



Red Hat OpenStack Platform 16.1

インスタンス作成のための Compute サービスの 設定

インスタンスを作成するための Red Hat OpenStack Platform Compute (nova) サービスの設定および管理に関するガイド

Red Hat OpenStack Platform 16.1 インスタンス作成のための Compute サービスの設定

インスタンスを作成するための Red Hat OpenStack Platform Compute (nova) サービスの設定および管理に関するガイド

OpenStack Team
rhos-docs@redhat.com

法律上の通知

Copyright © 2023 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux[®] is the registered trademark of Linus Torvalds in the United States and other countries.

Java[®] is a registered trademark of Oracle and/or its affiliates.

XFS[®] is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL[®] is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js[®] is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack[®] Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

概要

本ガイドでは、クラウド管理者が OpenStack クライアント CLI を使用して Red Hat OpenStack Platform Compute (nova) サービスを設定および管理するための概念および手順について説明します。

目次

多様性を受け入れるオープンソースの強化	4
RED HAT ドキュメントへのフィードバック (英語のみ)	5
第1章 COMPUTE サービス (NOVA) の機能	6
第2章 COMPUTE サービス (NOVA) の設定	8
2.1. オーバーコミットのためのメモリー設定	9
2.2. コンピュートノード上で確保するホストメモリーの計算	9
2.3. スワップサイズの計算	10
第3章 パフォーマンス改善のためのコンピュートノードの設定	11
3.1. コンピュートノードでの CPU ピニングの設定	11
3.2. エミュレータスレッドの設定	19
3.3. コンピュートノードでのヒュージページの設定	20
3.4. インスタンスにファイルベースのメモリーを使用するコンピュートノードの設定	25
第4章 COMPUTE サービスのストレージの設定	28
4.1. イメージキャッシュの設定オプション	28
4.2. インスタンスの一時ストレージ属性の設定オプション	30
4.3. 共有インスタンスストレージの設定	33
4.4. 関連情報	34
第5章 PCI パススルーの設定	36
5.1. PCI パススルー用コンピュートノードの指定	36
5.2. PCI パススルー用コンピュートノードの設定	38
5.3. PCI パススルーデバイス種別フィールド	41
5.4. NOVAPCIPASSTHROUGH 設定のガイドライン	42
第6章 ホストアグリゲートの作成と管理	44
6.1. ホストアグリゲートでのスケジューリングの有効化	44
6.2. ホストアグリゲートの作成	45
6.3. アベイラビリティゾーンの作成	47
6.4. ホストアグリゲートの削除	48
6.5. プロジェクト分離ホストアグリゲートの作成	48
第7章 インスタンスのスケジューリングと配置の設定	51
7.1. PLACEMENT サービスを使用した事前絞り込み	51
7.2. COMPUTE スケジューラーサービス用フィルターおよび重みの設定	58
7.3. COMPUTE スケジューラーのフィルター	59
7.4. COMPUTE スケジューラーの重み	65
第8章 インスタンス起動用のフレーバーの作成	73
8.1. フレーバーの作成	73
8.2. フレーバーの引数	75
8.3. フレーバーのメタデータ	76
第9章 インスタンスへのメタデータの追加	91
9.1. インスタンスメタデータの種別	91
9.2. 全インスタンスへのコンフィグドライブの追加	91
9.3. インスタンスへの動的メタデータの追加	93
第10章 インスタンスのメモリーを暗号化するための AMD SEV コンピュートノードの設定	95
10.1. SECURE ENCRYPTED VIRTUALIZATION (SEV)	95

10.2. メモリー暗号化用 AMD SEV コンピュートノードの指定	96
10.3. メモリー暗号化用 AMD SEV コンピュートノードの設定	99
10.4. メモリー暗号化用のイメージの作成	101
10.5. メモリー暗号化用のフレーバーの作成	102
10.6. メモリーが暗号化されたインスタンスの起動	102
第11章 インスタンスに永続メモリーを提供する NVDIMM コンピュートノードの設定	103
11.1. PMEM 用コンピュートノードの指定	104
11.2. PMEM 用コンピュートノードの設定	106
第12章 インスタンス用の仮想 GPU の設定	108
12.1. サポートされる設定および制限	108
12.2. コンピュートノードでの仮想 GPU の設定	109
12.3. カスタム GPU インスタンスイメージの作成	113
12.4. インスタンス用の仮想 GPU フレーバーの作成	114
12.5. 仮想 GPU インスタンスの起動	115
12.6. GPU デバイスの PCI パススルーの有効化	116
第13章 リアルタイムコンピュートの設定	120
13.1. リアルタイム処理用コンピュートノードの準備	120
13.2. リアルタイム COMPUTE ロールのデプロイメント	123
13.3. デプロイメントおよびテストシナリオの例	126
13.4. リアルタイムインスタンスの起動およびチューニング	127
第14章 インスタンスの管理	129
14.1. データベースのクリーニング	129
14.2. コンピュートノード間の仮想マシンインスタンスの移行	132

多様性を受け入れるオープンソースの強化

Red Hat では、コード、ドキュメント、Web プロパティにおける配慮に欠ける用語の置き換えに取り組んでいます。まずは、マスター (master)、スレーブ (slave)、ブラックリスト (blacklist)、ホワイトリスト (whitelist) の 4 つの用語の置き換えから始めます。この取り組みは膨大な作業を要するため、今後の複数のリリースで段階的に用語の置き換えを実施して参ります。詳細は、[Red Hat CTO である Chris Wright のメッセージ](#) を参照してください。

RED HAT ドキュメントへのフィードバック (英語のみ)

弊社ドキュメントに対するご意見をお聞かせください。ドキュメントの改善点があればお知らせください。

ドキュメントへのダイレクトフィードバック (DDF) 機能の使用 (英語版のみ)

特定の文章、段落、またはコードブロックに対して直接コメントを送付するには、DDF の **Add Feedback** 機能を使用してください。なお、この機能は英語版のドキュメントでのみご利用いただけます。

1. **Multi-page HTML** 形式でドキュメントを表示します。
2. ドキュメントの右上隅に **Feedback** ボタンが表示されていることを確認してください。
3. コメントするテキスト部分をハイライト表示します。
4. **Add Feedback** をクリックします。
5. **Add Feedback** フィールドにコメントを入力します。
6. (オプション) ドキュメントチームが連絡を取り問題についてお伺いできるように、ご自分のメールアドレスを追加します。
7. **Submit** をクリックします。

第1章 COMPUTE サービス (NOVA) の機能

Compute (nova) サービスを使用して、Red Hat OpenStack Platform (RHOSP) 環境で仮想マシンインスタンスおよびベアメタルサーバーを作成、プロビジョニング、および管理します。Compute サービスは、ベースとなるホストプラットフォームの詳細を公開するのではなく、Compute サービスを実行しているベースとなるハードウェアを抽象化します。たとえば、ホスト上で動作中の CPU の種別およびトポロジを公開するのではなく、Compute サービスは多数の仮想 CPU (vCPU) を公開し、これらの仮想 CPU のオーバーコミットに対応します。

Compute サービスは、KVM ハイパーバイザーを使用して Compute サービスのワークロードを実行します。libvirt ドライバーは QEMU と協調して KVM との相互のやり取りをすべて処理し、仮想マシンインスタンスの作成を可能にします。インスタンスを作成およびプロビジョニングするために、Compute サービスは以下の RHOSP サービスと協調します。

- 認証のための Identity (keystone) サービス
- リソースインベントリをトラッキングおよび選択するための Placement サービス
- ディスクおよびインスタンスイメージのための Image サービス (glance)
- インスタンスがブート時に接続する仮想ネットワークまたは物理ネットワークをプロビジョニングするための Networking (neutron) サービス

Compute サービスは、**nova-*** という名前のデーモンプロセスおよびサービスで設定されます。コアの Compute サービスを以下に示します。

Compute サービス (nova-compute)

このサービスは、KVM または QEMU ハイパーバイザー API の libvirt を使用してインスタンスを作成、管理、および削除し、インスタンスの状態をデータベースを更新します。

Compute コンダクター (nova-conductor)

このサービスは、Compute サービスとデータベースの協調を仲介します。これにより、コンピュータードをデータベースへの直接アクセスから隔離します。このサービスを **nova-compute** サービスが実行されているノードにデプロイしないでください。

Compute スケジューラー (nova-scheduler)

このサービスはキューからインスタンスの要求を受け取り、インスタンスをホストするコンピュータードを決定します。

Compute API (nova-api)

このサービスは、ユーザーに外部 REST API を提供します。

API データベース

このデータベースはインスタンスの場所の情報をトラッキングします。また、ビルドされているがスケジュールされていないインスタンスの一時的な場所を提供します。マルチセルのデプロイメントでは、このデータベースには各セルのデータベース接続を指定するセルのマッピングも含まれます。

セルデータベース

このデータベースには、インスタンスに関するほとんどの情報が含まれます。API データベース、コンダクター、および Compute サービスによって使用されます。

メッセージキュー

このメッセージングサービスは、セル内の各サービス間の通信およびグローバルなサービスとの通信のために、すべてのサービスによって使用されます。

Compute メタデータ

このサービスは、インスタンス固有のデータを保管します。インスタンス

は、<http://169.254.169.254> または IPv6 を通じてリンクローカルアドレス 80::a9fe:a9fe から、メタデータサービスにアクセスします。Networking (neutron) サービスは、要求をメタデータ API サーバーに転送するロールを持ちます。**NeutronMetadataProxySharedSecret** パラメーターを使用して、Networking サービスと Compute サービス両方の設定でシークレットキーワードを設定する必要があります。これにより、これらのサービスが通信を行うことができます。Compute メタデータサービスは、Compute API の一部としてグローバルに実行することや、それぞれのセルで実行することができます。

複数のコンピュートノードをデプロイすることができます。インスタンスを操作するハイパーバイザーは、それぞれのコンピュートノードで実行されます。それぞれのコンピュートノードには、少なくとも2つのネットワークインターフェイスが必要です。コンピュートノードでは、インスタンスを仮想ネットワークに接続し、セキュリティーグループを介してインスタンスにファイアウォールサービスを提供する Networking サービスエージェントも実行されます。

デフォルトでは、director はすべてのコンピュートノードを単一のセルとしてオーバークラウドをインストールします。このセルには、仮想マシンインスタンスを制御および管理するすべての Compute サービスおよびデータベース、ならびにすべてのインスタンスおよびインスタンスのメタデータが含まれます。大規模なデプロイメントでは、複数のセルでオーバークラウドをデプロイし、多数のコンピュートノードに対応することができます。新しいオーバークラウドをインストールする際に、またはその後の任意の時に、セルを環境に追加することができます。詳細は、[Scaling Deployments with Compute Cells](#)を参照してください。

