso in modo che si connetta invece a più cluster. Per ulteriori informazioni, consulta Connessione di un nodo di accesso autonomo a più cluster in PCS AWS.

1. Imposta il file di ambiente per il `sackd` servizio con il comando che segue. Prima di eseguire il comando, sostituite *`ip-address`* e *`port`* con i valori recuperati dagli endpoint nel passaggio 1.

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

2. Create un file `systemd` di servizio per la gestione del `sackd` processo.

```
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. Imposta la proprietà del file `sackd` di servizio.

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

4. Abilita il `sackd` servizio.

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

5. Avviare il servizio `sackd`.

```
sudo systemctl start sackd
```

Fase 6 — (Facoltativo) Verifica della connessione

Verificare che il `sackd` servizio sia in esecuzione. Di seguito è riportato un output di esempio. Se ci sono errori, di solito vengono visualizzati qui.

```
[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
```

Conferma che le connessioni al cluster funzionino utilizzando i comandi del client Slurm come `esinfo`. `squeue` Ecco un esempio di output da. `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]
```

Dovresti anche essere in grado di inviare offerte di lavoro. Ad esempio, un comando simile a questo esempio avvierebbe un processo interattivo su 1 nodo del cluster.

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

Connessione di un nodo di accesso autonomo a più cluster in PCS AWS

Lo `pcs-multi-cluster-login-configure.sh` script fornisce un modo automatico per configurare più `sackd` demoni Slurm su un singolo nodo di accesso autonomo. Consente al nodo di accesso di comunicare con più cluster. Lo script automatizza le seguenti operazioni:

- Utilizza le azioni dell'API AWS PCS per ottenere informazioni sul cluster
- Richiede la chiave di autenticazione Slurm con codifica base64
- Crea un file Slurm JWKS con chiave di autenticazione del cluster
- Configura il `sackd` servizio con endpoint e porte del cluster
- Crea un file di `systemd` servizio per un demone specifico del cluster `sackd`
- Genera uno script di attivazione per la configurazione dell'ambiente cluster
- Abilita e avvia il `sackd` servizio

> **Note**
>
> Questo script richiede la versione Slurm 25.05 o successiva.

Slurm deve essere già installato sull'istanza (equivalente al passaggio 3 del processo manuale). L'istanza deve essere in grado di raggiungere gli endpoint del cluster di destinazione. Lo script esegue le operazioni equivalenti ai passaggi 4 e 5 del processo di configurazione manuale. Ottiene automaticamente le informazioni sul cluster, configura il `sackd` servizio, crea i file di `systemd` servizio necessari e crea uno script di attivazione che gli utenti possono utilizzare per configurare il proprio ambiente shell per l'interazione con il cluster.

Argomenti

Prerequisiti per lo script di configurazione del nodo di accesso multicluster AWS PCS

Requisiti di sistema

- Sistema operativo Linux con supporto `systemd`
- Privilegi di root per la configurazione del sistema

Comandi e pacchetti richiesti

- `bash`— Interprete Shell (versione 4.0+)
- `curl`— Per il recupero dei metadati AWS IMDS v2
- `jq`— Processore JSON per l'analisi delle risposte API AWS
- `aws`— AWS CLI v2 per eseguire azioni API AWS PCS e per l'accesso a Secrets Manager
- `systemctl`— gestione `systemd` dei servizi
- `find`— Utilità di ricerca nel file system
- `grep`— Corrispondenza dei modelli di testo
- `sed`— Stream editor per la manipolazione del testo
- `sort`— Utilità di ordinamento del testo
- `tail`— Visualizza le ultime righe di un file
- `mkdir`— Creazione di cartelle
- `chmod`— Modifica le autorizzazioni dei file
- `chown`— Modifica la proprietà dei file
- `ldconfig`— Configurazione dinamica del linker

AWS requisiti

- Un cluster AWS PCS che esegue Slurm versione 25.05 o successiva
- AWS credenziali configurate (tramite un ruolo IAM, un file di credenziali o variabili di ambiente)
- Autorizzazioni per:
  - `pcs:GetCluster`
  - `secretsmanager:GetSecretValue`(se usi un segreto alternativo)

Utenti e gruppi di sistema

- L'`slurm`utente e il gruppo devono esistere nel sistema

Installazione di Slurm

- Slurm deve essere installato nella stessa posizione dei pacchetti di installazione di AWS PCS Slurm:

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

AWS Codice dello script di configurazione del nodo di accesso multicluster PCS

Salva il seguente codice sorgente in un file con il seguente nome:

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

Codice sorgente dello 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 "$@"
```

Utilizzo dello script di configurazione del nodo di accesso multicluster AWS PCS

Esecuzione dello script

Per eseguire lo script di configurazione

1. Salva il contenuto dello script in un file denominato:

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

2. Rendilo eseguibile:

   ```
   chmod +x pcs-multi-cluster-login-configure.sh
   ```

3. Esegui lo script :

   ```
   ./pcs-multi-cluster-login-configure.sh --cluster-identifier cluster-name
   ```

Ambienti di interazione con i cluster

Dopo una corretta configurazione, lo script genera uno script di attivazione specifico del cluster nella directory corrente. Lo script ha il nome. `activate-pcs-`*`cluster-name`* Lo script di attivazione configura le variabili e i percorsi di ambiente necessari per interagire con il cluster di destinazione.

Per attivare un ambiente cluster

- Utilizzare il `source` comando per eseguire lo script di attivazione

```
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
```

Cosa fa lo script di attivazione

- Imposta la variabile di `SLURM_CONF` ambiente in modo che punti alla configurazione del cluster.
- Aggiorna il `PATH` per includere i file binari Slurm del cluster.
- Configura altre variabili di ambiente Slurm necessarie (,). `MANPATH` `LD_LIBRARY_PATH`
- Imposta le variabili di AWS identificazione del cluster PCS.
- Consente un'interazione senza interruzioni con il cluster AWS PCS di destinazione.

Per disattivare un ambiente cluster

- Esegui il comando di disattivazione.

```
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
```

Cosa fa il comando di disattivazione

- Ripristina la variabile di PATH ambiente originale.
- Annulla le variabili di ambiente Slurm specifiche del cluster.
- Riporta l'ambiente della shell allo stato di preattivazione.

> **Note**
>
> L'attivazione è specifica della sessione e deve provenire dalla sessione di shell in cui si desidera interagire con il cluster.

AWS Rete PCS

Il cluster AWS PCS viene creato in un Amazon VPC. Questo capitolo include i seguenti argomenti sul networking per lo scheduler e i nodi del cluster.

Ad eccezione della scelta di una sottorete in cui avviare le istanze, è necessario utilizzare i modelli di EC2 avvio per configurare la rete per i gruppi di nodi di calcolo AWS PCS. Per ulteriori informazioni sui modelli di avvio, consulta Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS.

Argomenti

AWS Requisiti e considerazioni su PCS, VPC e sottorete

Quando si crea un cluster AWS PCS, si specifica un VPC, una sottorete in quel VPC. Questo argomento fornisce una panoramica dei requisiti e delle considerazioni specifici del AWS PCS per il VPC e le sottoreti utilizzate con il cluster. Se non disponi di un VPC da utilizzare con AWS PCS, puoi crearne uno utilizzando un modello fornito AWS. CloudFormation Per ulteriori informazioni VPCs, consulta Virtual private cloud (VPC) nella Amazon VPC User Guide.

Considerazioni e requisiti relativi al VPC

Durante la creazione di un cluster, il VPC specificato deve soddisfare i requisiti e le considerazioni seguenti:

- Il VPC deve disporre di un numero sufficiente di indirizzi IP disponibili per il cluster, tutti i nodi e le altre risorse del cluster che si desidera creare. Per ulteriori informazioni, consulta la sezione Indirizzamento IP per le tue sottoreti VPCs e subnet nella Amazon VPC User Guide.
- Se il tuo cluster utilizza: IPv6
  - Associa un blocco IPv6 CIDR al tuo VPC. Per ulteriori informazioni, consulta Crea un VPC nella Guida per l'utente di Amazon VPC.

> **Important**
>
> Sebbene sia possibile configurare il VPC con entrambi IPv4 e IPv6, è possibile scegliere un solo tipo di rete per il cluster.

  - Abilita l'assegnazione automatica degli IPv6 indirizzi per le tue sottoreti.
  - Per ulteriori informazioni, consulta:
    - IPv6 su AWS
    - Comprendere l' IPv6 indirizzamento su AWS e progettare un piano di indirizzamento scalabile
- Il VPC deve avere un nome host DNS e un supporto per la risoluzione DNS. In caso contrario, i nodi non possono registrare il cluster di clienti. Per ulteriori informazioni, consulta Attributi DNS per il VPC nella Guida per l'utente di Amazon VPC.
- Il VPC potrebbe richiedere l'utilizzo di endpoint VPC AWS PrivateLink per poter contattare l'API PCS. AWS Per ulteriori informazioni, consulta Connect your VPC ai servizi utilizzando AWS PrivateLink nella Amazon VPC User Guide.

> **Important**
>
> AWS PCS non supporta un VPC con tenancy di istanza dedicata. Il VPC che usi per AWS PCS deve utilizzare la tenancy dell'`default`istanza. Puoi modificare la tenancy dell'istanza per un VPC esistente. Per ulteriori informazioni, consulta Modificare la tenance dell'istanza di un VPC nella Amazon Elastic Compute Cloud User Guide.

Considerazioni e requisiti relativi alle sottoreti

Quando crei un cluster Slurm, AWS PCS crea un'interfaccia di rete elastica (ENI) nella sottorete specificata. Questa interfaccia di rete consente la comunicazione tra il controller dello scheduler e il VPC del cliente. L'interfaccia di rete consente inoltre a Slurm di comunicare con i componenti distribuiti nel tuo account. È possibile specificare la sottorete per un cluster solo al momento della creazione.

Requisiti relativi alla sottorete per i cluster

La sottorete specificata quando si crea un cluster deve soddisfare i seguenti requisiti:

- La sottorete deve avere almeno 1 indirizzo IP per l'utilizzo da parte AWS di PCS.
- Se il cluster utilizza IPv6, tutte le sottoreti del cluster devono utilizzarlo. IPv6

> **Important**
>
> I gruppi di nodi di calcolo configurati con l'esempio AWS PCS AMIs e le interfacce di rete multiple non funzioneranno attualmente se le sottoreti sono configurate solo per l'uso. IPv6 Utilizza invece sottoreti dual-stack (and) o solo sottoreti. IPv4 IPv6 IPv4 Per ulteriori informazioni, consulta Utilizzo di Amazon Machine Images (AMIs) di esempio con AWS PCS.

- La sottorete non può risiedere in, o in una zona locale. AWS Outposts AWS Wavelength AWS
- La sottorete può essere pubblica o privata. Si consiglia di specificare una sottorete privata, se possibile. Una sottorete pubblica è una sottorete con una tabella di routing che include un percorso verso un gateway Internet; una sottorete privata è una sottorete con una tabella di routing che non include un percorso verso un gateway Internet.

Requisiti relativi alla sottorete per i nodi

È possibile distribuire nodi e altre risorse del cluster nella sottorete specificata al momento della creazione del cluster AWS PCS e su altre sottoreti nello stesso VPC.

Qualsiasi sottorete in cui vengono distribuiti nodi e risorse del cluster deve soddisfare i seguenti requisiti:

- È necessario assicurarsi che la sottorete disponga di un numero sufficiente di indirizzi IP disponibili per distribuire tutti i nodi e le risorse del cluster.
- Se il tuo cluster utilizza IPv4 e intendi distribuire nodi in una sottorete pubblica, tale sottorete deve assegnare automaticamente IPv4 gli indirizzi pubblici.

> **Note**
>
> Le istanze in una sottorete pubblica devono utilizzare un gruppo di sicurezza con regole in entrata che consentano il traffico proveniente da indirizzi IP pubblici. A meno che non siano previste restrizioni specifiche relative all'indirizzo di origine, ciò significa un indirizzo di IPv4 origine 0.0.0.0/0 o un indirizzo di origine di: :/0. IPv6

- Se la sottorete in cui distribuisci i nodi è una sottorete privata e la relativa tabella di routing non include un percorso verso un dispositivo NAT (Network Address Translation) () ()IPv4, aggiungi gli endpoint VPC utilizzando il VPC del cliente. AWS PrivateLink Gli endpoint VPC sono necessari per tutti i AWS servizi contattati dai nodi. L'unico endpoint richiesto è che AWS PCS consenta al nodo di richiamare l'azione dell'`RegisterComputeNodeGroupInstance`API. Per ulteriori informazioni, vedere RegisterComputeNodeGroupInstancenel AWS PCS API Reference.
- Lo stato della sottorete pubblica o privata non influisce sul AWS PCS; gli endpoint richiesti devono essere raggiungibili.

Creazione di un VPC per il AWS cluster PCS

Puoi creare un Amazon Virtual Private Cloud (Amazon VPC) per i tuoi cluster all'interno del AWS Parallel Computing Service (AWS PCS).

Usa Amazon VPC per lanciare risorse VPC in una rete virtuale che hai definito. Questa rete virtuale è simile a una rete tradizionale da gestire all'interno del proprio data center, ma con i vantaggi dell'infrastruttura scalabile di Amazon Web Services. Ti consigliamo di avere una conoscenza approfondita del servizio Amazon VPC prima di distribuire cluster VPC di produzione. Per ulteriori informazioni, consulta Cos'è Amazon VPC? in modalità visuale d'autore. Guida per l'utente di Amazon VPC.

Un cluster PCS, nodi e risorse di supporto (come file system e servizi di directory) vengono distribuiti all'interno del tuo Amazon VPC. Se desideri utilizzare un Amazon VPC esistente con PCS, deve soddisfare i requisiti descritti in. AWS Requisiti e considerazioni su PCS, VPC e sottorete Questo argomento descrive come creare un VPC che soddisfi i requisiti PCS utilizzando un modello fornito AWS. CloudFormation Dopo l'implementazione di un modello, puoi visualizzare le risorse create dal modello per sapere esattamente quali risorse ha creato e la configurazione di tali risorse.

Prerequisiti

Per creare un Amazon VPC per PCS, devi disporre delle autorizzazioni IAM necessarie per creare risorse Amazon VPC. Queste risorse sono sottoreti VPCs, gruppi di sicurezza, tabelle e percorsi di routing e gateway Internet e NAT. Per ulteriori informazioni, consulta Creare un VPC con una sottorete pubblica nella Amazon VPC User Guide. Per esaminare l'elenco completo di Amazon EC2, consulta Azioni, risorse e chiavi di condizione per Amazon EC2 nel Service Authorization Reference.

Crea un Amazon VPC

Crea un VPC copiando e incollando l'URL appropriato per il Regione AWS luogo in cui utilizzerai PCS. Puoi anche scaricare il CloudFormation modello e caricarlo tu stesso sulla CloudFormation console.

- US East (N. Virginia) (Stati Uniti orientali (Virginia settentrionale)) (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
```

- US East (Ohio) (Stati Uniti orientali (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
```

- US West (Oregon) (Stati Uniti occidentali (Oregon)) (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 modello

```
https://aws-hpc-recipes.s3.us-east-1.amazonaws.com/main/recipes/net/hpc_large_scale/
assets/main.yaml
```

Per creare un Amazon VPC per PCS

1. Apri il modello nella CloudFormation console.

   > **Note**
   >
   > Questi sono precompilati nel modello in modo che tu possa semplicemente lasciarli come valori predefiniti.

2. In Fornisci un nome per lo stack, quindi per nome dello stack, inserisci. `hpc-networking`
3. In Parametri, inserisci i seguenti dettagli:

a. In VPC, quindi, inserisci CidrBlock`10.3.0.0/16`
b. In Sottoreti A:
    i. Quindi A, CidrPublicSubnetinserisci `10.3.0.0/20`
    ii. Poi CidrPrivateSubnetA, entra `10.3.128.0/20`
c. In Sottoreti B:
    i. Quindi CidrPublicSubnetB, inserisci `10.3.16.0/20`
    ii. Poi CidrPrivateSubnetA, entra `10.3.144.0/20`
d. In Sottoreti C:
    i. Per ProvisionSubnetsC, seleziona. `True`

> **Note**
>
> Se stai creando un VPC in una regione con meno di tre zone di disponibilità, questa opzione verrà ignorata se impostata su. `True`

    ii. Quindi CidrPublicSubnetB, inserisci `10.3.32.0/20`
    iii. Poi CidrPrivateSubnetA, entra `10.3.160.0/20`
4. In Capacità, seleziona la casella Riconosco che AWS CloudFormation potrebbe creare risorse IAM.

Monitora lo stato dello CloudFormation stack. Una volta raggiunta`CREATE_COMPLETE`, la risorsa VPC è pronta per l'uso.

> **Note**
>
> Per vedere tutte le risorse create dal CloudFormation modello, apri la CloudFormation console. Scegli lo stack `hpc-networking`, quindi la scheda Resources (Risorse).

Gruppi di sicurezza in AWS PCS

I gruppi di sicurezza in Amazon EC2 agiscono come firewall virtuali per controllare il traffico in entrata e in uscita verso le istanze. Utilizza un modello di avvio per un gruppo di nodi di calcolo AWS

PCS per aggiungere o rimuovere gruppi di sicurezza alle relative istanze. Se il modello di lancio non contiene interfacce di rete, utilizzalo `SecurityGroupIds` per fornire un elenco di gruppi di sicurezza. Se il modello di lancio definisce le interfacce di rete, è necessario utilizzare il `Groups` parametro per assegnare gruppi di sicurezza a ciascuna interfaccia di rete. Per ulteriori informazioni sui modelli di avvio, consulta Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS.

> **Note**
>
> Le modifiche alla configurazione del gruppo di sicurezza nel modello di avvio influiscono solo sulle nuove istanze avviate dopo l'aggiornamento del gruppo di nodi di calcolo.

Requisiti e considerazioni sui gruppi di sicurezza

AWS PCS crea un'interfaccia di rete elastica (ENI) tra account nella sottorete specificata durante la creazione di un cluster. Ciò fornisce allo scheduler HPC, che è in esecuzione in un account gestito da AWS, un percorso per comunicare con le istanze EC2 lanciate da PCS. AWS È necessario fornire un gruppo di sicurezza per tale ENI che consenta la comunicazione bidirezionale tra lo scheduler ENI e le istanze EC2 del cluster.

Un modo semplice per farlo è creare un gruppo di sicurezza autoreferenziato e permissivo che consenta il traffico su tutte le porte tra tutti i membri del gruppo. TCP/IP Puoi collegarlo sia al cluster che alle istanze EC2 del gruppo di nodi.

Esempio di configurazione permissiva del gruppo di sicurezza

IPv4

Tipo di regola	Protocolli	Porte	Origine	Destinazione
In entrata	Tutti	Tutti	Personale	
In uscita	Tutti	Tutti		0.0.0.0/0
In uscita	Tutti	Tutti		Personale

IPv6

Tipo di regola	Protocolli	Porte	Origine	Destinazione
In entrata	Tutti	Tutti	Personale	
In uscita	Tutti	Tutti		::/0
In uscita	Tutti	Tutti		Personale

Queste regole consentono a tutto il traffico di fluire liberamente tra il controller Slurm e i nodi, consentono tutto il traffico in uscita verso qualsiasi destinazione e abilitano il traffico EFA.

Esempio di configurazione restrittiva del gruppo di sicurezza

È inoltre possibile limitare le porte aperte tra il cluster e i relativi nodi di elaborazione. Per lo scheduler Slurm, il gruppo di sicurezza collegato al cluster deve consentire le seguenti porte:

- 6817: abilita le connessioni in entrata da istanze EC2 `slurmctld`
- 6818: abilita le connessioni in uscita e l'esecuzione su istanze EC2 `slurmctld` `slurmd`

Il gruppo di sicurezza collegato ai nodi di elaborazione deve consentire le seguenti porte:

- 6817: abilita le connessioni in uscita `slurmctld` da istanze EC2.
- 6818: abilita le connessioni in entrata e in uscita da e verso le istanze del gruppo di nodi `slurmd` `slurmctld` `slurmd`
- 60001—63000: connessioni in entrata e in uscita tra istanze di gruppi di nodi da supportare `srun`
- Traffico EFA tra istanze del gruppo di nodi. Per ulteriori informazioni, consulta Preparare un gruppo di sicurezza compatibile con EFA nella Guida per l'utente per le istanze Linux
- Qualsiasi altro traffico internodale richiesto dal carico di lavoro

Interfacce di rete multiple in AWS PCS

Alcune istanze EC2 dispongono di più schede di rete. Ciò consente loro di fornire prestazioni di rete più elevate, comprese capacità di larghezza di banda superiori a 100 Gbps e una migliore gestione

dei pacchetti. Per ulteriori informazioni sulle istanze con più schede di rete, consulta le interfacce di rete elastiche nella Amazon Elastic Compute Cloud User Guide.

Configura schede di rete aggiuntive per le istanze in un gruppo di nodi di calcolo AWS PCS aggiungendo interfacce di rete al relativo modello di lancio EC2. Di seguito è riportato un esempio di modello di avvio che abilita due schede di rete, ad esempio quelle disponibili su un'istanza. `hpc7a.96xlarge` Si notino i seguenti dettagli:

- La sottorete per ogni interfaccia di rete deve essere la stessa scelta durante la configurazione del gruppo di nodi di calcolo AWS PCS che utilizzerà il modello di avvio.
- Il dispositivo di rete primario, su cui avverranno le comunicazioni di rete di routine come il traffico SSH e HTTPS, viene stabilito impostando un di. `DeviceIndex` `0` Le altre interfacce di rete hanno un`DeviceIndex`. `1` Può esserci una sola interfaccia di rete principale, tutte le altre interfacce sono secondarie.
- Tutte le interfacce di rete devono avere un'interfaccia univoca. `NetworkCardIndex` Una pratica consigliata consiste nel numerarle in sequenza così come sono definite nel modello di avvio.
- I gruppi di sicurezza per ogni interfaccia di rete vengono impostati utilizzando`Groups`. In questo esempio, un gruppo di sicurezza SSH in ingresso (`sg-`*`SshSecurityGroupId`*) viene aggiunto all'interfaccia di rete principale, oltre al gruppo di sicurezza che abilita le comunicazioni all'interno del cluster (). `sg-`*`ClusterSecurityGroupId`* Infine, un gruppo di sicurezza che consente le connessioni in uscita a Internet (`sg-`*`InternetOutboundSecurityGroupId`*) viene aggiunto alle interfacce primarie e secondarie.

```
{
    "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"]
        }
    ]
}
```

Gruppi di posizionamento per istanze EC2 in PCS AWS

Puoi utilizzare un gruppo di collocamento per influenzare il posizionamento delle istanze EC2 in base alle esigenze del carico di lavoro che viene eseguito su di esse.

Tipi di gruppi di posizionamento

- Cluster: raggruppa le istanze in una zona di disponibilità per ottimizzare le comunicazioni a bassa latenza.
- Partizione: distribuisce le istanze su partizioni logiche per massimizzare la resilienza.
- Spread: impone rigorosamente l'avvio di un numero limitato di istanze su hardware distinto, il che può anche favorire la resilienza.

Per ulteriori informazioni, consulta i gruppi di posizionamento per le tue istanze Amazon EC2 nella Amazon Elastic Compute Cloud User Guide.

Ti consigliamo di includere un gruppo di posizionamento del cluster quando configuri un gruppo di nodi di calcolo AWS PCS per utilizzare Elastic Fabric Adapter (EFA).

Per creare un gruppo di posizionamento del cluster che funzioni con EFA

1. Crea un gruppo di posizionamento con il tipo cluster per il gruppo di nodi di calcolo.
   - Utilizzate il seguente AWS CLI comando:

     ```
     aws ec2 create-placement-group --strategy cluster --group-name PLACEMENT-GROUP-NAME
     ```

   - Potete anche utilizzare un CloudFormation modello per creare un gruppo di posizionamenti. Per ulteriori informazioni, consulta Lavorare con CloudFormation i modelli nella Guida AWS CloudFormation per l'utente. Scarica il modello dal seguente URL e caricalo nella CloudFormation console.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/enable_efa/assets/efa-placement-group.yaml
```

2. Includi il gruppo di posizionamento nel modello di lancio EC2 per il gruppo di nodi di calcolo AWS PCS.

Utilizzo di Elastic Fabric Adapter (EFA) con PCS AWS

Elastic Fabric Adapter (EFA) è un'interconnessione di rete avanzata ad alte prestazioni AWS che puoi collegare alla tua istanza EC2 per accelerare le applicazioni di High Performance Computing (HPC) e machine learning. L'abilitazione delle applicazioni in esecuzione su un cluster AWS PCS con EFA implica la configurazione delle istanze del gruppo di nodi di calcolo AWS PCS per utilizzare EFA come segue.

> **Note**
>
> Installa EFA su un'AMI AWS compatibile con PCS: sull'AMI utilizzata nel AWS gruppo di nodi di calcolo PCS deve essere installato e caricato il driver EFA. Per informazioni su come creare un'AMI personalizzata con il software EFA installato, consultaImmagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Indice

Identifica le istanze EC2 abilitate per EFA

Per utilizzare EFA, tutti i tipi di istanze consentiti per un gruppo di calcolo AWS PCS devono supportare EFA e devono avere lo stesso numero di v (e se appropriato). CPUs GPUs Per un elenco di istanze abilitate per EFA, consulta Elastic Fabric Adapter per carichi di lavoro HPC e ML su Amazon EC2 nella Amazon Elastic Compute Cloud User Guide. Puoi anche utilizzare il AWS CLI per visualizzare un elenco di tipi di istanze che supportano EFA. Sostituiscilo *region-code* con quello Regione AWS in cui usi AWS PCS, ad esempio`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**
>
> Determina quante interfacce di rete sono disponibili: alcune istanze EC2 dispongono di più schede di rete. Ciò consente loro di averne più di una. EFAs Per ulteriori informazioni, consulta Interfacce di rete multiple in AWS PCS.

Crea un gruppo di sicurezza per supportare le comunicazioni EFA

AWS CLI

È possibile utilizzare il seguente AWS CLI comando per creare un gruppo di sicurezza che supporti EFA. Il comando restituisce un ID del gruppo di sicurezza. Effettua le seguenti sostituzioni:

- *region-code*— Specificare Regione AWS dove si utilizza AWS PCS, ad esempio`us-east-1`.
- *vpc-id*— Specificare l'ID del VPC utilizzato per AWS PCS.
- *efa-group-name*— Fornisci il nome scelto per il gruppo di sicurezza.

```
aws ec2 create-security-group \
    --group-name efa-group-name \
    --description "Security group to enable EFA traffic" \
```

```
    --vpc-id vpc-id \
    --region region-code
```

Utilizzate i seguenti comandi per allegare le regole del gruppo di sicurezza in entrata e in uscita. Effettua la seguente sostituzione:

- *efa-secgroup-id*— Fornisci l'ID del gruppo di sicurezza EFA che hai appena creato.

```
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

È possibile utilizzare un CloudFormation modello per creare un gruppo di sicurezza che supporti EFA. Scarica il modello dal seguente URL, quindi caricalo nella AWS CloudFormation console.

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

Con il modello aperto nella AWS CloudFormation console, inserisci le seguenti opzioni.

- In Fornisci un nome per lo stack
  - In Nome dello stack, inserisci un nome come. `efa-sg-stack`
- In Parametri
  - In SecurityGroupName, inserisci un nome come`efa-sg`.
  - In VPC, seleziona il VPC in cui utilizzerai PCS. AWS

Completa la creazione dello CloudFormation stack e monitorane lo stato. Quando arriva, `CREATE_COMPLETE` il gruppo di sicurezza EFA è pronto per l'uso.

(Facoltativo) Crea un gruppo di collocamento

Ti consigliamo di avviare tutte le istanze che utilizzano EFA in un gruppo di posizionamento del cluster per ridurre al minimo la distanza fisica tra di esse. Crea un gruppo di posizionamento per ogni gruppo di nodi di calcolo in cui intendi utilizzare EFA. Vedi Gruppi di posizionamento per istanze EC2 in PCS AWS per creare un gruppo di posizionamento per il tuo gruppo di nodi di calcolo.

Crea o aggiorna un modello di lancio EC2

Le interfacce di rete EFA sono configurate nel modello di lancio EC2 per un gruppo di nodi di calcolo AWS PCS. Se sono presenti più schede di rete, è possibile configurarne più EFAs di una. Il gruppo di sicurezza EFA e il gruppo di collocamento opzionale sono inclusi anche nel modello di lancio.

Ecco un esempio di modello di avvio per istanze con due schede di rete, come hpc7a.96xlarge. Le istanze verranno avviate in un gruppo di collocamento del cluster. `subnet-`*`SubnetID1`* `pg-`*`PlacementGroupId1`*

I gruppi di sicurezza devono essere aggiunti in modo specifico a ciascuna interfaccia EFA. Ogni EFA ha bisogno del gruppo di sicurezza che abilita il traffico EFA (). `sg-`*`EfaSecGroupId`* Gli altri gruppi di sicurezza, in particolare quelli che gestiscono traffico regolare come SSH o HTTPS, devono essere collegati solo all'interfaccia di rete principale (indicata da un `DeviceIndex` di). `0` I modelli di avvio in cui sono definite le interfacce di rete non supportano l'impostazione di gruppi di sicurezza mediante il `SecurityGroupIds` parametro: è necessario impostare un valore per `Groups` ogni interfaccia di rete configurata.

```
{
    "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"]
        }
    ]
}
```

Crea o aggiorna gruppi di nodi di calcolo per EFA

I gruppi di nodi di calcolo AWS PCS devono contenere istanze con lo stesso numero di vCPUs, architettura di processore e supporto EFA. Configura il gruppo di nodi di calcolo per utilizzare l'AMI con il software EFA installato su di esso e per utilizzare il modello di avvio che configura le interfacce di rete abilitate per EFA.

(Facoltativo) Prova EFA

È possibile dimostrare la comunicazione abilitata all'EFA tra due nodi in un gruppo di nodi di calcolo eseguendo il `fi_pingpong` programma, incluso nell'installazione del software EFA. Se questo test ha esito positivo, è probabile che EFA sia configurato correttamente.

Per iniziare, sono necessarie due istanze in esecuzione nel gruppo di nodi di calcolo. Se il gruppo di nodi di calcolo utilizza una capacità statica, dovrebbero esserci già delle istanze disponibili. Per un gruppo di nodi di calcolo che utilizza capacità dinamica, puoi avviare due nodi utilizzando il comando. `salloc` Ecco un esempio tratto da un cluster con un gruppo di nodi dinamico denominato `hpc7g` associato a una coda denominata. `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
```

Scopri l'indirizzo IP per i due nodi allocati utilizzando. `scontrol` Nell'esempio che segue, gli indirizzi sono `10.3.140.69` for `hpc7g-1` e `10.3.132.211` for`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
```

Connect a uno dei nodi (in questo casohpc7g-1) utilizzando SSH (o SSM). Tieni presente che si tratta di un indirizzo IP interno, quindi potresti dover connetterti da uno dei tuoi nodi di accesso se usi SSH. Tieni inoltre presente che l'istanza deve essere configurata con una chiave SSH tramite il modello di avvio del gruppo di nodi di calcolo.

```
% ssh ec2-user@10.3.140.69
```

Ora, avvia `fi_pingpong` in modalità server.

```
/opt/amazon/efa/bin/fi_pingpong -p efa
```

Connect alla seconda istanza (`hpc7g-2`).

```
% ssh ec2-user@10.3.132.211
```

Esegui `fi_pingpong` in modalità client, con connessione al server attiva`hpc7g-1`. L'output dovrebbe essere simile a quello dell'esempio seguente.

```
% /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
```

(Facoltativo) Utilizzate un CloudFormation modello per creare un modello di lancio compatibile con EFA

Poiché esistono diverse dipendenze dalla configurazione di EFA, è stato fornito un CloudFormation modello che è possibile utilizzare per configurare un gruppo di nodi di calcolo. Supporta istanze con un massimo di quattro schede di rete. Per ulteriori informazioni sulle istanze con più schede di rete, consulta le interfacce di rete elastiche nella Amazon Elastic Compute Cloud User Guide.

Scarica il CloudFormation modello dal seguente URL, quindi caricalo sulla CloudFormation console in Regione AWS cui usi PCS. AWS

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/enable_efa/assets/pcs-lt-efa.yaml
```

Con il modello aperto nella CloudFormation console, inserisci i seguenti valori. Tieni presente che il modello fornirà alcuni valori di parametro predefiniti: puoi lasciarli come valori predefiniti.

- In Fornisci un nome per lo stack
  - In Nome dello stack, inserisci un nome descrittivo. Ti consigliamo di incorporare il nome che sceglierai per il tuo gruppo di nodi di calcolo AWS PCS, ad esempio. *NODEGROUPNAME*`-efa-lt`
- In Parametri
  - In NumberOfNetworkCards, scegli il numero di schede di rete nelle istanze che faranno parte del tuo gruppo di nodi.
  - In VpcId, scegli il VPC in cui è distribuito il cluster AWS PCS.
  - In NodeGroupSubnetId, scegli la sottorete nel VPC del cluster in cui verranno lanciate le istanze abilitate per EFA.
  - Sotto PlacementGroupName, lascia il campo vuoto per creare un nuovo gruppo di posizionamento del cluster per il gruppo di nodi. Se disponi di un gruppo di collocamento esistente che desideri utilizzare, inseriscine il nome qui.
  - In ClusterSecurityGroupId, scegli il gruppo di sicurezza che stai utilizzando per consentire l'accesso ad altre istanze del cluster e all'API AWS PCS. Molti clienti scelgono il gruppo di sicurezza predefinito dal proprio VPC del cluster.
  - In SshSecurityGroupId, fornisci l'ID di un gruppo di sicurezza che stai utilizzando per consentire l'accesso SSH in entrata ai nodi del cluster.
  - Per SshKeyName, seleziona la coppia di chiavi SSH per l'accesso ai nodi del cluster.
  - Per LaunchTemplateName, inserisci un nome descrittivo per il modello di lancio, ad esempio. *NODEGROUPNAME*`-efa-lt` Il nome deve essere univoco per il luogo Account AWS in Regione AWS cui utilizzerai AWS PCS.
- In Capacità
  - Seleziona la casella Riconosco che AWS CloudFormation potrebbe creare risorse IAM.

Monitora lo stato dello CloudFormation stack. Quando raggiunge `CREATE_COMPLETE` il modello di lancio è pronto per essere utilizzato. Usalo con un gruppo di nodi di calcolo AWS PCS, come descritto sopra inCrea o aggiorna gruppi di nodi di calcolo per EFA.

Utilizzo di file system di rete con AWS PCS

È possibile collegare i file system di rete ai nodi avviati in un gruppo di nodi di calcolo AWS Parallel Computing Service (AWS PCS) per fornire una posizione persistente in cui è possibile scrivere e accedere a dati e file. Puoi utilizzare i file system forniti da AWS servizi, tra cui Amazon Elastic File System (Amazon EFS), Amazon FSx for Lustre, Amazon FSx for NetApp ONTAP, Amazon FSx for OpenZFS e Amazon File Cache. Puoi anche utilizzare file system autogestiti, come i server NFS.

In questo argomento vengono fornite considerazioni ed esempi sull'utilizzo dei file system di rete con PCS. AWS

Considerazioni sull'utilizzo dei file system di rete

I dettagli di implementazione per i vari file system sono diversi, ma ci sono alcune considerazioni comuni.

- Il software del file system pertinente deve essere installato sull'istanza. Ad esempio, per utilizzare Amazon FSx for Lustre, deve essere presente il Lustre pacchetto appropriato. Ciò può essere ottenuto includendolo nell'AMI del gruppo di nodi di calcolo o utilizzando uno script che viene eseguito all'avvio dell'istanza.
- Deve esserci un percorso di rete tra il file system di rete condiviso e le istanze del gruppo di nodi di calcolo.
- Le regole del gruppo di sicurezza sia per il file system di rete condiviso che per le istanze del gruppo di nodi di calcolo devono consentire le connessioni alle porte pertinenti.
- È necessario mantenere uno spazio dei nomi POSIX utente e di gruppo coerente tra le risorse che accedono ai file system. In caso contrario, i lavori e i processi interattivi eseguiti sul cluster PCS potrebbero riscontrare errori di autorizzazione.
- I montaggi del file system vengono eseguiti utilizzando modelli di EC2 avvio. Errori o timeout nel montaggio di un file system di rete possono impedire la disponibilità di istanze per l'esecuzione dei job. Ciò, a sua volta, può comportare costi imprevisti. Per ulteriori informazioni sul debug dei modelli di avvio, consulta. Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS

Esempi di montaggi di rete

Puoi creare file system utilizzando Amazon EFS, Amazon FSx for Lustre, Amazon for NetApp ONTAP, Amazon FSx FSx for OpenZFS e Amazon File Cache. Espandi la sezione pertinente di seguito per vedere un esempio di ogni montaggio di rete.

Amazon EFS

Configurazione del file system

Crea un file system Amazon EFS. Assicurati che abbia un target di montaggio in ogni zona di disponibilità in cui lancerai le istanze del gruppo di nodi di calcolo PCS. Assicurati inoltre che ogni target di montaggio sia associato a un gruppo di sicurezza che consenta l'accesso in entrata e in uscita dalle istanze del gruppo di nodi di calcolo PCS. Per ulteriori informazioni, consulta Mount targets and security group nella Amazon Elastic File System User Guide.

Modello di lancio

Aggiungi i gruppi di sicurezza dalla configurazione del file system al modello di lancio che utilizzerai per il gruppo di nodi di calcolo.

Includi i dati utente che utilizzano un `cloud-config` meccanismo per montare il file system Amazon EFS. Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`mount-point-directory`*— Il percorso su ogni istanza in cui monterai Amazon EFS
- *`filesystem-id`*— L'ID del file system per il file system 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 per Lustre

Configurazione del file system

Crea un file system FSx for Lustre nel VPC dove utilizzerai AWS PCS. Per ridurre al minimo i trasferimenti tra zone, esegui la distribuzione in una sottorete nella stessa zona di disponibilità, dove lancerai la maggior parte delle istanze del gruppo di nodi di calcolo PCS. Assicurati che il file system sia associato a un gruppo di sicurezza che consenta l'accesso in entrata e in uscita dalle istanze del gruppo di nodi di calcolo PCS. Per ulteriori informazioni sui gruppi di sicurezza, consulta Controllo degli accessi al file system con Amazon VPC nella Guida per l'utente di Amazon FSx for Lustre.