第2章 COMPUTE サービス (NOVA) の設定

クラウド管理者は、環境ファイルを使用して Compute (nova) サービスをカスタマイズします。Puppet は、この設定を生成して `/var/lib/config-data/puppet-generated/<nova_container>/etc/nova/nova.conf` ファイルに保存します。Compute サービスの設定をカスタマイズするには、以下の設定方法を順番どおりに使用します。

1. **heat パラメーター**: 詳細は、[オーバークラウドのパラメーターの Compute \(nova\) パラメーター セクション](#)を参照してください。以下の例では、heat パラメーターを使用して、デフォルトのスケジューラーフィルターを設定し、Compute サービスの NFS バックエンドを設定します。

```
parameter_defaults:
  NovaSchedulerDefaultFilters:
    AggregateInstanceExtraSpecsFilter,ComputeFilter,ComputeCapabilitiesFilter,ImagePropertiesFilter
  NovaNfsEnabled: true
  NovaNfsShare: '192.0.2.254:/export/nova'
  NovaNfsOptions: 'context=system_u:object_r:nfs_t:s0'
  NovaNfsVersion: '4.2'
```

2. **Puppet パラメーター**: `/etc/puppet/modules/nova/manifests/*` で定義されます。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::compute::force_raw_images: True
```



注記

この方法は、等価な heat パラメーターが存在しない場合にのみ使用してください。

3. **手動での hieradata の上書き**: heat パラメーターまたは Puppet パラメーターが存在しない場合の、パラメーターカスタマイズ用。たとえば、以下の設定により Compute ロールの **[DEFAULT]** セクションに `timeout_nbd` が定義されます。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
      DEFAULT/timeout_nbd:
        value: '20'
```



警告

heat パラメーターが存在する場合は、Puppet パラメーターではなく heat パラメーターを使用します。Puppet パラメーターは存在するが heat パラメーターは存在しない場合は、手動で上書きする方法ではなく Puppet パラメーターを使用します。等価な heat パラメーターまたは Puppet パラメーターがない場合に限り、手動で上書きする方法を使用してください。

ヒント

特定の設定をカスタマイズする際に heat パラメーターまたは Puppet パラメーターが利用可能かどうかを判断するには、オーバークラウドの高度なカスタマイズの [変更するパラメーターの特定](#) のガイドンスに従ってください。

オーバークラウドサービスの設定方法の詳細は、[オーバークラウドの高度なカスタマイズ](#) の [Heat パラメーター](#) を参照してください。

2.1. オーバーコミットのためのメモリー設定

メモリーのオーバーコミットを使用する場合 (**NovaRAMAllocationRatio** ≥ 1.0)、割り当て率をサポートするのに十分なスワップ領域を確保してオーバークラウドをデプロイする必要があります。



注記

NovaRAMAllocationRatio パラメーターが 1 より小さい値に設定されている場合は、RHEL でのスワップサイズの推奨事項に従ってください。詳しくは、RHEL の [ストレージデバイスの管理](#) の [システムの推奨スワップ領域](#) を参照してください。

前提条件

- ノードに必要なスワップサイズが計算されている。詳細は、[スワップサイズの計算](#) を参照してください。

手順

1. **/usr/share/openstack-tripleo-heat-templates/environments/enable-swap.yaml** ファイルを環境ファイルのディレクトリーにコピーします。

```
$ cp /usr/share/openstack-tripleo-heat-templates/environments/enable-swap.yaml
/home/stack/templates/enable-swap.yaml
```

2. 以下のパラメーターを **enable-swap.yaml** ファイルに追加して、スワップサイズを設定します。

```
parameter_defaults:
  swap_size_megabytes: <swap size in MB>
  swap_path: <full path to location of swap, default: /swap>
```

3. その他の環境ファイルと共に **enable_swap.yaml** 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/enable-swap.yaml
```

2.2. コンピュートノード上で確保するホストメモリーの計算

ホストのプロセス用に確保する RAM 容量の合計を判断するには、以下の各項目に十分なメモリーを割り当てる必要があります。

- ホスト上で実行されるリソース (たとえば、OSD は 3 GB のメモリーを消費します)

- ホストインスタンスに必要なエミュレーターのオーバーヘッド
- 各インスタンスのハイパーバイザー

メモリーへの追加要求を計算したら、以下の式を使用して各ノードのホストプロセス用に確保するメモリーの容量を決定します。

```
NovaReservedHostMemory = total_RAM - ( (vm_no * (avg_instance_size + overhead)) + (resource1 * resource_ram) + (resourcen * resource_ram))
```

- **vm_no** は、インスタンスの数に置き換えてください。
- **avg_instance_size** は、各インスタンスが使用できるメモリーの平均容量に置き換えてください。
- **overhead** は、各インスタンスに必要なハイパーバイザーのオーバーヘッドに置き換えてください。
- **resource1** および **<resourcen>** までのすべてのリソースを、ノード上のリソース種別の数に置き換えてください。
- **resource_ram** は、この種別の各リソースに必要な RAM の容量に置き換えてください。

2.3. スワップサイズの計算

割り当てるスワップサイズは、メモリーのオーバーコミットを処理するのに十分な容量でなければなりません。以下の式を使用して、ノードに必要なスワップサイズを計算することができます。

- $\text{overcommit_ratio} = \text{NovaRAMAllocationRatio} - 1$
- $\text{最小スワップサイズ (MB)} = (\text{total_RAM} * \text{overcommit_ratio}) + \text{RHEL_min_swap}$
- $\text{推奨 (最大) スワップサイズ (MB)} = \text{total_RAM} * (\text{overcommit_ratio} + \text{percentage_of_RAM_to_use_for_swap})$

percentage_of_RAM_to_use_for_swap 変数は、QEMU のオーバーヘッドおよびオペレーティングシステムまたはホストのサービスが消費するその他のリソースに対応するバッファです。

たとえば、RAM 容量の合計が 64 GB で **NovaRAMAllocationRatio** が 1 に設定されている場合、利用可能な RAM の 25% をスワップに使用するには、

- $\text{推奨 (最大) スワップサイズ} = 64000 \text{ MB} * (0 + 0.25) = 16000 \text{ MB}$

NovaReservedHostMemory の値を計算する方法は、[コンピュートノード上で確保するホストメモリーの計算](#) を参照してください。

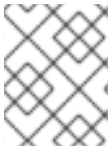
RHEL_min_swap の値を決定する方法は、RHEL のストレージデバイスの管理の [システムの推奨スワップ領域](#) を参照してください。

第3章 パフォーマンス改善のためのコンピュートノードの設定

クラウドユーザーは、インスタンスのスケジューリングおよび配置を設定して、最大のパフォーマンスを得ることができます。そのためには、NFV や高性能コンピューティング (HPC) などの特化されたワークロードを対象にするカスタムフレーバーを作成します。

以下の機能を使用して、最大のパフォーマンスを得るためにインスタンスを調整します。

- **CPU ピニング**: 仮想 CPU を物理 CPU に固定します。
- **エミュレータスレッド**: インスタンスに関連付けられたエミュレータスレッドを物理 CPU に固定します。
- **ヒュージページ**: 通常のメモリー (4 KB ページ) とヒュージページ (2 MB または 1 GB ページ) の両方について、インスタンスのメモリー割り当てポリシーを調整します。



注記

これらの機能のいずれかを設定すると、インスタンス上に NUMA トポロジが存在しない場合、暗黙的な NUMA トポロジが作成されます。

3.1. コンピュートノードでの CPU ピニングの設定

コンピュートノードで CPU ピニングを有効化することで、各インスタンスの CPU プロセスを専用のホスト CPU で実行するように設定することができます。インスタンスが CPU ピニングを使用する場合には、各インスタンスの仮想 CPU プロセスには、他のインスタンスの仮想 CPU プロセスが使用できない独自のホストの物理 CPU が割り当てられます。CPU ピニングが設定されたコンピュートノード上で動作するインスタンスには、NUMA トポロジがあります。インスタンスの NUMA トポロジの各 NUMA ノードは、ホストコンピュートノード上の NUMA ノードにマッピングされます。

専用の (ピンングされた) CPU を持つインスタンスと共有 (フローティング) の CPU を持つインスタンスを同じコンピュートノード上にスケジューリングするように、Compute のスケジューラーを設定することができます。NUMA トポロジを持つコンピュートノード上で CPU ピニングを設定するには、以下の手順を実施する必要があります。

1. CPU ピニング用のコンピュートノードを指定する。
2. ピニングされたインスタンス仮想 CPU プロセス、フローティングのインスタンス仮想 CPU プロセス、およびホストのプロセス用にホストコアを確保するようにコンピュートノードを設定する。
3. オーバークラウドをデプロイする。
4. CPU ピニングを要求するインスタンスを起動するためのフレーバーを作成する。
5. 共有 (あるいはフローティング) の CPU を使用するインスタンスを起動するためのフレーバーを作成する。

3.1.1. 前提条件

- コンピュートノードの NUMA トポロジを把握している。

3.1.2. CPU ピニング用コンピュートノードの指定

ピンングされた CPU を使用するインスタンス用にコンピュートノードを指定するには、CPU ピニング

ロールを設定するための新規ロールファイルを作成し、CPU ピニングのためにコンピュートノードをタグ付けするための新規オーバークラウドフレーバーおよび CPU ピニングリソースクラスを設定する必要があります。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. **roles_data_cpu_pinning.yaml** という名前で、**Controller**、**Compute**、および **ComputeCPUPinning** ロールが含まれる新しいロールデータファイルを生成します。

```
(undercloud)$ openstack overcloud roles \
generate -o /home/stack/templates/roles_data_cpu_pinning.yaml \
Compute:ComputeCPUPinning Compute Controller
```

4. **roles_data_cpu_pinning.yaml** を開き、以下のパラメーターおよびセクションを編集または追加します。

セクション/パラメーター	現在の値	新しい値
ロールのコメント	Role: Compute	Role: ComputeCPUPinning
ロール名	Compute	name: ComputeCPUPinning
description	Basic Compute Node role	CPU Pinning Compute Node role
HostnameFormatDefault	%stackname%-novacompute-%index%	%stackname%-novacomputepinning-%index%
deprecated_nic_config_name	compute.yaml	compute-cpu-pinning.yaml

5. オーバークラウド用の CPU ピニングコンピュートノードをノード定義のテンプレート **node.json** または **node.yaml** に追加して、そのノードを登録します。詳細は、**director** のインストールと使用方法の [オーバークラウドノードの登録](#) を参照してください。
6. ノードのハードウェアを検査します。

```
(undercloud)$ openstack overcloud node introspect \
--all-manageable --provide
```

詳細は、**Director Installation and Usage** ガイドの [Creating an inventory of the bare-metal node hardware](#) を参照してください。

7. CPU ピニングを使用するコンピュートノード用の **compute-cpu-pinning** オーバークラウドフレーバーを作成します。

```
(undercloud)$ openstack flavor create --id auto \
--ram <ram_size_mb> --disk <disk_size_gb> \
--vcpus <no_vcpus> compute-cpu-pinning
```

- **<ram_size_mb>** をベアメタルノードの RAM (MB 単位) に置き換えます。
- **<disk_size_gb>** をベアメタルノード上のディスク容量 (GB 単位) に置き換えます。
- **<no_vcpus>** をベアメタルノードの CPU 数に置き換えます。



注記

これらの属性は、インスタンスのスケジューリングには使用されません。ただし Compute スケジューラーは、ディスク容量を使用してルートパーティションのサイズを決定します。

8. ノード一覧を取得して UUID を把握します。

```
(undercloud)$ openstack baremetal node list
```

9. CPU ピニング用に指定する各ベアメタルノードに、カスタムの CPU ピニングリソースクラスをタグ付けします。

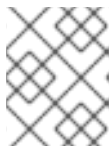
```
(undercloud)$ openstack baremetal node set \
--resource-class baremetal.CPU-PINNING <node>
```

<node> をベアメタルノードの ID に置き換えてください。

10. **compute-cpu-pinning** フレーバーをカスタムの CPU ピニングリソースクラスに関連付けます。

```
(undercloud)$ openstack flavor set \
--property resources:CUSTOM_BAREMETAL_CPU_PINNING=1 \
compute-cpu-pinning
```

Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、それぞれの句読点をアンダースコアに置き換え、**CUSTOM_** の接頭辞を追加します。



注記

フレーバーが要求できるのは、ベアメタルリソースクラスの1つのインスタンスだけです。

11. 以下のフレーバー属性を設定して、Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぎます。

```
(undercloud)$ openstack flavor set \
--property resources:VCPU=0 \
--property resources:MEMORY_MB=0 \
```

```
--property resources:DISK_GB=0 compute-cpu-pinning
```

12. (オプション) **ComputeCPUPinning** ロールのネットワークポロジが **Compute** ロールのネットワークポロジと異なる場合は、カスタムネットワークインターフェイステンプレートを作成します。詳細は、[オーバークラウドの高度なカスタマイズの カスタムネットワークインターフェイステンプレート](#) を参照してください。

ComputeCPUPinning ロールのネットワークポロジが **Compute** ロールと同じ場合は、**compute.yaml** で定義されるデフォルトのネットワークポロジを使用することができます。

13. **ComputeCPUPinning** ロールの **Net::SoftwareConfig** を **network-environment.yaml** ファイルに登録します。

```
resource_registry:
  OS::TripleO::Compute::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/compute.yaml
  OS::TripleO::ComputeCPUPinning::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/<cpu_pinning_net_top>.yaml
  OS::TripleO::Controller::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/controller.yaml
```

<cpu_pinning_net_top> を **ComputeCPUPinning** ロールのネットワークポロジが含まれるファイルの名前に置き換えます。たとえば、デフォルトのネットワークポロジを使用する場合は **compute.yaml** です。

14. 以下のパラメーターを **node-info.yaml** ファイルに追加して、CPU ピニングコンピュートノードの数および CPU ピニング対応コンピュートノード用に使用するフレーバーを指定します。

```
parameter_defaults:
  OvercloudComputeCPUPinningFlavor: compute-cpu-pinning
  ComputeCPUPinningCount: 3
```

15. ロールが作成されたことを確認するには、以下のコマンドを入力します。

```
(undercloud)$ openstack baremetal node list --long -c "UUID" \
-c "Instance UUID" -c "Resource Class" -c "Provisioning State" \
-c "Power State" -c "Last Error" -c "Fault" -c "Name" -f json
```

出力例:

```
[
  {
    "Fault": null,
    "Instance UUID": "e8e60d37-d7c7-4210-acf7-f04b245582ea",
    "Last Error": null,
    "Name": "compute-0",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "baremetal.CPU-PINNING",
    "UUID": "b5a9ac58-63a7-49ba-b4ad-33d84000ccb4"
  },
  {
    "Fault": null,
    "Instance UUID": "3ec34c0b-c4f5-4535-9bd3-8a1649d2e1bd",
    "Last Error": null,
```

```

    "Name": "compute-1",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "compute",
    "UUID": "432e7f86-8da2-44a6-9b14-dfacdf611366"
  },
  {
    "Fault": null,
    "Instance UUID": "4992c2da-adde-41b3-bef1-3a5b8e356fc0",
    "Last Error": null,
    "Name": "controller-0",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "controller",
    "UUID": "474c2fc8-b884-4377-b6d7-781082a3a9c0"
  }
]

```

3.1.3. CPU ピニング用コンピュートノードの設定

ノードの NUMA トポロジーに基づいて、コンピュートノードでの CPU ピニングを設定します。効率を高めるために、全 NUMA ノードにわたって、CPU コアの一部をホストのプロセス用に確保します。残りの CPU コアをインスタンスの管理に割り当てます。

以下の手順では、以下の NUMA トポロジー (8 つの CPU コアを 2 つの NUMA ノードに分散) を使用して、CPU ピニングの設定方法を説明します。

表3.1 NUMA トポロジーの例

NUMA ノード 0		NUMA ノード 1	
コア 0	コア 1	コア 2	コア 3
コア 4	コア 5	コア 6	コア 7

以下の手順では、コア 0 および 4 をホストのプロセス用に、コア 1、3、5、および 7 を CPU ピニングが必要なインスタンス用に、そしてコア 2 および 6 を CPU ピニングが不要なフローティングインスタンス用に、それぞれ確保します。

手順

1. ピニングされたインスタンス、フローティングのインスタンス、およびホストプロセス用にコアを確保するようにコンピュートノードを設定する環境ファイルを作成します (例: **cpu_pinning.yaml**)。
2. NUMA 対応コンピュートノードに NUMA トポロジーが設定されたインスタンスをスケジュールするには、Compute 環境ファイルの **NovaSchedulerDefaultFilters** パラメーターに **NUMATopologyFilter** がなければ、このフィルターを追加します。

```

parameter_defaults:
  NovaSchedulerDefaultFilters:
    ['AvailabilityZoneFilter','ComputeFilter','ComputeCapabilitiesFilter','ImagePropertiesFilter','ServerGroupAntiAffinityFilter','ServerGroupAffinityFilter','PciPassthroughFilter','NUMATopologyFilter']

```

NUMATopologyFilter の詳細は、[Compute scheduler filters](#) を参照してください。

3. 専用のインスタンス用に物理 CPU コアを確保するには、以下の設定を **cpu_pinning.yaml** に追加します。

```
parameter_defaults:
  ComputeCPUPinningParameters:
    NovaComputeCpuDedicatedSet: 1,3,5,7
```

4. 共有のインスタンス用に物理 CPU コアを確保するには、以下の設定を **cpu_pinning.yaml** に追加します。

```
parameter_defaults:
  ComputeCPUPinningParameters:
    ...
    NovaComputeCpuSharedSet: 2,6
```

5. ホストのプロセス用に確保する RAM 容量を指定するには、以下の設定を **cpu_pinning.yaml** に追加します。

```
parameter_defaults:
  ComputeCPUPinningParameters:
    ...
    NovaReservedHostMemory: <ram>
```

<ram> を、確保するメモリー容量 (MB 単位) に置き換えます。

6. インスタンス用に確保した CPU コアでホストプロセスが実行されないようにするには、**IsolCpusList** パラメーターに、インスタンス用に確保した CPU コアを設定します。

```
parameter_defaults:
  ComputeCPUPinningParameters:
    ...
    IsolCpusList: 1-3,5-7
```

コンマ区切りの CPU インデックスのリストまたは範囲を使用して、**IsolCpusList** パラメーターの値を指定します。

7. その他の環境ファイルと共に新しいロールファイルおよび環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
  -e [your environment files] \
  -r /home/stack/templates/roles_data_cpu_pinning.yaml \
  -e /home/stack/templates/network-environment.yaml \
  -e /home/stack/templates/cpu_pinning.yaml \
  -e /home/stack/templates/node-info.yaml
```

3.1.4. インスタンス用の専用 CPU フレーバーの作成

クラウドユーザーが専用の CPU を持つインスタンスを作成できるようにするには、インスタンス起動用の専用 CPU ポリシーが設定されたフレーバーを作成します。

前提条件

- ホストで同時マルチスレッド (SMT) が有効である。
- コンピュートノードが CPU ピニングを許可するように設定されている。詳しくは、[コンピュートノードでの CPU ピニングの設定](#) を参照してください。

手順

1. source コマンドで **overcloudrc** ファイルを読み込みます。

```
(undercloud)$ source ~/overcloudrc
```

2. CPU ピニングを要求するインスタンス用のフレーバーを作成します。

```
(overcloud)$ openstack flavor create --ram <size_mb> \
--disk <size_gb> --vcpus <no_reserved_vcpus> pinned_cpus
```

3. ピニングされた CPU を要求するには、フレーバーの **hw:cpu_policy** 属性を **dedicated** に設定します。

```
(overcloud)$ openstack flavor set \
--property hw:cpu_policy=dedicated pinned_cpus
```

4. それぞれの仮想 CPU をスレッドシブリングに配置するには、フレーバーの **hw:cpu_thread_policy** 属性を **require** に設定します。

```
(overcloud)$ openstack flavor set \
--property hw:cpu_thread_policy=require pinned_cpus
```



注記

- ホストに SMT アーキテクチャがない場合や、スレッドシブリングが利用可能な CPU コアが十分でない場合には、スケジューリングが失敗します。これを回避するには、**hw:cpu_thread_policy** を **require** ではなく **prefer** に設定します。**prefer** ポリシーは、スレッドシブリングが利用可能な場合に使用されるデフォルトのポリシーです。
- **hw:cpu_thread_policy=isolate** を使用する場合は、SMT を無効にするか、SMT をサポートしないプラットフォームを使用する必要があります。

検証

1. フレーバーにより専用の CPU を持つインスタンスが作成されることを確認するには、新しいフレーバーを使用してインスタンスを起動します。

```
(overcloud)$ openstack server create --flavor pinned_cpus \
--image <image> pinned_cpu_instance
```

2. 新規インスタンスが正しく配置されていることを確認するには、以下のコマンドを入力し、その出力で **OS-EXT-SRV-ATTR:hypervisor_hostname** の箇所を確認します。

```
(overcloud)$ openstack server show pinned_cpu_instance
```

3.1.5. インスタンス用の共有 CPU フレーバーの作成

クラウドユーザーが共有の (あるいはフローティング) CPU を使用するインスタンスを作成できるようにするには、インスタンス起動用の共有 CPU ポリシーが設定されたフレーバーを作成します。

前提条件

- コンピュートノードが共有 CPU 用に物理 CPU コアを確保するように設定されている。詳しくは、[コンピュートノードでの CPU ピニングの設定](#) を参照してください。

手順

1. source コマンドで **overcloudrc** ファイルを読み込みます。

```
(undercloud)$ source ~/overcloudrc
```