Modello di lancio

Includi i dati utente utilizzati `cloud-config` per montare il file system FSx for Lustre. Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`mount-point-directory`*— Il percorso su un'istanza in cui si desidera montare FSx Lustre
- *`filesystem-id`*— L'ID del file system per il file system FSx for Lustre
- *`mount-name`*— Il nome di montaggio per il file FSx system for Lustre
- *`region-code`*— Il Regione AWS luogo in cui è distribuito il file system FSx for Lustre (deve essere lo stesso del sistema AWS PCS in uso)
- (Facoltativo)*`latest`*: qualsiasi versione Lustre supportata da for Lustre FSx

```
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 per NetApp ONTAP

Configurazione del file system

Crea un file system Amazon FSx for NetApp ONTAP nel VPC dove utilizzerai AWS PCS. Per ridurre al minimo i trasferimenti tra zone, esegui la distribuzione in una sottorete nella stessa zona di disponibilità, dove lancerai la maggior parte delle istanze del gruppo di nodi di calcolo PCS. AWS Assicurati che il file system sia associato a un gruppo di sicurezza che consenta l'accesso in entrata e in uscita dalle istanze del gruppo di nodi di calcolo PCS. AWS Per ulteriori informazioni sui gruppi di sicurezza, consulta File System Access Control with Amazon VPC nella Guida FSx per l'utente di for ONTAP.

Modello di lancio

Includi i dati utente utilizzati `cloud-config` per montare il volume root per un file system FSx for ONTAP. Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`mount-point-directory`*— Il percorso su un'istanza in cui desideri montare il volume FSx for ONTAP
- *`svm-id`*— L'ID SVM per il file system FSx for ONTAP
- *`filesystem-id`*— L'ID del file system per il file system FSx for ONTAP
- *`region-code`*— Il Regione AWS luogo in cui viene distribuito il file system FSx for ONTAP (deve essere lo stesso del sistema AWS PCS in uso)
- *`volume-name`*— Il nome del volume FSx for 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 per OpenZFS

Configurazione del file system

Crea un file system FSx per OpenZFS nel VPC dove utilizzerai PCS. AWS Per ridurre al minimo i trasferimenti tra zone, esegui la distribuzione in una sottorete nella stessa zona di disponibilità, dove lancerai la maggior parte delle istanze del gruppo di nodi di calcolo PCS. AWS Assicurati che il file system sia associato a un gruppo di sicurezza che consenta l'accesso in entrata e in uscita dalle istanze del gruppo di nodi di calcolo PCS. AWS Per ulteriori informazioni sui gruppi di sicurezza, consulta Gestire l'accesso al file system con Amazon VPC nella Guida FSx per l'utente di OpenZFS.

Modello di lancio

Includi i dati utente utilizzati `cloud-config` per montare il volume root per un file system FSx per OpenZFS. Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`mount-point-directory`*— Il percorso su un'istanza in cui si desidera montare la condivisione FSx for OpenZFS
- *`filesystem-id`*— L'ID del file system FSx per il file system di OpenZFS
- *`region-code`*— Il Regione AWS luogo in cui è distribuito il file system FSx per OpenZFS (deve essere lo stesso del sistema PCS in uso) 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

Configurazione del file system

Crea un Amazon File Cache nel VPC dove AWS utilizzerai PCS. Per ridurre al minimo i trasferimenti tra zone, scegli una sottorete nella stessa zona di disponibilità in cui lancerai la maggior parte delle istanze del gruppo di nodi di calcolo PCS. Assicurati che File Cache sia associato a un gruppo di sicurezza che consenta il traffico in entrata e in uscita sulla porta 988 tra le istanze PCS e la File Cache. Per ulteriori informazioni sui gruppi di sicurezza, consulta la sezione Controllo dell'accesso alla cache con Amazon VPC nella Amazon File Cache User Guide.

Modello di lancio

Aggiungi i gruppi di sicurezza dalla configurazione del file system al modello di lancio che utilizzerai per il gruppo di nodi di calcolo.

Includi i dati utente utilizzati `cloud-config` per montare Amazon File Cache. Sostituisci i seguenti valori in questo script con i tuoi dati:

- *`mount-point-directory`*— Il percorso su un'istanza in cui si desidera montare FSx Lustre
- *`cache-dns-name`*— Il nome DNS (Domain Name System) per la File Cache
- *`mount-name`*— Il nome di montaggio per la File Cache

```
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) per AWS PCS

AWS PCS funziona con AMIs ciò che fornite, offrendo una grande flessibilità nel software e nella configurazione presenti sui nodi del cluster. Se stai provando AWS PCS, puoi usare un'AMI di esempio fornita e gestita da AWS. Se utilizzi AWS PCS in produzione, ti consigliamo di crearne uno tuo AMIs. Questo argomento spiega come scoprire e utilizzare l'esempio AMIs, nonché come crearne e utilizzarne uno personalizzato AMIs.

Argomenti

Utilizzo di Amazon Machine Images (AMIs) di esempio con AWS PCS

AWS fornisce esempi AMIs che puoi usare come punto di partenza per lavorare con AWS PCS.

> **Important**
>
> AMIs Gli esempi sono a scopo dimostrativo e non sono consigliati per carichi di lavoro di produzione.

> **Important**
>
> I gruppi di nodi di calcolo configurati con AWS PCS sample AMIs e interfacce di rete multiple non funzioneranno attualmente se le sottoreti sono configurate solo per l'uso. IPv6 Utilizza invece sottoreti dual-stack (and) o solo sottoreti. IPv4 IPv6 IPv4

Trova l'esempio AWS PCS attuale AMIs

Console di gestione AWS

Gli esempi di AWS PCS AMIs hanno la seguente convenzione di denominazione:

```
aws-pcs-sample_ami-OS-architecture-scheduler-scheduler-major-version
```

Valori accettati

- *OS* – amzn2
- *architecture* – x86_64 o arm64
- *scheduler* – slurm
- *scheduler-major-version* – 25.05

Per trovare un esempio di AWS PCS AMIs

1. Apri la EC2 console Amazon.
2. Accedi a AMIs.
3. Scegliere Immagini pubbliche.
4. In Trova AMI per attributo o tag, cerca un AMI utilizzando il nome del modello.

   Esempi

   - AMI di esempio per Slurm 25.05 su istanze Arm64

     ```
     aws-pcs-sample_ami-amzn2-arm64-slurm-25.05
     ```

   - AMI di esempio per Slurm 25.05 su istanze x86

     ```
     aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05
     ```

   > **Note**
   >
   > Se ce ne sono più AMIs, usa l'AMI con il timestamp più recente.

5. Usa l'ID AMI quando crei o aggiorni un gruppo di nodi di calcolo.

AWS CLI

Puoi trovare l'AMI di esempio AWS PCS più recente con i comandi seguenti. Sostituiscila *region-code* con quella Regione AWS in cui usi AWS PCS, ad esempious-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
```

- Arm 64

```
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
```

Usa l'ID AMI quando crei o aggiorni un gruppo di nodi di calcolo.

Scopri di più sull'esempio AWS PCS AMIs

Per visualizzare i contenuti e i dettagli di configurazione per le versioni correnti e precedenti dell'esempio AWS PCS AMIs, vedereNote di rilascio per l'esempio AWS PCS AMIs.

Creane uno tuo AMIs compatibile con AWS PCS

Per imparare a crearne di personalizzati AMIs che funzionino con AWS PCS, consultaImmagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS

AWS PCS è progettato per funzionare con Amazon Machine Images (AMI) che offri al servizio. Questi AMIs possono avere software e configurazioni arbitrari installati su di essi, purché abbiano l'agente AWS PCS e una versione compatibile di Slurm installati e configurati correttamente. È necessario utilizzare i programmi AWS di installazione forniti per installare il software AWS PCS

sull'AMI personalizzata. Ti consigliamo di utilizzare i programmi AWS di installazione forniti per installare Slurm sulla tua AMI personalizzata, ma puoi installare Slurm da solo se preferisci (non consigliato).

> **Note**
>
> Se vuoi provare AWS PCS senza creare un'AMI personalizzata, puoi usare un'AMI di esempio fornita da AWS. Per ulteriori informazioni, consulta Utilizzo di Amazon Machine Images (AMIs) di esempio con AWS PCS.

> **Important**
>
> AWS PCS attualmente richiede un kernel con IPv4 supporto per la comunicazione tra nodi locali, anche quando si utilizza AWS PCS in una rete di IPv6 sola rete.

Questo tutorial ti aiuta a creare un'AMI che può essere utilizzata con i gruppi di nodi di calcolo PCS per potenziare l'HPC e AI/ML i carichi di lavoro.

Argomenti

Fase 1: Avviare un'istanza temporanea

Avvia un'istanza temporanea che puoi utilizzare per installare e configurare il software AWS PCS e lo scheduler Slurm. Questa istanza viene utilizzata per creare un'AMI compatibile con AWS PCS.

Per avviare un'istanza temporanea

1. Aprire la console di Amazon EC2.

2. Nel riquadro di navigazione, scegli Istanze, quindi scegli Avvia istanze per aprire la nuova procedura guidata di avvio dell'istanza.
3. (Facoltativo) Nella sezione Nome e tag, fornisci un nome per l'istanza, ad esempio. `PCS-AMI-instance` Il nome viene assegnato all'istanza come tag di risorsa (`Name=PCS-AMI-instance`).
4. Nella sezione Application and OS Images (Immagini di applicazioni e sistema operativo), seleziona un'AMI per uno dei sistemi operativi supportati.
5. Nella sezione Instance type (Tipo di istanza), seleziona un tipo di istanza supportato.
6. Nella sezione Key pair (Coppia di chiavi), seleziona la coppia di chiavi da utilizzare per l'istanza.
7. Nella sezione Impostazioni di rete:
   - Per Firewall (gruppi di sicurezza), scegli Seleziona gruppo di sicurezza esistente, quindi seleziona un gruppo di sicurezza che consenta l'accesso SSH in entrata all'istanza.
8. Nella sezione Storage (Archiviazione), configura i volumi secondo necessità. Assicurati di configurare uno spazio sufficiente per installare le tue applicazioni e librerie.
9. Nel pannello Summary (Riepilogo), scegliere Launch instance (Avvia istanza).

Fase 2 — Installare l'agente AWS PCS

Installa l'agente che configura le istanze lanciate da AWS PCS per l'uso con Slurm. Per ulteriori informazioni sull'agente AWS PCS, vedere. AWS Versioni dell'agente PCS

Per installare l'agente AWS PCS

1. Connettersi all'istanza avviata. Per ulteriori informazioni, consulta Connect to your Linux istance.
2. (Facoltativo) Per assicurarti che tutti i pacchetti software siano aggiornati, esegui un rapido aggiornamento del software sull'istanza. Questo processo può richiedere alcuni minuti.
   - Amazon Linux 2, Amazon Linux 2023, RHEL 9, RHEL 8, Rocky Linux 9 e Rocky Linux 8

     ```
     sudo yum update -y
     ```

   - Ubuntu 22.04 e Ubuntu 24.04

     ```
     sudo apt-get update && sudo apt-get upgrade -y
     ```

3. Riavviare l'istanza e riconnettersi a essa.

4. Scarica i file di installazione dell'agente AWS PCS. I file di installazione sono impacchettati in un file tarball () `.tar.gz` compresso. Per scaricare l'ultima versione stabile, utilizzare il comando seguente. Sostituiteli *`region`* con il Regione AWS punto in cui avete lanciato l'istanza temporanea, ad esempio. `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
```

È inoltre possibile ottenere la versione più recente sostituendo il numero di versione con quello `latest` indicato nel comando precedente (ad esempio:`aws-pcs-agent-v1-latest.tar.gz`).

> **Note**
>
> Ciò potrebbe cambiare nelle future versioni del software dell'agente AWS PCS.

5. (Facoltativo) Verifica l'autenticità e l'integrità del tarball del software AWS PCS. È consigliabile eseguire questa operazione per verificare l'identità dell'autore del software e che il file non sia stato alterato o danneggiato dopo la pubblicazione.

   a. Scaricate la chiave GPG pubblica per AWS PCS e importatela nel vostro portachiavi. Sostituiscila *`region`* con il Regione AWS punto in cui hai lanciato l'istanza temporanea. Il comando dovrebbe restituire un valore di chiave. Registra il valore della chiave; lo utilizzerai nel passaggio successivo.

```
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. Eseguite il comando seguente per verificare l'impronta digitale della chiave GPG.

```
gpg --fingerprint 7EEF030EDDF5C21C
```

   Il comando dovrebbe restituire un'impronta digitale identica alla seguente:

```
1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
```

> **Important**
>
> Non eseguire lo script di installazione dell'agente AWS PCS se l'impronta digitale non corrisponde. Contatta il Supporto AWS.

c. Scarica il file della firma e verifica la firma del file tarball del software AWS PCS. Sostituiscilo *region* con il Regione AWS punto in cui hai lanciato l'istanza temporanea, ad esempio. `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
```

L'output visualizzato dovrebbe essere simile al seguente:

```
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
```

Se il risultato include `Good signature` e l'impronta digitale corrisponde all'impronta digitale restituita nel passaggio precedente, procedi al passaggio successivo.

> **Important**
>
> Non eseguire lo script di installazione del software AWS PCS se l'impronta digitale non corrisponde. Contatta il Supporto AWS.

6. Estrai i file dal file compresso `.tar.gz` e vai alla directory estratta.

```
tar -xf aws-pcs-agent-v1.3.2-1.tar.gz && \
    cd aws-pcs-agent
```

7. Installa il software AWS PCS.

```
sudo ./installer.sh
```

8. Controllate il file della versione del software AWS PCS per confermare l'avvenuta installazione.

```
cat /opt/aws/pcs/version
```

L'output visualizzato dovrebbe essere simile al seguente:

```
AGENT_INSTALL_DATE='Fri Dec 13 12:28:43 UTC 2024'
AGENT_VERSION='1.3.2'
AGENT_RELEASE='1'
```

Fase 3 — Installare Slurm

Installa una versione di Slurm compatibile con PCS. AWS Per ulteriori informazioni, consulta Versioni Slurm in PCS AWS.

> **Note**
>
> Se hai un'AMI su cui è installata una versione precedente del software Slurm, devi eseguire le seguenti operazioni per installare la nuova versione di Slurm. L'agente AWS PCS abilita la versione corretta dei binari Slurm in fase di esecuzione, in base alla versione Slurm configurata al momento della creazione del cluster.

Per installare Slurm

1. Connect alla stessa istanza temporanea in cui è stato installato il software AWS PCS.
2. Scarica il software di installazione Slurm. Il programma di installazione di Slurm è impacchettato in un file tarball () compresso. `.tar.gz` Per scaricare l'ultima versione stabile, utilizzare il comando seguente. Sostituitelo *`region`* con quello della vostra istanza temporanea, ad Regione AWS esempio. `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
```

È inoltre possibile ottenere la versione più recente sostituendo il numero di versione con `latest` il comando precedente (ad esempio:`aws-pcs-slurm-25.05-installer-latest.tar.gz`). Per un elenco completo delle versioni disponibili con checksum, consulta. Versioni Slurm in PCS AWS

> **Note**
>
> Questo potrebbe cambiare nelle future versioni del software di installazione Slurm.

3. (Facoltativo) Verifica l'autenticità e l'integrità del tarball del programma di installazione di Slurm. È consigliabile eseguire questa operazione per verificare l'identità dell'autore del software e che il file non sia stato alterato o danneggiato dopo la pubblicazione.

   a. Scarica la chiave GPG pubblica per AWS PCS e importala nel tuo portachiavi. Sostituiscila *`region`* con il Regione AWS punto in cui hai lanciato l'istanza temporanea. Il comando dovrebbe restituire un valore di chiave. Registra il valore della chiave; lo utilizzerai nel passaggio successivo.

   ```
   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. Eseguite il comando seguente per verificare l'impronta digitale della chiave GPG.

   ```
   gpg --fingerprint 7EEF030EDDF5C21C
   ```

   Il comando dovrebbe restituire un'impronta digitale identica alla seguente:

   ```
   1C24 32C1 862F 64D1 F90A  239A 7EEF 030E DDF5 C21C
   ```

   > **Important**
   >
   > Non eseguire lo script di installazione di Slurm se l'impronta digitale non corrisponde. Contatta il Supporto AWS.

   c. Scarica il file della firma e verifica la firma del file tarball del programma di installazione di Slurm. *`region`*Sostituiscilo con il Regione AWS punto in cui hai lanciato l'istanza temporanea, ad esempio. `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
```

L'output visualizzato dovrebbe essere simile al seguente:

```
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
```

Se il risultato include `Good signature` e l'impronta digitale corrisponde all'impronta digitale restituita nel passaggio precedente, procedi al passaggio successivo.

> **Important**
>
> Non eseguire lo script di installazione di Slurm se l'impronta digitale non corrisponde. Contatta il Supporto AWS.

4. Estrarre i file dal file `.tar.gz` compresso e andare alla directory estratta.

```
tar -xf aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz && \
    cd aws-pcs-slurm-25.05-installer
```

5. Installa Slurm. Il programma di installazione scarica, compila e installa Slurm e le sue dipendenze. L'operazione richiede alcuni minuti, a seconda delle specifiche dell'istanza temporanea selezionata.

```
sudo ./installer.sh -y
```

6. Controlla il file della versione dello scheduler per confermare l'installazione.

```
cat /opt/aws/pcs/scheduler/slurm-25.05/version
```

L'output visualizzato dovrebbe essere simile al seguente:

```
SLURM_INSTALL_DATE='Mon Nov 3 14:23:38 UTC 2025'
SLURM_VERSION='25.05.4'
PCS_SLURM_RELEASE='1'
```

Fase 4 — (Facoltativo) Installare driver, librerie e software applicativi aggiuntivi

Installa driver, librerie e software applicativi aggiuntivi sull'istanza temporanea. Le procedure di installazione variano a seconda delle applicazioni e delle librerie specifiche. Se non hai mai creato un'AMI personalizzata per AWS PCS, ti consigliamo di creare e testare prima un'AMI solo con il software AWS PCS e Slurm installato, quindi aggiungere in modo incrementale il tuo software e le tue configurazioni una volta confermato il successo iniziale.

Esempi

- Software Elastic Fabric Adapter (EFA). Per ulteriori informazioni, consulta Get started with EFA and MPI for HPC workload on Amazon EC2 nella Amazon Elastic Compute Cloud User Guide.
- Client Amazon Elastic File System (Amazon EFS). Per ulteriori informazioni, consulta Installazione manuale del client Amazon EFS nella Amazon Elastic File System User Guide.
- Client Lustre, per utilizzare Amazon FSx for Lustre e Amazon File Cache. Per ulteriori informazioni, consulta Installazione del client Lustre nella guida FSx per l'utente di for Lustre.
- CloudWatch Agente Amazon, per utilizzare CloudWatch Logs and Metrics. Per ulteriori informazioni, consulta Installa l' CloudWatch agente nella Amazon CloudWatch User Guide.
- AWS Neuron, per usare i tipi di istanza trn* e inf*. Per ulteriori informazioni, consultate la documentazione di Neuron.AWS
- NVIDIA Driver, CUDA e DCGM, per utilizzare i tipi di istanze p* o g*.

Fase 5 — Creare un'AMI compatibile con AWS PCS

Dopo aver installato i componenti software richiesti, crei un'AMI che puoi riutilizzare per avviare istanze nei gruppi di nodi di calcolo AWS PCS.

> **Important**
>
> AWS Attualmente PCS richiede un kernel con IPv4 supporto per la comunicazione tra nodi locali, anche quando si utilizza AWS PCS in una rete di sola rete. IPv6

Per creare un'AMI dall'istanza temporanea

1. Aprire la console di Amazon EC2.
2. Nel riquadro di navigazione, seleziona Instances (Istanze).
3. Seleziona l'istanza temporanea che hai creato. Scegli Azioni, Immagine, Crea immagine.
4. Per Create image (Crea immagine), effettua le seguenti operazioni:
    a. In Image name (Nome immagine), immettere un nome descrittivo per l'AMI.
    b. (Facoltativo) In Image description (Descrizione immagine), inserire una breve descrizione dell'AMI.
    c. Scegliere Create Image (Crea immagine).
5. Nel pannello di navigazione, scegli AMIs.
6. Individua l'AMI che hai creato nell'elenco. Attendi che il suo stato cambi da In sospeso a Disponibile, quindi usalo con un gruppo di nodi di calcolo AWS PCS.

Passaggio 6: utilizzare l'AMI personalizzata con un gruppo di nodi di calcolo AWS PCS

Puoi usare la tua AMI personalizzata con un gruppo di nodi di calcolo AWS PCS nuovo o esistente.

> **Important**
>
> AWS Attualmente PCS richiede un kernel con IPv4 supporto per la comunicazione tra nodi locali, anche quando si utilizza AWS PCS in una rete di IPv6 sola rete.

New compute node group

Per utilizzare l'AMI personalizzata

1. Aprire la console AWS PCS.

2. Nel pannello di navigazione scegliere Cluster.
3. Scegli il cluster in cui utilizzerai l'AMI personalizzata, quindi seleziona Gruppi di nodi di calcolo.
4. Crea un nuovo gruppo di nodi di calcolo. Per ulteriori informazioni, consulta Creazione di un gruppo di nodi di calcolo in AWS PCS. In ID AMI, cerca il nome o l'ID dell'AMI personalizzata che desideri utilizzare. Completa la configurazione del gruppo di nodi di calcolo, quindi scegli Crea gruppo di nodi di calcolo.
5. (Facoltativo) Conferma che l'AMI supporti l'avvio delle istanze. Avvia un'istanza nel gruppo di nodi di calcolo. Puoi farlo configurando il gruppo di nodi di calcolo in modo che abbia una singola istanza statica oppure puoi inviare un lavoro a una coda che utilizza il gruppo di nodi di calcolo.

    a. Controlla la console Amazon EC2 finché un'istanza non appare etichettata con il nuovo ID del gruppo di nodi di calcolo. Per ulteriori informazioni su questo argomento, consulta.. Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS
    b. Quando vedi un'istanza avviarsi e completare il processo di bootstrap, conferma che stia utilizzando l'AMI prevista. Per fare ciò, seleziona l'istanza, quindi controlla l'ID AMI in Dettagli. Dovrebbe corrispondere all'AMI configurato nelle impostazioni del gruppo di nodi di calcolo.
    c. (Facoltativo) Aggiorna la configurazione di ridimensionamento dei gruppi di nodi di calcolo ai tuoi valori preferiti.

Existing compute node group

Per utilizzare l'AMI personalizzata

1. Aprire la console AWS PCS.
2. Nel pannello di navigazione scegliere Cluster.
3. Scegli il cluster in cui utilizzerai l'AMI personalizzata, quindi seleziona Gruppi di nodi di calcolo.
4. Seleziona il gruppo di nodi che desideri configurare e scegli Modifica. In ID AMI, cerca il nome o l'ID dell'AMI personalizzata che desideri utilizzare. Completa la configurazione del gruppo di nodi di calcolo, quindi scegli Aggiorna. Le nuove istanze lanciate nel gruppo di nodi di calcolo utilizzeranno l'ID AMI aggiornato. Le istanze esistenti continueranno a utilizzare la vecchia AMI fino a quando AWS PCS non le sostituirà. Per ulteriori informazioni, consulta Aggiornamento di un gruppo di nodi di calcolo AWS PCS.

5. (Facoltativo) Conferma che l'AMI supporti l'avvio delle istanze. Avvia un'istanza nel gruppo di nodi di calcolo. Puoi farlo configurando il gruppo di nodi di calcolo in modo che abbia una singola istanza statica oppure puoi inviare un lavoro a una coda che utilizza il gruppo di nodi di calcolo.

   a. Controlla la console Amazon EC2 finché un'istanza non appare etichettata con il nuovo ID del gruppo di nodi di calcolo. Per ulteriori informazioni su questo argomento, consulta.. Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS

   b. Quando vedi un'istanza avviarsi e completare il processo di bootstrap, conferma che stia utilizzando l'AMI prevista. Per fare ciò, seleziona l'istanza, quindi controlla l'ID AMI in Dettagli. Dovrebbe corrispondere all'AMI configurato nelle impostazioni del gruppo di nodi di calcolo.

   c. (Facoltativo) Aggiorna la configurazione di ridimensionamento dei gruppi di nodi di calcolo ai tuoi valori preferiti.

Passaggio 7: terminare l'istanza temporanea

Dopo aver verificato che l'AMI funzioni come previsto con AWS PCS, puoi chiudere l'istanza temporanea per evitare di incorrere in addebiti.

Per terminare l'istanza temporanea

1. Aprire la console di Amazon EC2.
2. Nel riquadro di navigazione, seleziona Instances (Istanze).
3. Seleziona l'istanza temporanea che hai creato e scegli Azioni, Stato dell'istanza, Termina istanza.
4. Quando ti viene richiesto di confermare, scegli Termina.

Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS

AWS fornisce un file scaricabile che consente di installare il software AWS PCS su un'istanza. AWS fornisce inoltre software in grado di scaricare, compilare e installare le versioni pertinenti di Slurm e delle sue dipendenze. È possibile utilizzare queste istruzioni per crearne di personalizzate AMIs da utilizzare con AWS PCS oppure è possibile utilizzare metodi personalizzati.

Indice

AWS Programma di installazione del software PCS Agent

Il programma di installazione del software AWS PCS Agent configura un'istanza per funzionare con AWS PCS durante il processo di avvio dell'istanza. È necessario utilizzare i programmi AWS di installazione forniti per installare l'agente AWS PCS sull'AMI personalizzata.

Per ulteriori informazioni sul software AWS PCS Agent, vedere. AWS Versioni dell'agente PCS

Programma di installazione Slurm

Il programma di installazione di Slurm scarica, compila e installa le versioni pertinenti di Slurm e delle sue dipendenze. Puoi usare il programma di installazione Slurm per creare creazioni personalizzate per PC. AMIs AWS È inoltre possibile utilizzare i propri meccanismi se sono coerenti con la configurazione software fornita dal programma di installazione di Slurm. Per ulteriori informazioni sul supporto AWS PCS per Slurm, vedere. Versioni Slurm in PCS AWS

Il software AWS fornito installa quanto segue:

- Slurm alla versione principale e di manutenzione richiesta (attualmente versione 25.05.x) - Licenza GPL 2
  - Slurm è costruito con set to `--sysconfdir /etc/slurm`
  - Slurm è costruito con l'opzione e `--enable-pam --without-munge`
  - Slurm è costruito con l'opzione `--sharedstatedir=/run/slurm/`
  - Slurm è costruito con supporto PMIX e JWT
  - Slurm è installato su `/opt/aws/pcs/schedulers/slurm-25.05`
- OpenPMIX (versione 4.2.6) — Licenza
  - OpenPMIX è installato come sottodirectory di `/opt/aws/pcs/scheduler/`

- libjwt (versione 1.17.0) — Licenza MPL-2.0
  - libjwt è installato come sottodirectory di `/opt/aws/pcs/scheduler/`

Il software AWS fornito modifica la configurazione del sistema come segue:

- Il `systemd` file Slurm creato dalla build viene copiato con il nome del file. `/etc/systemd/system/ slurmd-25.05.service`
- Se non esistono, vengono creati un utente e un gruppo Slurm (`slurm:slurm`) con of. UID/GID `401`
- La cartella `/etc/aws/pcs/scheduler/slurm-25.05/plugstack.conf.d/` viene creata per memorizzare la configurazioneEstendi la funzionalità Slurm sui AWS PC con i plugin SPANK.
- Su Amazon Linux 2 e Rocky Linux 9 l'installazione aggiunge il repository EPEL per installare il software richiesto per creare Slurm o le sue dipendenze.
- Durante RHEL9 l'installazione abiliterà `codeready-builder-for-rhel-9-rhui-rpms` e `epel-release-latest-9` riavvierà l'installazione del software richiesto `fedoraproject` per creare Slurm o le sue dipendenze.

Sistemi operativi supportati

Per informazioni, consulta Sistemi operativi supportati in AWS PCS.

> **Note**
>
> AWS Deep Learning AMIs Le versioni (DLAMI) basate su Amazon Linux 2 e Ubuntu 22.04 dovrebbero essere compatibili con il software AWS PCS e i programmi di installazione Slurm. Per ulteriori informazioni, consulta Scelta del DLAMI nella Guida per gli AWS Deep Learning AMIs sviluppatori.

Tipi di istanze supportati

AWS Il software PCS e i programmi di installazione Slurm supportano qualsiasi tipo di istanza x86_64 o arm64 in grado di eseguire uno dei sistemi operativi supportati.

Versioni Slurm supportate

Per informazioni, consulta Versioni Slurm in PCS AWS.

Verifica gli installatori utilizzando un checksum

È possibile utilizzare i SHA256 checksum per verificare i file tarball (.tar.gz) del programma di installazione. È consigliabile eseguire questa operazione per verificare l'identità dell'autore del software e che l'applicazione non sia stata alterata o danneggiata dopo la pubblicazione.

Per verificare un tarball

Utilizzate l'utilità sha256sum per il SHA256 checksum e specificate il nome del file tarball. È necessario eseguire il comando dalla directory in cui è stato salvato il file tarball.

- SHA256

```
$ sha256sum tarball_filename.tar.gz
```

Il comando deve restituire un valore di checksum nel formato seguente.

```
checksum_value tarball_filename.tar.gz
```

Confrontate il valore di checksum restituito dal comando con il valore di checksum fornito nella tabella seguente. Se i checksum corrispondono, è sicuro eseguire lo script di installazione.

> **Important**
>
> Se i checksum non corrispondono, non eseguite lo script di installazione. Contattare Supporto.

Ad esempio, il comando seguente genera il SHA256 checksum per il tarball Slurm 25.05.4-1.

```
$ sha256sum aws-pcs-slurm-25.05-installer-25.05.4-1.tar.gz
```

Output di esempio:

```
3b0f93bce441d4f4f6935175f2c1e81cd961cb923adb416fa6689f5592047a7d aws-pcs-slurm-25.05-
installer-25.05.4-1.tar.gz
```

Nelle tabelle seguenti sono elencati i checksum per le versioni recenti dei programmi di installazione. *us-east-1*Sostituiscilo con quello Regione AWS in cui usi PCS. AWS

AWS Agente PCS

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
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`
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`	`2b784643ca01ccca1baa64fbfb34bb41efe8bdca69470998b74ce3962bc271d4`

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
	-agent-v1.2.1-1.tar.gz	
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
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

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
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`

programma di installazione Slurm

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
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-slurm-24.11-installer-24.11.6-2.tar.gz`	`f17cd78e0bc6b9c818b794d9d2685cceabdc73f4fbb12f7566ae5b86a5abc32b`
Slurm 24.11.6-1	`https://aws-pcs-repo-us-east-1.s3.us-`	`225de9fc18206f5f65f412effe1fd457614a`

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
	east-1 .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.11-insta ller-24.11.6-1.tar.gz	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
Slurm 24.05.7-1	https://aws-pcs-re po- us-east-1 .s3.us- east-1 .amazonaws.com/ aws-pcs-slurm/aws-pcs -slurm-24.05-insta ller-24.05.7-1.tar.gz	0b5ed7c81195de2628 c78f37c79e63fc4ae9 9132ca6b019b53a0d6 8792ee82c5

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
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`
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`

Installer (Programma di installazione)	Scarica il URL	SHA256 checksum
Slurm 23.11.10-1 (deprecato)	`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 (deprecato)	`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`

Note di rilascio per l'esempio AWS PCS AMIs

AMIs per le ultime versioni principali supportate dello scheduler ricevi aggiornamenti di sicurezza e correzioni di bug critici. Queste patch di sicurezza incrementali non sono incluse nelle note di rilascio ufficiali.

Important

Gli esempi AMIs relativi alle vecchie versioni dello scheduler non sono supportati e non ricevono aggiornamenti.

Important

AMIs Gli esempi sono a scopo dimostrativo e non sono consigliati per carichi di lavoro di produzione.

Indice

- AWS Esempio di PCS AMIs per x86_64 (Amazon Linux 2)
- AWS Esempio di PCS AMIs per Arm64 (Amazon Linux 2)

AWS Esempio di PCS AMIs per x86_64 (Amazon Linux 2)

Slurm 25.05

Nome AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-25.05`

Istanze EC2 supportate

- Tutte le istanze con processore x86 a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca Architectures=x86_64.

Contenuti di AMI

- Servizio AWS supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: x86_64
- Tipo di volume EBS: gp2
- Installatore EFA: 1.43.1
- GDRCopy: 2.5.1
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.11

> **Note**
>
> AWS PCS supporta la contabilità per Slurm 24.11 e versioni successive. Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

Nome AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.11`

Istanze EC2 supportate

- Tutte le istanze con processore x86 a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=x86_64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: x86_64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.05

Nome AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.05`

Istanze EC2 supportate

- Tutte le istanze con processore x86 a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=x86_64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2

- Architettura di calcolo: x86_64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 23.11

Nome AMI

- `aws-pcs-sample_ami-amzn2-x86_64-slurm-23.11`

Istanze EC2 supportate

- Tutte le istanze con processore x86 a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=x86_64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: x86_64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

AWS Esempio di PCS AMIs per Arm64 (Amazon Linux 2)

Slurm 25.05

Nome AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-25.05`

Istanze EC2 supportate

- Tutte le istanze con processore Arm a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca Architectures=ARM64.

Contenuti di AMI

- Servizio AWS supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: arm64
- Tipo di volume EBS: gp2
- Installatore EFA: 1.43.1
- GDRCopy: 2.5.1
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.11

> **Note**
>
> AWS PCS supporta la contabilità per Slurm 24.11 e versioni successive. Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

Nome AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-24.11`

Istanze EC2 supportate

- Tutte le istanze con processore Arm a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=arm64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: arm64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 24.05

Nome AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-24.05`

Istanze EC2 supportate

- Tutte le istanze con processore Arm a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=arm64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: arm64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4

- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Slurm 23.11

Nome AMI

- `aws-pcs-sample_ami-amzn2-arm64-slurm-23.11`

Istanze EC2 supportate

- Tutte le istanze con processore Arm a 64 bit. Per trovare istanze compatibili, accedi alla console Amazon EC2. Scegli Tipi di istanze, quindi cerca. `Architectures=arm64`

Contenuti di AMI

- AWS Servizio supportato: AWS PCS
- Sistema operativo: Amazon Linux 2
- Architettura di calcolo: arm64
- Tipo di volume EBS: gp2
- Programma di installazione EFA: 1.33.0
- GDRCopy: 2.4
- Driver NVIDIA: 550.127.08
- NVIDIA CUDA: 12.4.1_550.54.15

Sistemi operativi supportati in AWS PCS

AWS PCS utilizza l'Amazon Machine Image (AMI) configurata per un gruppo di nodi di calcolo per avviare EC2 istanze in quel gruppo di nodi di calcolo. L'AMI determina il sistema operativo utilizzato dalle EC2 istanze. Non è possibile modificare il sistema operativo nell'esempio AWS AMIs PCS. È necessario creare un'AMI personalizzata se si desidera utilizzare un sistema operativo diverso. Per ulteriori informazioni, consulta Amazon Machine Images (AMIs) per AWS PCS.

Sistemi operativi supportati

- Amazon Linux 2

  Questo è il sistema operativo nell'esempio AWS PCS AMIs.

  > **Important**
  >
  > AMIs I campioni sono a scopo dimostrativo e non sono consigliati per carichi di lavoro di produzione. È necessario creare e utilizzare un'AMI personalizzata per i carichi di lavoro di produzione, anche se si intende utilizzare Amazon Linux 2.

- Amazon Linux 2023
- RedHat Enterprise Linux 9 (RHEL 9)

  Il costo on-demand per RHEL, qualsiasi tipo di istanza, è superiore a quello di altri sistemi operativi supportati. Per ulteriori informazioni sui prezzi, consulta la sezione Prezzi On-Demand e In che modo viene offerto e prezzato Red Hat Enterprise Linux su Amazon Elastic Compute Cloud? .

- RedHat Linux 8 aziendale (RHEL 8)
- Rocky Linux 9

  Puoi usare Rocky Linux 9 ufficiale AMIs come base per un'AMI personalizzata. La compilazione dell'AMI personalizzata potrebbe fallire se l'AMI di base non dispone del kernel più recente.

  Per aggiornare il kernel

  1. Avvia un'istanza utilizzando un ID AMI rocky9 da qui: https://rockylinux.org/cloud-images/
  2. ssh nell'istanza ed esegui il seguente comando:

     ```
     sudo yum -y update
     ```

3. Crea un'immagine dall'istanza. Specifichi questa immagine come `ParentImage` per la tua AMI personalizzata.

- Rocky Linux 8
- Ubuntu 22.04

  Ubuntu 22.04 richiede chiavi più sicure per SSH e non supporta le chiavi RSA per impostazione predefinita. Ti consigliamo invece di generare e utilizzare una ED25519 chiave.
- Ubuntu 24.04

AWS Versioni dell'agente PCS

Il software AWS PCS agent configura le istanze EC2 lanciate da AWS PCS per l'uso con Slurm. Includi l'agente in un Amazon Machine Images (AMI) che specifichi quando crei gruppi di nodi di calcolo per il tuo cluster. Le istanze EC2 lanciate in questi gruppi di nodi di calcolo utilizzano l'AMI specificata e il software agente AWS PCS incluso. L'agente AWS PCS consente a un'istanza EC2 di registrarsi come parte del cluster. Per utilizzare il software AWS PCS Agent più recente, è necessario aggiornare il software personalizzato AMIs. Per ulteriori informazioni, consulta Fase 2 — Installare l'agente AWS PCS in Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

AWS Versione dell'agente PCS	Data di rilascio	Note di rilascio
v1.3.2-1	10 marzo 2026	• È stato risolto un problema a causa del quale i nodi di elaborazione che eseguivano RHEL 8.10 o Rocky Linux 8.10 non riuscivano ad avviarsi a causa di un backport SigV4 `curl` difettoso in tali sistemi operativi.
v1.3.1-1	7 novembre 2025	• È stata migliorata la disabilitazione dell'hyperthreading utilizzando il parametro sysfs `smt/control` quando disponibile. • Risolta una potenziale condizione di gara che si verificava quando la CPU veniva bloccata durante l'avvio mentre l'agente PCS tentava di disabilitare l'hyperthreading.

AWS Versione dell'agente PCS	Data di rilascio	Note di rilascio
		• È stato risolto il problema che causava il `InstanceType` riempimento dei campi `InstanceId` e dei nodi di calcolo Slurm rispettivamente con un timestamp e un trattino.
v1.3.0-1	3 novembre 2025	• Aggiunto il supporto per nuovi sistemi operativi: Amazon Linux 2023, Ubuntu 24, RHEL 8, Rocky 8.
v1.2.2-1	16 ottobre 2025	• Sono consentite le interrogazioni sui metadati delle istanze a un IPv6 endpoint se un IPv4 endpoint non è disponibile. • È stato risolto un problema che impediva la disabilitazione dell'hyperthreading se il kernel restituiva thread di pari livello come intervalli di ID CPU. • È stato risolto un problema che produceva falsi messaggi di errore nei log quando l'hyperthreading veniva disabilitato correttamente.

AWS Versione dell'agente PCS	Data di rilascio	Note di rilascio
v1.2.1-1	19 giugno 2025	• L'agente AWS PCS ora tenta di avviare slurmd per un massimo di 30 minuti se il controller non è disponibile. • È stato risolto un problema che produceva una configurazione slurmd errata se la risposta a RegisterComputeNodeGroupInstance conteneva un endpoint SLURMDBD.
v1.2.0-1	7 marzo 2025	• Supporto abilitato per in. IPv6 `slurmd.conf`
v1.1.1-1	13 dicembre 2024	• È stato risolto un problema per cui nella chiamata a veniva segnalata una versione di Slurm errata. RegisterComputeNodeGroupInstance • È stato risolto un problema per cui i metadati dell'istanza non venivano recuperati correttamente se veniva eseguito uno script personalizzato. `/opt/aws/pcs/etc/bootstrap_hooks/`

AWS Versione dell'agente PCS	Data di rilascio	Note di rilascio
v1.1.0-1	6 dicembre 2024	• Ha abilitato l'esecuzione degli script personalizzati prima dei passaggi di bootstrap. `/opt/aws/pcs/etc/bootstrap_hooks/`
v1.0.1-1	22 ottobre 2024	• Risolto un problema per cui i dispositivi NVIDIA non funzionavano quando `slurmd` venivano avviati su istanze abilitate per GPU.
v1.0.0-1	28 agosto 2024	• Versione iniziale.

programma di pianificazione Slurm in PCS AWS

Slurm è un gestore di carichi di lavoro open source progettato per i cluster Linux che fornisce funzionalità di pianificazione dei lavori, allocazione delle risorse e monitoraggio dei lavori per carichi di lavoro HPC. AWS PCS supporta lo scheduler Slurm per gestire i carichi di lavoro dei cluster.

Argomenti

Versioni Slurm in PCS AWS

SchedMD migliora continuamente Slurm con nuove funzionalità, ottimizzazioni e patch di sicurezza. SchedMD rilascia una nuova versione principale a intervalli regolari e prevede di supportare fino a 3 versioni alla volta. AWS PCS è progettato per aggiornare automaticamente il controller Slurm con versioni patch.

Quando SchedMD termina il supporto per una particolare versione principale, AWS PCS designa quella versione come End of Life (EOL). Dopo EOL, non è possibile creare nuovi cluster con quella versione, sebbene i cluster esistenti possano continuare a funzionare fino a 12 mesi senza supporto garantito. AWS PCS invia un avviso anticipato se una versione principale di Slurm è prossima alla fine del ciclo di vita, per aiutare i clienti a sapere quando aggiornare i propri cluster a una versione più recente supportata.

Ti consigliamo di utilizzare l'ultima versione supportata di Slurm per distribuire il tuo cluster, per accedere ai progressi e ai miglioramenti più recenti.

Versioni Slurm supportate in PCS AWS

La tabella seguente mostra le versioni di Slurm supportate e le date e le informazioni importanti per ciascuna versione.

Versione Slurm	Data di rilascio di SchedMD	AWS Data di rilascio del PCS	AWS Data PCS EOL	Versione minima compatibile dell'agente AWS PCS	Esempio di AWS PCS supportato AMIs
25.05	29/5/2025	16/10/2025	31/5/2027	1,0,0-1	• `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	31/5/2026	1,0,0-1	• `aws-pcs-sample_ami-amzn2-x86_64-slurm-24.11` • `aws-pcs-sample_ami-amzn2-arm64-slurm-24.11`

Versioni Slurm non supportate in PCS AWS

La tabella seguente mostra le versioni di Slurm che non sono supportate in PCS. AWS

Versione Slurm	Data di rilascio di SchedMD	AWS Data di rilascio del PCS	AWS Data PCS EOL		
24.05	30/05/2024	18/12/2024	30/11/2025		
23,11	21/11/2023	28/08/2024	31/5/2025		

Note di rilascio per le versioni Slurm in PCS AWS

Questo argomento descrive le modifiche importanti per ogni versione di Slurm attualmente supportata in PCS. AWS Ti consigliamo di rivedere le modifiche tra la vecchia e la nuova versione quando aggiorni il tuo cluster.

Slurm 25.05

Modifiche implementate in PCS AWS

- Lo Slurm requeue_on_resume_failure è ora abilitato per impostazione predefinita SchedulerParameter .
- «stderr» è stato rimosso come opzione per, poiché era disabilitato in Slurm 25.05. LogTimeFormat
- AWS PCS supporta la configurazione sackd a più cluster: il nodo di accesso può accedere a più cluster.

Per ulteriori informazioni su Slurm 25.05, consulta le seguenti pubblicazioni:

- Annuncio di rilascio di SchedMD: https://www.schedmd.com/slurm-version-25-05-0-is-now-available/
- Note di rilascio di SchedMD: _Notes.md https://github.com/SchedMD/ slurm/blob/slurm-25-05-0-1/RELEASE

Slurm 24.11

Modifiche implementate in PCS AWS

- AWS PCS supporta la contabilità Slurm. Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

Per ulteriori informazioni su Slurm 24.11, consulta le seguenti pubblicazioni:

- Annuncio di rilascio di SchedMD
- Note di rilascio di SchedMD

Slurm 24.05

Modifiche implementate in PCS AWS

- Il nuovo modulo Slurm Step Manager è ora abilitato di default in AWS PCS. Questo modulo offre vantaggi significativi trasferendo la gestione delle fasi dal controller centrale ai nodi di calcolo, migliorando notevolmente la concorrenza del sistema in ambienti con un utilizzo intensivo delle fasi. Per supportare questa configurazione e isolare `Prolog` ed `Epilog` elaborare meglio l'esecuzione, sono abilitati i nuovi flag prolog (,). `Contain` `Alloc`
- La comunicazione gerarchica dal controller ai nodi di calcolo è abilitata per ottimizzare la comunicazione tra nodi Slurm, migliorando la scalabilità e le prestazioni. Inoltre, la configurazione di routing ora utilizza elenchi di nodi di partizione per le comunicazioni dal controller, anziché l'algoritmo di routing predefinito del plug-in, migliorando la resilienza del sistema.
- Un nuovo plugin hash sostituisce il precedente. `HashPlugin=hash/sha3` `hash/k12 plugin` Questo è ora abilitato di default nei cluster AWS PCS.
- I log dei controller Slurm ora includono funzionalità di controllo avanzate per tutte le chiamate di procedura remota (RPC) in entrata verso. `slurmctld` I log includono l'indirizzo di origine, l'utente autenticato e il tipo di RPC prima dell'elaborazione della connessione.

Per ulteriori informazioni su Slurm 24.05, consultate le seguenti pubblicazioni:

- Annuncio di rilascio di SchedMD
- Note di rilascio di SchedMD

Slurm 23.11

Le impostazioni di Slurm possono essere modificate in PCS AWS

- L'impostazione `SuspendTime` predefinita è. `60` Utilizzate il parametro di `scaleDownIdleTimeInSeconds` configurazione AWS PCS per impostarlo. Per ulteriori informazioni, consulta il `scaleDownIdleTimeInSeconds`parametro del tipo di `ClusterSlurmConfiguration` dati nel AWS PCS API Reference.
- La `MaxJobCount` e `MaxArraySize` si basa sulla dimensione scelta per il cluster. Per ulteriori informazioni, consulta il `size`parametro dell'azione `CreateCluster` API nel AWS PCS API Reference.
- L'impostazione predefinita di `SelectTypeParameters` Slurm è. `CR_CPU` Puoi fornirlo come valore per `slurmCustomSettings` impostarlo quando crei un cluster. Per ulteriori informazioni, consulta il `slurmCustomSettings`parametro dell'azione `CreateCluster` API e SlurmCustomSettingnel AWS PCS API Reference.
- È possibile impostare `Prolog` e `Epilog` a livello di cluster. Puoi fornirlo come valore per `slurmCustomSettings` impostarlo quando crei un cluster. Per ulteriori informazioni, vedere `CreateCluster`e SlurmCustomSettingnel AWS PCS API Reference.
- È possibile impostare `Weight` e `RealMemory` a livello di gruppo di nodi di calcolo. Puoi fornirlo come valore per `slurmCustomSettings` impostarlo quando crei un gruppo di nodi di calcolo. Per ulteriori informazioni, vedere `CreateComputeNodeGroup`e SlurmCustomSettingnel AWS PCS API Reference.

Domande frequenti sulle versioni di Slurm in PCS AWS

AWS PCS mantiene il supporto per più versioni di Slurm. Quando viene introdotta una nuova versione di Slurm, AWS PCS fornisce supporto tecnico e patch di sicurezza fino al raggiungimento della fine del supporto (EOS) da SchedMD. AWS PCS fa riferimento alla data EOS per una versione di Slurm come fine del ciclo di vita (EOL) per coerenza con la terminologia. AWS

Per quanto tempo AWS PCS supporta una versione Slurm?

AWS Il supporto PCS per le versioni di Slurm è in linea con i cicli di supporto di SchedMD per le versioni principali. AWS PCS supporta la versione corrente e le 2 versioni principali precedenti più recenti. Quando SchedMD rilascia una nuova versione principale, AWS PCS termina il supporto per la versione più vecchia supportata. AWS PCS rilascia nuove versioni principali di Slurm il prima possibile, ma potrebbe esserci un ritardo tra il rilascio di SchedMD e la sua disponibilità in PCS. AWS

In che modo i miei cluster ottengono nuove versioni di patch per Slurm?

Per risolvere bug e correzioni di sicurezza, AWS PCS è progettato per applicare automaticamente le patch ai controller del cluster eseguiti in account interni di proprietà dei servizi. Per installare le patch sulle istanze EC2 nel tuo Account AWS, aggiorna l'Amazon Machine Image (AMI) per i tuoi gruppi di nodi di calcolo e aggiorna i gruppi di nodi di calcolo per utilizzare l'AMI aggiornata. Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

> **Note**
>
> I controller Slurm non sono disponibili durante l'aggiornamento. I lavori in esecuzione non sono influenzati. I lavori inviati prima che il controller del cluster diventasse non disponibile vengono mantenuti fino a quando il controller non è disponibile.

Come posso essere informato su un imminente evento EOL della versione Slurm?

Ti invieremo un messaggio e-mail 6 mesi prima della data EOL. Ti inviamo un messaggio e-mail ogni mese prima dell'EOL, con un messaggio e-mail finale 1 settimana prima della data EOL. Dopo la data EOL, inviamo messaggi e-mail mensili per 12 mesi ai clienti che utilizzano cluster AWS PCS con versioni EOL Slurm. Potremmo sospendere un cluster con una versione EOL Slurm se vengono identificate vulnerabilità di sicurezza per quella versione.

Come posso determinare se la versione Slurm utilizzata dal mio cluster esegue una versione EOL Slurm?

Ti inviamo un messaggio e-mail per informarti che hai un cluster in esecuzione con una versione EOL Slurm. Pubblichiamo un avviso negli Dashboard AWS Health avvisi che contiene i dettagli dei tuoi cluster con versioni EOL Slurm. È inoltre possibile utilizzare la console AWS PCS per identificare i cluster con versioni EOL Slurm.

Cosa devo fare se la mia versione di Slurm è prossima o superiore alla fine del ciclo di vita?

Crea un nuovo cluster con una versione più recente supportata di Slurm e aggiorna la versione Slurm nelle AMI del tuo gruppo di nodi di calcolo. La versione Slurm nelle tue AMI e nelle istanze EC2 in esecuzione non può differire di più di due versioni rispetto alla versione Slurm del cluster. Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Cosa succede se non passo a una versione più recente di Slurm entro la data di fine del ciclo di vita?

Non è possibile creare nuovi cluster con una versione EOL Slurm. I cluster esistenti possono funzionare fino a 12 mesi senza AWS supporto e non è richiesta alcuna azione immediata per mantenerne il funzionamento. Dopo la data di fine del ciclo di vita, il supporto, gli aggiornamenti di sicurezza e la disponibilità non sono garantiti. Potremmo sospendere un cluster per motivi di sicurezza. Ti consigliamo vivamente di utilizzare una versione Slurm supportata per mantenere la sicurezza e il supporto per i tuoi AWS cluster PCS.

Quali sono i rischi della gestione di un cluster con versioni EOL Slurm?

I cluster con versioni EOL Slurm presentano rischi operativi e di sicurezza significativi. Senza il monitoraggio attivo di SchedMD, le vulnerabilità di sicurezza potrebbero rimanere inosservate o non risolte. Se vengono scoperte vulnerabilità critiche, potremmo sospendere immediatamente i tuoi cluster.

Cosa succede ai miei lavori, alle risorse di calcolo, di archiviazione e di rete del cluster quando il mio cluster viene sospeso?

Tutte le risorse gestite da AWS PCS vengono interrotte. Ciò include il controller Slurm, i gruppi di nodi di calcolo e le istanze EC2. Tutti i processi in esecuzione su istanze di calcolo vengono immediatamente interrotti e il cluster entra in uno stato sospeso. Le risorse gestite dal cliente, come i file system esterni, rimangono intatte. È possibile utilizzare la console AWS PCS e le azioni API per accedere alla configurazione del cluster.

Posso riavviare un cluster sospeso per riprendere i lavori rimanenti?

No, non è possibile riavviare un cluster sospeso. Puoi utilizzare la configurazione del cluster sospeso per creare un nuovo cluster con una versione Slurm supportata. Puoi eseguire i lavori rimanenti se li hai salvati in un file system esterno.

Posso richiedere una proroga oltre il periodo di grazia di 12 mesi?

No, non puoi richiedere un'estensione per far funzionare il tuo cluster oltre il periodo di grazia di 12 mesi. Forniamo un periodo di tempo prolungato per aiutarti a passare a una versione Slurm supportata. Per evitare interruzioni delle operazioni del cluster, consigliamo di passare alla versione di Slurm prima che la versione di Slurm raggiunga l'EOL.

Contabilità Slurm in PCS AWS

È possibile abilitare la contabilità sui nuovi cluster AWS PCS per monitorare l'utilizzo del cluster, applicare i limiti delle risorse e gestire un controllo granulare degli accessi a code o gruppi di nodi

di calcolo specifici. AWS PCS crea e gestisce il database di contabilità per il cluster, eliminando la necessità di creare e gestire un proprio database di contabilità separato. AWS PCS utilizza la funzionalità di contabilità di Slurm. Per ulteriori informazioni sulla funzionalità di contabilità in Slurm, consulta la documentazione Slurm su SchedMD.

Per utilizzare la contabilità, abilitala quando crei un nuovo cluster e, facoltativamente, imposti i parametri contabili. Dopo aver impostato lo stato del cluster `Active` e aver impostato i gruppi di nodi di calcolo, puoi connetterti alla shell Linux di un nodo di accesso per eseguire funzioni di contabilità, come la visualizzazione dei dati dei lavori con il comando `sacct` Slurm.

> **Note**
>
> La contabilità è supportata per Slurm 24.11 o versioni successive.

AWS PCS console

Nella pagina Crea cluster, è necessario selezionare una versione valida di Slurm (versione 24.11 o successiva). Nelle impostazioni di Scheduler, abilita Accounting.

AWS PCS API

Fornisci la `accounting` configurazione nella chiamata all'azione `CreateCluster` API. Nell'`accounting`oggetto, imposta `mode` su`STANDARD`. Per ulteriori informazioni, vedere CreateClusterand Accounting nel AWS PCS API Reference.

L'esempio seguente utilizza l'azione AWS CLI per chiamare l'`CreateCluster`API. La sottostringa del valore del parametro `accounting='{mode=STANDARD}'` abilita la contabilità.

```
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**
>
> Se abiliti la contabilità, ti verranno addebitati costi di fatturazione aggiuntivi. Per ulteriori informazioni, consulta la pagina dei prezzi AWS PCS.

Modifica delle impostazioni contabili

È possibile abilitare o disabilitare la contabilità sui cluster esistenti senza ricostruire l'infrastruttura. Per ulteriori informazioni, consulta Aggiornamento di un cluster in AWS PCS.

Quando disabiliti la contabilità, la fatturazione per la funzionalità di contabilità si interrompe non appena il cluster entra nello stato. `UPDATING` Quando abiliti la contabilità, la fatturazione inizia quando il cluster torna correttamente allo `ACTIVE` stato.

Concetti chiave per la contabilità Slurm in PCS AWS

I seguenti concetti sono specifici del AWS PCS e controllano il modo in cui PCS implementa la contabilità AWS Slurm.

Database contabile

AWS PCS archivia i dati contabili in un database creato Account AWS in un AWS proprietario. Non hai accesso al `slurmdbd.conf`.

Tempo di eliminazione predefinito

Questa impostazione AWS PCS specifica il periodo di conservazione (in giorni) per tutti i tipi di record contabili (lavori, eventi, prenotazioni, fasi, sospensioni, transazioni, dati di utilizzo). Ad esempio, se il valore è 30, AWS PCS conserva i record contabili per 30 giorni. Fornisci questo valore quando crei il cluster. Se non fornisci un valore, AWS PCS conserva i record contabili nel database a tempo indeterminato.

AWS PCS console

L'ora di eliminazione predefinita viene specificata come parte dei passaggi per la creazione di un cluster. Nella pagina Crea cluster, è necessario selezionare una versione valida di Slurm (versione 24.11 o successiva) e abilitare la contabilità. Nelle impostazioni di Scheduler, fornisci un valore intero per il tempo di eliminazione predefinito (giorni).

AWS PCS API

Specificalo `defaultPurgeTimeInDays` come parte delle `accounting` informazioni fornite nella chiamata all'`CreateCluster`azione API. Per ulteriori informazioni, consulta CreateClusterand Accounting nel AWS PCS API Reference.

> **Note**
>
> Quando si utilizza l'API AWS PCS per creare un cluster, il valore predefinito di `defaultPurgeTimeInDays` è `-1` e `0` non è un valore valido.

Applicazione delle politiche contabili

Questa impostazione determina con che rigore Slurm applica le regole di invio dei lavori, i limiti delle risorse e le politiche contabili per il cluster. Questa impostazione corrisponde al `AccountingStorageEnforce` parametro nel file del cluster. `slurm.conf` È possibile selezionare qualsiasi combinazione di opzioni di applicazione. Se non si seleziona alcuna opzione, non vengono applicati vincoli contabili ai lavori nel cluster. AWS PCS supporta le seguenti opzioni:

- associazioni — job-to-account mappatura
- limiti — vincoli relativi alle risorse
- QoS: requisiti di qualità del servizio
- modalità sicura: completamento garantito entro limiti
- nosteps — disabilita la contabilità dei passaggi
- nojobs — disabilita la contabilità dei lavori

Per ulteriori informazioni su queste opzioni, consulta la documentazione Slurm su SchedMD.

AWS PCS console

Le opzioni vengono impostate come parte dei passaggi per creare un cluster. Nella pagina Crea cluster, è necessario selezionare una versione valida di Slurm (versione 24.11 o successiva) e abilitare la contabilità. Seleziona le opzioni che desideri dall'elenco a discesa per l'applicazione delle politiche contabili nelle impostazioni di Scheduler.

AWS PCS API

In Slurm, queste opzioni sono impostate nel file di un cluster. `slurm.conf` Non hai accesso diretto al cluster `slurm.conf` for your AWS PCS. Invece, fornisci `SlurmCustomSettings` all'`CreateCluster`API l'azione quando crei un cluster. Per ulteriori informazioni, vedere CreateClusternel AWS PCS API Reference.

Ottieni la configurazione contabile per un cluster AWS PCS esistente

La configurazione contabile Slurm è inclusa nella configurazione Slurm per il tuo cluster.

AWS PCS console

1. Scegli Clusters dal pannello di navigazione.
2. Scegli il nome del cluster dall'elenco.
3. Nella scheda Configurazione, trova la configurazione contabile in Configurazione Slurm

AWS PCS API

Usa l'azione `GetCluster` API per ottenere la configurazione del cluster. È possibile trovare la configurazione contabile in`slurmConfiguration`. L'impostazione `mode` e il valore di `defaultPurgeTimeInDays` sono inferiori `aaccounting`. Le opzioni selezionate per l'applicazione delle politiche contabili sono riportate di seguito`slurmCustomSettings`. Per ulteriori informazioni, vedere GetClusternel AWS PCS API Reference.

API REST Slurm in PCS AWS

AWS PCS fornisce supporto gestito per l'API REST nativa di Slurm tramite un'interfaccia HTTP per l'`slurmrestd`interazione programmatica dei cluster. È possibile inviare lavori, monitorare lo stato del cluster e gestire le risorse tramite richieste HTTP standard senza richiedere l'accesso diretto dalla shell al cluster.

Casi di utilizzo comune

L'API REST di Slurm supporta vari scenari di integrazione:

- Integrazione di applicazioni Web: crea frontend e applicazioni Web personalizzati che inviano e gestiscono direttamente i lavori.

- Jupyter Notebook Integration: consente ai ricercatori di inviare lavori da ambienti notebook senza abbandonare il flusso di lavoro di sviluppo.
- Integrazione delle soluzioni partner: Collega strumenti HPC e gestori di flussi di lavoro di terze parti ai tuoi cluster AWS PCS.
- Gestione programmatica dei cluster: automatizza i flussi di lavoro per l'invio dei lavori, il monitoraggio e la gestione delle risorse.
- Flussi di lavoro di Research Computing: Supporta ambienti di ricerca accademici e aziendali che richiedono una gestione del lavoro basata sulle API.

Requisiti e limitazioni

Prima di utilizzare l'API REST di Slurm, esamina questi dettagli:

- Il tuo cluster deve utilizzare la versione Slurm 25.05 o successiva.
- L'endpoint dell'API sarà accessibile solo tramite un indirizzo IP privato all'interno del VPC del cluster.
- Il gruppo di sicurezza del cluster deve consentire il traffico HTTP sulla porta 6820.
- L'autenticazione richiede token JWT con dichiarazioni di identità utente specifiche.

Le limitazioni attuali includono:

- I token generati da non `scontrol token` sono supportati.
- `X-SLURM-USER-NAME`la rappresentazione dell'intestazione non è disponibile.
- Alcune funzionalità richiedono l'abilitazione della contabilità Slurm.
- Non compatibile con il meccanismo del plugin di filtro Slurm CLI.
- Le connessioni all'endpoint dell'API REST non sono crittografate con TLS.

Argomenti

Abilitazione dell'API REST Slurm nei PCS AWS

Abilita l'API REST Slurm per accedere all'interfaccia HTTP del cluster per la gestione e il monitoraggio programmatici dei lavori. È possibile abilitare questa funzionalità durante la creazione del cluster o aggiornare un cluster esistente che soddisfi i requisiti.

Prerequisiti

Prima di abilitare l'API REST di Slurm, assicurati di avere:

- Versione cluster: Slurm versione 25.05 o successiva.
- Gruppo di sicurezza: regole che consentono il traffico HTTP sulla porta 6820 dalle sorgenti desiderate.

Procedura

Per abilitare l'API REST di Slurm su un nuovo cluster

Console di gestione AWS

1. Apri la console AWS PCS all'indirizzo. https://console.aws.amazon.com/pcs/
2. Scegli Crea cluster.
3. In Dettagli del cluster, scegli Slurm versione 25.05 o successiva.
4. Configura le altre impostazioni del cluster secondo necessità.
5. Nella sezione di configurazione Scheduler, imposta l'API REST su Enabled.
6. Configura il gruppo di sicurezza del cluster per consentire il traffico HTTP sulla porta 6820 dalle sorgenti desiderate.
7. Completa il processo di creazione del cluster.

AWS CLI

1. Aggiungi una configurazione Slurm REST durante la creazione del cluster.

```
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. Configura il gruppo di sicurezza del cluster per consentire il traffico HTTP sulla porta 6820 dalle sorgenti desiderate.

Per abilitare l'API REST di Slurm su un cluster esistente

Console di gestione AWS

1. Apri la console AWS PCS all'indirizzo. https://console.aws.amazon.com/pcs/
2. Scegli il tuo cluster dall'elenco.
3. Verifica che il tuo cluster utilizzi Slurm versione 25.05 o successiva nei dettagli del cluster.
4. Scegli Modifica cluster.
5. Nella sezione di configurazione Scheduler, imposta l'API REST su Enabled.
6. Scegli Aggiorna cluster per applicare le modifiche.
7. Configura il gruppo di sicurezza del cluster per consentire il traffico HTTP sulla porta 6820 dalle sorgenti desiderate.

AWS CLI

1. Aggiorna il tuo cluster con una configurazione Slurm REST, come in questo esempio.

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

2. Configura il gruppo di sicurezza del cluster per consentire il traffico HTTP sulla porta 6820 dalle sorgenti desiderate.

Cosa succede dopo l'attivazione

Quando abiliti l'API REST, AWS PCS automaticamente:

- Genera una chiave di firma JWT e la memorizza in AWS Secrets Manager.
- Espone l'endpoint dell'API all'`https://<clusterPrivateIpAddress>:6820`interno del tuo VPC.
- Aggiorna la configurazione del cluster per mostrare i dettagli dell'endpoint dell'API REST.

Ora puoi autenticare e utilizzare l'API REST per la gestione dei lavori e le operazioni del cluster.

Autenticazione con l'API REST di Slurm in PCS AWS

L'API Slurm REST in AWS PCS utilizza l'autenticazione JSON Web Token (JWT) per garantire un accesso sicuro alle risorse del cluster. AWS PCS fornisce una chiave di firma gestita archiviata in AWS Secrets Manager, che viene utilizzata per generare token JWT contenenti dichiarazioni di identità utente specifiche.

Prerequisiti

Prima di autenticarti con l'API REST di Slurm, assicurati di avere:

- Configurazione del cluster: cluster AWS PCS con Slurm 25.05+ e API REST abilitate.
- Autorizzazioni AWS: accesso a AWS Secrets Manager per la chiave di firma JWT.
- Informazioni utente: nome utente, ID utente POSIX e uno o più gruppi POSIX IDs per l'account del cluster.
- Accesso alla rete: connettività all'interno del VPC del cluster con gruppo di sicurezza che consente la porta 6820.

Procedura

Per recuperare l'indirizzo dell'endpoint dell'API REST di Slurm

Console di gestione AWS

1. Apri la console PCS all'indirizzo. AWS https://console.aws.amazon.com/pcs/
2. Scegli il tuo cluster dall'elenco.
3. Nei dettagli della configurazione del cluster, individua la sezione Endpoints.
4. Annota l'indirizzo IP e la porta privati per l'API REST di Slurm (slurmrestd).
5. È possibile effettuare chiamate API inviando richieste HTTP formattate correttamente a questo indirizzo.

AWS CLI

1. Interroga lo stato del cluster con`aws pcs get-cluster`. Cerca l'`SLURMREST`Dendpoint nel `endpoints` campo della risposta. Ecco un esempio:

```
"endpoints": [
      {
          "type": "SLURMCTLD",
          "privateIpAddress": "192.0.2.1",
          "port": "6817"
      },
      {
          "type": "SLURMRESTD",
          "privateIpAddress": "192.0.2.1",
          "port": "6820"
      }
  ]
```

2. Puoi effettuare chiamate API inviando richieste HTTP formattate correttamente a `http://<privateIpAddress>:<port>/`

Per recuperare la chiave di firma JWT

1. Apri la console AWS PCS all'indirizzo. https://console.aws.amazon.com/pcs/
2. Scegli il tuo cluster dall'elenco.
3. Nei dettagli di configurazione del cluster, individua la sezione Scheduler Authentication.
4. Nota l'ARN e la versione della chiave JSON Web Token (JWT).
5. Usa AWS CLI per recuperare la chiave di firma da Secrets Manager:

```
aws secretsmanager get-secret-value --secret-
id arn:aws:secretsmanager:region:account:secret:name --version-id version
```

Per generare un token JWT

1. Crea un JWT con le seguenti attestazioni obbligatorie:
   - `exp`— Tempo di scadenza in secondi dal 1970 per il JWT
   - `iat`— Ora attuale in secondi dal 1970
   - `sun`— Il nome utente per l'autenticazione
   - `uid`— L'ID utente POSIX
   - `gid`— L'ID del gruppo POSIX

- `id`— Proprietà di identità POSIX aggiuntive
  - `gecos`— Campo di commento dell'utente, spesso utilizzato per memorizzare un nome leggibile dall'uomo
  - `dir`— La home directory dell'utente
  - `shell`— shell predefinita dell'utente
  - `gids`— Elenco dei gruppi POSIX aggiuntivi in cui si trova IDs l'utente

2. Firma il JWT utilizzando la chiave di firma recuperata da Secrets Manager.
3. Imposta un orario di scadenza appropriato per il token.

> **Note**
>
> In alternativa al `sun` reclamo, puoi fornire una delle seguenti informazioni:
>
> - `username`
> - Un nome di campo personalizzato che definisci tramite `userclaimfield` in `AuthAltParameters Slurm custom settings`
> - Un `name` campo all'interno del `id` claim

Per autenticare le richieste API

1. Includi il token JWT nelle tue richieste HTTP utilizzando uno di questi metodi:
   - Token Bearer: aggiungi un'intestazione `Authorization: Bearer <jwt>`
   - Slurm header: aggiunge un'intestazione `X-SLURM-USER-TOKEN: <jwt>`
2. Effettua richieste HTTP all'endpoint dell'API REST:

   Ecco un esempio di accesso all'`/ping`API utilizzando curl e l'`Authorized: Bearer`header.

```
curl -X GET -H "Authorization: Bearer <jwt>" \
      http://<privateIpAddress>:6820/slurm/v0.0.43/ping
```

Esempio di generazione JWT

Recupera la chiave di firma JWT del cluster AWS PCS e memorizzala come file locale. Sostituisci i valori per aws-region, secret-arn e secret version con valori appropriati per il tuo cluster.

```
#!/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
```

Questo esempio in Python illustra come utilizzare la chiave di firma per generare 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)
```

Lo script stamperà un JWT sullo schermo.

```
abcdefgtjwttoken...
```

Utilizzo dell'API REST Slurm per la gestione dei lavori in PCS AWS

Panoramica dell'API REST Slurm

L'API REST di Slurm fornisce l'accesso programmatico alle funzioni di gestione dei cluster tramite richieste HTTP. La comprensione di queste caratteristiche chiave ti aiuterà a utilizzare efficacemente l'API con PCS: AWS

- Protocollo di accesso: l'API utilizza HTTP (non HTTPS) per la comunicazione all'interno della rete privata del cluster.
- Dettagli di connessione: accedi all'API utilizzando l'indirizzo IP privato del cluster e la `slurmrestd` porta (in genere 6820). Il formato URL di base completo è`http://<privateIpAddress>:6820`.
- Controllo delle versioni dell'API: La versione dell'API corrisponde all'installazione di Slurm. Per Slurm 25.05, usa la versione v0.0.43. Il numero di versione cambia con ogni versione di Slurm. Puoi trovare le versioni delle API attualmente supportate nelle note di rilascio di Slurm.
- Struttura dell'URL: La struttura dell'URL per l'API REST di Slurm è. `http://<privateIpAddress>:<port>/<api-version>/<endpoint>` Informazioni dettagliate sull'utilizzo degli endpoint dell'API REST sono disponibili nella documentazione di Slurm.

Prerequisiti

Prima di utilizzare l'API REST di Slurm, assicurati di avere:

- Configurazione del cluster: cluster AWS PCS con Slurm 25.05+ e API REST abilitate.
- Autenticazione: token JWT valido con affermazioni di identità utente corrette.
- Accesso alla rete: connettività all'interno del VPC del cluster con un gruppo di sicurezza che consente la porta 6820.

Procedura

Per inviare un lavoro utilizzando l'API REST

1. Crea una richiesta di invio di lavoro con i parametri richiesti:

```
{
  "job": {
    "name": "my-job",
    "partition": "compute",
    "nodes": 1,
    "tasks": 1,
    "script": "#!/bin/bash\necho 'Hello from Slurm REST API'"
  }
}
```

2. Invia il lavoro utilizzando una richiesta 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. Annota l'ID del lavoro restituito nella risposta a scopo di monitoraggio.

Per monitorare lo stato del lavoro

1. Ottieni informazioni su un lavoro specifico:

```
curl -X GET -H "Authorization: Bearer <jwt>" \
    https://<privateIpAddress>:6820/slurm/v0.0.43/job/<job-id>
```

2. Elenca tutti i lavori per l'utente autenticato:

```
curl -X GET -H "Authorization: Bearer <jwt>" \
    https://<privateIpAddress>:6820/slurm/v0.0.43/jobs
```

Come annullare un processo

- Invia una richiesta DELETE per annullare un lavoro specifico:

```
curl -X DELETE -H "Authorization: Bearer <jwt>" \
    https://<privateIpAddress>:6820/slurm/v0.0.43/job/<job-id>
```

Domande frequenti sull'API REST di Slurm in PCS AWS

Questa sezione risponde alle domande frequenti sull'API REST di Slurm in AWS PCS.

Cos'è l'API REST di Slurm?

L'API Slurm REST è un'interfaccia HTTP che consente di interagire con il gestore del carico di lavoro Slurm a livello di codice. È possibile utilizzare metodi HTTP standard come GET, POST e DELETE per inviare lavori, monitorare lo stato del cluster e gestire le risorse senza richiedere l'accesso da riga di comando al cluster.

Posso usare i token generati da? **`scontrol token`**

No, lo standard `scontrol token` output non è compatibile con AWS PCS. L'API REST PCS Slurm richiede token JWT arricchiti contenenti affermazioni di identità specifiche che includono username (`sun`), ID utente POSIX () e group (`uid`). IDs `gids` I token Slurm standard non dispongono di queste affermazioni obbligatorie e verranno rifiutati dall'API.

Posso accedere all'API dall'esterno del mio VPC?

No, l'endpoint dell'API REST è accessibile solo dall'interno del tuo VPC utilizzando l'indirizzo IP privato del controller Slurm. Per abilitare l'accesso esterno, implementa AWS servizi come Application Load Balancer with VPC Link, API Gateway o stabilisci connessioni VPC peering o VPN per una connettività sicura.

Perché l'API utilizza HTTP anziché HTTPS?

L'API REST di Slurm è pensata per essere un endpoint interno all'interno della rete privata del cluster. Per le implementazioni di produzione che richiedono la crittografia, è possibile

implementare la SSL/TLS terminazione a un livello superiore nell'architettura, ad esempio tramite un gateway API, un sistema di bilanciamento del carico o un proxy inverso.

Come posso controllare l'accesso all'API REST?

Configura le regole del gruppo di sicurezza del cluster per limitare l'accesso alla porta 6820 sul controller Slurm. Imposta regole in entrata per consentire le connessioni solo da intervalli IP affidabili o fonti specifiche all'interno del tuo VPC, bloccando l'accesso non autorizzato all'endpoint API.

Come faccio a ruotare la chiave di firma JWT?

Metti il cluster in modalità di manutenzione senza istanze attive, quindi avvia la rotazione delle chiavi tramite AWS Secrets Manager. Al termine della rotazione, riattiva le code. Tutti i token JWT esistenti non saranno più validi e dovranno essere rigenerati utilizzando la nuova chiave di firma di Secrets Manager.

Ho bisogno che la contabilità Slurm sia abilitata per utilizzare l'API REST?

No, la contabilità Slurm non è richiesta per le operazioni di base dell'API REST come l'invio e il monitoraggio dei lavori. Tuttavia, l'intero `/slurmdb` endpoint richiede che la contabilità sia attiva.

Quali strumenti di terze parti funzionano con l'API REST AWS PCS?

Molti client Slurm REST API esistenti dovrebbero funzionare con AWS PCS, tra cui Slurm Exporter for Prometheus, e applicazioni personalizzate che seguono il formato API REST standard di Slurm. SlurmWeb Tuttavia, gli strumenti che si basano sull'autenticazione dovranno essere modificati `scontrol token` per funzionare con i requisiti PCS JWT. AWS

Sono previsti costi aggiuntivi per l'utilizzo dell'API REST?

No, non ci sono costi aggiuntivi per l'attivazione o l'utilizzo della funzionalità API REST di Slurm. Come di consueto, paghi solo per le risorse del cluster sottostanti.

Come posso risolvere i problemi relativi all'API REST?

- Problemi di connettività di rete

  Se non riesci a raggiungere l'endpoint dell'API, vedrai dei timeout di connessione o degli errori di «connessione rifiutata» quando effettui richieste HTTP al controller del cluster.

  Cosa fare: verifica che il client si trovi nello stesso VPC o disponga del routing di rete corretto e conferma che il gruppo di sicurezza consenta il traffico HTTP sulla porta 6820 dall'IP o dalla sottorete di origine.