2. CPU ピニングを要求しないインスタンス用のフレーバーを作成します。

```
(overcloud)$ openstack flavor create --ram <size_mb> \
--disk <size_gb> --vcpus <no_reserved_vcpus> floating_cpus
```

3. フローティング CPU を要求するには、フレーバーの **hw:cpu_policy** 属性を **shared** に設定します。

```
(overcloud)$ openstack flavor set \
--property hw:cpu_policy=shared floating_cpus
```

検証

1. フレーバーにより共有 CPU を使用するインスタンスが作成されることを確認するには、新しいフレーバーを使用してインスタンスを起動します。

```
(overcloud)$ openstack server create --flavor floating_cpus \
--image <image> floating_cpu_instance
```

2. 新規インスタンスが正しく配置されていることを確認するには、以下のコマンドを入力し、その出力で **OS-EXT-SRV-ATTR:hypervisor_hostname** の箇所を確認します。

```
(overcloud)$ openstack server show floating_cpu_instance
```

3.1.6. 同時マルチスレッド (SMT) 対応のコンピュートノードでの CPU ピニングの設定

コンピュートノードが同時マルチスレッド (SMT) をサポートする場合、スレッドシブリングを専用または共有セットのいずれかにグルーピングします。スレッドシブリングは共通のハードウェアを共有するため、あるスレッドシブリング上で動作しているプロセスが、他のスレッドシブリングのパフォーマンスに影響を与える可能性があります。

たとえば、ホストは、SMT 対応のデュアルコア CPU に 4 つの論理 CPU コア (0、1、2、および 3) を認識します。この 4 つの CPU に対して、スレッドシブリングのペアが 2 つあります。

- スレッドシブリング 1: 論理 CPU コア 0 および 2
- スレッドシブリング 2: 論理 CPU コア 1 および 3

このシナリオでは、論理 CPU コア 0 および 1 を専用として、2 および 3 を共有として割り当てないでください。そうではなく、0 および 2 を専用として、1 および 3 を共有として割り当てます。

/sys/devices/system/cpu/cpuN/topology/thread_siblings_list のファイル。N は論理 CPU 番号で、スレッドペアが含まれます。以下のコマンドを使用して、スレッドシブリングである論理 CPU コアを特定できます。

```
# grep -H . /sys/devices/system/cpu/cpu*/topology/thread_siblings_list | sort -n -t ':' -k 2 -u
```

以下の出力は、論理 CPU コア 0 と論理 CPU コア 2 が同じコア上のスレッドであることを示しています。

```
/sys/devices/system/cpu/cpu0/topology/thread_siblings_list:0,2
/sys/devices/system/cpu/cpu2/topology/thread_siblings_list:1,3
```

3.1.7. 関連情報

- ネットワーク機能仮想化 (NFV) のプランニングおよび設定ガイドの [NUMA ノードのトポロジについての理解](#)
- ネットワーク機能仮想化 (NFV) の製品ガイドの [CPU と NUMA ノード](#)

3.2. エミュレータスレッドの設定

コンピュートノードには、エミュレータスレッドと呼ばれる各インスタンスのハイパーバイザーとリンクするオーバーヘッドタスクがあります。デフォルトでは、エミュレータスレッドはインスタンスと同じ CPU で実行され、インスタンスのパフォーマンスに影響を及ぼします。

エミュレータスレッドポリシーを設定して、インスタンスが使用する CPU とは別の CPU でエミュレータスレッドを実行することができます。



注記

パケットロス为了避免のために、NFV デプロイメントでは絶対に仮想 CPU のプリエンプシオンを行わないでください。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. Compute 環境ファイルを開きます。
3. CPU ピニングを必要とするインスタンス用に物理 CPU コアを確保するには、Compute 環境ファイルで **NovaComputeCpuDedicatedSet** パラメーターを設定します。たとえば、以下の設定では、32 コア CPU を持つコンピュートノードに専用の CPU を設定します。

```
parameter_defaults:
...
NovaComputeCpuDedicatedSet: 2-15,18-31
...
```

詳しくは、[コンピュートノードでの CPU ピニングの設定](#) を参照してください。

4. エミュレータスレッド用に物理 CPU コアを確保するには、Compute 環境ファイルで **NovaComputeCpuSharedSet** パラメーターを設定します。たとえば、以下の設定では、32 コア CPU を持つコンピュートノードに共有の CPU を設定します。

```
parameter_defaults:
...
NovaComputeCpuSharedSet: 0,1,16,17
...
```



注記

Compute スケジューラーは、共有 (またはフローティング) CPU 上で動作するインスタンス用にも共有セット内の CPU を使用します。詳しくは、[コンピュートノードでの CPU ピニングの設定](#)を参照してください。

5. **NovaSchedulerDefaultFilters** パラメーターにまだ **NUMATopologyFilter** がなければ、この Compute スケジュールフィルターを追加します。
6. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

7. インスタンスのエミュレータスレッドを **NovaComputeCpuSharedSet** を使用して設定した共有 CPU から選択した専用の CPU 上で実行するフレーバーを設定します。

```
(overcloud)$ openstack flavor set --property hw:cpu_policy=dedicated \
--property hw:emulator_threads_policy=share \
dedicated_emulator_threads
```

hw:emulator_threads_policy の設定オプションについての詳しい情報は、[Flavor metadata の Emulator threads policy](#) を参照してください。

3.3. コンピュートノードでのヒュージページの設定

クラウド管理者は、インスタンスがヒュージページを要求できるようにコンピュートノードを設定することができます。

手順

1. Compute 環境ファイルを開きます。
2. インスタンスではないプロセス用に各 NUMA ノードで確保するヒュージページのメモリー量を設定します。

```
parameter_defaults:
  ComputeParameters:
    NovaReservedHugePages: ["node:0,size:1GB,count:1","node:1,size:1GB,count:1"]
```

- 各ノードの **size** の値を、割り当てられたヒュージページのサイズに置き換えます。以下の有効な値のいずれかに設定します。

- 2048 (2 MB 用)
- 1GB
- 各ノードの **count** の値を、NUMA ノードごとに OVS が使用するヒュージページの数に置き換えます。たとえば、Open vSwitch が 4096 のソケットメモリーを使用する場合、この属性を 2 に設定します。

3. コンピュートノードで Huge Page を設定します。

```
parameter_defaults:
  ComputeParameters:
    ...
    KernelArgs: "default_hugepagesz=1GB hugepagesz=1G hugepages=32"
```



注記

複数の Huge Page サイズを設定する場合、初回の起動時に Huge Page フォルダーもマウントする必要があります。詳細は、[Mounting multiple huge page folders during first boot](#) を参照してください。

- ### 4. (オプション) インスタンスが 1GB のヒュージページを割り当てるのを許可するには、CPU 機能フラグ **NovaLibvirtCPUModelExtraFlags** を設定して **pdpe1gb** を指定します。

```
parameter_defaults:
  ComputeParameters:
    NovaLibvirtCPUModel: 'custom'
    NovaLibvirtCPUModels: 'Haswell-noTSX'
    NovaLibvirtCPUModelExtraFlags: 'vmx, pdpe1gb'
```



注記

- インスタンスが 2 MB のヒュージページしか要求しない場合、CPU 機能フラグを設定する必要はありません。
- ホストが 1GB ヒュージページの割り当てをサポートする場合に限り、インスタンスに 1GB のヒュージページを割り当てることができます。
- **NovaLibvirtCPUModel** が **host-model** または **custom** に設定されている場合にのみ、**NovaLibvirtCPUModelExtraFlags** を **pdpe1gb** に設定する必要があります。
- ホストが **pdpe1gb** をサポートし、**NovaLibvirtCPUModel** に **host-passthrough** が使用される場合、**NovaLibvirtCPUModelExtraFlags** に **pdpe1gb** を設定する必要はありません。**pdpe1gb** フラグは Opteron_G4 および Opteron_G5 CPU モデルにのみ含まれ、QEMU がサポートする Intel CPU モデルには含まれません。
- Microarchitectural Data Sampling (MDS) 等の CPU ハードウェアの問題を軽減するには、他の CPU フラグを設定しなければならない場合があります。詳しくは、[RHOS Mitigation for MDS \("Microarchitectural Data Sampling"\) Security Flaws](#) を参照してください。

5. Meltdown に対する保護の適用後にパフォーマンスが失われないようにするには、CPU 機能フラグ **NovaLibvirtCPUModelExtraFlags** を設定して **+pcid** を指定します。

```
parameter_defaults:
  ComputeParameters:
    NovaLibvirtCPUModel: 'custom'
    NovaLibvirtCPUModels: 'Haswell-noTSX'
    NovaLibvirtCPUModelExtraFlags: 'vmx, pdpe1gb, +pcid'
```

ヒント

詳しくは、[Reducing the performance impact of Meltdown CVE fixes for OpenStack guests with "PCID" CPU feature flag](#) を参照してください。

6. **NovaSchedulerDefaultFilters** パラメーターにまだ **NUMATopologyFilter** がなければ、このフィルターを追加します。
7. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

3.3.1. インスタンス用のヒュージページフレイバーの作成

クラウドユーザーがヒュージページを使用するインスタンスを作成できるようにするには、インスタンス起動用の **hw:mem_page_size** 追加スペックキーが設定されたフレイバーを作成します。

前提条件

- コンピュートノードがヒュージページに対応する設定である。詳しくは、[Configuring huge pages on Compute nodes](#) を参照してください。

手順

1. ヒュージページを要求するインスタンス用のフレイバーを作成します。

```
$ openstack flavor create --ram <size_mb> --disk <size_gb> \
--vcpus <no_reserved_vcpus> huge_pages
```

2. ヒュージページを要求するには、フレイバーの **hw:mem_page_size** 属性を必要なサイズに設定します。

```
$ openstack flavor set huge_pages --property hw:mem_page_size=1GB
```

hw:mem_page_size を以下の有効な値のいずれかに設定します。

- **large**: ホストでサポートされる最大のページサイズを選択します。x86_64 システムでは 2 MB または 1 GB です。
- **small**: (デフォルト) ホストでサポートされる最小のページサイズを選択します。X86_64 システムでは、4 kB (通常のページ) です。

- **any**: libvirt ドライバーで決定される、利用可能な最大のヒュージページサイズを選択します。
 - `<pagesize>`: (文字列) ワークロードに具体的な要件がある場合、ページサイズを明示的に設定します。ページサイズには整数値を使用し、kB またはその他の標準単位で指定します。(例: 4kB、2MB、2048、1GB)。
3. フレーバーによりヒュージページを使用するインスタンスが作成されることを確認するには、新しいフレーバーを使用してインスタンスを起動します。

```
$ openstack server create --flavor huge_pages \
  --image <image> huge_pages_instance
```

Compute スケジューラーは、インスタンスのメモリーをサポートするのに十分なサイズの空きヒュージページを持つホストを特定します。スケジューラーが十分なページを持つホストおよび NUMA ノードを検出できない場合、リクエストは失敗して **NoValidHost** エラーが報告されます。