- Problemi di autenticazione Slurm REST

  Se il token JWT non è valido, è scaduto o firmato in modo errato, le richieste API restituiranno «Errore di autenticazione del protocollo» nel campo degli errori della risposta.

  Esempio di messaggio di errore:

  ```
  {
  "errors": [
      {
      "description": "Batch job submission failed",
      "error_number": 1007,
      "error": "Protocol authentication error",
      "source": "slurm_submit_batch_job()"
      }
    ]
  }
  ```

  Cosa fare: verifica che il token JWT sia formattato correttamente, non scaduto e firmato con la chiave corretta di Secrets Manager. Verifica che il token sia formato correttamente e includa le attestazioni richieste e che stia utilizzando il formato di intestazione di autenticazione corretto.

- Job non riuscito a eseguire dopo l'invio

  Se il token JWT è valido ma contiene una struttura o un contenuto interni errati, è possibile che i job abbiano inserito uno stato paused (PD) con codice motivo. `JobAdminHead scontrol show job <job-id>`Usalo per ispezionare il lavoro: vedrai`JobState=PENDING, Reason=JobHeldAdmin`, e. `SystemComment=slurm_cred_create failure, holding job`

  Cosa fare: la causa principale potrebbe essere rappresentata da valori errati in JWT. Verifica che il token sia strutturato correttamente e includa le attestazioni richieste secondo la documentazione PCS.

- Problemi di autorizzazione alla directory di lavoro

  Se l'identità utente specificata nel JWT non dispone dei permessi di scrittura nella directory di lavoro del lavoro, il lavoro avrà esito negativo con errori di autorizzazione, simili all'utilizzo `sbatch --chdir` con una directory inaccessibile.

  Cosa fare: assicurati che l'utente specificato nel tuo token JWT disponga delle autorizzazioni appropriate per la directory di lavoro del lavoro.

Riavvio dei nodi di calcolo con Slurm in PCS AWS

AWS PCS supporta il comando nativo di Slurm. `scontrol reboot` Usa questo comando per riavviare i nodi di calcolo senza sostituire l'istanza EC2. Altri metodi di riavvio (console Amazon EC2 AWS CLI, patch automatiche o manutenzione del sistema) AWS fanno sì che PCS consideri l'istanza EC2 non integra e la sostituisca.

Vantaggi del riavvio di Slurm

Il riavvio di Slurm offre diversi vantaggi per la manutenzione del cluster:

- Conserva la capacità: evita di perdere istanze EC2 con limiti di capacità a favore di altri clienti.
- Riduzione dei costi: elimina i cicli di sostituzione delle istanze non necessari e la fatturazione continua per i nodi inattivi.
- Ripristino più rapido: nessun ritardo nel provisioning rispetto alla sostituzione delle istanze.
- Flessibilità operativa: elimina le perdite di memoria, rimuove i file temporanei e ripristina i nodi da stati degradati.

Quando usare Slurm reboot

Usa Slurm reboot per scenari di manutenzione operativa comuni:

- Risoluzione dei problemi: risolvi i problemi di prestazioni o i processi che non rispondono, in particolare per i nodi GPU.
- Pulizia delle risorse: elimina le perdite di memoria, i file temporanei o i processi bloccati che `/tmp` influiscono sulle prestazioni lavorative.
- Ripristino: ripristina i nodi da stati bloccati o degradati prima di richiedere la sostituzione completa dei nodi.

Limitazioni

- Solo gli utenti Slurm Admin (utenti root) possono eseguire comandi di riavvio.
- Il supporto per il riavvio è limitato a. `scontrol reboot`
- RebootProgram la configurazione non è supportata.
- Nessuna interfaccia di console, solo riga di comando.

Argomenti

Riavvia un nodo di elaborazione utilizzando Slurm in PCS AWS

Usa il comando di riavvio nativo di Slurm per risolvere problemi di prestazioni, eliminare problemi relativi alle risorse o ripristinare da stati degradati senza perdita della capacità delle istanze EC2.

Prerequisiti

- Privilegi di amministratore di Slurm (accesso utente root)
- Accesso a un nodo di accesso nel cluster PCS AWS

Procedura

1. Connect a un nodo di accesso tramite la console EC2.
   a. Nella console EC2, scegli Instances (Istanze).
   b. Seleziona l'istanza del tuo nodo di accesso.
   c. Scegli Connetti.
2. Identifica il nome del nodo di calcolo di destinazione utilizzando `sinfo` `oscontrol show node`.

```
sinfo
# or
scontrol show node
```

3. Esegui il comando reboot utilizzando una di queste opzioni:

> **Warning**
>
> Non utilizzare `nextstate=DOWN` con il `scontrol reboot` comando. Questo parametro contrassegna il nodo come non integro e attiva la sostituzione dell'istanza.

- Riavvio di base (attende che il nodo diventi inattivo):

```
scontrol reboot nodename
```

- Riavvio immediato (drena il nodo e si riavvia al termine dei processi):

```
scontrol reboot ASAP nodename
```

- Riavvia con motivo:

```
scontrol reboot ASAP reason="troubleshooting" nodename
```

- Riavvia con lo stato di ripristino:

```
scontrol reboot ASAP nextstate=RESUME nodename
```

4. Monitora l'avanzamento del riavvio utilizzando. `scontrol show node`

```
scontrol show node nodename
```

5. Verifica che il nodo ritorni in servizio dopo il completamento del riavvio.

Annulla un riavvio in sospeso in PCS AWS

Annulla un riavvio in sospeso per evitare tempi di inattività non necessari quando il problema è stato risolto o quando il riavvio non è più necessario.

Prerequisiti

- Privilegi di amministratore di Slurm
- Il nodo deve avere un riavvio in sospeso (che mostra lo stato «riavvio emesso»)
- Accesso al nodo di login per l'esecuzione dei comandi

Procedura

1. Connect al nodo di accesso.
2. Verifica che il nodo abbia un riavvio in sospeso utilizzando. `scontrol show node`

```
scontrol show node nodename
```

Cerca «riavvio emesso» nello stato del nodo.

3. Esegui il comando cancel.

```
scontrol cancel_reboot nodename
```

4. Verifica l'annullamento del riavvio e il ripristino dello stato del nodo alla normalità.

```
scontrol show node nodename
```

Domande frequenti sul riavvio di Slurm in PCS AWS

Trova le risposte alle domande più comuni sull'utilizzo di Slurm reboot in PCS. AWS

Cos'è il supporto per il riavvio di Slurm?

Support per il comando nativo Slurm. `scontrol reboot` Utilizza questo comando per riavviare i nodi di calcolo senza la sostituzione automatica delle istanze, in modo da preservare la capacità delle istanze EC2 e ridurre i costi operativi.

Chi può usare i comandi di riavvio di Slurm?

Solo gli utenti Slurm Admin (utenti root) possono eseguire i comandi di riavvio. Gli utenti normali che tentano di utilizzare `scontrol reboot` riceveranno un errore di autorizzazione negata da Slurm senza influire sul nodo.

Cosa succede ai job in esecuzione durante un riavvio?

Per impostazione predefinita, i processi vengono completati normalmente prima del riavvio. Con l'opzione ASAP, il nodo viene svuotato per evitare nuovi lavori e il riavvio avviene al termine dei processi correnti. I lavori possono essere annullati o richiesti per riavvii immediati.

In che modo è diverso dal riavvio della console EC2?

Slurm reboot preserva l'istanza EC2 ed evita la sostituzione, mentre i riavvii della console EC2 attivano PCS per sostituire l'istanza a causa dei controlli di integrità falliti durante il processo di riavvio.

Posso configurare script di riavvio personalizzati?

No, la RebootProgram configurazione non è supportata nella versione iniziale. La funzionalità utilizza il comportamento di riavvio standard di Slurm senza supporto di script personalizzati.

Quanto tempo richiede il riavvio di Slurm?

Il tempo di riavvio varia in base al tipo di istanza, ai processi di avvio del cliente, alla configurazione AMI e al fatto che i job debbano essere completati o meno. Il processo include l'attesa del completamento dei lavori, il riavvio fisico, i controlli di integrità e la registrazione del demone slurmd.

Posso vedere una cronologia dei riavvii?

Gli eventi di riavvio vengono registrati nei log di Slurm (slurmctld e slurmd) che possono essere monitorati. CloudWatch Il campo motivo nello stato del nodo mostra il motivo del riavvio durante il processo.

Cosa succede se un nodo si blocca durante il riavvio?

Se un nodo non completa il processo di riavvio all'interno ResumeTimeout, verrà contrassegnato come DOWN. Controlla la presenza di errori CloudWatch nei log, verifica la connettività di rete ed esamina i log slurmd. Contatta l' AWS assistenza se il problema persiste.

Posso riavviare più nodi contemporaneamente?

Sì, puoi specificare più nodi nel comando reboot:

```
scontrol reboot ASAP node1,node2,node3
```

Come posso riavviare un nodo senza attendere il completamento dei processi?

Per il riavvio immediato dei nodi in caso di problemi quali nodi problematici che influiscono su processi multinodo, un significativo peggioramento delle prestazioni o un comportamento instabile della GPU, sono disponibili due opzioni:

- Annulla e riavvia: innanzitutto, annulla i lavori interessati utilizzando, quindi avvia un riavvio immediato utilizzando`scancel <job_id>`. `scontrol reboot ASAP <nodename>` I processi in esecuzione verranno interrotti e dovranno essere nuovamente inviati dopo il ripristino del nodo.
- Drain and Requeue (meno impattante): inizia avviando un drenaggio e riavvia con, quindi richiedi i lavori interessati utilizzando. `scontrol reboot ASAP <nodename>` `scontrol requeue <job_id>` In questo modo i lavori tornano in sospeso invece di annullarli.

Cosa succede se specifico nextState=DOWN?

Se si specifica`nextstate=DOWN`, il nodo verrà contrassegnato come non integro dopo il riavvio e attiverà la sostituzione dell'istanza. Per evitare la sostituzione dell'istanza, non specificare nextstate o use. `nextstate=RESUME`

Risorse aggiuntive

- Per le procedure di riavvio di base, vedere. Riavvia un nodo di elaborazione utilizzando Slurm in PCS AWS
- Per la risoluzione dei problemi di riavvio, vedere. Risoluzione dei problemi di riavvio di Slurm nei PCS AWS
- Per la documentazione sul riavvio di Slurm, consulta la documentazione di Slurm scontrol.

Risoluzione dei problemi di riavvio di Slurm nei PCS AWS

Quando riscontri problemi di riavvio del nodo, controlla innanzitutto lo stato del nodo utilizzando. `scontrol show node` *`nodename`* Quindi esamina CloudWatch i log di Slurm (slurmctld e slurmd) e i log di sistema per identificare potenziali errori.

Per la risoluzione dei problemi di base, verifica la connettività di rete, controlla le impostazioni del gruppo di sicurezza e assicurati che tutti i servizi richiesti siano in esecuzione dopo il riavvio. Se i problemi persistono dopo i passaggi di base per la risoluzione dei problemi, contatta l' AWS assistenza. Quando contattate l'assistenza, fornite gli estratti dei log pertinenti, le informazioni sullo stato del nodo e una cronologia del tentativo di riavvio per velocizzare il processo di risoluzione.

Risorse aggiuntive

- Per il monitoraggio delle istanze AWS PCS tramite Amazon CloudWatch, consulta Monitoraggio delle istanze AWS PCS tramite Amazon. CloudWatch
- Per una risoluzione generale dei problemi, consulta. Risoluzione dei problemi in Parallel Computing AWS Service
- Per la documentazione su Slurm, consulta la Guida alla risoluzione dei problemi di Slurm.

Configurazione delle impostazioni Slurm personalizzate in PCS AWS

Utilizza le impostazioni Slurm personalizzate per configurare parametri Slurm aggiuntivi tra le risorse Cluster, Queue e Compute Node Group. Questa versione aggiunge il supporto per le impostazioni Slurm sulle risorse Queue, fornendo un controllo granulare sui comportamenti specifici delle partizioni.

Vantaggi delle impostazioni Slurm personalizzate

Le impostazioni Slurm personalizzate offrono un controllo sofisticato sull'ambiente HPC basato su PC AWS . È possibile implementare una contabilità dettagliata, applicare i controlli di accesso e ottimizzare l'esecuzione del carico di lavoro attraverso configurazioni e politiche di priorità. quality-of-service Queste funzionalità garantiscono che i lavori critici ricevano le risorse necessarie, mantenendo al contempo un utilizzo efficiente del cluster. Che si tratti di gestire carichi di lavoro accelerati da GPU, implementare una pianificazione equa o controllare i cicli di vita dei lavori, le impostazioni personalizzate aiutano ad allineare l'infrastruttura HPC ai requisiti operativi e agli obiettivi di ricerca.

Configurazione delle impostazioni personalizzate

Le impostazioni personalizzate di Slurm possono essere configurate tramite AWS Console, CLI o SDKs durante la creazione di risorse o modificate in un secondo momento tramite operazioni di aggiornamento.

Console di gestione AWS

Passa alle impostazioni aggiuntive dello scheduler nella pagina di creazione o modifica per qualsiasi tipo di risorsa (cluster, coda o gruppo di nodi di calcolo).

Per aggiungere una nuova impostazione

1. Scegli Aggiungi nuova impostazione.
2. Seleziona il nome di un parametro dal menu a discesa (che include brevi descrizioni dei parametri).
3. Fornisci il valore corrispondente.

Per annullare l'impostazione personalizzata

1. Scegli Rimuovi accanto alla parameter/value coppia pertinente.
2. Crea o aggiorna la risorsa.

AWS CLI

Per la gestione programmatica delle impostazioni personalizzate, utilizzate il `SlurmCustomSettings` campo nelle operazioni di creazione o aggiornamento.

Example— Aggiornamento del Prolog parametro su un cluster

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

Example— Impostazione di una coda da inserire Default in un cluster

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

Example— Impostazione personalizzata Features su un gruppo di nodi di calcolo

```
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"}]'
```

Convalida e gestione degli errori

AWS PCS implementa un processo di convalida a più livelli per le impostazioni Slurm personalizzate. Durante le operazioni di creazione e aggiornamento, eseguiamo convalide sincrone che includono:

- Controlli a livello di campo: convalidiamo le impostazioni individuali per verificare la correttezza dei tipi di dati, dei valori consentiti e dei requisiti di formato. Ad esempio, ci assicuriamo che i

valori temporali siano nel formato Slurm corretto e che i valori booleani utilizzino rappresentazioni booleane Slurm accettate.

- Convalide sensibili al contesto: alcune impostazioni vengono verificate rispetto al contesto di configurazione più ampio. Ad esempio, alcuni parametri sono validi solo quando la contabilità Slurm è abilitata.
- Coerenza tra le impostazioni: verifichiamo che le opzioni che si escludono a vicenda non siano impostate insieme e che le impostazioni interdipendenti siano configurate correttamente.

Se la convalida fallisce, riceverai un messaggio `ValidationException` con un codice di errore specifico (ad esempio InvalidInput), un messaggio di errore chiaro che descrive il problema e un elenco dei campi non validi con i rispettivi dettagli di errore.

Sebbene durante questa convalida iniziale vengano rilevati molti problemi, alcune interazioni complesse tra le impostazioni possono diventare evidenti solo quando si applica la configurazione. In questi casi, l'operazione avrà esito negativo e verrà visualizzato un messaggio di errore informativo e tutte le modifiche parziali verranno annullate.

Limitazioni

AWS PCS implementa un approccio basato sulla lista delle autorizzazioni per proteggere la sicurezza del servizio e la stabilità operativa. Le impostazioni che potrebbero compromettere la sicurezza dell'account di servizio o interferire con le funzionalità dei servizi gestiti sono limitate. Tuttavia, valutiamo continuamente le esigenze dei clienti e possiamo aggiungere supporto per impostazioni aggiuntive in base al feedback dei clienti.

Argomenti

- Impostazioni Slurm personalizzate per AWS cluster PCS
- Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS
- Impostazioni Slurm personalizzate per le code PCS AWS
- Risoluzione dei problemi relativi alle impostazioni Slurm personalizzate nei PCS AWS

Impostazioni Slurm personalizzate per AWS cluster PCS

Le seguenti impostazioni Slurm personalizzate sono supportate a livello di cluster:

- AccountingStorageEnforce

> **Important**
>
> AWS PCS supporta un sottoinsieme delle opzioni per. `AccountingStorageEnforce` Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.

- AccountingStorageTRES
- AccountingStoreFlags
- DefMemPerCPU
- Epilog
- EnforcePartLimits
- FairShareDampeningFactor
- HealthCheckInterval
- HealthCheckNodeState
- HealthCheckProgram
- JobRequeue
- LaunchParameters
- Licenses
- MinJobAge

> **Note**
>
> AWS PCS supporta un valore minimo di 5 secondi per`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

Impostazioni Slurm personalizzate per gruppi di nodi di calcolo AWS PCS

Le seguenti impostazioni Slurm personalizzate sono supportate a livello di gruppo di nodi di calcolo:

- CpuSpecList

- Features
- MemSpecLimit
- RealMemory
- Weight

Impostazioni Slurm personalizzate per le code PCS AWS

Le seguenti impostazioni Slurm personalizzate sono supportate a livello di coda:

- AllowAccounts
- AllowQoS
- Default
- DefaultTime
- DenyAccounts
- DenyQoS
- ExclusiveUser
- GraceTime
- MaxTime
- OverSubscribe
- OverTimeLimit
- PreemptMode
- PriorityJobFactor
- PriorityTier
- QOS
- TRESBillingWeights

Risoluzione dei problemi relativi alle impostazioni Slurm personalizzate nei PCS AWS

Se riscontri errori durante la creazione o l'aggiornamento delle risorse AWS PCS con le impostazioni personalizzate di Slurm, puoi utilizzare la registrazione per diagnosticare e risolvere i problemi.

Risoluzione dei problemi delle impostazioni personalizzate Slurm incompatibili

Problema: durante l'esecuzione di operazioni su cluster, gruppi di nodi di calcolo o code, viene visualizzato un messaggio di errore simile al seguente:

```
{OPERATION} failed. The Slurm custom settings of the cluster might be incompatible.
 Check the settings and try again.
```

Questo errore può verificarsi con le seguenti operazioni:

- CreateCluster
- CreateComputeNodeGroup
- UpdateComputeNodeGroup
- CreateQueue
- UpdateQueue

Soluzione: abilitare la registrazione per comprendere il problema specifico e risolvere le impostazioni incompatibili.

Per risolvere i problemi relativi alle impostazioni personalizzate di Slurm incompatibili

1. Crea il cluster se non esiste ancora o assicurati che il cluster esistente sia in uno stato in cui sia possibile abilitare la registrazione.
2. Abilita la registrazione per il tuo cluster. Per istruzioni dettagliate, vedi Registrazione e monitoraggio per AWS PCS.

   > **Note**
   >
   > La registrazione può essere abilitata una volta che il cluster è in fase di creazione.

3. Esamina i log per identificare lo specifico problema di configurazione di Slurm che causa l'incompatibilità.
4. Correggi le impostazioni personalizzate incompatibili in base alle informazioni di registro e riprova l'operazione.

Per informazioni sulle impostazioni personalizzate Slurm supportate, consulta:

Estendi la funzionalità Slurm sui AWS PC con i plugin SPANK

Usa i plugin SPANK (Slurm Plug-in Architecture for Node and job Kontrol) per estendere e modificare il comportamento di Slurm durante l'avvio e l'esecuzione dei job su cluster PCS. AWS I plugin SPANK forniscono un'interfaccia generica per intercettare e modificare le fasi di avvio del lavoro.

Installa i plugin SPANK sull'AMI del tuo nodo di calcolo e configurali per personalizzare il comportamento del tuo cluster Slurm in base ai requisiti del tuo carico di lavoro. Per ulteriori informazioni su SPANK, consulta la documentazione SPANK sul sito Web SchedMD.

Indice

Installa i plugin SPANK sui PC AWS

Segui la documentazione del plugin per installare i plugin SPANK sul tuo AMI.

Compila i plugin SPANK per la versione Slurm specifica sul tuo cluster. Il programma di installazione Slurm fornito da PCS memorizza Slurm in. AWS `/opt/aws/pcs/scheduler/slurm-version` Quando compili il plugin, specifica la versione di Slurm.

L'esempio seguente mostra come specificare la versione Slurm per alcuni plugin:

```
export CFLAGS="-I/opt/aws/pcs/scheduler/slurm-version/include"
```

Se hai più versioni di Slurm nell'AMI, compila il plugin per ogni versione. Memorizza i plugin compilati in cartelle con versione.

L'esempio seguente mostra come specificare la cartella di destinazione per alcuni plugin:

```
export DESTDIR="your-preferred-versioned-path"
```

> **Important**
>
> I plugin potrebbero richiedere variabili diverse. Consulta la documentazione ufficiale del plugin che stai installando.

Configura i plugin SPANK su PCS AWS

Per impostazione predefinita, memorizza i file di configurazione in. `/etc/aws/pcs/scheduler/slurm-version/plugstack.conf.d/`

Per memorizzare la configurazione di SPANK in una posizione diversa, aggiungi le tue posizioni a un file di configurazione nella directory predefinita.

L'esempio seguente mostra come includere file di configurazione da altre directory:

```
# 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
```

Memorizza ogni configurazione in un file dedicato o in un file comune. È possibile utilizzare più file di configurazione.

Gli esempi seguenti mostrano file di configurazione di esempio:

```
# 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
```

Per ulteriori informazioni su come configurare i plugin, consulta la documentazione di configurazione SPANK sul sito Web SchedMD.

> **Important**
>
> Imposta i permessi delle cartelle per impedire modifiche non autorizzate alla configurazione del plugin.

> **Note**
>
> AWS PCS non gestisce i tuoi plugin SPANK. Se riscontri errori relativi ai plugin, controlla i log degli errori sui tuoi nodi di calcolo.

> **Note**
>
> Slurm registra erroneamente un errore simile al seguente quando carica la configurazione SPANK:
>
> ```
> error: "Include" failed in file /etc/slurm/plugstack.conf line 3
> ```
>
> Puoi ignorare questo errore. Non influisce sul funzionamento dei plugin SPANK.

Domande frequenti sui plugin SPANK su PC AWS

Questa sezione affronta le domande più comuni sull'installazione e la configurazione dei plugin SPANK sui cluster PCS. AWS

Devo installare i plugin SPANK sia sui nodi di accesso che sui nodi di calcolo?

Alcuni plugin SPANK non richiedono l'installazione su tutti i nodi; ma per una migliore compatibilità, ti consigliamo di installare tutti i plugin SPANK su ogni nodo.

Quale configurazione aggiuntiva è necessaria per l'uso in produzione dei plugin SPANK?

Oltre all'installazione e alla configurazione di base mostrate negli esempi, le implementazioni di produzione richiedono in genere una configurazione aggiuntiva. I plugin basati su container come Pyxis potrebbero richiedere l'impostazione di variabili di ambiente per Enroot, l'abilitazione dell'interfaccia PMI (Process Management Interface) e la configurazione delle autorizzazioni per il runtime del contenitore. Consulta la documentazione del plug-in specifico per i requisiti di distribuzione in produzione dettagliati.

Come posso risolvere i problemi relativi al plug-in SPANK?

AWS PCS non gestisce i plugin SPANK. Esamina i log degli errori sui nodi di calcolo per risolvere i problemi.

Usa i plugin di filtro CLI Slurm per personalizzare l'invio dei lavori in PCS AWS

AWS PCS supporta i plugin Slurm CLI Filter per eseguire script Lua personalizzati che convalidano e modificano i parametri di invio dei lavori sui nodi di accesso e di calcolo. Per informazioni dettagliate sui plug-in di filtro CLI, consulta la documentazione dell'API del plug-in cli_filter sul sito Web SchedMD.

Requisiti

I plugin di filtro CLI richiedono la versione Slurm 24.11 o successiva e uno script Lua distribuito su tutti i nodi di accesso e di calcolo.

> **Important**
>
> Per le versioni Slurm 24.11 e 25.05, i plugin di filtro CLI richiedono l'installazione di Slurm AWS utilizzando il programma di installazione PCS Slurm (versione 24.11.6-2+ o 25.05.4-1+). Per ulteriori informazioni sull'installazione Fase 3 — Installare Slurm di Slurm, consulta.

Limitazioni e considerazioni sulla sicurezza

- Applicazione della sicurezza: i plugin del filtro CLI possono essere facilmente aggirati da qualsiasi utente e non devono essere utilizzati per politiche critiche per la sicurezza. Gli utenti possono disabilitare i plugin del filtro CLI fornendo una configurazione personalizzata che è stata `CLIFilterPlugins` disabilitata durante l'invio dei lavori.
- Solo implementazione Lua: l'implementazione dello script Lua è supportata. L'implementazione C non è supportata.

Argomenti

Configurare i plugin del filtro CLI Slurm su un cluster PCS AWS

Configura i plugin di filtro CLI quando crei un nuovo AWS cluster PCS. Puoi abilitare o disabilitare i plugin di filtro CLI sui cluster esistenti utilizzando l'API o la console di aggiornamento senza ricreare il cluster.

Prerequisiti

Prima di configurare i plugin di filtro CLI, completa queste attività:

- Scrivi e testa uno script Lua che implementa l'API CLI Filter Plugin
- Assegna un nome esatto al tuo script Lua `cli_filter.lua`
- Scegli un metodo per distribuire lo script su tutte le istanze del cluster (AMI, S3 o file system)
- Verifica di utilizzare Slurm versione 24.11 o successiva

Abilita i plugin del filtro CLI su un nuovo cluster

AWS PCS console

1. Apri la console AWS PCS all'indirizzo. https://console.aws.amazon.com/pcs/
2. Nel pannello di navigazione scegliere Cluster.
3. Scegli Crea cluster.
4. Seleziona una versione valida di Slurm (versione 24.11 o successiva).
5. In Impostazioni di pianificazione, espandi Impostazioni di pianificazione aggiuntive.
6. Aggiungi una nuova impostazione personalizzata Slurm con il nome del parametro impostato su `CliFilterPlugins` e il valore del parametro impostato su. `cli_filter/lua`
7. Completa la configurazione rimanente del cluster e scegli Crea cluster.

AWS PCS API

Fornisci la `slurmCustomSettings` configurazione nella chiamata all'azione `CreateCluster` API. Imposta «`parameterName`to» `CliFilterPlugins` e «`parameterValue`to`cli_filter/lua`». Per ulteriori informazioni, vedere CreateClusternel AWS PCS API Reference.

L'esempio seguente utilizza AWS CLI per chiamare l'azione `CreateCluster` API. L'impostazione personalizzata `CliFilterPlugins=cli_filter/lua` abilita i plugin di 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"}]'
```

Implementa gli script del plugin CLI Filter

Per distribuire gli script del CLI Filter Plugin nel tuo cluster

1. Assicurati che tutti i gruppi di nodi AMIs utilizzati nei gruppi di nodi di calcolo abbiano Slurm installato tramite il programma di installazione PCS Slurm. AWS

   > **Note**
   >
   > Se utilizzi l'AMI AWS PCS Sample per tutti i gruppi di nodi di calcolo, salta questo passaggio. Slurm è già installato.

2. Implementa `cli_filter.lua` lo script `/etc/aws/pcs/scheduler/slurm-<version>/cli_filter.lua` su tutte le istanze del cluster.

   Ad esempio, per la versione 24.11 di Slurm:

   ```
   /etc/aws/pcs/scheduler/slurm-24.11/cli_filter.lua
   ```

3. Avvia tutti i nodi di accesso e calcolo utilizzando i tuoi predisposti. AMIs
4. Verifica l'invio del lavoro per verificare che il plugin CLI Filter venga eseguito correttamente.

Abilita o disabilita i plugin di filtro CLI su cluster esistenti

Puoi abilitare o disabilitare i plugin di filtro CLI sui cluster esistenti senza ricostruire l'infrastruttura. Per ulteriori informazioni, consulta Aggiornamento di un cluster in AWS PCS.

AWS PCS console

1. Apri la console PCS all'indirizzo. AWS https://console.aws.amazon.com/pcs/

2. Nel pannello di navigazione scegliere Cluster.
3. Seleziona il cluster da aggiornare.
4. Scegli Modifica azione.
5. Nella pagina Modifica cluster, in Impostazioni aggiuntive dello scheduler:
   - Per abilitare i plugin del filtro CLI: aggiungi una nuova impostazione personalizzata Slurm con il nome del parametro impostato su **`CliFilterPlugins`** e il valore del parametro impostato su. `cli_filter/lua`
   - Per disabilitare i plugin del filtro CLI: rimuovi l'impostazione esistente. `CliFilterPlugins`
6. Scegli Aggiorna cluster per inviare le modifiche.
7. Monitora lo stato del cluster, che viene visualizzato come «Aggiornamento» durante il processo e «Attivo» quando l'aggiornamento è completo.

AWS PCS API

Utilizza l'azione `UpdateCluster` API per abilitare o disabilitare i plugin del filtro CLI. Per ulteriori informazioni, vedere UpdateClusternel AWS PCS API Reference.

Per abilitare i plugin di filtro CLI su un cluster esistente:

```
aws pcs update-cluster --cluster-identifier my-cluster \
--slurm-configuration \
'slurmCustomSettings=[{parameterName=CliFilterPlugins,parameterValue="cli_filter/
lua"}]'
```

Per disabilitare i plugin del filtro CLI su un cluster esistente:

```
aws pcs update-cluster --cluster-identifier my-cluster \
--slurm-configuration \
'slurmCustomSettings=[]'
```

Risultato previsto

Dopo aver completato la configurazione:

- Il cluster viene creato con il plug-in CLI Filter attivato

- Gli invii di lavoro attivano la tua logica di convalida personalizzata prima di raggiungere il controller Slurm
- I lavori non conformi vengono rifiutati con i tuoi messaggi di errore personalizzati
- I lavori conformi procedono normalmente tramite lo scheduler Slurm

Risoluzione dei problemi

Script CLI Filter Plugin mancante su qualsiasi nodo

Sintomi: l'invio del Job fallisce immediatamente con un errore di caricamento del plugin.

Causa probabile: lo script non è stato distribuito su tutte le istanze o il percorso o il nome del file non sono corretti.

Risoluzione: verifica che lo script esista nel percorso corretto su tutti i nodi di accesso e di calcolo con il nome file esatto. `cli_filter.lua`

Configurazione del plugin del filtro CLI non valida

Sintomi: la creazione del cluster non riesce e causa un errore di convalida.

Probabile causa: `CliFilterPlugins` parametro non impostato sul `cli_filter/lua` formato.

Risoluzione: utilizza il valore esatto del parametro `cli_filter/lua` in`slurmCustomSettings`.

Usa Amazon S3 per distribuire uno script CLI Filter Plugin in PCS AWS

Usa S3 per distribuire lo script del CLI Filter Plugin quando desideri aggiornare la logica di invio dei lavori su un cluster live senza ricostruirlo. AMIs Questo approccio scarica lo script da S3 durante l'avvio dell'istanza utilizzando i dati dell'utente.

Prerequisiti

Prima di distribuire lo script utilizzando S3, completa queste attività:

- Crea un bucket S3 con lo script Lua del plugin CLI Filter
- Configura il profilo dell'istanza IAM con accesso in lettura al bucket S3
- Configura l'endpoint S3 VPC Gateway per l'accesso diretto senza Internet
- Prepara lo script dei dati utente da scaricare da S3

Per distribuire lo script CLI Filter Plugin utilizzando S3

1. Carica `cli_filter.lua` lo script nel tuo bucket S3.
2. Configura il tuo profilo di istanza IAM con le autorizzazioni di lettura S3 per il bucket.
3. Aggiungi il codice della shell ai dati utente del modello di lancio per scaricare lo 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. Implementa gruppi di nodi di calcolo con i tuoi modelli di lancio aggiornati.
5. Testa l'invio del lavoro per verificare la funzionalità dello script.

Risultato previsto

Dopo aver completato la distribuzione di S3:

- Lo script CLI Filter Plugin viene scaricato automaticamente in tutte le istanze durante l'avvio
- Gli aggiornamenti degli script in S3 si riflettono sulle istanze appena lanciate
- Le politiche di invio dei lavori vengono applicate in modo coerente in tutto il cluster

Risoluzione dei problemi

Accesso S3 negato

Sintomi: l'avvio dell'istanza non riesce o lo script non viene scaricato.

Causa probabile: permessi IAM o endpoint VPC S3 mancanti.

Risoluzione: verifica che il profilo dell'istanza IAM `s3:GetObject` disponga dell'autorizzazione e che l'endpoint VPC S3 sia configurato.

Traduci uno script del plug-in Slurm Job Submit per utilizzare il plug-in di filtro CLI in PCS AWS

Traduci lo script Lua del Job Submit Plugin esistente in CLI Filter Plugin quando esegui la migrazione da altri ambienti Slurm. Il processo di traduzione prevede l'aggiornamento dei nomi delle funzioni e dei modelli di accesso ai campi per funzionare con l'API CLI Filter Plugin.

Prerequisiti

Prima di tradurre lo script, completa queste attività:

- Rivedi lo script Lua del Job Submit Plugin esistente
- Comprendi le differenze tra Job Submit e CLI Filter Plugin APIs
- Accedi alla documentazione del plugin Slurm CLI Filter

Per tradurre lo script Job Submit Plugin in CLI Filter Plugin

1. Rivedi le funzioni dello script Job Submit Plugin esistenti (`slurm_job_submit`,`slurm_job_modify`).
2. Identifica le funzioni equivalenti del plugin di filtro CLI:
   - `slurm_job_submit`diventa `slurm_cli_pre_submit`
   - Aggiungi `slurm_cli_setup_defaults` per l'impostazione dei parametri di default
   - Aggiungi `slurm_cli_post_submit` per azioni successive all'invio
3. Traduci la logica di convalida del lavoro dai `job_desc` campi all'accesso agli `options` array:
   - `job_desc.account`diventa `options["account"]`
   - `job_desc.partition`diventa `options["partition"]`
   - `job_desc.features`diventa `options["constraint"]`
4. Aggiorna la registrazione delle chiamate da `slurm.log_user()` a. `slurm.log_error()`
5. Testa lo script tradotto su un cluster di sviluppo.
6. Esegui la distribuzione nel tuo cluster di produzione seguendo il processo di distribuzione del plug-in CLI Filter standard.

Risultato previsto

Dopo aver completato la traduzione:

- Lo script tradotto fornisce una convalida equivalente per l'invio del lavoro
- Gli utenti visualizzano messaggi di errore e prompt simili a quelli del plugin Job Submit originale
- Le politiche di invio dei lavori vengono mantenute durante la migrazione a AWS PCS

Risoluzione dei problemi

Errori di traduzione degli script

Sintomi: gli invii di lavoro falliscono con errori di esecuzione Lua.

Causa probabile: accesso errato ai campi o chiamate di funzione nello script tradotto.

Risoluzione: rivedi la documentazione dell'API CLI Filter Plugin e confronta le mappature dei campi tra le interfacce Job Submit e CLI Filter.

Domande frequenti sui plugin di filtro CLI Slurm in PCS AWS

Consulta queste domande frequenti sui plugin del filtro CLI.

Qual è la differenza tra CLI Filter Plugin e Job Submit Plugin?

CLI Filter Plugin viene eseguito lato client sui nodi di accesso e calcolo prima che l'invio del lavoro raggiunga il controller, mentre Job Submit Plugin viene eseguito lato server sul controller dopo l'invio del lavoro. Il CLI Filter Plugin può essere aggirato dagli utenti ma non mantiene i blocchi del controller, mentre Job Submit è sicuro ma può influire sulle prestazioni del cluster durante l'esecuzione.

AWS PCS supporta il plugin Slurm Job Submit?

No, il Job Submit Plugin non è supportato in AWS PCS. Utilizza invece il plug-in CLI Filter per la convalida e la modifica dell'invio dei lavori.

Posso usare il plugin CLI Filter per l'applicazione della sicurezza?

No, il CLI Filter Plugin può essere aggirato da determinati utenti e non deve essere utilizzato per l'applicazione della sicurezza. Usalo per migliorare l'esperienza utente, impostare i parametri predefiniti e orientare le politiche piuttosto che per politiche critiche per la sicurezza.

Perché lo script deve essere presente su tutti i nodi di elaborazione, non solo sui nodi di accesso?

Comandi Slurm come `srun` possono essere eseguiti all'interno di script di lavoro su nodi di calcolo, il che attiva anche l'esecuzione del CLI Filter Plugin. Lo script deve essere disponibile ovunque vengano eseguiti i comandi Slurm.

Posso modificare lo script CLI Filter Plugin su un cluster live?

Sì, se utilizzi l'approccio di implementazione di S3 o del file system. Le nuove istanze riceveranno lo script aggiornato, ma le istanze esistenti richiedono l'aggiornamento dello script manualmente o tramite il metodo di distribuzione scelto.

Posso usare diversi script del CLI Filter Plugin su diversi gruppi di nodi di calcolo?

Sì, ma questo non è consigliato. Puoi fornire script con logica diversa a diversi gruppi di nodi di calcolo, ma sei responsabile della gestione delle interdipendenze e della prevenzione delle sovrapposizioni logiche. La maggior parte dei clienti fornisce un set di logica per un intero cluster.

Posso usare CLI Filter Plugin con implementazione C anziché Lua?

L'implementazione C non è supportata. In AWS PCS è supportata solo l'implementazione dello script Lua. SchedMD consiglia ai clienti di utilizzare Lua su C per facilità d'uso durante l'implementazione dei plugin di filtro CLI.

Posso attivare o disattivare il plug-in CLI Filter su un cluster esistente?

Sì, puoi abilitare o disabilitare il plug-in di filtro CLI sui cluster esistenti utilizzando l'API di aggiornamento senza ricreare il cluster.

Risoluzione dei problemi del plug-in Slurm CLI Filter in PCS AWS

Utilizza queste informazioni di risoluzione dei problemi per risolvere i problemi più comuni del CLI Filter Plugin.

L'invio del Job fallisce immediatamente con un errore di caricamento del plugin

Sintomi: gli utenti ricevono messaggi di errore relativi al plug-in di filtro CLI mancante o fallito durante l'invio dei lavori.

Possibili cause:

- Script CLI Filter Plugin mancante da uno o più nodi
- Nome del file di script errato (deve essere esattamente) `cli_filter.lua`
- Script distribuito in un percorso di directory errato
- Lo script ha autorizzazioni di file errate

Risoluzione:

- Verifica che lo script esista `/etc/aws/pcs/scheduler/slurm-<version>/cli_filter.lua` in tutti i nodi di accesso e di calcolo
- Controlla che il nome del file dello script sia esattamente `cli_filter.lua`
- Assicurati che lo script abbia autorizzazioni leggibili (644 o simili)
- Verifica la distribuzione degli script su un singolo nodo di accesso prima di distribuirli all'intero cluster

La creazione del cluster non riesce a causa dell'errore di convalida del plugin CLI Filter

Sintomi: la creazione del cluster non riesce e viene generato un errore relativo a un parametro non valido`CliFilterPlugins`.

Possibili cause:

- Formato errato del valore del parametro in `slurmCustomSettings`
- Digitare il nome o il valore del parametro

Risoluzione:

- Usa il nome esatto del parametro: `CliFilterPlugins`
- Usa il valore esatto del parametro: `cli_filter/lua`
- Verifica la sintassi JSON nell'array `slurmCustomSettings`

Lo script CLI Filter Plugin viene eseguito ma la convalida del lavoro non funziona come previsto

Sintomi: i lavori vengono inviati correttamente, ma la logica di convalida personalizzata non si attiva o produce risultati imprevisti.

Possibili cause:

- Errori di sintassi dello script Lua
- Schemi di accesso ai campi errati (utilizzo della sintassi Job Submit Plugin anziché del CLI Filter Plugin)
- Errori logici nelle condizioni di convalida

Risoluzione:

- Controlla lo script Lua per eventuali errori di sintassi
- Verify Field Access utilizza `options["field_name"]` format anziché `job_desc.field_name`
- Aggiungi istruzioni di registrazione al flusso di esecuzione degli script di debug

- Verifica prima la logica degli script con semplici casi di convalida

La distribuzione degli script S3 non riesce

Sintomi: le istanze vengono avviate ma lo script CLI Filter Plugin non viene scaricato da S3.

Possibili cause:

- Il profilo dell'istanza IAM non dispone delle autorizzazioni di lettura S3
- Endpoint VPC S3 non configurato
- Il bucket S3 o il percorso dell'oggetto nei dati utente non sono corretti

Risoluzione:

- Verifica che il profilo dell'istanza IAM disponga `s3:GetObject` dell'autorizzazione per il tuo bucket
- Configura l'endpoint S3 VPC Gateway per l'accesso diretto
- Controlla il nome del bucket S3 e il percorso dell'oggetto nello script dei dati utente
- Esamina i log dei dati utente dell'istanza per verificare la presenza di errori di download di S3

Servizio di sicurezza nel servizio AWS Parallel Computing

La sicurezza del cloud AWS è la massima priorità. In qualità di AWS cliente, puoi beneficiare di data center e architetture di rete progettati per soddisfare i requisiti delle organizzazioni più sensibili alla sicurezza.

La sicurezza è una responsabilità condivisa tra te AWS e te. Il modello di responsabilità condivisa descrive questo aspetto come sicurezza del cloud e sicurezza nel cloud:

- Sicurezza del cloud: AWS è responsabile della protezione dell'infrastruttura che gestisce AWS i servizi in Cloud AWS. AWS fornisce inoltre servizi che è possibile utilizzare in modo sicuro. I revisori esterni testano e verificano regolarmente l'efficacia della nostra sicurezza nell'ambito dei AWS Programmi di AWS conformità dei Programmi di conformità dei di . Per ulteriori informazioni sui programmi di conformità che si applicano al servizio di elaborazione AWS parallela, vedere AWS Servizi nell'ambito del programma di conformitàAWS .
- Sicurezza nel cloud: la tua responsabilità è determinata dal AWS servizio che utilizzi. L'utente è anche responsabile di altri fattori, tra cui la riservatezza dei dati, i requisiti della propria azienda e le leggi e normative vigenti.

Questa documentazione aiuta a capire come applicare il modello di responsabilità condivisa quando si utilizza AWS PCS. I seguenti argomenti mostrano come configurare AWS PCS per soddisfare gli obiettivi di sicurezza e conformità. Imparerai anche come utilizzare altri AWS servizi che ti aiutano a monitorare e proteggere le tue risorse AWS PCS.

Argomenti

Protezione dei dati in AWS Parallel Computing Service

Il modello di responsabilità AWS condivisa modello si applica alla protezione dei dati in AWS Parallel Computing Service. Come descritto in questo modello, AWS è responsabile della protezione dell'infrastruttura globale che gestisce tutti i Cloud AWS. L'utente è responsabile del controllo dei contenuti ospitati su questa infrastruttura. L'utente è inoltre responsabile della configurazione della protezione e delle attività di gestione per i Servizi AWS utilizzati. Per maggiori informazioni sulla privacy dei dati, consulta le Domande frequenti sulla privacy dei dati. Per informazioni sulla protezione dei dati in Europa, consulta il post del blog relativo al AWS Modello di responsabilità condivisa e GDPR nel AWS Blog sulla sicurezza.

Ai fini della protezione dei dati, consigliamo di proteggere Account AWS le credenziali e configurare i singoli utenti con AWS IAM Identity Center or AWS Identity and Access Management (IAM). In tal modo, a ogni utente verranno assegnate solo le autorizzazioni necessarie per svolgere i suoi compiti. Suggeriamo, inoltre, di proteggere i dati nei seguenti modi:

- Utilizza l'autenticazione a più fattori (MFA) con ogni account.
- SSL/TLS Da utilizzare per comunicare con AWS le risorse. È richiesto TLS 1.2 ed è consigliato TLS 1.3.
- Configura l'API e la registrazione delle attività degli utenti con AWS CloudTrail. Per informazioni sull'utilizzo dei CloudTrail percorsi per acquisire AWS le attività, consulta Lavorare con i CloudTrail percorsi nella Guida per l'AWS CloudTrail utente.
- Utilizza soluzioni di AWS crittografia, insieme a tutti i controlli di sicurezza predefiniti all'interno Servizi AWS.
- Utilizza i servizi di sicurezza gestiti avanzati, come Amazon Macie, che aiutano a individuare e proteggere i dati sensibili archiviati in Amazon S3.
- Se hai bisogno di moduli crittografici convalidati FIPS 140-3 per accedere AWS tramite un'interfaccia a riga di comando o un'API, usa un endpoint FIPS. Per ulteriori informazioni sugli endpoint FIPS disponibili, consulta il Federal Information Processing Standard (FIPS) 140-3.

Ti consigliamo di non inserire mai informazioni riservate o sensibili, ad esempio gli indirizzi e-mail dei clienti, nei tag o nei campi di testo in formato libero, ad esempio nel campo Nome. Ciò include quando lavori con AWS PCS o altri dispositivi Servizi AWS utilizzando la console, l'API o. AWS CLI AWS SDKs I dati inseriti nei tag o nei campi di testo in formato libero utilizzati per i nomi possono essere utilizzati per i la fatturazione o i log di diagnostica. Quando si fornisce un URL a un

server esterno, suggeriamo vivamente di non includere informazioni sulle credenziali nell'URL per convalidare la richiesta al server.

Crittografia dei dati a riposo

La crittografia è abilitata per impostazione predefinita per i dati inattivi quando si crea un cluster AWS PCS (AWS Parallel Computing Service) con Console di gestione AWS AWS CLI,, AWS PCS API o AWS SDKs. AWS PCS utilizza una chiave KMS AWS di proprietà per crittografare i dati inattivi. Per ulteriori informazioni, consulta Customer keys and AWS keys nella AWS KMS Developer Guide. Puoi anche utilizzare una chiave gestita dal cliente. Per ulteriori informazioni, consulta Politica delle chiavi KMS richiesta per l'uso con volumi EBS crittografati in PCS AWS.

Il segreto del cluster viene archiviato Gestione dei segreti AWS e crittografato con la chiave KMS gestita da Secrets Manager. Per ulteriori informazioni, consulta Utilizzo dei segreti del cluster in AWS PCS.

In un cluster AWS PCS, i seguenti dati sono inattivi:

- Stato dell'utilità di pianificazione: include i dati sui processi in esecuzione e sui nodi a cui è stato assegnato il provisioning nel cluster. Questi sono i dati in cui Slurm persiste nei dati definiti nel tuo. `StateSaveLocation` `slurm.conf` Per ulteriori informazioni, consulta la descrizione contenuta StateSaveLocationnella documentazione di Slurm. AWS PCS elimina i dati del lavoro dopo il completamento di un lavoro.
- Segreto di autenticazione dello scheduler: AWS PCS lo utilizza per autenticare tutte le comunicazioni dello scheduler nel cluster.

Per quanto riguarda le informazioni sullo stato dello scheduler, AWS PCS crittografa automaticamente i dati e i metadati prima di scriverli nel file system. Il file system crittografato utilizza l'algoritmo di crittografia AES-256 standard del settore per i dati inattivi.

Crittografia dei dati in transito

Le tue connessioni all'API AWS PCS utilizzano la crittografia TLS con il processo di firma Signature Version 4, indipendentemente dal fatto che utilizzi AWS Command Line Interface ()AWS CLI o. AWS SDKs Per ulteriori informazioni, consulta Firmare le richieste AWS API nella Guida per l'AWS Identity and Access Management utente. AWS gestisce il controllo degli accessi tramite l'API con le politiche IAM per le credenziali di sicurezza utilizzate per la connessione.

AWS PCS utilizza TLS per connettersi ad altri AWS servizi.

All'interno di un cluster Slurm, lo scheduler è configurato con il plug-in di autenticazione che fornisce l'`auth/slurm`autenticazione per tutte le comunicazioni dello scheduler. Slurm non fornisce la crittografia a livello di applicazione per le sue comunicazioni, tutti i dati che fluiscono tra le istanze del cluster rimangono locali sul VPC EC2 e pertanto sono soggetti alla crittografia VPC se tali istanze supportano la crittografia in transito. Per ulteriori informazioni, consulta Encryption in transit nella Amazon Elastic Compute Cloud User Guide. La comunicazione è crittografata tra il controller (fornito in un account di servizio) e i nodi del cluster del tuo account.

Gestione delle chiavi

AWS PCS utilizza una chiave KMS AWS di proprietà per crittografare i dati. Per ulteriori informazioni, consulta Customer keys and AWS keys nella AWS KMS Developer Guide. Puoi anche utilizzare una chiave gestita dal cliente. Per ulteriori informazioni, consulta Politica delle chiavi KMS richiesta per l'uso con volumi EBS crittografati in PCS AWS.

Il segreto del cluster viene archiviato Gestione dei segreti AWS e crittografato con la chiave KMS gestita da Secrets Manager. Per ulteriori informazioni, consulta Utilizzo dei segreti del cluster in AWS PCS.

Riservatezza del traffico inter-rete

AWS Le risorse di calcolo PCS per un cluster risiedono all'interno di 1 VPC nell'account del cliente. Pertanto, tutto il traffico del servizio AWS PCS interno all'interno di un cluster rimane all'interno della AWS rete e non viaggia su Internet. La comunicazione tra l'utente e i nodi AWS PCS può viaggiare su Internet e consigliamo di utilizzare SSH o Systems Manager per connettersi ai nodi. Per ulteriori informazioni, consulta Cos'è AWS Systems Manager? nella Guida AWS Systems Manager per l'utente.

Puoi anche utilizzare le seguenti offerte per connettere la tua rete locale a: AWS

- AWS Site-to-Site VPN. Per ulteriori informazioni, vedi Cos'è AWS Site-to-Site VPN? nella Guida AWS Site-to-Site VPN per l'utente.
- Un AWS Direct Connect. Per ulteriori informazioni, vedi Cos'è AWS Direct Connect? nella Guida AWS Direct Connect per l'utente.

Si accede all'API AWS PCS per eseguire attività amministrative per il servizio. Tu e i tuoi utenti accedete alle porte degli endpoint Slurm per interagire direttamente con lo scheduler.

Crittografia del traffico API

Per accedere all'API AWS PCS, i client devono supportare Transport Layer Security (TLS) 1.2 o versione successiva. È richiesto TLS 1.2 ed è consigliato TLS 1.3. I client devono inoltre supportare le suite di cifratura con PFS (Perfect Forward Secrecy), ad esempio Ephemeral Diffie-Hellman (DHE) o Elliptic Curve Diffie-Hellman Ephemeral (ECDHE). La maggior parte dei sistemi moderni, come Java 7 e versioni successive, supporta tali modalità. Inoltre, le richieste devono essere firmate utilizzando un ID chiave di accesso e una chiave di accesso segreta associata a un principale IAM. È inoltre possibile utilizzare AWS Security Token Service (AWS STS) per generare credenziali di sicurezza temporanee per firmare le richieste.

Crittografia del traffico dati

La crittografia dei dati in transito è abilitata dalle istanze EC2 supportate che accedono all'endpoint dello scheduler e tra ComputeNodeGroup le istanze dall'interno di. Cloud AWS Per ulteriori informazioni, consulta Crittografia dei dati in transito.

Politica delle chiavi KMS richiesta per l'uso con volumi EBS crittografati in PCS AWS

AWS PCS utilizza ruoli collegati ai servizi per delegare le autorizzazioni ad altri. Servizi AWS Il ruolo collegato al servizio AWS PCS è predefinito e include le autorizzazioni richieste da AWS PCS per chiamare altri utenti per conto dell'utente. Servizi AWS Le autorizzazioni predefinite includono anche l'accesso alle chiavi gestite dai clienti, Chiavi gestite da AWS ma non a quelle gestite dai clienti.

Questo argomento descrive come configurare la politica delle chiavi richiesta per avviare le istanze quando si specifica una chiave gestita dal cliente per la crittografia Amazon EBS.

> **Note**
>
> AWS PCS non richiede un'autorizzazione aggiuntiva per utilizzare l'impostazione predefinita Chiave gestita da AWS per proteggere i volumi crittografati nel tuo account.

Indice

- Esempio 1: sezioni delle policy delle chiavi che permettono l'accesso alla chiave gestita dal cliente
- Esempio 2: sezioni delle policy delle chiavi che permettono l'accesso multiaccount alla chiave gestita dal cliente
- Modifica le politiche chiave nella console AWS KMS

Panoramica di

È possibile utilizzare quanto segue AWS KMS keys per la crittografia Amazon EBS quando AWS PCS avvia istanze:

- Chiave gestita da AWS— Una chiave di crittografia nel tuo account che Amazon EBS crea, possiede e gestisce. Questa è la chiave di crittografia di default per un nuovo account. Amazon EBS utilizza la Chiave gestita da AWS crittografia a meno che non venga specificata una chiave gestita dal cliente.
- Chiave gestita dal cliente: una chiave di crittografia personalizzata che l'utente crea, possiede e gestisce. Per ulteriori informazioni, consulta Creare una chiave KMS nella Guida per gli AWS Key Management Service sviluppatori.

> **Note**
>
> La chiave deve essere simmetrica. Amazon EBS non supporta chiavi asimmetriche gestite dai clienti.

Le chiavi gestite dal cliente vengono configurate quando si creano istantanee crittografate o un modello di avvio che specifica i volumi crittografati o quando si sceglie di abilitare la crittografia per impostazione predefinita.

Configurare le policy chiave

Le tue chiavi KMS devono avere una politica chiave che consenta a AWS PCS di avviare istanze con volumi Amazon EBS crittografati con una chiave gestita dal cliente.

Utilizza gli esempi in questa pagina per configurare una politica chiave che consenta a AWS PCS di accedere alla chiave gestita dal cliente. È possibile modificare la politica chiave della chiave gestita dal cliente al momento della creazione della chiave o in un secondo momento.

La politica chiave deve contenere le seguenti dichiarazioni:

- Un'istruzione che consente all'identità IAM specificata nell'`Principal`elemento di utilizzare direttamente la chiave gestita dal cliente. Include le autorizzazioni per eseguire AWS KMS `Encrypt`Decrypt, `ReEncrypt*`GenerateDataKey*, e `DescribeKey` le operazioni sulla chiave.
- Un'istruzione che consente all'identità IAM specificata nell'`Principal`elemento di utilizzare l'`CreateGrant`operazione per generare concessioni che delegano un sottoinsieme delle proprie autorizzazioni a quelle integrate con o con un Servizi AWS altro principale. AWS KMS Questo permette di utilizzare la chiave per creare le risorse crittografate per te.

Non modificate alcuna dichiarazione esistente nella policy quando aggiungete le nuove dichiarazioni politiche alla vostra policy chiave.

Per ulteriori informazioni, consulta:

- create-key nel Command Reference AWS CLI
- put-key-policy in Riferimento ai comandi AWS CLI
- Trova l'ID chiave e l'ARN della chiave nella Guida per gli sviluppatori AWS Key Management Service
- Ruoli collegati ai servizi per PCS AWS
- Crittografia Amazon EBS nella Guida per l'utente di Amazon EBS
- AWS Key Management Servicenella Guida per gli sviluppatori AWS Key Management Service

Esempio 1: sezioni delle policy delle chiavi che permettono l'accesso alla chiave gestita dal cliente

Aggiungi le seguenti dichiarazioni politiche alla politica chiave della chiave gestita dal cliente. Sostituisci l'ARN di esempio con l'ARN del tuo ruolo collegato al servizio. `AWSServiceRoleForPCS` Questa politica di esempio fornisce al ruolo collegato al servizio AWS PCS (`AWSServiceRoleForPCS`) le autorizzazioni per utilizzare la chiave gestita dal 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
        }
     }
}
```

Esempio 2: sezioni delle policy delle chiavi che permettono l'accesso multiaccount alla chiave gestita dal cliente

Se si crea una chiave gestita dal cliente in un account diverso da quello del cluster AWS PCS, è necessario utilizzare una concessione in combinazione con la politica chiave per consentire l'accesso alla chiave da più account.

Per concedere l'accesso alla chiave

1. Aggiungi le seguenti dichiarazioni politiche alla politica chiave della chiave gestita dal cliente. Sostituisci l'ARN di esempio con l'ARN dell'altro account. Sostituiscilo *111122223333* con l'ID effettivo dell'account in Account AWS cui desideri creare il cluster AWS PCS. Ciò permette di dare a un utente o ruolo IAM dell'account specificato l'autorizzazione a creare una concessione per la chiave utilizzando il comando CLI che segue. Per impostazione predefinita, gli utenti non hanno accesso alla chiave.

```
{.
   "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. Dall'account in cui desideri creare il cluster AWS PCS, crea una concessione che deleghi le autorizzazioni pertinenti al ruolo collegato al servizio AWS PCS. Il valore di `grantee-principal` è l'ARN del ruolo collegato al servizio. Il valore di `key-id` è l'ARN della chiave.

   Il comando CLI create-grant di esempio seguente fornisce al ruolo collegato al servizio indicato `AWSServiceRoleForPCS` nelle *111122223333* autorizzazioni dell'account l'utilizzo della chiave gestita dal cliente nell'account. *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**
   >
   > L'utente che effettua la richiesta deve disporre delle autorizzazioni per utilizzare l'azione. `kms:CreateGrant`

   L'esempio seguente di policy IAM consente a un'identità IAM (utente o ruolo) in un account di *111122223333* creare una concessione per l'account *444455556666* key in gestito dal cliente.

   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"
    }
  ]
```

```
}
```

Per ulteriori informazioni sulla creazione di una concessione per una chiave KMS in un diverso Account AWS, consulta Concessioni in AWS KMS nella Guida per gli sviluppatori AWS Key Management Service .

> **Important**
>
> Il nome del ruolo collegato al servizio specificato come principale assegnatario deve essere il nome di un ruolo esistente. Dopo aver creato la concessione, per assicurarti che la concessione consenta a AWS PCS di utilizzare la chiave KMS specificata, non eliminare e ricreare il ruolo collegato al servizio.

Modifica le politiche chiave nella console AWS KMS

Gli esempi nelle seguenti sezioni mostrano solo come aggiungere le istruzioni alla policy di una chiave, che è solo uno dei modi per modificare questo tipo di policy. Il modo più semplice per modificare una policy chiave consiste nell'utilizzare la visualizzazione predefinita della AWS KMS console per le policy chiave e rendere un'identità IAM (utente o ruolo) uno degli utenti chiave per la policy chiave appropriata. Per ulteriori informazioni, consulta Using the Console di gestione AWS default view nella AWS Key Management Service Developer Guide.

> **Warning**
>
> Le dichiarazioni sulla politica di visualizzazione predefinita della console includono le autorizzazioni per eseguire AWS KMS `Revoke` operazioni sulla chiave gestita dal cliente. Se revochi una concessione che consentiva Account AWS l'accesso a una chiave gestita dal cliente nel tuo account, gli utenti in tale account Account AWS perdono l'accesso ai dati crittografati e alla chiave.

Accesso AWS Parallel Computing Service tramite un'interfaccia endpoint ()AWS PrivateLink

Puoi usarlo AWS PrivateLink per creare una connessione privata tra il tuo VPC e AWS Parallel Computing Service ()AWS PCS. Puoi accedere AWS PCS come se fosse nel tuo VPC, senza l'uso di

un gateway Internet, un dispositivo NAT, una connessione VPN o una connessione. Direct Connect Le istanze del tuo VPC non necessitano di indirizzi IP pubblici per accedervi. AWS PCS

Stabilisci questa connessione privata creando un endpoint di interfaccia attivato da AWS PrivateLink. In ciascuna sottorete viene creata un'interfaccia di rete endpoint da abilitare per l'endpoint di interfaccia. Queste sono interfacce di rete gestite dal richiedente che fungono da punto di ingresso per il traffico destinato a AWS PCS.

Per ulteriori informazioni, consulta Access Servizi AWS through AWS PrivateLink nella AWS PrivateLink Guida.

Considerazioni per AWS PCS

Prima di configurare un endpoint di interfaccia per AWS PCS, consulta Accedere a un servizio AWS utilizzando un endpoint VPC di interfaccia nella Guida.AWS PrivateLink

AWS PCS supporta l'esecuzione di chiamate a tutte le sue azioni API tramite l'endpoint dell'interfaccia.

Se il tuo VPC non dispone di accesso diretto a Internet, devi configurare un endpoint VPC per consentire alle istanze del gruppo di nodi di calcolo di richiamare l'azione API. AWS PCS `RegisterComputeNodeGroupInstance`

Crea un endpoint di interfaccia per AWS PCS

Puoi creare un endpoint di interfaccia per AWS PCS utilizzare la console Amazon VPC o AWS Command Line Interface ().AWS CLI Per ulteriori informazioni, consulta la sezione Creazione di un endpoint di interfaccia nella Guida per l'utente di AWS PrivateLink .

Crea un endpoint di interfaccia per AWS PCS utilizzare il seguente nome di servizio:

```
com.amazonaws.region.pcs
```

Sostituisci *`region`* con l'ID del dispositivo in Regione AWS cui creare l'endpoint, ad esempio. `us-east-1`

Se abiliti il DNS privato per l'endpoint dell'interfaccia, puoi effettuare richieste API AWS PCS utilizzando il nome DNS regionale predefinito. Ad esempio, `pcs.us-east-1.amazonaws.com`.

Creazione di una policy dell' endpoint per l'endpoint dell'interfaccia

Una policy dell'endpoint è una risorsa IAM che è possibile allegare all'endpoint dell'interfaccia. La policy predefinita per gli endpoint consente l'accesso completo AWS PCS tramite l'endpoint dell'interfaccia. Per controllare l'accesso consentito AWS PCS dal tuo VPC, collega una policy endpoint personalizzata all'endpoint di interfaccia.

Una policy di endpoint specifica le informazioni riportate di seguito:

- I principali che possono eseguire azioni (Account AWS, utenti IAM e ruoli IAM).
- Le azioni che possono essere eseguite.
- Le risorse in cui è possibile eseguire le operazioni.

Per ulteriori informazioni, consulta la sezione Controllo dell'accesso ai servizi con policy di endpoint nella Guida di AWS PrivateLink .

Esempio: policy degli endpoint VPC per le azioni AWS PCS

Di seguito è riportato l'esempio di una policy dell'endpoint personalizzata. Quando si allega questa policy all'endpoint di interfaccia, si concede l'accesso alle AWS PCS azioni elencate per tutti i principali attori del cluster con le specifiche. *`cluster-id`* Sostituisci *`region`* con l'ID Regione AWS del cluster, ad esempio. `us-east-1` Sostituisci *`account-id`* con il Account AWS numero del cluster.

```
{
    "Statement": [
            {
                "Action": [
                "pcs:CreateCluster",
                "pcs:ListClusters",
                "pcs:DeleteCluster",
                "pcs:GetCluster",
                ],
                "Effect": "Allow",
                "Principal": "*",
                "Resource": [
                    "arn:aws:pcs:region:account-id:cluster/cluster-id*"
                ]
            }
        ]
```

```
}
```

Servizio di Identity and Access Management per AWS Parallel Computing

AWS Identity and Access Management (IAM) è un software Servizio AWS che aiuta un amministratore a controllare in modo sicuro l'accesso alle AWS risorse. Gli amministratori IAM controllano chi può essere autenticato (effettuato l'accesso) e autorizzato (disporre delle autorizzazioni) a utilizzare le risorse PCS. AWS IAM è uno Servizio AWS strumento che puoi utilizzare senza costi aggiuntivi.

Argomenti

Destinatari

Il modo in cui utilizzi AWS Identity and Access Management (IAM) varia in base al tuo ruolo:

- Utente del servizio: richiedi le autorizzazioni all'amministratore se non riesci ad accedere alle funzionalità (consulta Risoluzione dei problemi di identità e accesso al AWS Parallel Computing Service)
- Amministratore del servizio: determina l'accesso degli utenti e invia le richieste di autorizzazione (consulta Come funziona AWS Parallel Computing Service con IAM)

- Amministratore IAM: scrivi policy per gestire l'accesso (consulta Esempi di policy basate sull'identità per Parallel Computing Service AWS)

Autenticazione con identità

L'autenticazione è il modo in cui accedi AWS utilizzando le tue credenziali di identità. Devi autenticarti come utente IAM o assumendo un ruolo IAM. Utente root dell'account AWS

Puoi accedere come identità federata utilizzando credenziali provenienti da una fonte di identità come AWS IAM Identity Center (IAM Identity Center), autenticazione Single Sign-On o credenziali. Google/ Facebook Per ulteriori informazioni sull'accesso, consulta Come accedere all' Account AWS nella Guida per l'utente di Accedi ad AWS .

Per l'accesso programmatico, AWS fornisce un SDK e una CLI per firmare crittograficamente le richieste. Per ulteriori informazioni, consulta AWS Signature Version 4 per le richieste API nella Guida per l'utente di IAM.

Account AWS utente root

Quando si crea un Account AWS, si inizia con un'identità di accesso denominata utente Account AWS root che ha accesso completo a tutte Servizi AWS le risorse. Consigliamo vivamente di non utilizzare l'utente root per le attività quotidiane. Per le attività che richiedono le credenziali dell'utente root, consulta Attività che richiedono le credenziali dell'utente root nella Guida per l'utente IAM.

Identità federata

Come procedura ottimale, richiedi agli utenti umani di utilizzare la federazione con un provider di identità per accedere Servizi AWS utilizzando credenziali temporanee.

Un'identità federata è un utente della directory aziendale, del provider di identità Web o Directory Service che accede Servizi AWS utilizzando le credenziali di una fonte di identità. Le identità federate assumono ruoli che forniscono credenziali temporanee.

Per la gestione centralizzata degli accessi, si consiglia di utilizzare AWS IAM Identity Center. Per ulteriori informazioni, consulta Che cos'è il Centro identità IAM? nella Guida per l'utente di AWS IAM Identity Center .

Utenti e gruppi IAM

Un utente IAM è una identità che dispone di autorizzazioni specifiche per una singola persona o applicazione. Ti consigliamo di utilizzare credenziali temporanee invece di utenti IAM con credenziali a lungo termine. Per ulteriori informazioni, consulta Richiedere agli utenti umani di utilizzare la federazione con un provider di identità per accedere AWS utilizzando credenziali temporanee nella Guida per l'utente IAM.

Un gruppo IAM specifica una raccolta di utenti IAM e semplifica la gestione delle autorizzazioni per gestire gruppi di utenti di grandi dimensioni. Per ulteriori informazioni, consulta Casi d'uso per utenti IAM nella Guida per l'utente di IAM.

Ruoli IAM

Un ruolo IAM è un'identità con autorizzazioni specifiche che fornisce credenziali temporanee. Puoi assumere un ruolo passando da un ruolo utente a un ruolo IAM (console) o chiamando un'operazione AWS CLI o AWS API. Per ulteriori informazioni, consulta Metodi per assumere un ruolo nella Guida per l'utente di IAM.

I ruoli IAM sono utili per l'accesso degli utenti federati, le autorizzazioni utente IAM temporanee, l'accesso multi-account, l'accesso multi-servizio e le applicazioni in esecuzione su Amazon EC2. Per maggiori informazioni, consultare Accesso a risorse multi-account in IAM nella Guida per l'utente IAM.

Gestione dell'accesso tramite policy

Puoi controllare l'accesso AWS creando policy e associandole a AWS identità o risorse. Una policy definisce le autorizzazioni quando è associata a un'identità o a una risorsa. AWS valuta queste politiche quando un preside effettua una richiesta. La maggior parte delle politiche viene archiviata AWS come documenti JSON. Per maggiori informazioni sui documenti delle policy JSON, consulta Panoramica delle policy JSON nella Guida per l'utente IAM.

Utilizzando le policy, gli amministratori specificano chi ha accesso a cosa definendo quale principale può eseguire azioni su quali risorse e in quali condizioni.

Per impostazione predefinita, utenti e ruoli non dispongono di autorizzazioni. Un amministratore IAM crea le policy IAM e le aggiunge ai ruoli, che gli utenti possono quindi assumere. Le policy IAM definiscono le autorizzazioni indipendentemente dal metodo utilizzato per eseguirle.

Policy basate sull'identità

Le policy basate su identità sono documenti di policy di autorizzazione JSON che è possibile collegare a un'identità (utente, gruppo o ruolo). Tali policy controllano le operazioni autorizzate per l'identità, nonché le risorse e le condizioni in cui possono essere eseguite. Per informazioni su come creare una policy basata su identità, consultare Definizione di autorizzazioni personalizzate IAM con policy gestite dal cliente nella Guida per l'utente IAM.

Le policy basate su identità possono essere policy in linea (con embedding direttamente in una singola identità) o policy gestite (policy autonome collegate a più identità). Per informazioni su come scegliere tra una policy gestita o una policy inline, consulta Scegliere tra policy gestite e policy in linea nella Guida per l'utente di IAM.

Policy basate sulle risorse

Le policy basate su risorse sono documenti di policy JSON che è possibile collegare a una risorsa. Gli esempi includono le policy di trust dei ruoli IAM e le policy dei bucket di Amazon S3. Nei servizi che supportano policy basate sulle risorse, gli amministratori dei servizi possono utilizzarli per controllare l'accesso a una risorsa specifica. In una policy basata sulle risorse è obbligatorio specificare un'entità principale.

Le policy basate sulle risorse sono policy inline che si trovano in tale servizio. Non è possibile utilizzare le policy AWS gestite di IAM in una policy basata sulle risorse.

Altri tipi di policy

AWS supporta tipi di policy aggiuntivi che possono impostare le autorizzazioni massime concesse dai tipi di policy più comuni:

- Limiti delle autorizzazioni: imposta il numero massimo di autorizzazioni che una policy basata su identità ha la possibilità di concedere a un'entità IAM. Per ulteriori informazioni, consulta Limiti delle autorizzazioni per le entità IAM nella Guida per l'utente di IAM.
- Politiche di controllo del servizio (SCPs): specificano le autorizzazioni massime per un'organizzazione o un'unità organizzativa in. AWS Organizations Per ulteriori informazioni, consultare Policy di controllo dei servizi nella Guida per l'utente di AWS Organizations .
- Politiche di controllo delle risorse (RCPs): imposta le autorizzazioni massime disponibili per le risorse nei tuoi account. Per ulteriori informazioni, consulta Politiche di controllo delle risorse (RCPs) nella Guida per l'AWS Organizations utente.

- Policy di sessione: policy avanzate passate come parametro quando si crea una sessione temporanea per un ruolo o un utente federato. Per maggiori informazioni, consultare Policy di sessione nella Guida per l'utente IAM.

Più tipi di policy

Quando a una richiesta si applicano più tipi di policy, le autorizzazioni risultanti sono più complicate da comprendere. Per scoprire come si AWS determina se consentire o meno una richiesta quando sono coinvolti più tipi di policy, consulta Logica di valutazione delle policy nella IAM User Guide.

Come funziona AWS Parallel Computing Service con IAM

Prima di utilizzare IAM per gestire l'accesso ai AWS PCS, scopri quali funzionalità IAM sono disponibili per l'uso con AWS PCS.

Funzionalità IAM che puoi utilizzare con AWS Parallel Computing Service

Funzionalità IAM	AWS Supporto PCS
Policy basate sull'identità	Sì
Policy basate su risorse	No
Operazioni di policy	Sì
Risorse relative alle policy	Sì
Chiavi di condizione della policy (specifica del servizio)	Sì
ACLs	No
ABAC (tag nelle policy)	Sì
Credenziali temporanee	Sì
Autorizzazioni del principale	Sì
Ruoli di servizio	No
Ruoli collegati al servizio	Sì

Per avere una visione di alto livello di come AWS PCS e altri AWS servizi funzionano con la maggior parte delle funzionalità IAM, consulta AWS i servizi che funzionano con IAM nella IAM User Guide.

Politiche basate sull'identità per i PC AWS

Supporta le policy basate sull'identità: sì

Le policy basate sull'identità sono documenti di policy di autorizzazione JSON che è possibile allegare a un'identità (utente, gruppo di utenti o ruolo IAM). Tali policy definiscono le operazioni che utenti e ruoli possono eseguire, su quali risorse e in quali condizioni. Per informazioni su come creare una policy basata su identità, consulta Definizione di autorizzazioni personalizzate IAM con policy gestite dal cliente nella Guida per l'utente di IAM.

Con le policy basate sull'identità di IAM, è possibile specificare quali operazioni e risorse sono consentite o respinte, nonché le condizioni in base alle quali le operazioni sono consentite o respinte. Per informazioni su tutti gli elementi utilizzabili in una policy JSON, consulta Guida di riferimento agli elementi delle policy JSON IAM nella Guida per l'utente IAM.

Esempi di policy basate sull'identità per PCS AWS

Per visualizzare esempi di politiche AWS PCS basate sull'identità, vedere. Esempi di policy basate sull'identità per Parallel Computing Service AWS

Politiche basate sulle risorse all'interno di PCS AWS

Supporta le policy basate su risorse: no

Le policy basate su risorse sono documenti di policy JSON che è possibile collegare a una risorsa. Esempi di policy basate sulle risorse sono le policy di attendibilità dei ruoli IAM e le policy di bucket Amazon S3. Nei servizi che supportano policy basate sulle risorse, gli amministratori dei servizi possono utilizzarli per controllare l'accesso a una risorsa specifica. Quando è collegata a una risorsa, una policy definisce le operazioni che un principale può eseguire su tale risorsa e a quali condizioni. In una policy basata sulle risorse è obbligatorio specificare un'entità principale. I principali possono includere account, utenti, ruoli, utenti federati o. Servizi AWS

Per consentire l'accesso multi-account, è possibile specificare un intero account o entità IAM in un altro account come entità principale in una policy basata sulle risorse. Per ulteriori informazioni, consulta Accesso a risorse multi-account in IAM nella Guida per l'utente IAM.

Azioni politiche per PCS AWS

Supporta le operazioni di policy: si

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale entità principale può eseguire operazioni su quali risorse e in quali condizioni.

L'elemento `Action` di una policy JSON descrive le operazioni che è possibile utilizzare per consentire o negare l'accesso in una policy. Includere le operazioni in una policy per concedere le autorizzazioni a eseguire l'operazione associata.

Per visualizzare un elenco delle azioni AWS PCS, vedere Azioni definite da AWS Parallel Computing Service nel Service Authorization Reference.

Le azioni politiche in AWS PCS utilizzano il seguente prefisso prima dell'azione:

```
pcs
```

Per specificare più operazioni in una sola istruzione, occorre separarle con la virgola.

```
"Action": [
      "pcs:action1",
      "pcs:action2"
         ]
```

Risorse politiche per AWS PCS

Supporta le risorse relative alle policy: sì

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale entità principale può eseguire operazioni su quali risorse e in quali condizioni.

L'elemento JSON `Resource` della policy specifica l'oggetto o gli oggetti ai quali si applica l'operazione. Come best practice, specifica una risorsa utilizzando il suo nome della risorsa Amazon (ARN). Per le azioni che non supportano le autorizzazioni a livello di risorsa, si utilizza un carattere jolly (*) per indicare che l'istruzione si applica a tutte le risorse.

```
"Resource": "*"
```

Per visualizzare un elenco dei tipi di risorse AWS PCS e relativi ARNs, vedere Resources Defined by AWS Parallel Computing Service nel Service Authorization Reference. Per sapere con quali azioni è possibile specificare l'ARN di ogni risorsa, vedere Azioni definite dal servizio AWS Parallel Computing.

Per visualizzare esempi di politiche basate sull'identità AWS PCS, vedere. Esempi di policy basate sull'identità per Parallel Computing Service AWS

Chiavi relative alle condizioni delle policy per PCS AWS

Supporta le chiavi di condizione delle policy specifiche del servizio: sì

Gli amministratori possono utilizzare le policy AWS JSON per specificare chi ha accesso a cosa. In altre parole, quale entità principale può eseguire operazioni su quali risorse e in quali condizioni.