3.3.2. 初回起動時に複数の Huge Page フォルダーをマウント

Compute サービス (nova) を設定して、初回起動プロセスの一環として複数のページサイズを処理することができます。初回起動プロセスでは、初めてノードを起動する際に heat テンプレート設定をすべてのノードに追加します。これ以降は (たとえば、オーバークラウドスタックの更新時)、これらのテンプレートを追加してもこれらのスクリプトは実行されません。

手順

1. Huge Page フォルダーのマウントを作成するためにスクリプトを実行する初回起動テンプレートファイル **hugepages.yaml** を作成します。**OS::TripleO::MultipartMime** リソースタイプを使用して、この設定スクリプトを送信することができます。

```
heat_template_version: <version>

description: >
  Huge pages configuration

resources:
  userdata:
    type: OS::Heat::MultipartMime
    properties:
      parts:
        - config: {get_resource: hugepages_config}

  hugepages_config:
    type: OS::Heat::SoftwareConfig
    properties:
      config: |
        #!/bin/bash
        hostname | grep -qiE 'co?mp' || exit 0
        systemctl mask dev-hugepages.mount || true
        for pagesize in 2M 1G;do
          if ! [ -d "/dev/hugepages${pagesize}" ]; then
            mkdir -p "/dev/hugepages${pagesize}"
            cat << EOF > /etc/systemd/system/dev-hugepages${pagesize}.mount
[Unit]
```

```

Description=${pagesize} Huge Pages File System
Documentation=https://www.kernel.org/doc/Documentation/vm/hugetlbpage.txt
Documentation=https://www.freedesktop.org/wiki/Software/systemd/APIFileSystems
DefaultDependencies=no
Before=sysinit.target
ConditionPathExists=/sys/kernel/mm/hugepages
ConditionCapability=CAP_SYS_ADMIN
ConditionVirtualization=!private-users

```

```

[Mount]
What=hugetlbfs
Where=/dev/hugepages${pagesize}
Type=hugetlbfs
Options=pagesize=${pagesize}

```

```

[Install]
WantedBy = sysinit.target
EOF
fi
done
systemctl daemon-reload
for pagesize in 2M 1G;do
    systemctl enable --now dev-hugepages${pagesize}.mount
done

```

```

outputs:
  OS::stack_id:
    value: {get_resource: userdata}

```

このテンプレートの **config** スクリプトは、以下のタスクを実行します。

- a. **'co?mp'** に一致するホスト名を指定することで、ホストをフィルターリングして、huge page フォルダーのマウントを作成します。必要に応じて、特定のコンピュートの filter grep パターンを更新できます。
 - b. デフォルトの **dev-hugepages.mount systemd** ユニットファイルをマスクし、ページサイズを使用して新規マウントを作成できるようにします。
 - c. フォルダーが最初に作成されていることを確認します。
 - d. **pagesize** ごとに **systemd** マウントユニットを作成します。
 - e. 最初のループ後に **systemd daemon-reload** を実行し、新たに作成されたユニットファイルを追加します。
 - f. 2M および 1G のページサイズのマウントを有効にします。このループを更新して、必要に応じて追加のページサイズを含めることができます。
2. オプション: **/dev** フォルダーは、自動的に **nova_compute** および **nova_libvirt** のコンテナにバインドマウントされます。Huge Page マウントに異なる宛先を使用した場合は、そのマウントを **nova_compute** および **nova_libvirt** コンテナに渡す必要があります。

```

parameter_defaults
  NovaComputeOptVolumes:
    - /opt/dev:/opt/dev

```

```
NovaLibvirtOptVolumes:
  - /opt/dev:/opt/dev
```

3. ~/templates/firstboot.yaml 環境ファイルの **OS::TripleO::NodeUserData** リソースタイプとして heat テンプレートを登録します。

```
resource_registry:
  OS::TripleO::NodeUserData: ./hugepages.yaml
```



重要

NodeUserData リソースを登録することができるのは、1つのリソースにつき1つの heat テンプレートだけです。別の heat テンプレートに登録すると、使用する heat テンプレートがそのテンプレートに変わります。

4. 初回起動環境ファイルを他の環境ファイルと一緒にスタックに追加し、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
  -e [your environment files] \
  -e /home/stack/templates/firstboot.yaml \
  ...
```

3.4. インスタンスにファイルベースのメモリーを使用するコンピュートノードの設定

ファイルベースのメモリーを使用して、コンピュートノードのメモリー容量を拡張することができます。この場合、libvirt メモリーバックングディレクトリー内のファイルをインスタンスのメモリーとして割り当てます。インスタンスのメモリーとして使用できるホストディスクの容量、およびインスタンスメモリーファイルのディスク上の場所を設定することができます。

Compute サービスは、ファイルベースのメモリーに設定された容量をシステムメモリーの合計容量として Placement サービスに報告します。これにより、コンピュートノードは通常システムメモリー内で対応できるよりも多くのインスタンスをホストすることができます。

インスタンスにファイルベースのメモリーを使用するには、コンピュートノードでファイルベースのメモリーを有効にする必要があります。

制限

- ファイルベースのメモリーが有効なコンピュートノードとファイルベースのメモリーが有効ではないコンピュートノード間で、インスタンスのライブマイグレーションを行うことはできません。
- ファイルベースのメモリーとヒュージページとの間に互換性はありません。ファイルベースのメモリーが有効なコンピュートノード上で、ヒュージページを使用するインスタンスを起動することはできません。ホストアグリゲートを使用して、ヒュージページを使用するインスタンスがファイルベースのメモリーが有効なコンピュートノードに配置されないようにします。
- ファイルベースのメモリーとメモリーのオーバーコミットとの間に互換性はありません。
- **NovaReservedHostMemory** を使用してホストのプロセス用にメモリーを確保することはできません。ファイルベースのメモリーが使用されている場合、確保されるメモリーはファイル

ベースのメモリー用に用意されないディスク領域を表します。ファイルベースのメモリーは、キャッシュメモリーとして使用される RAM と共に、システムメモリーの合計として Placement サービスに報告されます。

前提条件

- ノードおよびノードが追加されたすべてのホストアグリゲートで、**NovaRAMAllocationRatio** を 1.0 に設定する必要があります。
- **NovaReservedHostMemory** を 0 に設定する必要があります。

手順

1. Compute 環境ファイルを開きます。
2. Compute 環境ファイルに以下のパラメーターを追加して、インスタンスの RAM 用に利用可能なホストディスク容量を MiB 単位で設定します。

```
parameter_defaults:
  NovaLibvirtFileBackedMemory: 102400
```

3. オプション: メモリーバックアップファイルを保存するディレクトリを設定するには、Compute 環境ファイルに **QemuMemoryBackingDir** パラメーターを設定します。設定しなければ、メモリーバックアップディレクトリはデフォルトの `/var/lib/libvirt/qemu/ram/` に設定されます。



注記

デフォルトのディレクトリ (`/var/lib/libvirt/qemu/ram/`) またはそれより上の階層のディレクトリに、バックアップストアを配置する必要があります。

バックアップストアのホストディスクを変更することもできます。詳細は、[Changing the memory backing directory host disk](#) を参照してください。

4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
  -e [your environment files] \
  -e /home/stack/templates/<compute_environment_file>.yaml
```

3.4.1. メモリーバックアップディレクトリのホストディスクの変更

デフォルトのプライマリーディスクから別のディスクに、メモリーバックアップディレクトリを変更することができます。

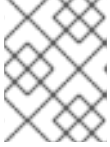
手順

1. 代替のバックアップデバイスにファイルシステムを作成します。たとえば、`/dev/sdb` に **ext4** ファイルシステムを作成するには、以下のコマンドを入力します。

```
# mkfs.ext4 /dev/sdb
```

2. バックアップデバイスをマウントします。たとえば、**/dev/sdb** をデフォルトの libvirt メモリーバックアップディレクトリーにマウントするには、以下のコマンドを入力します。

```
# mount /dev/sdb /var/lib/libvirt/qemu/ram
```



注記

マウントポイントは、**QemuMemoryBackingDir** パラメーターの値と一致する必要があります。

第4章 COMPUTE サービスのストレージの設定

Compute サービスが Image (glance) サービスからコピーしてコンピュートノード上でローカルにキャッシュしたベースイメージから、インスタンスを作成します。インスタンスのバックエンドであるインスタンスディスクも、ベースイメージに基づいています。

Compute サービスを設定して、一時インスタンスのディスクデータをホストのコンピュートノード上にローカルに保存するか、または NFS 共有または Ceph クラスタでリモートで保存することができます。あるいは、Compute サービスを設定して、Block Storage (Cinder) サービスが提供する永続ストレージにインスタンスディスクデータを保存することもできます。


環境のイメージキャッシュを設定し、インスタンスディスクのパフォーマンスおよびセキュリティーを設定することができます。

4.1. イメージキャッシュの設定オプション

以下の表で詳細を説明するパラメーターを使用して、Compute サービスがどのようにコンピュートノードでのイメージのキャッシュを実装して管理するかを設定します。

表4.1 Compute (nova) サービスのイメージキャッシュパラメーター

設定方法	パラメーター	説明
Puppet	nova::compute::image_cache::manager_interval	<p>コンピュートノードでのベースイメージのキャッシュを管理するイメージキャッシュマネージャーの実行間で待機する時間を秒単位で指定します。 nova::compute::image_cache::remove_unused_base_images が True に設定されている場合、Compute サービスはこの期間を使用して使用されていないキャッシュイメージの自動削除を実行します。</p> <p>デフォルトのメトリック間隔である 60 秒で実行するには、0 に設定します (推奨されません)。イメージキャッシュマネージャーを無効にするには、-1 に設定します。</p> <p>デフォルト: 2400</p>

設定方法	パラメーター	説明
Puppet	nova::compute::image_cache::precache_concurrency	<p>イメージを並行して事前キャッシュできるコンピュートノードの数の最大値を指定します。</p> <div>  <div> 注記 <ul style="list-style-type: none"> このパラメーターに大きな数値を設定すると、事前キャッシュの処理が遅くなり、Image サービスで DDoS が発生する場合があります。 このパラメーターに小さな数値を設定すると、Image サービスの負荷は軽減されますが、事前キャッシュがより連続的な操作で実行されるため、完了までの実行時間が長くなる場合があります。 </div> </div> <p>デフォルト: 1</p>
Puppet	nova::compute::image_cache::remove_unused_base_images	<p>manager_interval を使用して設定された間隔で使用されていないベースイメージをキャッシュから自動的に削除するには、True に設定します。NovalImageCacheTTL を使用して指定された期間アクセスされなかった場合、イメージは使用されていないと定義されます。</p> <p>デフォルト: True</p>
Puppet	nova::compute::image_cache::remove_unused_resized_minimum_age_seconds	<p>未使用のサイズ変更されたベースイメージをキャッシュから削除する最短の期間を指定します (秒単位)。未使用のサイズ変更されたベースイメージは、この期間削除されません。無効にする場合は undef に設定します。</p> <p>デフォルト: 3600</p>

設定方法	パラメーター	説明
Puppet	nova::compute::image_cache::subdirectory_name	<p>キャッシュされたイメージを保存するフォルダーの名前を、\$instances_path との相対パスで指定します。</p> <p>デフォルト: _base</p>
heat	NovalImageCacheTTL	<p>コンピュートノード上のどのインスタンスもイメージを使用しなくなった場合に、Compute サービスがイメージをキャッシュし続ける期間を秒単位で指定します。Compute サービスは、コンピュートノードにキャッシュされたイメージのうち、ここで設定したライフタイムを過ぎたイメージを、再度必要になるまでキャッシュディレクトリーから削除します。</p> <p>デフォルト: 86400 (24 時間)</p>

4.2. インスタンスの一時ストレージ属性の設定オプション

以下の表で詳細を説明するパラメーターを使用して、インスタンスが使用する一時ストレージのパフォーマンスおよびセキュリティーを設定します。



注記


Red Hat OpenStack Platform (RHOSP) は、インスタンスディスクの LVM イメージ種別をサポートしません。したがって、インスタンスの削除時に一時ディスクを消去する **[libvirt]/volume_clear** 設定オプションは、インスタンスディスクイメージ種別が LVM の場合にのみ適用されるため、サポートされません。

表4.2 Compute (nova) サービスのインスタンス一時ストレージパラメーター

設定方法	パラメーター	説明
Puppet	nova::compute::default_ephemeral_format	<p>新規の一時ボリュームに使用されるデフォルトの形式を指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● ext2 ● ext3 ● ext4 <p>ext4 形式では、ext3 に比べ、サイズの大きい新規ディスクを初期化する時間が大幅に短縮されます。</p> <p>デフォルト: ext4</p>

設定方法	パラメーター	説明
Puppet	nova::compute::force_raw_images	<p>raw 以外の形式でキャッシュされたベースイメージを raw 形式に変換するには、True に設定します。raw イメージ形式は、qcow2 等の他のイメージ形式よりも多くの領域を使用します。raw 以外のイメージ形式は、圧縮により多くの CPU パワーを使用します。False に設定すると、CPU のボトルネックを防ぐために、Compute サービスは圧縮時にベースイメージからすべての圧縮を削除します。システムの I/O 処理が遅い場合や空き容量が少ない場合は、入力の帯域幅を削減するために False に設定します。</p> <p>デフォルト: True</p>
Puppet	nova::compute::use_cow_images	<p>インスタンスのディスクに qcow2 形式の CoW (Copy on Write) イメージを使用するには、True に設定します。CoW を使用する場合には、バックングストアとホストのキャッシュによっては、各インスタンスが独自のコピー上で稼働することで、並行処理が改善される場合があります。</p> <p>raw 形式を使用するには、False に設定します。raw 形式は、ディスクイメージの共通の部分により多くの領域を使用します。</p> <p>デフォルト: True</p>
Puppet	nova::compute::libvirt::preallocate_images	<p>インスタンスディスクに対する事前割り当てモードを指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● none: インスタンスの起動時にはストレージがプロビジョニングされません。 ● space: インスタンスのディスクイメージで fallocate(1) を実行することで、Compute サービスはインスタンスの起動時にストレージを完全に割り当てます。これにより CPU のオーバーヘッドおよびファイルの断片化が軽減され、I/O パフォーマンスが向上し、要求されたディスク領域を確保するのに役立ちます。 <p>デフォルト: none</p>

設定方法	パラメーター	説明
hieradata override	DEFAULT/resize_fs_using_block_device	<p>ブロックデバイスを使用してイメージにアクセスすることで、ベースイメージのサイズを直接変更できるようにするには、True に設定します。これは、cloud-init のバージョンがより古いイメージ (それ自体ではサイズの変更ができない) でのみ必要です。</p> <p>このパラメーターによりイメージの直接マウントが可能になるため (セキュリティ上の理由により無効にされる場合がある)、このパラメーターはデフォルトでは有効化されていません。</p> <p>デフォルト: False</p>

設定方法	パラメーター	説明
hieradata override	[libvirt]/images_type	<p>インスタンスのディスクに使用するイメージ種別を指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● raw ● qcow2 ● flat ● rbd ● default <div>  <div> <p>注記</p> <p>RHOSP は、インスタンスディスクの LVM イメージ種別をサポートしません。</p> </div> </div> <p>default 以外の有効な値に設定すると、イメージ種別は use_cow_images の設定よりも優先されます。default が指定されると、use_cow_images の設定によりイメージ種別が決定されます。</p> <ul style="list-style-type: none"> ● use_cow_images が True(デフォルト) に設定されると、イメージ種別は qcow2 になります。 ● use_cow_images を False に設定すると、イメージ種別は Flat になります。 <p>デフォルト値は NovaEnableRbdBackend の設定により決定されます。</p> <ul style="list-style-type: none"> ● NovaEnableRbdBackend: False デフォルト: default ● NovaEnableRbdBackend: True デフォルト: rbd

4.3. 共有インスタンスストレージの設定

デフォルトでは、インスタンスの起動時に、インスタンスのディスクはインスタンスディレクトリー **/var/lib/nova/instances** にファイルとして保存されます。Compute サービスの NFS ストレージバックエンドを設定して、これらのインスタンスファイルを共有 NFS ストレージに保存することができます。

前提条件

- NFSv4 以降を使用している。Red Hat OpenStack Platform (RHOSP) は、以前のバージョンの NFS をサポートしません。詳細は、Red Hat ナレッジベースのソリューション [RHOS NFSv4-Only Support Notes](#) を参照してください。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. 共有インスタンスストレージを設定するための環境ファイルを作成します (例: **nfs_instance_disk_backend.yaml**)。
4. インスタンスファイル用に NFS バックエンドを設定するには、以下の設定を **nfs_instance_disk_backend.yaml** に追加します。

```
parameter_defaults:
...
NovaNfsEnabled: True
NovaNfsShare: <nfs_share>
```

<nfs_share> をインスタンスのファイルストレージ用にマウントする NFS 共有ディレクトリーに置き換えます (例: **'192.168.122.1:/export/nova'** または **'192.168.24.1:/var/nfs'**)。IPv6 を使用している場合は、二重と単一引用符の両方を使用してください (例: **""[fdd0::1]:/export/nova""**)。

5. (オプション) NFS バックエンドストレージが有効な場合、NFS ストレージ用のデフォルトのマウント SELinux コンテキストは **'context=system_u:object_r:nova_var_lib_t:s0'** です。以下のパラメーターを追加して、NFS インスタンスファイルストレージのマウントポイントのマウントオプションを変更します。

```
parameter_defaults:
...
NovaNfsOptions: 'context=system_u:object_r:nova_var_lib_t:s0,
<additional_nfs_mount_options>'
```

<additional_nfs_mount_options> を、NFS インスタンスファイルストレージに使用するマウントオプションのコンマ区切りリストに置き換えます。利用可能なマウントオプションの詳細は、**mount** の man ページを参照してください。

```
$ man 8 mount.
```

6. 更新を環境ファイルに保存します。
7. その他の環境ファイルと共に新しい環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/nfs_instance_disk_backend.yaml
```

4.4. 関連情報

- [Compute サービス \(nova\) の設定](#)

第5章 PCI パススルーの設定

PCI パススルーを使用して、グラフィックカードまたはネットワークデバイス等の物理 PCI デバイスをインスタンスにアタッチすることができます。デバイスに PCI パススルーを使用する場合、インスタンスはタスクを実行するためにデバイスへの排他的アクセスを確保し、ホストはデバイスを利用することができません。

重要

ルーティング対応プロバイダーネットワークでの PCI パススルーの使用

Compute サービスは、複数のプロバイダーネットワークにまたがる単一のネットワークをサポートしません。ネットワークに複数の物理ネットワークが含まれる場合、Compute サービスは最初の物理ネットワークだけを使用します。したがって、ルーティング対応プロバイダーネットワークを使用する場合は、すべてのコンピュートノードで同じ **physical_network** 名を使用する必要があります。

VLAN またはフラットネットワークのルーティング対応プロバイダーネットワークを使用する場合は、すべてのセグメントで同じ **physical_network** 名を使用する必要があります。その後、ネットワークに複数のセグメントを作成し、そのセグメントを適切なサブネットにマッピングします。

クラウドユーザーが PCI デバイスがアタッチされたインスタンスを作成できるようにするには、以下の手順を実施する必要があります。

1. PCI パススルー用のコンピュートノードを指定する。
2. 必要な PCI デバイスを持つ PCI パススルー用のコンピュートノードを設定する。
3. オーバークラウドをデプロイする。
4. PCI デバイスがアタッチされたインスタンスを起動するためのフレーバーを作成する。

前提条件

- 必要な PCI デバイスを持つコンピュートノード

5.1. PCI パススルー用コンピュートノードの指定

物理 PCI デバイスがアタッチされたインスタンス用にコンピュートノードを指定するには、PCI パススルーロールを設定するための新規ロールファイルを作成し、PCI パススルーのためにコンピュートノードをタグ付けするための新規オーバークラウドフレーバーおよび PCI パススルーリソースクラスを設定する必要があります。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. **roles_data_pci_passthrough.yaml** という名前で、**Controller**、**Compute**、および **ComputePCI** ロールが含まれる新しいロールデータファイルを生成します。

```
(undercloud)$ openstack overcloud roles \
generate -o /home/stack/templates/roles_data_pci_passthrough.yaml \
Compute:ComputePCI Compute Controller
```

4. **roles_data_pci_passthrough.yaml** を開き、以下のパラメーターおよびセクションを編集または追加します。

セクション/パラメーター	現在の値	新しい値
ロールのコメント	Role: Compute	Role: ComputePCI
ロール名	Compute	name: ComputePCI
description	Basic Compute Node role	PCI パススルー用コンピュートノードロール
HostnameFormatDefault	%stackname%-novacompute-%index%	%stackname%-novacomputepci-%index%
deprecated_nic_config_name	compute.yaml	compute-pci-passthrough.yaml

5. オーバークラウドの PCI パススルー用コンピュートノードをノード定義のテンプレート **node.json** または **node.yaml** に追加して、そのノードを登録します。詳細は、**director** のインストールと使用方法の [オーバークラウドノードの登録](#) を参照してください。
6. ノードのハードウェアを検査します。

```
(undercloud)$ openstack overcloud node introspect \
--all-manageable --provide
```

詳細は、**Director Installation and Usage** ガイドの [Creating an inventory of the bare-metal node hardware](#) を参照してください。

7. PCI パススルーを使用するコンピュートノード用の **compute-pci-passthrough** オーバークラウドフレーバーを作成します。