L'elemento `Condition` specifica quando le istruzioni vengono eseguite in base a criteri definiti. È possibile compilare espressioni condizionali che utilizzano operatori di condizione, ad esempio uguale a o minore di, per soddisfare la condizione nella policy con i valori nella richiesta. Per visualizzare tutte le chiavi di condizione AWS globali, consulta le chiavi di contesto delle condizioni AWS globali nella Guida per l'utente IAM.

Per visualizzare un elenco delle chiavi di condizione AWS PCS, consulta Condition Keys for AWS Parallel Computing Service nel Service Authorization Reference. Per sapere con quali azioni e risorse è possibile utilizzare una chiave di condizione, consulta Actions Defined by AWS Parallel Computing Service.

Per visualizzare esempi di politiche basate sull'identità AWS PCS, vedere. Esempi di policy basate sull'identità per Parallel Computing Service AWS

ACLs AWS in PCS

Supporti ACLs: no

Le liste di controllo degli accessi (ACLs) controllano quali principali (membri dell'account, utenti o ruoli) dispongono delle autorizzazioni per accedere a una risorsa. ACLs sono simili alle politiche basate sulle risorse, sebbene non utilizzino il formato del documento di policy JSON.

ABAC con PCS AWS

Supporta ABAC (tag nelle policy): sì

Il controllo degli accessi basato su attributi (ABAC) è una strategia di autorizzazione che definisce le autorizzazioni in base ad attributi chiamati tag. Puoi allegare tag a entità e AWS risorse IAM, quindi

progettare politiche ABAC per consentire operazioni quando il tag del principale corrisponde al tag sulla risorsa.

Per controllare l'accesso basato su tag, fornire informazioni sui tag nell'elemento condizione di una policy utilizzando le chiavi di condizione `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` o `aws:TagKeys`.

Se un servizio supporta tutte e tre le chiavi di condizione per ogni tipo di risorsa, il valore per il servizio è Sì. Se un servizio supporta tutte e tre le chiavi di condizione solo per alcuni tipi di risorsa, allora il valore sarà Parziale.

Per maggiori informazioni su ABAC, consulta Definizione delle autorizzazioni con autorizzazione ABAC nella Guida per l'utente di IAM. Per visualizzare un tutorial con i passaggi per l'impostazione di ABAC, consulta Utilizzo del controllo degli accessi basato su attributi (ABAC) nella Guida per l'utente di IAM.

Utilizzo di credenziali temporanee con PCS AWS

Supporta le credenziali temporanee: sì

Le credenziali temporanee forniscono un accesso a breve termine alle AWS risorse e vengono create automaticamente quando si utilizza la federazione o si cambia ruolo. AWS consiglia di generare dinamicamente credenziali temporanee anziché utilizzare chiavi di accesso a lungo termine. Per ulteriori informazioni, consulta Credenziali di sicurezza temporanee in IAM e Servizi AWS compatibili con IAM nella Guida per l'utente IAM.

Autorizzazioni principali multiservizio per PCS AWS

Supporta l'inoltro delle sessioni di accesso (FAS): sì

Le sessioni di accesso inoltrato (FAS) utilizzano le autorizzazioni del principale chiamante an Servizio AWS, combinate con la richiesta di effettuare richieste Servizio AWS ai servizi downstream. Per i dettagli delle policy relative alle richieste FAS, consulta Forward access sessions.

Ruoli di servizio per PCS AWS

Supporta i ruoli di servizio: no

Un ruolo di servizio è un ruolo IAM che un servizio assume per eseguire operazioni per tuo conto. Un amministratore IAM può creare, modificare ed eliminare un ruolo di servizio dall'interno di IAM. Per

ulteriori informazioni, consulta Create a role to delegate permissions to an Servizio AWS nella Guida per l'utente IAM.

> **Warning**
>
> La modifica delle autorizzazioni per un ruolo di servizio potrebbe interrompere la funzionalità AWS PCS. Modifica i ruoli di servizio solo quando AWS PCS fornisce indicazioni in tal senso.

Ruoli collegati ai servizi per PCS AWS

Supporta i ruoli collegati ai servizi: sì

Un ruolo collegato al servizio è un tipo di ruolo di servizio collegato a un. Servizio AWS Il servizio può assumere il ruolo per eseguire un'operazione per tuo conto. I ruoli collegati al servizio vengono visualizzati nel tuo account Account AWS e sono di proprietà del servizio. Un amministratore IAM può visualizzare le autorizzazioni per i ruoli collegati al servizio, ma non modificarle.

Per informazioni dettagliate sulla creazione o la gestione di ruoli collegati ai servizi AWS PCS, consulta. Ruoli collegati ai servizi per PCS AWS

Esempi di policy basate sull'identità per Parallel Computing Service AWS

Per impostazione predefinita, gli utenti e i ruoli non dispongono dell'autorizzazione per creare o modificare AWS risorse PCS. Per concedere agli utenti l'autorizzazione a eseguire azioni sulle risorse di cui hanno bisogno, un amministratore IAM può creare policy IAM.

Per informazioni su come creare una policy basata su identità IAM utilizzando questi documenti di policy JSON di esempio, consulta Creazione di policy IAM (console) nella Guida per l'utente di IAM.

Per informazioni dettagliate sulle azioni e sui tipi di risorse definiti da AWS PCS, incluso il formato di ARNs per ciascun tipo di risorsa, vedere Actions, Resources and Condition Keys for AWS Parallel Computing Service nel Service Authorization Reference.

Argomenti

- Best practice per le policy
- Utilizzo della console PCS AWS
- Consentire agli utenti di visualizzare le loro autorizzazioni

Best practice per le policy

Le politiche basate sull'identità determinano se qualcuno può creare, accedere o eliminare le risorse AWS PCS nel tuo account. Queste azioni possono comportare costi aggiuntivi per l' Account AWS. Quando si creano o modificano policy basate sull'identità, seguire queste linee guida e raccomandazioni:

- Inizia con le policy AWS gestite e passa alle autorizzazioni con privilegi minimi: per iniziare a concedere autorizzazioni a utenti e carichi di lavoro, utilizza le politiche gestite che concedono le autorizzazioni per molti casi d'uso comuni.AWS Sono disponibili nel tuo. Account AWS Ti consigliamo di ridurre ulteriormente le autorizzazioni definendo politiche gestite dai AWS clienti specifiche per i tuoi casi d'uso. Per maggiori informazioni, consulta Policy gestite da AWS o Policy gestite da AWS per le funzioni dei processi nella Guida per l'utente di IAM.
- Applicazione delle autorizzazioni con privilegio minimo - Quando si impostano le autorizzazioni con le policy IAM, concedere solo le autorizzazioni richieste per eseguire un'attività. È possibile farlo definendo le azioni che possono essere intraprese su risorse specifiche in condizioni specifiche, note anche come autorizzazioni con privilegio minimo. Per maggiori informazioni sull'utilizzo di IAM per applicare le autorizzazioni, consulta Policy e autorizzazioni in IAM nella Guida per l'utente di IAM.
- Condizioni d'uso nelle policy IAM per limitare ulteriormente l'accesso - Per limitare l'accesso ad azioni e risorse è possibile aggiungere una condizione alle policy. Ad esempio, è possibile scrivere una condizione di policy per specificare che tutte le richieste devono essere inviate utilizzando SSL. Puoi anche utilizzare le condizioni per concedere l'accesso alle azioni del servizio se vengono utilizzate tramite uno specifico Servizio AWS, ad esempio CloudFormation. Per maggiori informazioni, consultare la sezione Elementi delle policy JSON di IAM: condizione nella Guida per l'utente di IAM.
- Utilizzo dello strumento di analisi degli accessi IAM per convalidare le policy IAM e garantire autorizzazioni sicure e funzionali - Lo strumento di analisi degli accessi IAM convalida le policy nuove ed esistenti in modo che aderiscano al linguaggio (JSON) della policy IAM e alle best practice di IAM. Lo strumento di analisi degli accessi IAM offre oltre 100 controlli delle policy e consigli utili per creare policy sicure e funzionali. Per maggiori informazioni, consultare Convalida delle policy per il Sistema di analisi degli accessi IAM nella Guida per l'utente di IAM.
- Richiedi l'autenticazione a più fattori (MFA): se hai uno scenario che richiede utenti IAM o un utente root nel Account AWS tuo, attiva l'MFA per una maggiore sicurezza. Per richiedere la MFA quando vengono chiamate le operazioni API, aggiungere le condizioni MFA alle policy. Per maggiori informazioni, consultare Protezione dell'accesso API con MFA nella Guida per l'utente di IAM.

Per maggiori informazioni sulle best practice in IAM, consulta Best practice di sicurezza in IAM nella Guida per l'utente di IAM.

Utilizzo della console PCS AWS

Per accedere alla console di AWS Parallel Computing Service, è necessario disporre di un set minimo di autorizzazioni. Queste autorizzazioni devono consentirti di elencare e visualizzare i dettagli sulle risorse AWS PCS presenti nel tuo. Account AWS Se crei una policy basata sull'identità più restrittiva rispetto alle autorizzazioni minime richieste, la console non funzionerà nel modo previsto per le entità (utenti o ruoli) associate a tale policy.

Non è necessario concedere autorizzazioni minime per la console agli utenti che effettuano chiamate solo verso AWS CLI o l' AWS API. Al contrario, è opportuno concedere l'accesso solo alle azioni che corrispondono all'operazione API che stanno cercando di eseguire.

Per ulteriori informazioni sulle autorizzazioni minime richieste per utilizzare la console AWS PCS, consulta. Autorizzazioni minime per AWS PCS

Consentire agli utenti di visualizzare le loro autorizzazioni

Questo esempio mostra in che modo è possibile creare una policy che consente agli utenti IAM di visualizzare le policy inline e gestite che sono collegate alla relativa identità utente. Questa politica include le autorizzazioni per completare questa azione sulla console o utilizzando l'API o a livello di codice. AWS CLI 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 politiche gestite per AWS Parallel Computing Service

Una politica AWS gestita è una politica autonoma creata e amministrata da AWS. AWS le politiche gestite sono progettate per fornire autorizzazioni per molti casi d'uso comuni, in modo da poter iniziare ad assegnare autorizzazioni a utenti, gruppi e ruoli.

Tieni presente che le policy AWS gestite potrebbero non concedere le autorizzazioni con il privilegio minimo per i tuoi casi d'uso specifici, poiché sono disponibili per tutti i clienti. AWS Si consiglia pertanto di ridurre ulteriormente le autorizzazioni definendo policy gestite dal cliente specifiche per i propri casi d'uso.

Non è possibile modificare le autorizzazioni definite nelle politiche gestite. AWS Se AWS aggiorna le autorizzazioni definite in una politica AWS gestita, l'aggiornamento ha effetto su tutte le identità principali (utenti, gruppi e ruoli) a cui è associata la politica. AWS è più probabile che aggiorni una policy AWS gestita quando ne Servizio AWS viene lanciata una nuova o quando diventano disponibili nuove operazioni API per i servizi esistenti.

Per ulteriori informazioni, consultare Policy gestite da AWS nella Guida per l'utente di IAM.

AWS politica gestita: AWSPCSCompute NodePolicy

Puoi collegarti AWSPCSCompute NodePolicy alle tue entità IAM. Puoi collegare questa policy a un ruolo IAM del nodo di calcolo AWS PCS da te specificato per consentire ai nodi che utilizzano quel ruolo di connettersi a un cluster AWS PCS.

AWS PCS associa questa politica a un ruolo di gruppo di nodi di calcolo quando si utilizza la console per creare un gruppo di nodi di calcolo.

Dettagli delle autorizzazioni

Questa policy include le seguenti autorizzazioni:

- `pcs:RegisterComputeNodeGroupInstance`— Consenti a un nodo di calcolo AWS PCS (istanza EC2) di registrarsi in un cluster PCS. AWS

Per vedere le autorizzazioni per questa policy, consulta AWSPCSComputeNodePolicy nella Guida di riferimento sulle policy gestite da AWS .

AWS politica gestita: AWSPCSService RolePolicy

Non puoi collegarti AWSPCSService RolePolicy alle tue entità IAM. Questa policy è associata a un ruolo collegato al servizio che consente a AWS PCS di eseguire azioni per tuo conto. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per PCS AWS.

Dettagli delle autorizzazioni

Questa policy include le seguenti autorizzazioni:

- `ec2`— Consente a AWS PCS di creare e gestire risorse Amazon EC2.
- `iam`— Consente a AWS PCS di creare un ruolo collegato ai servizi per la flotta Amazon EC2 e di passare il ruolo ad Amazon EC2.
- `cloudwatch`— Consente a AWS PCS di pubblicare i parametri del servizio su Amazon CloudWatch.
- `secretsmanager`— Consente a AWS PCS di gestire i segreti per le risorse del cluster AWS PCS.

Per vedere le autorizzazioni per questa policy, consulta AWSPCSServiceRolePolicy nella Guida di riferimento sulle policy gestite da AWS .

AWS Aggiornamenti PCS alle politiche AWS gestite

Visualizza i dettagli sugli aggiornamenti delle politiche AWS gestite per AWS PCS da quando questo servizio ha iniziato a tenere traccia di queste modifiche. Per avvisi automatici sulle modifiche a questa pagina, iscriviti al feed RSS nella pagina della cronologia dei documenti AWS PCS.

Modifica	Descrizione	Data
AWSPCSServiceRolePolicy: aggiornamento a una policy esistente	AWS PCS ha aggiunto nuove autorizzazioni per supportar e Capacity Blocks per una capacità di calcolo prevedibile. È stata aggiunta `ec2:DescribeCapacityReservations` l'autorizzazione per consentire a AWS PCS di rilevare e utilizzare le	11 settembre 2025

Modifica	Descrizione	Data
	prenotazioni di Capacity Block per i gruppi di nodi di calcolo.	
AWSPCSComputeNodePolicy: nuova policy	AWS PCS ha aggiunto una nuova politica per concedere l'autorizzazione ai nodi di calcolo AWS PCS per la connessione ai cluster AWS PCS. AWS PCS associa questa policy a un ruolo IAM quando si crea un gruppo di nodi di calcolo nella console PCS. AWS	23 giugno 2025
È stato aggiornato il codice JSON in questo documento	È stato corretto il codice JSON in questo documento per includerlo. `"arn:aws:ec2:*:*:spot-instances-request/*"`	5 settembre 2024
AWS PCS ha iniziato a tracciare le modifiche	AWS PCS ha iniziato a tenere traccia delle modifiche per le sue politiche AWS gestite.	28 agosto 2024

Ruoli collegati ai servizi per PCS AWS

AWS Parallel Computing Service utilizza ruoli AWS Identity and Access Management collegati ai servizi (IAM). Un ruolo collegato ai servizi è un tipo unico di ruolo IAM collegato direttamente al PCS. AWS I ruoli collegati ai servizi sono predefiniti da AWS PCS e includono tutte le autorizzazioni richieste dal servizio per chiamare altri servizi per tuo conto. AWS

Un ruolo collegato al servizio semplifica la configurazione di AWS PCS perché non è necessario aggiungere manualmente le autorizzazioni necessarie. AWS PCS definisce le autorizzazioni dei suoi ruoli collegati ai servizi e, se non diversamente definito, solo AWS PCS può assumerne i ruoli. Le

autorizzazioni definite includono la policy di attendibilità e la policy delle autorizzazioni. Una policy delle autorizzazioni specifica non può essere collegata a un'altra entità IAM.

È possibile eliminare un ruolo collegato ai servizi solo dopo aver eliminato le risorse correlate. Questo protegge le tue risorse AWS PCS perché non puoi rimuovere accidentalmente l'autorizzazione ad accedere alle risorse.

Per informazioni su altri servizi che supportano i ruoli collegati ai servizi, consulta AWS i servizi che funzionano con IAM e cerca i servizi con Sì nella colonna Ruoli collegati ai servizi. Scegli Sì in corrispondenza di un link per visualizzare la documentazione relativa al ruolo collegato al servizio per tale servizio.

Autorizzazioni di ruolo collegate ai servizi per PC AWS

AWS PCS utilizza il ruolo collegato al servizio denominato AWSServiceRoleForPCS: concede l'autorizzazione a AWS PCS per gestire le risorse Amazon EC2.

Il ruolo AWSService RoleFor PCS collegato al servizio prevede che i seguenti servizi assumano il ruolo:

- `pcs.amazonaws.com`

La politica di autorizzazione dei ruoli denominata AWSPCSServiceRolePolicyconsente a AWS PCS di completare azioni su risorse specifiche.

Per consentire a utenti, gruppi o ruoli di creare, modificare o eliminare un ruolo orientato ai servizi, devi configurare le autorizzazioni. Per ulteriori informazioni, consulta Autorizzazioni del ruolo collegato ai servizi nella Guida per l'utente IAM.

Creazione di un ruolo collegato al servizio per PCS AWS

Non è necessario creare manualmente un ruolo collegato al servizio. AWS PCS crea automaticamente un ruolo collegato al servizio quando crei un cluster.

Modifica di un ruolo collegato al servizio per PCS AWS

AWS PCS non consente di modificare il ruolo collegato al servizio AWSService RoleFor PCS. Dopo avere creato un ruolo collegato al servizio, non sarà possibile modificarne il nome perché varie entità potrebbero farvi riferimento. È possibile tuttavia modificarne la descrizione utilizzando IAM. Per ulteriori informazioni, consulta Modifica di un ruolo collegato al servizio nella Guida per l'utente di IAM.

Eliminazione di un ruolo collegato al servizio per PCS AWS

Se non è più necessario utilizzare una funzionalità o un servizio che richiede un ruolo collegato al servizio, ti consigliamo di eliminare il ruolo. In questo modo non sarà più presente un'entità non utilizzata che non viene monitorata e gestita attivamente. Tuttavia, è necessario effettuare la pulizia delle risorse associate al ruolo collegato al servizio prima di poterlo eliminare manualmente.

> **Note**
>
> Se il servizio AWS PCS utilizza il ruolo quando si tenta di eliminare le risorse, l'eliminazione potrebbe non riuscire. In questo caso, attendi alcuni minuti e quindi ripeti l'operazione.

Per rimuovere le risorse AWS PCS utilizzate dal AWSService RoleFor PCS

È necessario eliminare tutti i cluster per eliminare il ruolo collegato al servizio AWSService RoleFor PCS. Per ulteriori informazioni, consulta Eliminare un cluster.

Per eliminare manualmente il ruolo collegato ai servizi mediante IAM

Utilizza la console IAM AWS CLI, o l' AWS API per eliminare il ruolo collegato al servizio AWSService RoleFor PCS. Per ulteriori informazioni, consulta Eliminazione del ruolo collegato ai servizi nella Guida per l'utente IAM.

Regioni supportate per i ruoli collegati ai servizi AWS PCS

AWS PCS supporta l'utilizzo di ruoli collegati al servizio in tutte le regioni in cui il servizio è disponibile. Per maggiori informazioni, consulta Regioni ed endpoint di AWS.

Ruolo Spot di Amazon EC2 per PCS AWS

Se desideri creare un gruppo di nodi di elaborazione AWS PCS che utilizzi Spot come opzione di acquisto, devi avere anche il ruolo collegato al servizio AWSServiceRoleForEC2Spot. Account AWSÈ possibile utilizzare il seguente AWS CLI comando per creare il ruolo. Per ulteriori informazioni, consulta Creare un ruolo collegato al servizio e Creare un ruolo per delegare le autorizzazioni a un AWS servizio nella Guida per l'utente.AWS Identity and Access Management

```
aws iam create-service-linked-role --aws-service-name spot.amazonaws.com
```

> **Note**
>
> Riceverai il seguente errore se disponi Account AWS già di un ruolo IAM. `AWSServiceRoleForEC2Spot`
>
> ```
> An error occurred (InvalidInput) when calling the CreateServiceLinkedRole
> operation: Service role name AWSServiceRoleForEC2Spot has been taken in this
> account, please try a different suffix.
> ```

Autorizzazioni minime per AWS PCS

Questa sezione descrive le autorizzazioni IAM minime richieste per un'identità IAM (utente, gruppo o ruolo) per utilizzare il servizio.

Indice

Autorizzazioni minime per utilizzare le azioni API

Azione API	Autorizzazioni minime	Autorizzazioni aggiuntive per la console
CreateCluster	`ec2:CreateNetworkInterface, ec2:DescribeVpcs, ec2:DescribeSubnets, ec2:DescribeSecurityGroups, ec2:GetSecurityGroupsForVpc,`	

Azione API	Autorizzazioni minime	Autorizzazioni aggiuntive per la console
	`iam:CreateServiceLinkedRole`, `secretsmanager:CreateSecret`, `secretsmanager:TagResource`, `secretsmanager:RotateSecret`, `pcs:CreateCluster`	
ListClusters	`pcs:ListClusters`	
GetCluster	`pcs:GetCluster`	`ec2:DescribeSubnets`
DeleteCluster	`pcs:DeleteCluster`	
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`

Azione API	Autorizzazioni minime	Autorizzazioni aggiuntive per la console
ListComputerNodeGroups	`pcs:ListComputeNodeGroups`	`pcs:GetCluster`
GetComputeNodeGroup	`pcs:GetComputeNodeGroup`	`ec2:DescribeSubnets`
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`

Azione API	Autorizzazioni minime	Autorizzazioni aggiuntive per la console
ListQueues	`pcs:ListQueues`	`pcs:GetCluster`
GetQueue	`pcs:GetQueue`	
UpdateQueue	`pcs:UpdateQueue`	`pcs:ListComputeNodeGroups, pcs:GetQueue`
DeleteQueue	`pcs:DeleteQueue`	

Autorizzazioni minime per l'utilizzo dei tag

Le seguenti autorizzazioni sono necessarie per utilizzare i tag con le risorse in AWS PCS.

```
pcs:ListTagsForResource,
pcs:TagResource,
pcs:UntagResource
```

Autorizzazioni minime per supportare i log

AWS PCS invia i dati di registro ad Amazon CloudWatch Logs (CloudWatch Logs). Devi assicurarti che la tua identità disponga delle autorizzazioni minime per utilizzare Logs. CloudWatch Per ulteriori informazioni, consulta Panoramica della gestione delle autorizzazioni di accesso alle risorse CloudWatch Logs nella Amazon CloudWatch Logs User Guide.

Per informazioni sulle autorizzazioni richieste a un servizio per inviare log a CloudWatch Logs, consulta Enabling logging from services AWS nella Amazon CloudWatch Logs User Guide.

Autorizzazioni minime per utilizzare Capacity Blocks

Amazon EC2 Capacity Blocks for ML è un'opzione di acquisto di Amazon EC2 che consente di pagare in anticipo per prenotare istanze di elaborazione accelerata basate su GPU entro un intervallo

di data e ora specifico per supportare carichi di lavoro di breve durata. Per ulteriori informazioni, consulta Utilizzo dei blocchi di capacità di Amazon EC2 per ML con PCS AWS.

Scegli di utilizzare Capacity Blocks quando crei o aggiorni un gruppo di nodi di calcolo. L'identità IAM che usi per creare o aggiornare il gruppo di nodi di calcolo deve avere le seguenti autorizzazioni:

```
ec2:DescribeCapacityReservations
```

Autorizzazioni minime per un amministratore del servizio

La seguente policy IAM specifica le autorizzazioni minime richieste per un'identità IAM (utente, gruppo o ruolo) per configurare e gestire il AWS servizio PCS.

> **Note**
>
> Gli utenti che non configurano e gestiscono il servizio non richiedono queste autorizzazioni. Gli utenti che eseguono solo processi utilizzano Secure Shell (SSH) per connettersi al cluster. AWS Identity and Access Management (IAM) non gestisce l'autenticazione o l'autorizzazione per 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": "*"
    }
  ]
}
```

Profili di istanza IAM per AWS Parallel Computing Service

Le applicazioni eseguite su un'istanza EC2 devono includere AWS le credenziali in tutte le richieste AWS API effettuate. Ti consigliamo di utilizzare un ruolo IAM per gestire le credenziali temporanee sull'istanza EC2. A tale scopo, puoi definire un profilo di istanza e collegarlo alle tue istanze. Per ulteriori informazioni, consulta i ruoli IAM per Amazon EC2 nella Amazon Elastic Compute Cloud User Guide.

> **Note**
>
> Quando si utilizza per Console di gestione AWS creare un ruolo IAM per Amazon EC2, la console crea automaticamente un profilo di istanza e gli assegna lo stesso nome del ruolo IAM. Se utilizzi le AWS CLI azioni AWS API o un AWS SDK per creare il ruolo IAM, crei il profilo dell'istanza come azione separata. Per ulteriori informazioni, consulta Profili di istanza nella Guida per l'utente di Amazon Elastic Compute Cloud.

È necessario specificare l'Amazon Resource Name (ARN) di un profilo di istanza quando si creano gruppi di nodi di calcolo. Puoi scegliere diversi profili di istanza per alcuni o tutti i gruppi di nodi di calcolo.

Requisiti

Ruolo IAM del profilo dell'istanza

Il ruolo IAM associato al profilo dell'istanza deve essere presente `/aws-pcs/` nel percorso, oppure il nome deve iniziare con`AWSPCS`.

Esempio di ruolo IAM ARNs

- `arn:aws:iam::*:role/AWSPCS-example-role-1`

- `arn:aws:iam::*:role/aws-pcs/example-role-2`

Permissions

Il ruolo IAM associato al profilo di istanza per AWS PCS deve includere la seguente policy.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

Politiche aggiuntive

Prendi in considerazione l'aggiunta di politiche gestite al profilo dell'istanza. Esempio:

- AmazonS3 ReadOnlyAccess fornisce accesso in sola lettura a tutti i bucket S3.
- Amazon SSMManaged InstanceCore abilita le funzionalità principali del servizio AWS Systems Manager, come l'accesso remoto direttamente dalla Console di gestione Amazon.
- CloudWatchAgentServerPolicycontiene le autorizzazioni necessarie per l'uso AmazonCloudWatchAgent sui server.

Puoi anche includere le tue policy IAM che supportano il tuo caso d'uso specifico.

Crea un profilo di istanza per AWS PCS

AWS PCS console

Seleziona Crea un profilo di base quando crei un gruppo di nodi di calcolo per fare in modo che AWS PCS ne crei uno per te con la policy minima richiesta.

Amazon EC2 console

Puoi creare un profilo di istanza direttamente dalla console Amazon EC2. Per ulteriori informazioni, consulta Usare i profili di istanza nella Guida per l'AWS Identity and Access Management utente.

> **Important**
>
> Assicurati di utilizzare il prefisso richiesto `AWSPCS` nel nome del ruolo IAM.

AWS CLI

Configurazione del profilo di istanza Basic tramite AWS CLI

> **Note**
>
> Sostituiscilo *`example-role`* negli esempi seguenti con il nome del tuo ruolo IAM.

1. Crea il ruolo IAM con `/aws-pcs/` come attributo path o con un nome che inizia con`AWSPCS`.

   a. Copia e incolla il seguente contenuto in un nuovo file di testo denominato`trust_policy.json`.

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
             "Principal": {
                "Service": [
                    "ec2.amazonaws.com"
                ]
            },
            "Action": [
                "sts:AssumeRole"
            ]
        }
    ]
```

```
}
```

b. Utilizza 1 dei seguenti comandi per creare il ruolo IAM.

```
aws iam create-role --path /aws-pcs/ --role-name example-role --assume-role-
policy-document file://trust_policy.json
```

or

```
aws iam create-role --role-name AWSPCS-example-role --assume-role-policy-
document file://trust_policy.json
```

2. Allega autorizzazioni.

a. Copia e incolla il seguente contenuto in un nuovo file di testo denominato`policy_document.json`.

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Action": [
                "pcs:RegisterComputeNodeGroupInstance"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

b. Allega il documento di policy al ruolo. Questo comando allega la politica come politica in linea.

```
aws iam put-role-policy \
    --role-name example-role \
    --policy-name pcsRegisterInstancePolicy \
    --policy-document file://policy_document.json
```

3. Crea un profilo di istanza. Sostituiscilo ***example-profile*** con il nome del tuo profilo di istanza.

```
aws iam create-instance-profile --instance-profile-name example-profile
```

4. Associa il ruolo IAM al profilo dell'istanza.

```
aws iam add-role-to-instance-profile \
   --instance-profile-name example-profile \
   --role-name example-role
```

Trova i profili di istanza utilizzati con AWS PCS

1. Se non conosci i nomi esatti dei tuoi ruoli IAM per AWS PCS, utilizza il seguente AWS CLI comando per elencare i ruoli IAM che soddisfano i requisiti relativi ai nomi AWS PCS.

```
aws iam list-roles --query "Roles[?starts_with(RoleName, 'AWSPCS') ||
 contains(Path, '/aws-pcs/')].[RoleName]" --output text
```

2. Utilizza il AWS CLI comando seguente per elencare i profili di istanza associati a uno specifico ruolo IAM. Sostituiscilo *role-name* con il nome di un ruolo IAM che soddisfa i requisiti relativi ai nomi AWS PCS.

```
aws iam list-instance-profiles-for-role --role-name role-name
```

Risoluzione dei problemi di identità e accesso al AWS Parallel Computing Service

Utilizza le seguenti informazioni per aiutarti a diagnosticare e risolvere i problemi più comuni che potresti riscontrare quando lavori con AWS PCS e IAM.

Argomenti

- Non sono autorizzato a eseguire un'azione in AWS PCS
- Non sono autorizzato a eseguire iam: PassRole
- Desidero consentire a persone esterne Account AWS a me di accedere alle mie risorse AWS PCS

Non sono autorizzato a eseguire un'azione in AWS PCS

Se ricevi un errore che indica che non sei autorizzato a eseguire un'operazione, le tue policy devono essere aggiornate per poter eseguire l'operazione.

L'errore di esempio seguente si verifica quando l'utente IAM `mateojackson` prova a utilizzare la console per visualizzare i dettagli relativi a una risorsa *`my-example-widget`* fittizia ma non dispone di autorizzazioni `pcs:`*`GetWidget`* fittizie.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 pcs:GetWidget on resource: my-example-widget
```

In questo caso, la policy per l'utente `mateojackson` deve essere aggiornata per consentire l'accesso alla risorsa *`my-example-widget`* utilizzando l'azione `pcs:`*`GetWidget`*.

Se hai bisogno di aiuto, contatta il tuo AWS amministratore. L'amministratore è la persona che ti ha fornito le credenziali di accesso.

Non sono autorizzato a eseguire iam: PassRole

Se ricevi un messaggio di errore indicante che non sei autorizzato a eseguire l'`iam:PassRole`azione, le tue politiche devono essere aggiornate per consentirti di trasferire un ruolo a AWS PCS.

Alcuni Servizi AWS consentono di trasferire un ruolo esistente a quel servizio anziché creare un nuovo ruolo di servizio o un ruolo collegato al servizio. Per eseguire questa operazione, è necessario disporre delle autorizzazioni per trasmettere il ruolo al servizio.

Il seguente errore di esempio si verifica quando un utente IAM denominato `marymajor` tenta di utilizzare la console per eseguire un'azione in AWS PCS. Tuttavia, l'azione richiede che il servizio disponga delle autorizzazioni concesse da un ruolo di servizio. Mary non dispone delle autorizzazioni per trasmettere il ruolo al servizio.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

In questo caso, le policy di Mary devono essere aggiornate per poter eseguire l'operazione `iam:PassRole`.

Se hai bisogno di aiuto, contatta il tuo AWS amministratore. L'amministratore è la persona che ti ha fornito le credenziali di accesso.

Desidero consentire a persone esterne Account AWS a me di accedere alle mie risorse AWS PCS

È possibile creare un ruolo con il quale utenti in altri account o persone esterne all'organizzazione possono accedere alle tue risorse. È possibile specificare chi è attendibile per l'assunzione del ruolo. Per i servizi che supportano politiche basate sulle risorse o liste di controllo degli accessi (ACLs), puoi utilizzare tali politiche per concedere alle persone l'accesso alle tue risorse.

Per maggiori informazioni, consulta gli argomenti seguenti:

- Per sapere se AWS PCS supporta queste funzionalità, consulta. Come funziona AWS Parallel Computing Service con IAM
- Per scoprire come fornire l'accesso alle tue risorse attraverso Account AWS le risorse di tua proprietà, consulta Fornire l'accesso a un utente IAM in un altro Account AWS di tua proprietà nella IAM User Guide.
- Per scoprire come fornire l'accesso alle tue risorse a terze parti Account AWS, consulta Fornire l'accesso a soggetti Account AWS di proprietà di terze parti nella Guida per l'utente IAM.
- Per informazioni su come fornire l'accesso tramite la federazione delle identità, consulta Fornire l'accesso a utenti autenticati esternamente (federazione delle identità) nella Guida per l'utente IAM.
- Per informazioni sulle differenze di utilizzo tra ruoli e policy basate su risorse per l'accesso multi-account, consulta Accesso a risorse multi-account in IAM nella Guida per l'utente di IAM.

Convalida della conformità per il servizio AWS Parallel Computing

Per sapere se un Servizio AWS programma rientra nell'ambito di specifici programmi di conformità, consulta Servizi AWS la sezione Scope by Compliance Program Servizi AWS e scegli il programma di conformità che ti interessa. Per informazioni generali, consulta Programmi di AWS conformità Programmi di di .

È possibile scaricare report di audit di terze parti utilizzando AWS Artifact. Per ulteriori informazioni, consulta Scaricamento dei report in AWS Artifact .

La vostra responsabilità di conformità durante l'utilizzo Servizi AWS è determinata dalla sensibilità dei dati, dagli obiettivi di conformità dell'azienda e dalle leggi e dai regolamenti applicabili. Per

ulteriori informazioni sulla responsabilità di conformità durante l'utilizzo Servizi AWS, consulta AWS la documentazione sulla sicurezza.

Resilienza nel servizio di elaborazione AWS parallela

L'infrastruttura AWS globale è costruita attorno a Regioni AWS zone di disponibilità. Regioni AWS forniscono più zone di disponibilità fisicamente separate e isolate, collegate con reti a bassa latenza, ad alto throughput e altamente ridondanti. Con le zone di disponibilità è possibile progettare e gestire applicazioni e database che eseguono automaticamente il failover tra zone di disponibilità senza interruzioni. Le zone di disponibilità sono più disponibili, tolleranti ai guasti e scalabili rispetto alle infrastrutture a data center singolo o multiplo tradizionali.

Per ulteriori informazioni sulle zone di disponibilità, vedere Global Regioni AWS Infrastructure.AWS

Servizio di sicurezza dell'infrastruttura nel servizio di elaborazione AWS parallela

In quanto servizio gestito, AWS Parallel Computing Service è protetto dalla sicurezza di rete AWS globale. Per informazioni sui servizi AWS di sicurezza e su come AWS protegge l'infrastruttura, consulta AWS Cloud Security. Per progettare il tuo AWS ambiente utilizzando le migliori pratiche per la sicurezza dell'infrastruttura, vedi Infrastructure Protection in Security Pillar AWS Well-Architected Framework.

Utilizzate chiamate API AWS pubblicate per accedere a AWS PCS attraverso la rete. I client devono supportare quanto segue:

- Transport Layer Security (TLS). È richiesto TLS 1.2 ed è consigliato TLS 1.3.
- Suite di cifratura con Perfect Forward Secrecy (PFS), ad esempio Ephemeral Diffie-Hellman (DHE) o Elliptic Curve Ephemeral Diffie-Hellman (ECDHE). La maggior parte dei sistemi moderni, come Java 7 e versioni successive, supporta tali modalità.

Quando AWS PCS crea un cluster, il servizio avvia il controller Slurm in un account di proprietà del servizio, separato dai nodi di elaborazione dell'account. Per collegare la comunicazione tra il controller e i nodi di elaborazione, AWS PCS crea un'interfaccia di rete elastica (ENI) tra account nel tuo VPC. Il controller Slurm utilizza l'ENI per gestire e comunicare con i nodi di elaborazione tra diversi nodi Account AWS, mantenendo la sicurezza e l'isolamento delle risorse e facilitando al contempo l'HPC e le operazioni efficienti. AI/ML

Analisi e gestione delle vulnerabilità in Parallel Computing Service AWS

La configurazione e i controlli IT sono una responsabilità condivisa tra voi AWS e l'utente. Per ulteriori informazioni, consulta il modello di responsabilitàAWS condivisa. AWS gestisce le attività di sicurezza di base per l'infrastruttura sottostante nell'account di servizio, come l'applicazione di patch al sistema operativo sulle istanze del controller, la configurazione del firewall e il ripristino di emergenza AWS dell'infrastruttura. Queste procedure sono state riviste e certificate dalle terze parti appropriate. Per ulteriori dettagli, consulta Best practice per la sicurezza, l'identità e la conformità.

> **Note**
>
> I controller Slurm non sono disponibili durante l'aggiornamento. I lavori in esecuzione non sono influenzati. I lavori inviati quando il controller del cluster non è disponibile vengono mantenuti finché il controller non è disponibile.

Sei responsabile della sicurezza dell'infrastruttura sottostante nel tuo Account AWS:

- Mantieni il codice, inclusi aggiornamenti e patch di sicurezza.
- Aggiorna e aggiorna il sistema operativo in Amazon Machine Image (AMI) per i tuoi gruppi di nodi di calcolo e aggiorna i tuoi gruppi di nodi di calcolo per utilizzare l'AMI aggiornata.
- Aggiorna lo scheduler per mantenerlo all'interno delle versioni supportate. Aggiorna l'AMI per i tuoi gruppi di nodi di calcolo e aggiorna il tuo gruppo di nodi di calcolo per utilizzare l'AMI aggiornata.
- Autentica e crittografa le comunicazioni tra i client utente e i nodi a cui si connettono.

Per ulteriori informazioni sull'aggiornamento dell'AMI per i gruppi di nodi di calcolo, consultaAmazon Machine Images (AMIs) per AWS PCS.

Prevenzione del problema "confused deputy" tra servizi

Il confused deputy è un problema di sicurezza in cui un'entità che non dispone dell'autorizzazione per eseguire un'azione può costringere un'entità dotata di privilegi maggiori a eseguire l'azione. Nel frattempo AWS, l'impersonificazione tra servizi può portare al confuso problema del vicesceriffo. La rappresentazione tra servizi può verificarsi quando un servizio (il servizio chiamante) effettua una chiamata a un altro servizio (il servizio chiamato). Il servizio chiamante può essere manipolato per

utilizzare le proprie autorizzazioni e agire sulle risorse di un altro cliente, a cui normalmente non avrebbe accesso. Per evitare ciò, AWS fornisce strumenti per poterti a proteggere i tuoi dati per tutti i servizi con entità di servizio a cui è stato concesso l'accesso alle risorse dell'account.

Si consiglia di utilizzare le `aws:SourceArn`chiavi di contesto della condizione `aws:SourceAccount`globale nelle politiche delle risorse per limitare le autorizzazioni che il AWS Parallel Computing Service (AWS PCS) concede a un altro servizio alla risorsa. Utilizzare `aws:SourceArn` se si desidera consentire l'associazione di una sola risorsa all'accesso tra servizi. Utilizzare `aws:SourceAccount` se si desidera consentire l'associazione di qualsiasi risorsa in tale account all'uso tra servizi.

Il modo più efficace per proteggersi dal problema "confused deputy" è quello di utilizzare la chiave di contesto della condizione globale `aws:SourceArn` con l'ARN completo della risorsa. Se non si conosce l'ARN completo della risorsa o si scelgono più risorse, utilizzare la chiave di contesto della condizione globale `aws:SourceArn` con caratteri jolly (*) per le parti sconosciute dell'ARN. Ad esempio, `arn:aws:`*`servicename`*`:*:`*`123456789012`*`:*`.

Se il valore `aws:SourceArn` non contiene l'ID account, ad esempio un ARN di un bucket Amazon S3, è necessario utilizzare entrambe le chiavi di contesto delle condizioni globali per limitare le autorizzazioni.

Il valore di `aws:SourceArn` deve essere un ARN del cluster.

L'esempio seguente mostra come utilizzare le chiavi di contesto `aws:SourceArn` e `aws:SourceAccount` global condition in AWS PCS per prevenire il confuso problema del vice.

```
{
"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"
      }
    }
  }
}
```

Ruolo IAM per le istanze Amazon EC2 fornite come parte di un gruppo di nodi di calcolo

AWS PCS orchestra automaticamente la capacità di Amazon EC2 per ciascuno dei gruppi di nodi di calcolo configurati in un cluster. Quando creano un gruppo di nodi di calcolo, gli utenti devono fornire un profilo di istanza IAM tramite il campo. `iamInstanceProfileArn` Il profilo dell'istanza specifica le autorizzazioni associate alle istanze EC2 fornite. AWS PCS accetta qualsiasi ruolo che abbia `AWSPCS` come prefisso del nome del ruolo o `/aws-pcs/` come parte del percorso del ruolo. L'`iam:PassRole`autorizzazione è richiesta sull'identità IAM (utente o ruolo) che crea o aggiorna un gruppo di nodi di calcolo. Quando un utente richiama le azioni `CreateComputeNodeGroup` o `UpdateComputeNodeGroup` API, AWS PCS verifica se l'utente è autorizzato a eseguire l'`iam:PassRole`azione.

La seguente policy di esempio concede le autorizzazioni per passare solo i ruoli IAM il cui nome inizia con `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"
                    ]
                }
            }
        }
    ]
}
```

Best practice di sicurezza per AWS Parallel Computing Service

Questa sezione descrive le migliori pratiche di sicurezza specifiche di AWS Parallel Computing Service (AWS PCS). Per ulteriori informazioni sulle best practice di sicurezza in AWS, consulta Best practice for Security, Identity and Compliance.

Sicurezza relativa all'AMI

- Non utilizzare AWS PCS sample AMIs per carichi di lavoro di produzione. I campioni non AMIs sono supportati e sono destinati esclusivamente ai test.
- Aggiorna regolarmente il sistema operativo e il software nell'AMI per i tuoi gruppi di nodi di calcolo per mitigare le vulnerabilità.
- Utilizza solo pacchetti AWS PCS ufficiali autenticati scaricati da fonti ufficiali. AWS
- Aggiorna regolarmente i pacchetti AWS PCS nell'AMI per i gruppi di nodi di calcolo e aggiorna i nodi di calcolo per utilizzare l'AMI aggiornata. Valuta la possibilità di automatizzare questo processo per ridurre al minimo le vulnerabilità.

Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Sicurezza di Slurm Workload Manager

- Implementa i controlli di accesso e le restrizioni di rete per proteggere i nodi di controllo e calcolo di Slurm. Consenti solo a utenti e sistemi affidabili di inviare lavori e accedere ai comandi di gestione Slurm.
- Utilizza le funzionalità di sicurezza integrate di Slurm, come l'autenticazione Slurm, per garantire che gli invii di lavori e le comunicazioni siano autenticati.
- Aggiorna le versioni di Slurm per mantenere operazioni fluide e supporto per i cluster.

> **Important**
>
> Qualsiasi cluster che utilizza una versione di Slurm che ha raggiunto la fine del ciclo di vita del supporto (EOSL) viene interrotto immediatamente. Usa il link nella parte superiore delle pagine della guida per l'utente per iscriverti al feed RSS della documentazione AWS PCS per ricevere una notifica quando una versione di Slurm si avvicina a EOSL.

Per ulteriori informazioni, consulta Versioni Slurm in PCS AWS.

- Ruota regolarmente i segreti del cluster per mantenere la conformità alla sicurezza e correggere potenziali compromessi in materia di sicurezza. Ciò è necessario per la conformità a HIPAA e FedRAMP.

  Per ulteriori informazioni, consulta Segreti dei cluster rotanti in AWS PCS.

Monitoraggio e registrazione dei log

- Usa Amazon CloudWatch Logs e AWS CloudTrail per monitorare e registrare le azioni nei tuoi cluster e. Account AWS Usa i dati per la risoluzione dei problemi e il controllo.

Sicurezza di rete

- Implementa i cluster AWS PCS in un VPC separato per isolare l'ambiente HPC dal resto del traffico di rete.
- Utilizza i gruppi di sicurezza e gli elenchi di controllo degli accessi alla rete (ACLs) per controllare il traffico in entrata e in uscita verso le istanze e le sottoreti PCS. AWS
- Usa i AWS PrivateLink nostri endpoint VPC per mantenere il traffico di rete tra i tuoi cluster e altri AWS servizi all'interno della rete. AWS Per ulteriori informazioni, consulta Accesso AWS Parallel Computing Service tramite un'interfaccia endpoint ()AWS PrivateLink.

Registrazione e monitoraggio per AWS PCS

Il monitoraggio è una parte importante per mantenere l'affidabilità, la disponibilità e le prestazioni dei AWS PCS e delle altre risorse AWS. AWS fornisce i seguenti strumenti di monitoraggio per monitorare i AWS PCS, segnalare quando qualcosa non va e intraprendere azioni automatiche quando necessario:

- Amazon CloudWatch monitora AWS le tue risorse e le applicazioni su cui esegui AWS in tempo reale. È possibile raccogliere e tenere traccia dei parametri, creare pannelli di controllo personalizzati e impostare allarmi per inviare una notifica o intraprendere azioni quando un parametro specificato raggiunge una determinata soglia. Ad esempio, puoi tenere CloudWatch traccia dell'utilizzo della CPU o di altri parametri delle tue istanze Amazon EC2 e avviare automaticamente nuove istanze quando necessario. Per ulteriori informazioni, consulta la Amazon CloudWatch User Guide.
- Amazon CloudWatch Logs ti consente di monitorare, archiviare e accedere ai tuoi file di log da istanze Amazon EC2 e altre CloudTrail fonti. CloudWatch I log possono monitorare le informazioni nei file di registro e avvisarti quando vengono raggiunte determinate soglie. Puoi inoltre archiviare i dati del log in storage estremamente durevole. Per ulteriori informazioni, consulta la Amazon CloudWatch Logs User Guide.
- AWS CloudTrailacquisisce le chiamate API e gli eventi correlati effettuati da o per conto del tuo AWS account e invia i file di log a un bucket Amazon S3 da te specificato. Puoi identificare quali utenti e account hanno chiamato AWS, l'indirizzo IP di origine da cui sono state effettuate le chiamate e quando sono avvenute le chiamate. Per ulteriori informazioni, consulta la Guida per l'utente AWS CloudTrail.

Registri di completamento dei lavori in AWS PCS

I log di completamento dei lavori forniscono dettagli chiave sui lavori del AWS Parallel Computing Service (AWS PCS) una volta completati, senza costi aggiuntivi. Puoi utilizzare altri AWS servizi per accedere ed elaborare i tuoi dati di log, come Amazon CloudWatch Logs, Amazon Simple Storage Service (Amazon S3) e Amazon Data Firehose AWS ; PCS registra i metadati sui tuoi lavori, come i seguenti.

- Job ID e nome
- Informazioni su utenti e gruppi

- Stato del lavoro (ad esempio`COMPLETED`,`FAILED`,`CANCELLED`)
- Partizione utilizzata
- Limiti di tempo
- Orari di inizio, fine, invio e tempi di idoneità
- Elenco e conteggio dei nodi
- Numero di processori
- Directory di lavoro
- Utilizzo delle risorse (CPU, memoria)
- Codici di uscita
- Dettagli dei nodi (nomi, istanze IDs, tipi di istanza)

Indice

Prerequisiti

Il principale IAM che gestisce il cluster AWS PCS deve consentire l'`pcs:AllowVendedLogDeliveryForResource`azione.

Il seguente esempio di politica IAM concede le autorizzazioni richieste.

JSON

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

```
            "Effect": "Allow",
            "Action": ["pcs:AllowVendedLogDeliveryForResource"],
            "Resource": [
               "arn:aws:pcs:*::cluster/*"
            ]
         }
      ]
}
```

Imposta i registri di completamento dei lavori

È possibile configurare i registri di completamento dei lavori per il cluster AWS PCS con o. Console di gestione AWS AWS CLI

Console di gestione AWS

Per configurare i registri di completamento dei lavori con la console

1. Aprire la console AWS PCS.
2. Nel pannello di navigazione scegliere Cluster.
3. Scegli il cluster in cui desideri aggiungere i registri di completamento dei lavori.
4. Nella pagina dei dettagli del cluster, scegli la scheda Registri.
5. In Job Completion Logs, scegli Aggiungi per aggiungere fino a 3 destinazioni di consegna dei log tra CloudWatch Logs, Amazon S3 e Firehose.
6. Scegli Aggiorna le consegne dei log.

AWS CLI

Per impostare i registri di completamento dei lavori con il AWS CLI

1. Crea una destinazione di consegna dei log:

```
aws logs put-delivery-destination --region region \
  --name pcs-logs-destination \
  --delivery-destination-configuration \
  destinationResourceArn=resource-arn
```

Sostituire:

- *region*— Il Regione AWS luogo in cui si desidera creare la destinazione, ad esempio `us-east-1`
- *pcs-logs-destination*— Un nome per la destinazione
- *resource-arn*— L'Amazon Resource Name (ARN) di un gruppo di log CloudWatch Logs, un bucket S3 o un flusso di distribuzione Firehose.

Per ulteriori informazioni, PutDeliveryDestinationconsulta Amazon CloudWatch Logs API Reference.

2. Imposta il cluster PCS come fonte di consegna dei log:

```
aws logs put-delivery-source --region region \
  --name cluster-logs-source-name \
  --resource-arn cluster-arn \
  --log-type PCS_JOBCOMP_LOGS
```

Sostituire:

- *region*— Il Regione AWS nome del tuo cluster, ad esempio `us-east-1`
- *cluster-logs-source-name*— Un nome per la fonte
- *cluster-arn*— l'ARN del tuo AWS cluster PCS

Per ulteriori informazioni, PutDeliverySourceconsulta Amazon CloudWatch Logs API Reference.

3. Connect l'origine di consegna alla destinazione di consegna:

```
aws logs create-delivery --region region \
  --delivery-source-name cluster-logs-source \
  --delivery-destination-arn destination-arn
```

Sostituire:

- *region*— I Regione AWS, ad esempio `us-east-1`
- *cluster-logs-source*— Il nome della fonte di consegna
- *destination-arn*— L'ARN della destinazione di consegna

Per ulteriori informazioni, CreateDeliveryconsulta Amazon CloudWatch Logs API Reference.

Come trovare i log di completamento dei lavori

Puoi configurare le destinazioni dei log in CloudWatch Logs e Amazon S3. AWS PCS utilizza i seguenti nomi di percorso e nomi di file strutturati.

CloudWatch Registri

AWS PCS utilizza il seguente formato di nome per il flusso CloudWatch Logs:

```
AWSLogs/PCS/cluster-id/jobcomp.log
```

Ad esempio: `AWSLogs/PCS/pcs_abc123de45/jobcomp.log`

Simple Storage Service (Amazon S3)

AWS PCS utilizza il seguente formato di nome per il percorso S3:

```
AWSLogs/account-id/PCS/region/cluster-id/jobcomp/year/month/day/hour/
```

Ad esempio: `AWSLogs/111122223333/PCS/us-east-1/pcs_abc123de45/jobcomp/2025/06/19/11/`

AWS PCS utilizza il seguente formato di nome per i file di registro:

```
PCS_jobcomp_year-month-day-hour_cluster-id_random-id.log.gz
```

Ad esempio: `PCS_jobcomp_2025-06-19-11_pcs_abc123de45_04be080b.log.gz`

Campi del registro di completamento del Job

AWS PCS scrive i dati del registro di completamento del lavoro come oggetti JSON. Il contenitore JSON `jobcomp` contiene i dettagli del lavoro. La tabella seguente descrive i campi all'interno del `jobcomp` contenitore. Alcuni campi sono presenti solo in circostanze specifiche, ad esempio per lavori di matrice o lavori eterogenei.

Campi del registro di completamento del Job

Name	Valore di esempio	Richiesto	Note
`job_id`	`11`	sì	Sempre presenti con valore
`user`	`"root"`	sì	Sempre presente con valore
`user_id`	`0`	sì	Sempre presente con valore
`group`	`"root"`	sì	Sempre presente con valore
`group_id`	`0`	sì	Sempre presente con valore
`name`	`"wrap"`	sì	Sempre presente con valore
`job_state`	`"COMPLETED"`	sì	Sempre presente con valore
`partition`	`"Hydra-MpiQueue-abcdef01-7"`	sì	Sempre presente con valore
`time_limit`	`"UNLIMITED"`	sì	Sempre presente, ma potrebbe esserlo `"UNLIMITED"`
`start_time`	`"2025-06-19T10:58:57"`	sì	Sempre presente, però potrebbe esserlo `"Unknown"`
`end_time`	`"2025-06-19T10:58:57"`	sì	Sempre presente, però potrebbe esserlo `"Unknown"`
`node_list`	`"Hydra-MpiNG-abcdef01-2345-1"`	sì	Sempre presente con valore
`node_cnt`	`1`	sì	Sempre presente con valore

Name	Valore di esempio	Richiesto	Note
`proc_cnt`	`1`	sì	Sempre presente con valore
`work_dir`	`"/root"`	sì	Sempre presente, ma potrebbe esserlo `"Unknown"`
`reservation_name`	`"weekly_maintenance"`	sì	Sempre presente, ma potrebbe essere una stringa vuota `""`
`tres.cpu`	`1`	sì	Sempre presente con valore
`tres.mem.val`	`600`	sì	Sempre presente con valore
`tres.mem.unit`	`"M"`	sì	Può essere `"M"` o `"bb"`
`tres.node`	`1`	sì	Sempre presente con valore
`tres.billing`	`1`	sì	Sempre presente con valore
`account`	`"finance"`	sì	Sempre presente, ma potrebbe essere una stringa vuota `""`
`qos`	`"normal"`	sì	Sempre presente, ma potrebbe essere una stringa vuota `""`
`wc_key`	`"project_1"`	sì	Sempre presente, ma potrebbe essere una stringa vuota `""`
`cluster`	`"unknown"`	sì	Sempre presente, ma potrebbe esserlo `"unknown"`
`submit_time`	`"2025-06-19T10:55:46"`	sì	Sempre presente, però potrebbe esserlo `"Unknown"`

Name	Valore di esempio	Richiesto	Note
`eligible_time`	`"2025-06-19T10:55:46"`	sì	Sempre presente, però potrebbe esserlo `"Unknown"`
`array_job_id`	`12`	no	Presente solo se il lavoro è un lavoro di matrice
`array_task_id`	`1`	no	Presente solo se il lavoro è un lavoro di matrice
`het_job_id`	`10`	no	Presente solo se il lavoro è eterogeneo
`het_job_offset`	`0`	no	Presente solo se la mansione è eterogenea
`derived_exit_code_status`	`0`	sì	Sempre presente con valore
`derived_exit_code_signal`	`0`	sì	Sempre presente con valore
`exit_code_status`	`0`	sì	Sempre presente con valore
`exit_code_signal`	`0`	sì	Sempre presente con valore
`node_details[0].name`	`"Hydra-MpiNG-abcdef01-2345-1"`	no	Sempre presente, ma `node_details` potrebbe esserlo `"[]"`

Name	Valore di esempio	Richiesto	Note
`node_details[0].instance_id`	`"i-0abcdef01234567a"`	no	Sempre presente, però `node_details` potrebbe esserlo `"[]"`
`node_details[0].instance_type`	`"t4g.micro"`	no	Sempre presente, però `node_details` potrebbe esserlo `"[]"`

Esempi di registri di completamento dei lavori

Gli esempi seguenti mostrano i registri di completamento dei lavori per vari tipi e stati dei lavori:


```
{ "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" } ] } }
```

Registri dell'utilità di pianificazione in PCS AWS

Puoi configurare AWS PCS per inviare dati di registrazione dettagliati dal tuo programma di pianificazione del cluster ad Amazon CloudWatch Logs, Amazon Simple Storage Service (Amazon S3) e Amazon Data Firehose. Questo può aiutare nel monitoraggio e nella risoluzione dei problemi.

Indice

Prerequisiti

Il responsabile IAM che gestisce il cluster AWS PCS deve consentire l'`pcs:AllowVendedLogDeliveryForResource`azione.

Il seguente esempio di politica IAM concede le autorizzazioni richieste.

JSON

```
{
   "Version":"2012-10-17",
   "Statement": [
      {
         "Sid": "PcsAllowVendedLogsDelivery",
         "Effect": "Allow",
         "Action": ["pcs:AllowVendedLogDeliveryForResource"],
         "Resource": [
            "arn:aws:pcs:*::cluster/*"
         ]
      }
   ]
}
```

Configura i registri dello scheduler

È possibile configurare i registri dello scheduler per il cluster AWS PCS con o. Console di gestione AWS AWS CLI

Console di gestione AWS

Per configurare i log dello scheduler con la console

1. Aprire la console AWS PCS.
2. Nel pannello di navigazione scegliere Cluster.
3. Scegli il cluster in cui desideri aggiungere i log dello scheduler.
4. Nella pagina dei dettagli del cluster, scegli la scheda Registri.
5. In Scheduler Logs, scegli Aggiungi per aggiungere fino a 3 destinazioni di consegna dei log tra CloudWatch Logs, Amazon S3 e Firehose.
6. Scegli Aggiorna le consegne dei log.

AWS CLI

Per configurare i registri dello scheduler con AWS CLI

1. Crea una destinazione di consegna dei log:

```
aws logs put-delivery-destination --region region \
  --name pcs-logs-destination \
  --delivery-destination-configuration \
  destinationResourceArn=resource-arn
```

Sostituire:

- *region*— Il Regione AWS luogo in cui si desidera creare la destinazione, ad esempio `us-east-1`
- *pcs-logs-destination*— Un nome per la destinazione
- *resource-arn*— L'Amazon Resource Name (ARN) di un gruppo di log CloudWatch Logs, un bucket S3 o un flusso di distribuzione Firehose.

Per ulteriori informazioni, PutDeliveryDestinationconsulta Amazon CloudWatch Logs API Reference.

2. Imposta il cluster PCS come fonte di consegna dei log:

```
aws logs put-delivery-source --region region \
  --name cluster-logs-source-name \
```

```
  --resource-arn cluster-arn \
  --log-type PCS_SCHEDULER_LOGS
```

Sostituire:

- *region*— Il Regione AWS nome del tuo cluster, ad esempio `us-east-1`
- *cluster-logs-source-name*— Un nome per la fonte
- *cluster-arn*— l'ARN del tuo AWS cluster PCS

Per ulteriori informazioni, PutDeliverySourceconsulta Amazon CloudWatch Logs API Reference.

3. Connect l'origine di consegna alla destinazione di consegna:

```
aws logs create-delivery --region region \
  --delivery-source-name cluster-logs-source \
  --delivery-destination-arn destination-arn
```

Sostituire:

- *region*— I Regione AWS, ad esempio `us-east-1`
- *cluster-logs-source*— Il nome della fonte di consegna
- *destination-arn*— L'ARN della destinazione di consegna

Per ulteriori informazioni, CreateDeliveryconsulta Amazon CloudWatch Logs API Reference.

I percorsi e i nomi dei flussi di log di Scheduler

Il percorso e il nome dei log dello scheduler AWS PCS dipendono dal tipo di destinazione.

- CloudWatch Log
  - Uno stream CloudWatch Logs segue questa convenzione di denominazione.

```
AWSLogs/PCS/${cluster_id}/${log_name}_${scheduler_major_version}.log
```

Example

```
AWSLogs/PCS/abcdef0123/slurmctld_24.05.log
```

- Bucket S3
  - Un percorso di output del bucket S3 segue questa convenzione di denominazione:

```
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.
```

  - Il nome di un oggetto S3 segue questa convenzione:

```
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
```

Esempio di record di log dello scheduler

AWS I log dello scheduler PCS sono strutturati. Includono campi come l'identificatore del cluster, il tipo di scheduler, le versioni principali e di patch, oltre al messaggio di registro emesso dal processo del controller Slurm. Ecco un esempio.

```
{
    "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"
}
```

Servizio di monitoraggio del calcolo AWS parallelo con Amazon CloudWatch

Amazon CloudWatch fornisce il monitoraggio dello stato e delle prestazioni del cluster AWS Parallel Computing Service (AWS PCS) raccogliendo parametri dal cluster a intervalli regolari. Queste metriche vengono mantenute, consentendoti di accedere ai dati storici e ottenere informazioni dettagliate sulle prestazioni del cluster nel tempo.

CloudWatch consente inoltre di monitorare le istanze EC2 lanciate da AWS PCS per soddisfare i requisiti di scalabilità. Sebbene sia possibile controllare i log sulle istanze in esecuzione, le CloudWatch metriche e i dati di registrazione vengono in genere eliminati una volta terminate le istanze. Tuttavia, puoi configurare l' CloudWatch agente sulle istanze utilizzando un modello di avvio EC2 per mantenere le metriche e i log anche dopo la chiusura dell'istanza, abilitando il monitoraggio e l'analisi a lungo termine.

Esplora gli argomenti di questa sezione per saperne di più sul monitoraggio tramite PCS. AWS CloudWatch

Argomenti

- Monitoraggio delle metriche AWS PCS tramite CloudWatch
- Monitoraggio delle istanze AWS PCS tramite Amazon CloudWatch

Monitoraggio delle metriche AWS PCS tramite CloudWatch

Puoi monitorare lo stato del cluster AWS PCS utilizzando Amazon CloudWatch, che raccoglie i dati dal cluster e li trasforma in metriche quasi in tempo reale. Queste statistiche vengono conservate per un periodo di 15 mesi, in modo da poter accedere alle informazioni storiche e avere una prospettiva migliore sulle prestazioni del cluster. Le metriche del cluster vengono inviate a CloudWatch intervalli di 1 minuto. Per ulteriori informazioni su CloudWatch, consulta What Is Amazon CloudWatch? nella Amazon CloudWatch User Guide.

AWS PCS pubblica le seguenti metriche nello spazio dei nomi AWS/PCS in. CloudWatch Hanno un'unica dimensione,. `ClusterId`

Nome	Description	unità
ActualCapacity	IdleCapacity + UtilizedCapacity	Conteggio
CapacityUtilization	UtilizedCapacity / ActualCapacity	Conteggio
DesiredCapacity	ActualCapacity + PendingCapacity	Conteggio
IdleCapacity	Numero di istanze in esecuzione ma non assegnate ai job	Conteggio
UtilizedCapacity	Numero di istanze in esecuzione e assegnate ai job	Conteggio

Monitoraggio delle istanze AWS PCS tramite Amazon CloudWatch

AWS PCS lancia le istanze Amazon EC2 in base alle esigenze per soddisfare i requisiti di scalabilità definiti nei gruppi di nodi di calcolo PCS. Puoi monitorare queste istanze mentre sono in esecuzione utilizzando Amazon CloudWatch. Puoi controllare i log delle istanze in esecuzione accedendovi e utilizzando strumenti interattivi da riga di comando. Tuttavia, per impostazione predefinita, i dati CloudWatch delle metriche vengono conservati solo per un periodo limitato dopo la chiusura di un'istanza e i log delle istanze vengono generalmente eliminati insieme ai volumi EBS che supportano l'istanza. Per conservare le metriche o i dati di registrazione delle istanze avviate da PCS dopo la loro chiusura, puoi configurare l' CloudWatch agente sulle tue istanze con un modello di avvio EC2. Questo argomento fornisce una panoramica del monitoraggio delle istanze in esecuzione e fornisce esempi su come configurare i parametri e i log delle istanze persistenti.

Monitoraggio delle istanze in esecuzione

Ricerca di istanze AWS PCS

Per monitorare le istanze lanciate da PCS, trova le istanze in esecuzione associate a un cluster o a un gruppo di nodi di calcolo. Quindi, nella console EC2 per una determinata istanza, ispeziona le sezioni Stato e allarmi e Monitoraggio. Se l'accesso di accesso è configurato per tali istanze,

puoi connetterti ad esse e controllare i vari file di registro sulle istanze. Per ulteriori informazioni sull'identificazione delle istanze gestite da PCS, vedere. Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS

Abilitazione di metriche dettagliate

Per impostazione predefinita, le metriche delle istanze vengono raccolte a intervalli di 5 minuti. Per raccogliere le metriche a intervalli di un minuto, abilita il CloudWatch monitoraggio dettagliato nel modello di lancio del gruppo di nodi di calcolo. Per ulteriori informazioni, consulta Attiva il monitoraggio dettagliato CloudWatch.

Configurazione di metriche e log persistenti delle istanze

Puoi conservare i parametri e i log delle tue istanze installando e configurando l'agente CloudWatch Amazon su di esse. Si compone di tre passaggi principali:

1. Creare una configurazione CloudWatch dell'agente.
2. Archivia la configurazione dove può essere recuperata dalle istanze PCS.
3. Scrivi un modello di avvio EC2 che installi il software dell' CloudWatch agente, recuperi la configurazione e avvii l'agente utilizzando la CloudWatch configurazione.

Per ulteriori informazioni, consulta Raccogli metriche, log e tracce con l' CloudWatch agente nella Amazon CloudWatch User Guide e. Utilizzo dei modelli di lancio di Amazon EC2 con PCS AWS

Crea una configurazione dell'agente CloudWatch

Prima di distribuire l' CloudWatch agente sulle istanze, è necessario generare un file di configurazione JSON che specifichi le metriche, i log e le tracce da raccogliere. I file di configurazione possono essere creati utilizzando una procedura guidata o manualmente, utilizzando un editor di testo. Il file di configurazione verrà creato manualmente per questa dimostrazione.

Su un computer in cui è installata la CLI AWS, crea un file di CloudWatch configurazione denominato config.json con i contenuti seguenti. Puoi anche utilizzare il seguente URL per scaricare una copia del file.

```
https://aws-hpc-recipes.s3.amazonaws.com/main/recipes/pcs/cloudwatch/assets/config.json
```

Note

- I percorsi di log nel file di esempio sono per Amazon Linux 2. Se le tue istanze utilizzeranno un sistema operativo di base diverso, modifica i percorsi in modo appropriato.
- Per acquisire altri registri, aggiungi altre voci in. `collect_list`
- I valori in `{brackets}` sono variabili basate su modelli. Per l'elenco completo delle variabili supportate, consulta Creare o modificare manualmente il file di configurazione dell' CloudWatch agente nella Amazon CloudWatch User Guide.
- Puoi scegliere di omettere `logs` o `metrics` di non raccogliere questi tipi di informazioni.

```
{
    "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
            }
        }
    }
}
```

Questo file indica all' CloudWatch agente di monitorare diversi file che possono essere utili per diagnosticare errori relativi, ad esempio, al bootstrap, all'autenticazione e all'accesso e ad altri domini di risoluzione dei problemi. Ciò include:

- `/var/log/cloud-init.log`— Output dalla fase iniziale della configurazione dell'istanza
- `/var/log/cloud-init-output.log`— Output dei comandi eseguiti durante la configurazione dell'istanza
- `/var/log/amazon/pcs/bootstrap.log`— Output da operazioni specifiche per PC eseguite durante la configurazione dell'istanza
- `/var/log/slurmd.log`— Output dal demone slurmd del gestore del carico di lavoro Slurm
- `/var/log/messages`— Messaggi di sistema dal kernel, dai servizi di sistema e dalle applicazioni
- `/var/log/secure`— Registri relativi ai tentativi di autenticazione, come SSH, sudo e altri eventi di sicurezza

I file di registro vengono inviati a un gruppo di CloudWatch log denominato. `/PCSLogs/instances` I flussi di registro sono una combinazione dell'ID dell'istanza e del nome di base del file di registro. Il gruppo di log ha un tempo di conservazione di 30 giorni.

Inoltre, il file indica all' CloudWatch agente di raccogliere diverse metriche comuni, aggregandole per ID di istanza.

Memorizza la configurazione

Il file di configurazione dell' CloudWatch agente deve essere archiviato dove possono accedervi le istanze del nodo di calcolo PCS. Esistono due modi comuni per eseguire questa operazione. Puoi caricarlo in un bucket Amazon S3 a cui le tue istanze del gruppo di nodi di calcolo avranno accesso tramite il loro profilo di istanza. In alternativa, puoi archiviarlo come parametro SSM in Amazon Systems Manager Parameter Store.

Carica in un bucket S3

Per archiviare il file in S3, utilizza i comandi CLI di AWS riportati di seguito. Prima di eseguire il comando, effettua queste sostituzioni:

- *`amzn-s3-demo-bucket`*Sostituiscilo con il tuo nome di bucket S3

Innanzitutto, (questo è facoltativo se hai un bucket esistente), crea un bucket per contenere i tuoi file di configurazione.

```
aws s3 mb s3://amzn-s3-demo-bucket
```

Quindi, carica il file nel bucket.

```
aws s3 cp ./config.json s3://amzn-s3-demo-bucket/
```

Archivia come parametro SSM

Per memorizzare il file come parametro SSM, usa il comando che segue. Prima di eseguire il comando, effettuate le seguenti sostituzioni:

- Sostituisci *region-code* con la regione AWS in cui lavori con AWS PCS.
- (Facoltativo) Sostituisci il parametro *AmazonCloudWatch-PCS* con il tuo nome. Tieni presente che se modifichi il prefisso del nome da, `AmazonCloudWatch-` dovrai aggiungere specificamente l'accesso in lettura al parametro SSM nel profilo dell'istanza del gruppo di nodi.

```
aws ssm put-parameter \
   --region region-code \
   --name "AmazonCloudWatch-PCS" \
   --type String \
   --value file://config.json
```

Scrivi un modello di lancio EC2

I dettagli specifici del modello di lancio dipendono dal fatto che il file di configurazione sia archiviato in S3 o SSM.

Usa una configurazione archiviata in S3

Questo script installa CloudWatch l'agente, importa un file di configurazione da un bucket S3 e avvia l'agente con esso. CloudWatch Sostituisci i seguenti valori in questo script con i tuoi dati:

- *amzn-s3-demo-bucket*— Il nome di un bucket S3 da cui il tuo account può leggere
- */config.json*— Percorso relativo alla radice del bucket S3 in cui è archiviata la configurazione

```
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==--
```

Il profilo di istanza IAM per il gruppo di nodi deve avere accesso al bucket. Ecco un esempio di policy IAM per il bucket nello script di dati utente riportato sopra.

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/*"
            ]
        }
    ]
}
```

Tieni inoltre presente che le istanze devono consentire il traffico in uscita verso S3 e gli endpoint. CloudWatch Ciò può essere ottenuto utilizzando gruppi di sicurezza o endpoint VPC, a seconda dell'architettura del cluster.

Utilizza una configurazione archiviata in SSM

Questo script installa CloudWatch l'agente, importa un file di configurazione da un parametro SSM e avvia l' CloudWatch agente con esso. Sostituisci i seguenti valori in questo script con i tuoi dati:

- (Facoltativo) Sostituire il parametro *AmazonCloudWatch-PCS* con il proprio nome.

```
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 policy dell'istanza IAM per il gruppo di nodi deve avere il codice CloudWatchAgentServerPolicyallegato.

Se il nome del parametro non inizia con, `AmazonCloudWatch-` dovrai aggiungere specificamente l'accesso in lettura al parametro SSM nel profilo dell'istanza del gruppo di nodi. Ecco un esempio di policy IAM che illustra questo principio per il prefisso. *DOC-EXAMPLE-PREFIX*

JSON

```
{
  "Version":"2012-10-17",
  "Statement" : [
    {
      "Sid" : "CustomCwSsmMParamReadOnly",
      "Effect" : "Allow",
      "Action" : [
        "ssm:GetParameter"
      ],
      "Resource" : "arn:aws:ssm:*:*:parameter/DOC-EXAMPLE-PREFIX*"
    }
  ]
}
```

Tieni inoltre presente che le istanze devono consentire il traffico in uscita verso l'SSM e gli endpoint. CloudWatch Ciò può essere ottenuto utilizzando gruppi di sicurezza o endpoint VPC, a seconda dell'architettura del cluster.

Registrazione delle chiamate API di AWS Parallel Computing Service utilizzando AWS CloudTrail

AWS PCS è integrato con AWS CloudTrail, un servizio che fornisce una registrazione delle azioni intraprese da un utente, ruolo o AWS servizio in AWS PCS. CloudTrail acquisisce tutte le chiamate API per AWS PCS come eventi. Le chiamate acquisite includono chiamate dalla console AWS PCS e chiamate di codice alle operazioni dell'API AWS PCS. Se crei un trail, puoi abilitare la distribuzione continua di CloudTrail eventi a un bucket Amazon S3, inclusi gli eventi per PCS. AWS Se non configuri un percorso, puoi comunque visualizzare gli eventi più recenti nella CloudTrail console nella cronologia degli eventi. Utilizzando le informazioni raccolte da CloudTrail, è possibile determinare la richiesta effettuata a AWS PCS, l'indirizzo IP da cui è stata effettuata la richiesta, chi ha effettuato la richiesta, quando è stata effettuata e ulteriori dettagli.

Per ulteriori informazioni CloudTrail, consulta la Guida AWS CloudTrail per l'utente.

AWS Informazioni PCS in CloudTrail

CloudTrail è abilitato sul tuo account al Account AWS momento della creazione dell'account. Quando si verifica un'attività in AWS PCS, tale attività viene registrata in un CloudTrail evento insieme ad altri eventi AWS di servizio nella cronologia degli eventi. Puoi visualizzare, cercare e scaricare eventi recenti nel tuo Account AWS. Per ulteriori informazioni, consulta Visualizzazione degli eventi con la cronologia degli CloudTrail eventi.

Per una registrazione continua degli eventi del tuo Account AWS, inclusi gli eventi per AWS PCS, crea un percorso. Un trail consente di CloudTrail inviare file di log a un bucket Amazon S3. Per impostazione predefinita, quando si crea un percorso nella console, questo sarà valido in tutte le Regioni AWS. Il trail registra gli eventi di tutte le regioni della AWS partizione e consegna i file di log al bucket Amazon S3 specificato. Inoltre, puoi configurare altri AWS servizi per analizzare ulteriormente e agire in base ai dati sugli eventi raccolti nei log. CloudTrail Per ulteriori informazioni, consulta gli argomenti seguenti:

- Panoramica della creazione di un percorso
- CloudTrail servizi e integrazioni supportati
- Configurazione delle notifiche Amazon SNS per CloudTrail

- Ricezione di file di CloudTrail registro da più regioni e ricezione di file di CloudTrail registro da più account

Tutte le azioni AWS PCS vengono registrate CloudTrail e documentate nel AWS Parallel Computing Service API Reference. Ad esempio, le chiamate alle `CreateComputeNodeGroup` `DeleteCluster` azioni e generano voci nei file di CloudTrail registro. `UpdateQueue`

Ogni evento o voce di log contiene informazioni sull'utente che ha generato la richiesta. Le informazioni di identità consentono di determinare quanto segue:

- Se la richiesta è stata effettuata con credenziali utente root o AWS Identity and Access Management (IAM).
- Se la richiesta è stata effettuata con le credenziali di sicurezza temporanee per un ruolo o un utente federato.
- Se la richiesta è stata effettuata da un altro AWS servizio.

Per ulteriori informazioni, consulta Elemento CloudTrail userIdentity.

Comprensione delle voci dei file di CloudTrail registro da AWS PCS

Un trail è una configurazione che consente la consegna di eventi come file di registro a un bucket S3 specificato dall'utente. CloudTrail i file di registro contengono una o più voci di registro. Un evento rappresenta una singola richiesta proveniente da qualsiasi fonte e include informazioni sull'azione richiesta, la data e l'ora dell'azione, i parametri della richiesta e così via. CloudTrail i file di registro non sono una traccia ordinata dello stack delle chiamate API pubbliche, quindi non vengono visualizzati in un ordine specifico.

L'esempio seguente mostra una voce di CloudTrail registro per un'`CreateQueue`azione.

```
{
    "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"
}
```

Endpoint e quote di servizio per PCS AWS

Le sezioni seguenti descrivono gli endpoint e le quote di servizio per AWS Parallel Computing Service (PCS).AWS Le quote di servizio, precedentemente denominate limiti, rappresentano il numero massimo di risorse o operazioni di servizio per l'utente. Account AWS

Hai Account AWS delle quote predefinite per ogni servizio. AWS Salvo diversa indicazione, ogni quota si applica a una Regione specifica. Se per alcune quote è possibile richiedere aumenti, altre quote non possono essere modificate.