```
(undercloud)$ openstack flavor create --id auto \
--ram <ram_size_mb> --disk <disk_size_gb> \
--vcpus <no_vcpus> compute-pci-passthrough
```

- **<ram_size_mb>** をベアメタルノードの RAM (MB 単位) に置き換えます。
- **<disk_size_gb>** をベアメタルノード上のディスク容量 (GB 単位) に置き換えます。
- **<no_vcpus>** をベアメタルノードの CPU 数に置き換えます。



注記

これらの属性は、インスタンスのスケジューリングには使用されません。ただし Compute スケジューラーは、ディスク容量を使用してルートパーティションのサイズを決定します。

8. ノード一覧を取得して UUID を把握します。

```
(undercloud)$ openstack baremetal node list
```

9. PCI パススルー用に指定する各ベアメタルノードに、カスタムの PCI パススルーリソースクラスをタグ付けします。

```
(undercloud)$ openstack baremetal node set \
--resource-class baremetal.PCI-PASSTHROUGH <node>
```

<node> をベアメタルノードの ID に置き換えてください。

10. **compute-pci-passthrough** フレーバーをカスタムの PCI パススルーリソースクラスに関連付けます。

```
(undercloud)$ openstack flavor set \
--property resources:CUSTOM_BAREMETAL_PCI_PASSTHROUGH=1 \
compute-pci-passthrough
```

Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、すべての句読点をアンダースコアに置き換え、**CUSTOM_** の接頭辞を追加します。



注記

フレーバーが要求できるのは、ベアメタルリソースクラスの1つのインスタンスだけです。

11. 以下のフレーバー属性を設定して、Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぎます。

```
(undercloud)$ openstack flavor set \
--property resources:VCPU=0 --property resources:MEMORY_MB=0 \
--property resources:DISK_GB=0 compute-pci-passthrough
```

12. 以下のパラメーターを **node-info.yaml** ファイルに追加して、PCI パススルー用コンピュートノードの数および PCI パススルー対応コンピュートノードに使用するフレーバーを指定します。

```
parameter_defaults:
  OvercloudComputePCIFlavor: compute-pci-passthrough
  ComputePCICount: 3
```

13. ロールが作成されたことを確認するには、以下のコマンドを入力します。

```
(undercloud)$ openstack overcloud profiles list
```

5.2. PCI パススルー用コンピュートノードの設定

クラウドユーザーが PCI デバイスがアタッチされたインスタンスを作成できるようにするには、PCI デバイスを持つコンピュートノードとコントローラーノードの両方を設定する必要があります。

手順

1. PCI パススルー用にオーバークラウド上のコントローラーノードを設定するには、環境ファイル (例: **pci_passthrough_controller.yaml**) を作成します。
2. **pci_passthrough_controller.yaml** の **NovaSchedulerDefaultFilters** パラメーターに **PciPassthroughFilter** を追加します。

```
parameter_defaults:
  NovaSchedulerDefaultFilters:
    ['AvailabilityZoneFilter','ComputeFilter','ComputeCapabilitiesFilter','ImagePropertiesFilter','ServerGroupAntiAffinityFilter','ServerGroupAffinityFilter','PciPassthroughFilter','NUMATopologyFilter']
```

3. コントローラーノード上のデバイスの PCI エイリアスを指定するには、以下の設定を **pci_passthrough_controller.yaml** に追加します。

```
parameter_defaults:
  ...
  ControllerExtraConfig:
    nova::pci::aliases:
      - name: "a1"
        product_id: "1572"
        vendor_id: "8086"
        device_type: "type-PF"
```

device_type フィールドの設定に関する詳細は、[PCI パススルーデバイス種別フィールド](#) を参照してください。



注記

nova-api サービスが **Controller** 以外のロールで実行されている場合は、**ControllerExtraConfig** を **<Role>ExtraConfig** の形式でユーザーロールに置き換えます。

4. (オプション): PCI パススルーデバイスにデフォルトの NUMA アフィニティーポリシーを設定するには、ステップ 3 の **numa_policy** 設定に **nova::pci::aliases:** を追加します。

```
parameter_defaults:
  ...
  ControllerExtraConfig:
    nova::pci::aliases:
      - name: "a1"
        product_id: "1572"
        vendor_id: "8086"
        device_type: "type-PF"
        numa_policy: "preferred"
```

5. PCI パススルー用にオーバークラウド上のコンピュートノードを設定するには、環境ファイル (例: **pci_passthrough_compute.yaml**) を作成します。
6. コンピュートノード上のデバイスの利用可能な PCI を指定するには、**vendor_id** および **product_id** オプションを使用して、インスタンスへのパススルーに使用できる PCI デバイスのプールに、一致するすべての PCI デバイスを追加します。たとえば、すべての Intel® Ethernet Controller X710 デバイスをインスタンスへのパススルーに使用できる PCI デバイスのプールに追加するには、以下の設定を **pci_passthrough_compute.yaml** に追加します。

```
parameter_defaults:
...
ComputePCIParameters:
  NovaPCIPassthrough:
    - vendor_id: "8086"
      product_id: "1572"
```

NovaPCIPassthrough の設定方法の詳細は、[NovaPCIPassthrough 設定のガイドライン](#) を参照してください。

7. インスタンスの移行およびサイズ変更の操作を行うために、コンピュートノードの PCI エイリアスのコピーを作成する必要があります。PCI パススルー用コンピュートノード上のデバイスの PCI エイリアスを指定するには、以下の設定を **pci_passthrough_compute.yaml** に追加します。

```
parameter_defaults:
...
ComputePCIExtraConfig:
  nova::pci::aliases:
    - name: "a1"
      product_id: "1572"
      vendor_id: "8086"
      device_type: "type-PF"
```



注記

コンピュートノードのエイリアスは、コントローラーノードのエイリアスと同じでなければなりません。したがって、**pci_passthrough_controller.yaml** の **nova::pci::aliases** に **numa_affinity** を追加した場合は、**pci_passthrough_compute.yaml** の **nova::pci::aliases** にも追加する必要があります。

8. PCI パススルーをサポートするためにコンピュートノードのサーバー BIOS で IOMMU を有効にするには、**pci_passthrough_compute.yaml** に **KernelArgs** パラメーターを追加します。たとえば、Intel IOMMU を有効にするには、以下の **KernelArgs** 設定を使用します。

```
parameter_defaults:
...
ComputePCIParameters:
  KernelArgs: "intel_iommu=on iommu=pt"
```

AMD IOMMU を有効にするには、**KernelArgs** を **"amd_iommu=on iommu=pt"** に設定します。

9. その他の環境ファイルと共にこれらのカスタム環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/pci_passthrough_controller.yaml \
-e /home/stack/templates/pci_passthrough_compute.yaml \
```

10. ソフトウェアがPCIデバイスを要求するのに使用するフレーバーを作成および設定します。以下の例では、ステップ7で定義したエイリアスを使用して、それぞれベンダー ID および製品 ID が **8086** および **1572** の2つのデバイスを要求します。

```
(overcloud)# openstack flavor set \
--property "pci_passthrough:alias"="a1:2" device_passthrough
```

11. (オプション) フレーバーまたはイメージに NUMA アフィニティポリシーの属性キーを追加して、PCI パススルーデバイスのデフォルト NUMA アフィニティポリシーをオーバーライドすることができます。

- フレーバーを使用してデフォルトの NUMA アフィニティポリシーをオーバーライドするには、**hw:pci_numa_affinity_policy** 属性キーを追加します。

```
(overcloud)# openstack flavor set \
--property "hw:pci_numa_affinity_policy"="required" \
device_passthrough
```

hw:pci_numa_affinity_policy の有効な値についての詳しい情報は、[フレーバーのメタデータ](#) を参照してください。

- イメージを使用してデフォルトの NUMA アフィニティポリシーをオーバーライドするには、**hw_pci_numa_affinity_policy** 属性キーを追加します。

```
(overcloud)# openstack image set \
--property hw_pci_numa_affinity_policy=required \
device_passthrough_image
```



注記

イメージとフレーバーの両方で NUMA アフィニティポリシーを設定する場合には、属性の値が一致している必要があります。フレーバーの設定は、イメージおよびデフォルトの設定よりも優先されます。したがって、イメージの NUMA アフィニティポリシーの設定は、フレーバーで属性が設定されていない場合に限り効果を持ちます。

検証

1. PCI パススルーデバイスを設定してインスタンスを作成します。

```
# openstack server create --flavor device_passthrough \
--image <image> --wait test-pci
```

2. クラウドユーザーとしてインスタンスにログインします。詳細は、[Connecting to an instance](#) を参照してください。
3. インスタンスが PCI デバイスにアクセスできることを確認するには、インスタンスから以下のコマンドを入力します。

```
$ lspci -nn | grep <device_name>
```

5.3. PCI パススルーデバイス種別フィールド

Compute サービスでは、デバイスが報告する機能に応じて、PCI デバイスは 3 つの種別のいずれかに分類されます。**device_type** フィールドに設定することのできる有効な値を、以下の一覧に示します。

type-PF

デバイスは、SR-IOV をサポートする親またはルートデバイスです。SR-IOV を完全にサポートするデバイスをパススルーするには、このデバイス種別を指定します。

type-VF

デバイスは、SR-IOV をサポートするデバイスの子デバイスです。

type-PCI

デバイスは SR-IOV をサポートしません。**device_type** フィールドが設定されていない場合は、これがデフォルトのデバイス種別です。



注記

コンピュートノードとコントローラーノードの **device_type** を、同じ値に設定する必要があります。

5.4. NOVA PCIPASSTHROUGH 設定のガイドライン

- NIC のデバイス名は変更される可能性があるため、PCI パススルーを設定する場合は **devname** パラメーターを使用しないでください。代わりに、**vendor_id** と **product_id** の方が安定しているため使用するか、NIC の **address** を使用してください。
- 特定の Physical Function (PF) をパススルーするには、PCI アドレスが各デバイスに固有であるので、**address** パラメーターを使用できます。または、**product_id** パラメーターを使用して PF をパススルーすることもできますが、同じ種別の PF が複数ある場合には、PF の **address** も指定する必要があります。
- すべての Virtual Function (VF) をパススルーするには、PCI パススルーに使用する VF の **product_id** および **vendor_id** のみを指定します。NIC の分割に SRIOV を使用し、VF 上で OVS を実行している場合は、VF の **address** も指定する必要があります。
- PF の VF のパススルーだけを設定し、PF そのもののパススルーは設定しない場合は、**address** パラメーターを使用して PF の PCI アドレスを指定し、**product_id** を使用して VF の製品 ID を指定することができます。

address パラメーターの設定

address パラメーターは、デバイスの PCI アドレスを指定します。String または **dict** マッピングのいずれかを使用して、**address** パラメーターの値を設定できます。

文字列形式

文字列を使用してアドレスを指定する場合は、以下の例のようにワイルドカード (*) を含めることができます。

```
NovaPCIPassthrough:
-
  address: "*:0a:00.*"
  physical_network: physnet1
```