Per ulteriori informazioni, consulta AWS Service Quotas in Riferimenti generali di AWS .

Indice

Endpoint del servizio

Nome Regione	Regione	Endpoint	Protocollo
Stati Uniti orientali (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
Stati Uniti orientali (Virginia settentrionale)	us-east-1	pcs.us-east-1.amazonaws.com	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
		pcs-fips.us-east-1.amazonaws.com pcs-fips.us-east-1.api.aws pcs.us-east-1.api.aws	
Stati Uniti occidentali (Oregon)	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 Pacifico (Singapore)	ap-southeast-1	pcs.ap-southeast-1.amazonaws.com pcs.ap-southeast-1.api.aws	HTTPS
Asia Pacifico (Sydney)	ap-southeast-2	pcs.ap-southeast-2.amazonaws.com pcs.ap-southeast-2.api.aws	HTTPS
Asia Pacifico (Tokyo)	ap-northeast-1	pcs.ap-northeast-1.amazonaws.com pcs.ap-northeast-1.api.aws	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
Europa (Francoforte)	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
Europe (London)	eu-west-2	pcs.eu-west-2.amazonaws.com pcs.eu-west-2.api.aws	HTTPS
Europa (Stoccolma)	eu-north-1	pcs.eu-north-1.amazonaws.com pcs.eu-north-1.api.aws	HTTPS
AWS GovCloud (Stati Uniti orientali)	us-gov-east-1	pz. us-gov-east-1. amazonaws.com pezzi-fips. us-gov-east-1. amazonaws.com pezzi-fips. us-gov-east-1. api.aws pz. us-gov-east-1. api.aws	HTTPS

Nome Regione	Regione	Endpoint	Protocollo
AWS GovCloud (Stati Uniti occidentali)	us-gov-west-1	pz. us-gov-west-1. amazonaws.com pezzi-fips. us-gov-west-1. amazonaws.com pezzi-fips. us-gov-west-1. api.aws pz. us-gov-west-1. api.aws	HTTPS

Service Quotas

Nome	Impostazione predefinita	Regolabile	Descrizione
Cluster	5	Sì	Il numero massimo di cluster per. Regione AWS

> **Note**
>
> I valori predefiniti sono le quote iniziali impostate da. AWS Questi valori predefiniti sono separati dai valori effettivi delle quote applicate e dalle quote massime possibili del servizio. Per ulteriori informazioni, consulta Terminologia di Service Quotas nella Guida per l'utente di Service Quotas.

Queste quote di servizio sono elencate in AWS Parallel Computing Service (PCS) nel. Console di gestione AWS Per richiedere un aumento della quota per i valori indicati come regolabili, vedere Requesting a Quote Acrease nella Service Quotas User Guide.

> **Important**
>
> Ricordati di controllare l' Regione AWS impostazione corrente in. Console di gestione AWS

Quote interne

Le seguenti quote sono interne e non regolabili.

Nome	Impostazione predefinita	Regolabile	Descrizione
Creazione simultanea di cluster	1	No	Il numero massimo di cluster nello `Creating` stato per. Regione AWS
Gruppi di nodi di calcolo per cluster	10	No	Il numero massimo di gruppi di nodi di calcolo per cluster.
Code per cluster	10	No	Il numero massimo di code per cluster.

Quote pertinenti per altri servizi AWS

AWS PCS utilizza altri AWS servizi. Le quote di servizio per tali servizi influiscono sull'utilizzo di AWS PCS.

Quote di servizio Amazon EC2 che influiscono sui PCS AWS

- Richieste di istanze Spot
- Esecuzione di istanze su richiesta
- Modelli di avvio
- Versioni del modello di avvio
- Richieste API Amazon EC2

Per ulteriori informazioni, consulta le quote dei servizi Amazon EC2 nella Guida per l'utente di Amazon Elastic Compute Cloud.

Risoluzione dei problemi in Parallel Computing AWS Service

I seguenti argomenti forniscono indicazioni per risolvere alcuni problemi che potrebbero verificarsi in AWS PCS.

Un'istanza EC2 in AWS PCS viene terminata e sostituita dopo il riavvio

panoramica dei problemi

Dopo il riavvio di un'istanza EC2 in un gruppo di nodi di calcolo, AWS PCS termina e sostituisce automaticamente l'istanza.

Perché questo accade

AWS PCS non supporta il riavvio delle istanze. Se un'istanza EC2 viene riavviata, AWS PCS considera l'istanza non integra e la sostituisce. Se AWS PCS termina e sostituisce continuamente le istanze, potrebbe essere perché qualcosa riavvia le istanze dopo il loro avvio. Alcuni esempi includono i riavvii automatizzati sull'istanza EC2 (come il riavvio automatico dopo l'applicazione di patch), l'automazione esterna all'istanza EC2 (come un'applicazione per la gestione della rete), un altro AWS servizio (ad esempio) o il riavvio manuale da parte di una persona. AWS Systems Manager

Cosa fare

Puoi controllare i `slurmd` log del sistema operativo per vedere se l'`slurmctld`istanza è stata riavviata. Per ulteriori informazioni, consultare Registri dell'utilità di pianificazione in PCS AWS e Monitoraggio delle istanze AWS PCS tramite Amazon CloudWatch. La seguente voce di `slurmctld` registro di esempio indica che l'istanza è stata riavviata:

Example

```
[2024-09-12T06:42:50.393+00:00] validate_node_specs: Node Login-1 unexpectedly rebooted
 boot_time=1726123354 last response=1726123285
```

Riavvio a causa dell'applicazione di una patch

Spesso è necessario un riavvio dopo l'applicazione delle patch. Non applicare le patch direttamente a un'istanza EC2 che fa parte di un gruppo di nodi di calcolo AWS PCS. Se devi applicare le patch alle tue istanze EC2, devi applicare le patch a un'Amazon Machine Image (AMI) aggiornata e aggiornare i gruppi di nodi di calcolo per utilizzare l'AMI aggiornata. Le nuove istanze EC2 lanciate AWS da PCS per quei gruppi di nodi di calcolo utilizzeranno l'AMI aggiornata (con patch). Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Risolvi i problemi relativi al bootstrap e alla registrazione dei nodi di calcolo in PCS AWS

Quando i nodi di elaborazione non riescono ad avviarsi o a registrarsi correttamente nel cluster AWS PCS, potrebbero verificarsi i seguenti sintomi:

- I lavori non iniziano
- Non puoi connetterti alle istanze in AWS Systems Manager
- Le istanze si sono chiuse in modo imprevisto
- Le istanze vengono sostituite continuamente

Questi errori possono essere causati da problemi durante l'avvio dell'istanza EC2 o durante il processo di bootstrap del nodo di calcolo AWS PCS. Questo argomento descrive le procedure per aiutarti a risolvere i problemi durante il processo di bootstrap del nodo PCS. AWS Per ulteriori informazioni sulla risoluzione dei problemi di avvio delle istanze EC2, consulta Risoluzione dei problemi di avvio delle istanze Amazon EC2 nella Amazon Elastic Compute Cloud User Guide.

Gli errori di bootstrap si verificano quando un'istanza EC2 viene avviata correttamente, ma fallisce durante il processo di adesione al cluster PCS. AWS Il processo di bootstrap include due fasi principali:

- Registrazione del nodo: l'istanza EC2 richiama l'azione dell'API RegisterComputeNodeGroupInstance AWS PCS per registrarsi al servizio AWS PCS. I guasti possono verificarsi a causa di problemi quali:
  - Permissions
    - Profilo di istanza errato
  - Rete
    - Impossibile connettersi agli endpoint AWS PCS
    - Endpoint PCS non configurato correttamente AWS
    - Istanza in una sottorete pubblica senza IP pubblico
    - Istanza multi-NIC in una sottorete pubblica
  - Segreto del cluster
    - Il segreto del cluster è stato eliminato o contrassegnato per l'eliminazione
- Integrazione Slurm: l'istanza viene eseguita `slurmd` e si unisce al cluster Slurm. I guasti possono verificarsi a causa di problemi nei seguenti casi:
  - Permissions
    - Configurazione del gruppo di sicurezza
    - Slurmctld non è in grado di eseguire il ping del nodo di calcolo
  - Configurazione AMI personalizzata
    - Driver NVIDIA mancanti
    - ResumeTimeout raggiunto

Come funziona Slurm su PCS AWS

Potrebbe aiutarti a confrontare il modo standard di funzionamento di Slurm con il modo in cui Slurm funziona su PCS. AWS

Elaborazione standard dei lavori Slurm

Nell'elaborazione standard dei job Slurm si verificano i seguenti passaggi:

1. Quando invii un lavoro, lo `slurmctld` convalida e lo mette in coda.
2. Quando le risorse diventano disponibili, `slurmctld` alloca i nodi esistenti.
3. `slurmd` demoni eseguono i job sui nodi allocati.

Elaborazione dei job Slurm su PCS AWS

Nell'elaborazione dei lavori AWS PCS si verificano i seguenti passaggi:

1. Quando invii un lavoro, `slurmctld` lo convalida e lo mette in coda.
2. Quando è necessaria una capacità aggiuntiva, AWS PCS utilizza il modello di avvio per il gruppo di nodi di calcolo per lanciare nuove istanze EC2.
3. Le nuove istanze vengono avviate nel cluster:
   a. Le istanze vengono registrate con PCS. AWS
   b. Le istanze si uniscono al cluster Slurm.
4. Quando le risorse sono pronte, `slurmctld` alloca i nodi (compresi quelli appena avviati).
5. `slurmd`i demoni eseguono i job sui nodi allocati.

Recupera i log delle istanze

Il primo passo per risolvere i problemi di bootstrap dei nodi di calcolo consiste nel recuperare i log delle istanze. È possibile utilizzare uno dei seguenti metodi:

AWS CLI

Recupera l'output della console dal nodo di calcolo utilizzando il seguente comando:

```
aws ec2 get-console-output --region us-east-1 --instance-id i-1234567890abcdef0 --
output text
```

*us-east-1*Sostituiscilo con la tua AWS regione e *i-1234567890abcdef0* con l'ID dell'istanza.

AWS Systems Manager

Se è possibile connettersi all'istanza utilizzando Systems Manager, è possibile visualizzare direttamente il file di registro di bootstrap:

1. Connect all'istanza utilizzando Systems Manager. Per ulteriori informazioni, vedere Avvio di una sessione nella Guida per l'utente di Systems Manager.
2. Visualizza il file di registro di bootstrap:

```
sudo cat /var/log/amazon/pcs/bootstrap.log
```

> **Note**
>
> Se si verifica un problema durante la fase di inizializzazione, potrebbe essere necessario attendere circa 20 minuti prima di poterti connettere all'istanza. I servizi Systems Manager e SSH si avviano solo dopo il completamento dell'inizializzazione o quando l'esecuzione del bootstrap raggiunge un timeout in caso di errore.

Recupera gruppi da VPC/Subnet/Security un ID di istanza

Per risolvere i problemi con i nodi di elaborazione, potrebbe essere necessario recuperare informazioni sul VPC, sulla sottorete e sui gruppi di sicurezza associati alle istanze. Se non conosci la tua istanza, consulta. IDs Ricerca di istanze di gruppi di nodi di calcolo in PCS AWS

Console di gestione AWS

Per ottenere VPC, sottorete e gruppi di sicurezza

1. Aprire la console di Amazon EC2.
2. Seleziona Instances (Istanze).
3. Nella tabella Istanze, scegli l'ID dell'istanza.
4. Trova l'ID VPC e l'ID di sottorete nel riepilogo dell'istanza visualizzato per l'istanza.
5. Nel riepilogo dell'istanza, scegli la scheda Sicurezza.
6. Trova i gruppi di sicurezza nella scheda Sicurezza.

AWS CLI

Usa il comando seguente per recuperare le informazioni su VPC, sottorete e gruppo di sicurezza per la tua istanza:

```
aws ec2 describe-instances --instance-ids i-1234567890abcdef0 --query
 'Reservations[*].Instances[*].
{InstanceId:InstanceId,VpcId:VpcId,SubnetId:SubnetId,SecurityGroups:SecurityGroups[*].GroupI
 --output table
```

Problemi di registrazione dei nodi

La registrazione dei nodi è la prima azione eseguita da un nodo di calcolo durante il bootstrap. Il nodo chiama l'endpoint dell'API AWS PCS per registrarsi con PCS. AWS Gli errori di registrazione in genere mostrano messaggi di errore simili ai seguenti:

```
<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
```

Profilo di istanza errato

Se l'istanza non è in grado di registrarsi, verifica che il profilo di istanza associato al nodo di calcolo disponga dell'`pcs:RegisterComputeNodeGroupInstance`autorizzazione.

Per ulteriori informazioni su come creare un profilo di istanza valido, consultaCreare un profilo di istanza per AWS PCS.

Impossibile connettersi agli endpoint AWS PCS

Se i nodi di elaborazione si trovano in una sottorete privata, assicurati di aver configurato gli endpoint VPC per AWS PCS o che la sottorete abbia un percorso verso un gateway NAT per l'accesso a Internet. Per ulteriori informazioni, consulta gli argomenti seguenti:

- Accedi a un AWS servizio utilizzando un endpoint VPC di interfaccia nella guida Amazon Virtual Private Cloud. AWS PrivateLink
- Endpoint e quote di servizio per PCS AWS.
- Connetti il tuo VPC ad altre reti nella Guida per l'utente di Amazon Virtual Private Cloud
- AWS Rete PCS

Endpoint PCS non configurato correttamente AWS

Se viene visualizzato un messaggio di errore simile al seguente, verifica la policy associata all'endpoint AWS VPC PCS:

```
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
```

Per ulteriori informazioni su come configurare gli endpoint dell'interfaccia VPC per AWS PCS, vedere. Accesso AWS Parallel Computing Service tramite un'interfaccia endpoint ()AWS PrivateLink

Istanza in una sottorete pubblica senza IP pubblico

Se nella sottorete non è abilitata l'assegnazione automatica dell'IP pubblico e la configurazione del percorso utilizza un gateway Internet, le istanze non possono comunicare con l'API PCS. AWS

Le istanze in una sottorete con un gateway Internet devono avere un indirizzo IP pubblico. Per risolvere questo problema, scegli una delle seguenti opzioni:

- Aggiungi un endpoint VPC per AWS PCS al tuo VPC del cluster. Ciò consente alle istanze di comunicare con AWS PCS senza la necessità che un indirizzo IP pubblico passi attraverso il gateway Internet.
- Utilizza una sottorete privata con un gateway NAT, in modo che non sia richiesto un indirizzo IP pubblico.
- Abilita l'assegnazione automatica degli indirizzi IP pubblici tramite la sottorete o il modello di avvio in modo che le istanze possano contattare l'API tramite il gateway Internet. Tieni presente che questa opzione non è valida per le istanze di interfaccia multi-rete.

Istanza multi-NIC in una sottorete pubblica

È necessario utilizzare una sottorete privata se si utilizza un tipo di istanza con più interfacce di rete (). NICs

AWS gli indirizzi IP pubblici possono essere assegnati solo alle istanze avviate con un'unica interfaccia di rete. Per ulteriori informazioni sugli indirizzi IP, consulta Assegnare un IPv4 indirizzo pubblico durante il lancio dell'istanza nella Amazon EC2 User Guide for Linux Instances.

I tipi di istanze multi-NIC richiedono un gateway NAT o un proxy interno nella sottorete per accedere all'endpoint PCS. AWS In alternativa, puoi aggiungere un endpoint VPC per AWS PCS al tuo VPC del cluster.

Il segreto del cluster è stato eliminato o contrassegnato per l'eliminazione

Se il segreto condiviso di Slurm in AWS Secrets Manager è stato eliminato o contrassegnato per l'eliminazione, i nodi di calcolo non riusciranno a registrarsi e il cluster verrà danneggiato.

AWS PCS crea automaticamente un segreto condiviso Slurm in AWS Secrets Manager (con il formato del nome:`pcs!slurm-secret-<cluster-id>`) quando si crea un cluster. Questo segreto è necessario per comunicazioni sicure nel cluster. Per ulteriori informazioni, consulta Utilizzo dei segreti del cluster in AWS PCS.

Se questo segreto viene eliminato o contrassegnato per l'eliminazione, i nuovi nodi non potranno entrare a far parte del cluster e il controller o altri demoni del cluster (come `slurmd` and`slurmdbd`) potrebbero non essere in grado di ricongiungersi al cluster se riavviato.

Per risolvere questo problema, puoi ripristinare il segreto eliminato se è ancora all'interno della finestra di ripristino. Per istruzioni dettagliate, consulta Restore an AWS Secrets Manager secret.

Se la finestra di ripristino scade, il segreto non può essere ripristinato e il cluster AWS PCS interessato non può essere ripristinato. È necessario creare un nuovo cluster con la stessa configurazione. AWS PCS crea automaticamente un nuovo segreto dello scheduler.

Problemi di unione del cluster Slurm

Dopo una corretta registrazione del nodo, il nodo di calcolo tenta di unirsi al cluster Slurm. Il `slurmd` demone sul nodo contatta il controller Slurm per registrarsi nel cluster. Gli errori di Slurm join di solito mostrano messaggi di errore simili ai seguenti:

```
<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"] ----
```

Configurazione del gruppo di sicurezza

Verifica che i tuoi gruppi di sicurezza siano configurati correttamente per consentire la comunicazione tra i nodi di calcolo e il controller Slurm. I gruppi di sicurezza devono consentire il seguente traffico:

- Porta 6817 con `slurmd` cui comunicare `slurmctld`
- Porta 6818 per eseguire il ping `slurmctld` `slurmd`

Per ulteriori informazioni sui requisiti dei gruppi di sicurezza, consulta i seguenti argomenti:

- Creare gruppi di sicurezza per AWS PCS
- Crea modelli di lancio per AWS PCS
- Requisiti e considerazioni sui gruppi di sicurezza

> **Important**
>
> Il gruppo di sicurezza del cluster associato al cluster durante la creazione del cluster deve essere configurato anche nei gruppi di sicurezza del gruppo di nodi di calcolo per consentire ai nodi di elaborazione di comunicare con il controller.

Driver NVIDIA mancanti

Se l'istanza si avvia correttamente ma i processi non vengono avviati e nei log dell'istanza vengono visualizzati messaggi di errore simili ai seguenti, è possibile che manchino i driver 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!
```

Se ti connetti all'istanza e controlli lo stato del `slurmd` daemon, potresti visualizzare un errore simile al seguente:

```
$ systemctl status slurmd
...
fatal: can't stat gres.conf file /dev/nvidia0: No such file or directory
```

Per risolvere questo problema, installa i driver NVIDIA sulla tua AMI personalizzata. Per ulteriori informazioni, consulta Fase 4 — (Facoltativo) Installare driver, librerie e software applicativi aggiuntivi.

ResumeTimeout raggiunto

Se un nodo di calcolo e la relativa istanza EC2 vengono terminati perché il nodo non è integro, AWS PCS potrebbe non supportare l'AMI o potrebbero esserci problemi di rete. L'istanza EC2 viene eseguita per circa 30 minuti fino a quando non viene raggiunta quella ResumeTimeout di Slurm e contrassegna il nodo come. `DOWN`

Se l'istanza non si avvia correttamente e non è registrata con AWS PCS (nessuna `RegisterComputeNodeGroupInstance` chiamata per l'istanza EC2), controlla i log dell'istanza per verificare la presenza di messaggi di errore simili ai seguenti:

```
/opt/aws/pcs/bin/pcs_bootstrap_init.sh: No such file or directory
```

Questo errore indica che il software di bootstrap AWS PCS non fa parte dell'AMI. Per risolvere questo problema, assicurati che l'AMI personalizzata includa il software di bootstrap AWS PCS. Per ulteriori informazioni, consulta Immagini di macchine Amazon personalizzate (AMIs) per AWS PCS.

Slurmctld non è in grado di eseguire il ping del nodo di calcolo

Se l'istanza esegue correttamente la procedura di bootstrap ed è registrata con AWS PCS, ma non `slurmctld` è in grado di visualizzarla e di inviarle lavori, l'istanza viene impostata come dopo un certo periodo di tempo e quindi terminata. `DOWN`

Ciò potrebbe essere causato da gruppi di sicurezza configurati in modo errato. Ad esempio, se la porta 6817 è abilitata `slurmd` per consentire la comunicazione`slurmctld`, ma manca la porta 6818 per consentire `slurmctld` il ping. `slurmd`

Verifica che i tuoi gruppi di sicurezza includano tutte le regole richieste, come documentato in. Requisiti e considerazioni sui gruppi di sicurezza

Cronologia dei documenti per la AWS PCS User Guide

La tabella seguente descrive le modifiche importanti alla documentazione per AWS PCS.

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
10 marzo 2026	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.3.2-1. È stato risolto un problema relativo al bootstrap del nodo di calcolo RHEL 8.10 e Rocky Linux 8.10. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e AWS Versioni dell'agente PCS.	N/D
11 febbraio 2026	AWS PCS è stato rilasciato in Asia Pacifico (Mumbai) ed Europa (Parigi)	AWS Il PCS è ora disponibile in Asia Pacifico (Mumbai) (ap-south-1) e in Europa (Parigi) (eu-west-3). CloudFormation sono disponibili modelli per iniziare in Asia Pacifico (Mumbai) e in Europa (Parigi). Regione AWS Regione AWS Per ulteriori informazioni, consultare Utilizzare CloudFormation per	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
		creare un cluster AWS PCS di esempio e CloudFormation modelli per creare un cluster AWS PCS di esempio.	
18 novembre 2025	Nuova funzionalità: Slurm REST API	L'API REST di Slurm è ora supportata per Slurm 25.05 o versioni successive. Per ulteriori informazioni, consulta API REST Slurm in PCS AWS.	SDK AWS: 18-11-2025

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
10 novembre 2025	Nuova funzionalità: supporto del plug-in del filtro CLI Slurm	AWS PCS ora supporta i plugin di filtro Slurm CLI per eseguire script Lua personalizzati che convalidano e modificano i parametri di invio dei lavori prima che raggiungano il controller Slurm. Utilizza i filtri CLI per applicare politiche personalizzate, impostare parametri predefiniti e fornire indicazioni all'utente durante l'invio del lavoro. Questa funzionalità richiede la versione Slurm 25.05 o successiva. Per ulteriori informazioni, consulta Usa i plugin di filtro CLI Slurm per personalizzare l'invio dei lavori in PCS AWS.	N/D
7 novembre 2025	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.3.1-1. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e AWS Versioni dell'agente PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
3 novembre 2025	Agent PCS e programmi di installazione Slurm aggiornati	Aggiornato l'argomento AMI per l'agente AWS PCS 1.3.0-1 e gli installatori Slurm 24.11.6-2, 24.05.8-2 e 23.11.10-4. Elenco aggiornato dei sistemi operativi supportati. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e AWS Versioni dell'agente PCS.	N/D
23 ottobre 2025	Contenuto aggiornato: - configure.sh pcs-multi-cluster-login	Sono stati corretti alcuni errori nello script di configurazione del nodo di accesso multicluster. Per ulteriori informazioni, consulta AWS Codice dello script di configurazione del nodo di accesso multicluster PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
21 ottobre 2025	Nuova funzionalità: rotazione segreta del cluster	AWS PCS ora supporta la rotazione segreta del cluster per migliorare la sicurezza. Per ulteriori informazioni, consulta Segreti dei cluster rotanti in AWS PCS. Autorizzazioni minime di amministratore aggiornate per supportare la rotazione segreta del cluster. Per ulteriori informazioni, consulta Autorizzazioni minime per AWS PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
17 ottobre 2025	Nuovo argomento: script di configurazione del nodo di accesso multicluster	È stato aggiunto un nuovo argomento che fornisce uno script per configurare un nodo di accesso autonomo per la connessione a più cluster AWS PCS. Lo script automatizza la configurazione di più `sackd` demoni Slurm e crea script di attivazione per l'interazione con il cluster. Per ulteriori informazioni, consulta Connessione di un nodo di accesso autonomo a più cluster in PCS AWS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
16 ottobre 2025	Aggiornato per Slurm 25.05	Aggiornata la guida per l'utente per il supporto di Slurm 25.05. Slurm 25.05 è ora la versione predefinita. Per ulteriori informazioni, consulta gli argomenti seguenti: • Versioni Slurm in PCS AWS • Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS • Note di rilascio per l'esempio AWS PCS AMIs	N/D
16 ottobre 2025	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.2.2-1. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e AWS Versioni dell'agente PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
2 ottobre 2025	Nuove funzionalità: riavvio del nodo Slurm, aggiornamenti del cluster e impostazioni Slurm personalizzate	AWS PCS aggiunge il supporto per diverse nuove funzionalità: • Riavvio del nodo Slurm: utilizza il `scontrol reboot` comando nativo di Slurm per riavviare i nodi di calcolo senza la sostituzione dell'istanza. Per ulteriori informazioni, consulta Riavvio dei nodi di calcolo con Slurm in PCS AWS. • Aggiornamenti del cluster: modifica le configurazioni del cluster dopo la creazione senza ricostruzioni. Per ulteriori informazioni, consulta Aggiornamento di un cluster in AWS PCS. • Impostazioni personalizzate Slurm: configura i parametri Slurm avanzati tra le risorse Cluster, Queue e Compute Node Group. Per ulteriori informazi	2025-10-01

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
		oni, consulta Configurazione delle impostazioni Slurm personalizzate in PCS AWS.	
23 settembre 2025	Nuovo argomento per la risoluzione dei problemi: problemi di bootstrap del nodo di calcolo	È stata aggiunta una guida alla risoluzione dei problemi per la diagnosi e la risoluzione dei problemi di bootstrap del nodo di calcolo. Per ulteriori informazioni, consulta Risolvi i problemi relativi al bootstrap e alla registrazione dei nodi di calcolo in PCS AWS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
17 settembre 2025	Nuova funzionalità: Capacity Blocks for ML	AWS PCS ora supporta Amazon EC2 Capacity Blocks for ML, che ti consentono di riservare istanze di elaborazione accelerata basate su GPU per i tuoi cluster. Per ulteriori informazioni, consulta Utilizzo dei blocchi di capacità di Amazon EC2 per ML con PCS AWS. Le autorizzazioni minime per supportare i Capacity Blocks fanno ora parte delle autorizzazioni minime per un amministratore del servizio. Per ulteriori informazioni, consulta Autorizzazioni minime per AWS PCS.	17/09/2025
11 settembre 2025	Aggiornamento delle policy gestite da AWS	AWS PCS è stato aggiornato AWSPCSServiceRolePolicy per supportare Capacity Blocks. Per ulteriori informazioni, consulta AWS politiche gestite per AWS Parallel Computing Service.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
14 agosto 2025	Documentazione aggiornata del profilo dell'istanza	È stata migliorata la documentazione del profilo dell'istanza con istruzioni CLI complete per la creazione di ruoli IAM e profili di istanza. Sono state aggiunte step-by-step procedure per la configurazione dei profili di istanza utilizzando AWS CLI e sono state migliorate le linee guida per la ricerca dei profili di istanza utilizzati con AWS PCS. Per ulteriori informazioni, consulta Profili di istanza IAM per AWS Parallel Computing Service.	2025-08-14

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
1 agosto 2025	Nuovo argomento: plugin SPANK	È stata aggiunta documentazione per i plugin SPANK (Slurm Plug-in Architecture for Node and job Kontrol) che è possibile utilizzare per estendere e modificare il comportamento di Slurm durante l'avvio e l'esecuzione dei job su cluster PCS. AWS Per ulteriori informazioni, consulta Estendi la funzionalità Slurm sui AWS PC con i plugin SPANK.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
1 agosto 2025	IPv6 supporto di rete	È stato aggiunto il supporto per il IPv6 networking durante la creazione di cluster AWS PCS. Ora puoi scegliere IPv6 come tipo di rete per il tuo cluster, con gli aggiornamenti corrispon denti ai requisiti VPC, alla configurazione della sottorete, alle impostazi oni dei gruppi di sicurezza e alle procedure di creazione del cluster. Per ulteriori informazi oni, consultare AWS Requisiti e considerazioni su PCS, VPC e sottorete e Creazione di un cluster in AWS PCS.	01/08/2025

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
3 luglio 2025	AWS PCS rilasciato in Europa (Londra)	AWS Il PCS è ora disponibile in Europa (Londra) (eu-west-2). CloudFormation sono disponibili modelli per iniziare in Europa (Londra). Regione AWS Per ulteriori informazioni, consultare Utilizzare CloudFormation per creare un cluster AWS PCS di esempio e CloudFormation modelli per creare un cluster AWS PCS di esempio.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
1 luglio 2025	Istruzioni aggiornate per la console	Ora puoi fare in modo che AWS PCS crei automaticamente un profilo di istanza di base e un gruppo di sicurezza quando crei un cluster e un gruppo di nodi di calcolo nella console. Per ulteriori informazioni, consulta: • Creazione di un cluster in AWS PCS • Creazione di un gruppo di nodi di calcolo in AWS PCS • Profili di istanza IAM per AWS Parallel Computing Service	N/D
23 giugno 2025	Nuova politica gestita: AWSPCSCompute NodePolicy	È stata aggiunta una nuova politica gestita che concede l'autorizzazione ai nodi di elaborazione AWS PCS per connetter si ai cluster AWS PCS. Per ulteriori informazioni, consulta AWS politica gestita: AWSPCSCom pute NodePolicy.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
19 giugno 2025	Nuovo argomento: registri di completamento dei lavori	Utilizza i registri di completamento dei lavori per registrare i dettagli sui lavori una volta completati, senza costi aggiuntivi. Per ulteriori informazioni, consulta Registri di completamento dei lavori in AWS PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
18 giugno 2025	AWS Rilascio PCS in AWS GovCloud (US)	AWS Il PCS è ora disponibile in AWS GovCloud (Stati Uniti orientali) (us-gov-east-1) e AWS GovCloud (Stati Uniti occidentali) (us-gov-west-1). CloudFormation sono disponibili modelli per iniziare in. AWS GovCloud (US) Regions Per ulteriori informazioni, consultare Utilizzare CloudFormation per creare un cluster AWS PCS di esempio e CloudFormation modelli per creare un cluster AWS PCS di esempio. Per ulteriori informazioni sugli endpoint del servizio AWS PCS in AWS GovCloud (US) Regions, vedereEndpoint e quote di servizio per PCS AWS. Per ulteriori informazioni sulle differenze tra AWS GovCloud (US) Regions, consulta AWS PCS AWS GovCloud	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
		(US) nella Guida per l'AWS GovCloud (US) utente.	
18 giugno 2025	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.2.1-1. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D
15 maggio 2025	Nuova funzionalità: contabilità	La contabilità Slurm è ora supportata per Slurm 24.11 o versioni successive. Per ulteriori informazioni, consulta Contabilità Slurm in PCS AWS.	SDK AWS: 15/05/2020

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
15 maggio 2025	Aggiornato per Slurm 24.11	Aggiornata la guida per l'utente per il supporto di Slurm 24.11.5. Per ulteriori informazioni, consulta gli argomenti seguenti: • Versioni Slurm in PCS AWS • Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS • Note di rilascio per l'esempio AWS PCS AMIs	N/D
5 maggio 2025	Domande frequenti sulle versioni aggiornate di Slurm	Domande frequenti (FAQ) aggiornate sulle versioni di Slurm sulle versioni di Slurm prossime o oltre la fine del ciclo di vita (EOL). Per ulteriori informazioni, consulta Domande frequenti sulle versioni di Slurm in PCS AWS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
17 aprile 2025	Nuovo argomento: come ottenere i dettagli del gruppo di nodi di calcolo	Scopri come ottenere dettagli per un gruppo di nodi di calcolo AWS PCS, come ID, ARN e ID AMI. Per ulteriori informazioni, consulta Ottieni i dettagli del gruppo di nodi di calcolo in AWS PCS.	N/D
2 aprile 2025	Programma di installazione Slurm aggiornato	Aggiornato l'argomento AMI per l'installatore Slurm 24.05.7-1. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D
28 marzo 2025	Sono state aggiunte quote per il numero massimo di gruppi e code di nodi di calcolo	Sono state aggiunte quote interne non regolabili per il numero massimo di gruppi di nodi di calcolo per cluster e il numero massimo di code per cluster. Per ulteriori informazioni, consulta Quote interne.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
14 marzo 2025	È stata modificata una chiave di proprietà nel modello CloudFormation	`Idora è TemplateId per la CustomLaunchTemplate` proprietà nel CloudFormation modello. Per ulteriori informazioni, consulta Resources in Parti di un CloudFormation modello per AWS PCS.	N/D
13 marzo 2025	Sono state aggiunte informazioni sulla versione per l'agente AWS PCS e Slurm	È stato aggiunto un nuovo argomento che descrive le modifiche per ogni versione dell'agente AWS PCS. Per ulteriori informazioni, consulta AWS Versioni dell'agente PCS. Sono state aggiunte ulteriori informazioni all'argomento sulle versioni di Slurm che descrive le date di supporto importanti e le note di rilascio dettagliate per il supporto AWS PCS per Slurm. Per ulteriori informazioni, consulta Versioni Slurm in PCS AWS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
7 marzo 2025	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.2.0-1. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D
3 febbraio 2025	È stato aggiunto un argomento sull'utilizzo con AWS CloudFormation PCS AWS	È stato aggiunto un argomento alla guida per l'utente che fornisce un esempio di utilizzo CloudFormation con AWS PCS. L'argomento fornisce una procedura per utilizzare un CloudFormation modello di esempio per creare il cluster AWS PCS di esempio e descrive brevemente le sezioni di tale modello. Per ulteriori informazioni, consulta Inizia con CloudFormation e AWS PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
18 dicembre 2024	Aggiornato per Slurm 24.05	Aggiornata la guida per l'utente per il supporto di Slurm 24.05. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e Note di rilascio per l'esempio AWS PCS AMIs.	N/D
18 dicembre 2024	Versioni NVIDIA aggiornate per Slurm 23.11 sample AMIs	Versioni aggiornate dei driver NVIDIA e CUDA nell'esempio Slurm 23.11. AMIs Per ulteriori informazioni, consulta Note di rilascio per l'esempio AWS PCS AMIs.	N/D
17 dicembre 2024	Programma di installazione Slurm aggiornato	Aggiornato l'argomento AMI per il programma di installazione Slurm 23.11.10-3. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
13 dicembre 2024	Agente PCS aggiornato	Aggiornato l'argomento AMI per l'agente AWS PCS 1.1.1-1. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D
6 dicembre 2024	Agente PCS e programma di installazione Slurm aggiornati	Aggiornato l'argomento AMI per l'agente AWS PCS 1.1.0-1 e il programma di installazione Slurm 23.11.10-2. Per ulteriori informazioni, consulta Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS.	N/D
6 dicembre 2024	È stato aggiunto un argomento sul supporto del sistema operativo	Per ulteriori informazioni, consulta Sistemi operativi supportati in AWS PCS.	N/D
8 novembre 2024	Guida per l'utente riorganizzata	Abbiamo riorganizzato la guida per l'utente per portare gli argomenti al livello più alto, spostato alcuni argomenti nelle rispettive pagine e raggruppato argomenti simili.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
7 novembre 2024	Argomenti AMI aggiornati	Aggiornato l'argomento AMI per Slurm 23.11.10 e libjwt 17.0. Per ulteriori informazioni, consultar e Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e Fase 3 — Installare Slurm. Sono state semplific ate e corrette le note di rilascio per. AMIs Per ulteriori informazioni, consulta Note di rilascio per l'esempio AWS PCS AMIs.	N/D
7 novembre 2024	È stato aggiunto un nuovo argomento sull'util izzo di volumi EBS crittografati con PCS AWS	È stato aggiunto un argomento che descrive la politica delle chiavi KMS richiesta per i volumi EBS crittografati in PCS. AWS Per ulteriori informazioni, consulta Politica delle chiavi KMS richiesta per l'uso con volumi EBS crittografati in PCS AWS.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
18 ottobre 2024	AWS È stato rilasciato l'agente PCS 1.0.1-1	Documentazione relativa all'AMI aggiornata per fare riferimento alla versione AWS 1.0.1-1 dell'agente PCS. Per ulteriori informazioni, consultare Programmi di installazione software per creare soluzioni personalizzate AMIs per AWS PCS e Fase 2 — Installare l'agente AWS PCS.	N/D
10 ottobre 2024	È stato aggiunto un capitolo sulla risoluzione dei problemi	È stato aggiunto un capitolo sulla risoluzione dei problemi con un argomento sulla sostituzione automatica delle istanze EC2 dopo il riavvio. Per ulteriori informazioni, consulta Risoluzione dei problemi in Parallel Computing AWS Service.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
23 settembre 2024	Sono state aggiornate le autorizzazioni minime per utilizzare le azioni API e per un amministratore del servizio	L'`ec2:DescribeInstanceTypeOfferings` autorizzazione è ora richiesta per le azioni `CreateComputeNodeGroup` e `UpdateComputeNodeGroup` API. Per ulteriori informazioni, consulta Autorizzazioni minime per AWS PCS.	N/D
5 settembre 2024	È stata aggiornata la policy IAM di esempio per le autorizzazioni minime per un amministratore del servizio	Per ulteriori informazioni, consulta Autorizzazioni minime per un amministratore del servizio.	N/D
5 settembre 2024	È stata aggiunta un'autorizzazione mancante al JSON nella pagina delle politiche gestite	Questa è stata solo una correzione alla documentazione. La politica gestita effettiva non è stata modificata. Per ulteriori informazioni, consulta AWS politiche gestite per AWS Parallel Computing Service.	N/D
28 agosto 2024	È stata aggiunta la pagina delle politiche gestite	Per ulteriori informazioni, consulta AWS politiche gestite per AWS Parallel Computing Service.	N/D

Data	Modifica	Aggiornamenti della documentazione	Versioni API aggiornate
28 agosto 2024	AWS Versione PCS	Versione iniziale della guida per l'utente del AWS PCS.	AWS SDK: 2024-08-28

AWS Glossario

Per la AWS terminologia più recente, consultate il AWS glossario nella sezione Reference.Glossario AWS

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.