ディクショナリー形式

ディクショナリー形式を使用してアドレスを指定する場合は、以下の例のように正規表現構文を含めることができます。

NovaPCIPassthrough:

-

address:

domain: ".*"

bus: "02"

slot: "01"

function: "[0-2]"

physical_network: net1

第6章 ホストアグリゲートの作成と管理

クラウド管理者は、パフォーマンスおよび管理目的で、コンピュートのデプロイメントを論理グループに分割することができます。Red Hat OpenStack Platform (RHOSP) では、論理グループへの分割に以下のメカニズムを使用することができます。

ホストアグリゲート

ホストアグリゲートとは、ハードウェアやパフォーマンス特性などの属性に基づいてコンピュートノードを論理的なユニットにグループ化したものです。コンピュートノードを1つまたは複数のホストアグリゲートに割り当てることができます。

ホストアグリゲートでメタデータを設定してフレーバーおよびイメージをホストアグリゲートにマッピングし、続いてフレーバーの追加スペックまたはイメージのメタデータ属性をホストアグリゲートのメタデータにマッチングすることができます。必要なフィルターが有効な場合、Compute スケジューラーはこのメタデータを使用してインスタンスのスケジューリングを行うことができます。ホストアグリゲートで指定するメタデータは、ホストの使用をフレーバーまたはイメージで指定するメタデータと同じメタデータのインスタンスに限定します。

ホストアグリゲートのメタデータで **xxx_weight_multiplier** 設定オプションを定義することで、それぞれのホストアグリゲートに重みの乗数を設定することができます。

ホストアグリゲートを使用して、負荷分散の処理、物理的な分離または冗長性の適用、共通の属性を持つサーバーのグループ化、ハードウェアのクラス分け等を行うことができます。

ホストアグリゲートを作成する際に、ゾーン名を指定することができます。この名前は、クラウドユーザーが選択することのできるアベイラビリティゾーンとして提示されます。

アベイラビリティゾーン

アベイラビリティゾーンは、ホストアグリゲートのクラウドユーザー側のビューです。クラウドユーザーは、アベイラビリティゾーンに属するコンピュートノードやアベイラビリティゾーンのメタデータを把握することはできません。クラウドユーザーは、アベイラビリティゾーンの名前を見ることしかできません。

それぞれのコンピュートノードは、1つのアベイラビリティゾーンにしか割り当てることができます。デフォルトのアベイラビリティゾーンを設定することができます。クラウドユーザーがゾーンを指定しない場合、インスタンスはこのアベイラビリティゾーンにスケジューリングされます。特定の機能を持つアベイラビリティゾーンを使用するように、クラウドユーザーに指示することができます。

6.1. ホストアグリゲートでのスケジューリングの有効化

特定の属性を持つホストアグリゲートにインスタンスをスケジュールするには、Compute スケジューラーの設定を更新し、ホストアグリゲートのメタデータに基づく絞り込みを有効にします。

手順

1. Compute 環境ファイルを開きます。
2. **NovaSchedulerDefaultFilters** パラメーターにまだ以下の値がなければ、値を追加します。
 - **AggregateInstanceExtraSpecsFilter**: フレーバーの追加スペックに一致するホストアグリゲートメタデータでコンピュートノードを絞り込むには、この値を追加します。



注記

このフィルターが想定どおりに機能するには、**extra_specs** キーに **aggregate_instance_extra_specs**: 名前空間の接頭辞を指定して、フレーバーの追加スペックのスコープを定義する必要があります。

- **AggregateImagePropertiesIsolation**: イメージメタデータ属性に一致するホストアグリゲートメタデータでコンピュートノードを絞り込むには、この値を追加します。



注記

イメージメタデータ属性を使用してホストアグリゲートのメタデータを絞り込むには、ホストアグリゲートメタデータキーが有効なイメージメタデータ属性と一致する必要があります。有効なイメージメタデータプロパティの詳細は、**Creating and Managing Images** ガイドの [Image metadata](#) を参照してください。

- **AvailabilityZoneFilter**: インスタンスの起動時にアベイラビリティゾーンで絞り込むには、この値を追加します。



注記

Compute スケジューラーサービスのフィルター **AvailabilityZoneFilter** を使用する代わりに、Placement サービスを使用してアベイラビリティゾーンの要求を処理することができます。詳細は、[Filtering by availability zone using the Placement service](#) を参照してください。

3. 更新内容を Compute 環境ファイルに保存します。
4. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

6.2. ホストアグリゲートの作成

クラウド管理者は、ホストアグリゲートを必要なだけ作成することができます。

手順

1. ホストアグリゲートを作成するには、以下のコマンドを入力します。

```
(overcloud)# openstack aggregate create <aggregate_name>
```

<aggregate_name> をホストアグリゲートに割り当てる名前に置き換えてください。

2. ホストアグリゲートにメタデータを追加します。

```
(overcloud)# openstack aggregate set \
--property <key=value> \
--property <key=value> \
```

```
<aggregate_name>
```

- **<key=value>** はメタデータのキー/値のペアに置き換えてください。**AggregateInstanceExtraSpecsFilter** フィルターを使用している場合、キーは任意の文字列 (例: **ssd=true**) にすることができます。**AggregateImagePropertiesIsolation** フィルターを使用している場合は、キーは有効なイメージメタデータ属性と一致する必要があります。有効なイメージメタデータ属性に関する詳細は、[Image metadata](#) を参照してください。
- **<aggregate_name>** をホストアグリゲートの名前に置き換えてください。

3. コンピュートノードをホストアグリゲートに追加します。

```
(overcloud)# openstack aggregate add host \
  <aggregate_name> \
  <host_name>
```

- **<aggregate_name>** は、コンピュートノードを追加するホストアグリゲートの名前に置き換えます。
- **<host_name>** は、ホストアグリゲートに追加するコンピュートノードの名前に置き換えてください。

4. ホストアグリゲート用のフレーバーまたはイメージを作成します。

- フレーバーの作成

```
(overcloud)$ openstack flavor create \
  --ram <size_mb> \
  --disk <size_gb> \
  --vcpus <no_reserved_vcpus> \
  host-agg-flavor
```

- イメージの作成

```
(overcloud)$ openstack image create host-agg-image
```

5. フレーバーまたはイメージに、ホストアグリゲートのキー/値のペアに一致するキー/値のペアを1つまたは複数設定します。

- フレーバーにキー/値のペアを設定するには、スコープ **aggregate_instance_extra_specs** を使用します。

```
(overcloud)# openstack flavor set \
  --property aggregate_instance_extra_specs:ssd=true \
  host-agg-flavor
```

- イメージにキー/値のペアを設定するには、有効なイメージメタデータ属性をキーとして使用します。

```
(overcloud)# openstack image set \
  --property os_type=linux \
  host-agg-image
```

6.3. アベイラビリティーゾーンの作成

クラウド管理者は、クラウドユーザーがインスタンスを作成する際に選択できるアベイラビリティーゾーンを作成することができます。

手順

1. アベイラビリティーゾーンを作成するには、新しいアベイラビリティーゾーンホストアグリゲートを作成するか、既存のホストアグリゲートをアベイラビリティーゾーンにすることができます。

- a. 新しいアベイラビリティーゾーンホストアグリゲートを作成するには、以下のコマンドを入力します。

```
(overcloud)# openstack aggregate create \
--zone <availability_zone> \
<aggregate_name>
```

- **<availability_zone>** をアベイラビリティーゾーンに割り当てる名前に置き換えてください。
- **<aggregate_name>** をホストアグリゲートに割り当てる名前に置き換えてください。

- b. 既存のホストアグリゲートをアベイラビリティーゾーンにするには、以下のコマンドを入力します。

```
(overcloud)# openstack aggregate set --zone <availability_zone> \
<aggregate_name>
```

- **<availability_zone>** をアベイラビリティーゾーンに割り当てる名前に置き換えてください。
- **<aggregate_name>** をホストアグリゲートの名前に置き換えてください。

2. オプション: アベイラビリティーゾーンにメタデータを追加します。

```
(overcloud)# openstack aggregate set --property <key=value> \
<aggregate_name>
```

- **<key=value>** をメタデータのキー/値のペアに置き換えてください。キー/値の属性は、必要なだけ追加することができます。
- **<aggregate_name>** をアベイラビリティーゾーンホストアグリゲートの名前に置き換えてください。

3. コンピュートノードをアベイラビリティーゾーンホストアグリゲートに追加します。

```
(overcloud)# openstack aggregate add host <aggregate_name> \
<host_name>
```

- **<aggregate_name>** は、コンピュートノードを追加するアベイラビリティーゾーンホストアグリゲートの名前に置き換えてください。
- **<host_name>** は、アベイラビリティーゾーンに追加するコンピュートノードの名前に置き換えてください。

6.4. ホストアグリゲートの削除

ホストアグリゲートを削除するには、まずホストアグリゲートからすべてのコンピュートノードを削除します。

手順

1. ホストアグリゲートに割り当てられたすべてのコンピュートノードの一覧を表示するには、以下のコマンドを入力します。

```
(overcloud)# openstack aggregate show <aggregate_name>
```

2. ホストアグリゲートから割り当てられたすべてのコンピュートノードを削除するには、それぞれのコンピュートノードで以下のコマンドを入力します。

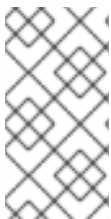
```
(overcloud)# openstack aggregate remove host <aggregate_name> \
<host_name>
```

- **<aggregate_name>** は、コンピュートノードを削除するホストアグリゲートの名前に置き換えてください。
 - **<host_name>** は、ホストアグリゲートから削除するコンピュートノードの名前に置き換えてください。
3. ホストアグリゲートからすべてのコンピュートノードを削除したら、以下のコマンドを入力してホストアグリゲートを削除します。

```
(overcloud)# openstack aggregate delete <aggregate_name>
```

6.5. プロジェクト分離ホストアグリゲートの作成

特定のプロジェクトでのみ利用可能なホストアグリゲートを作成することができます。ホストアグリゲートに割り当てたプロジェクトだけが、ホストアグリゲートでインスタンスを起動することができます。



注記

プロジェクト分離では、Placement サービスを使用して各プロジェクトのホストアグリゲートを絞り込みます。このプロセスは、**AggregateMultiTenancyIsolation** フィルターの機能に優先します。したがって、**AggregateMultiTenancyIsolation** フィルターを使用する必要はありません。

手順

1. Compute 環境ファイルを開きます。
2. プロジェクト分離ホストアグリゲートでプロジェクトインスタンスをスケジュールするには、Compute 環境ファイルの **NovaSchedulerLimitTenantsToPlacementAggregate** パラメーターを **True** に設定します。
3. オプション: ホストアグリゲートに割り当てたプロジェクトだけがクラウド上でインスタンスを作成できるようにするには、**NovaSchedulerPlacementAggregateRequiredForTenants** パラメーターを **True** に設定します。



注記

NovaSchedulerPlacementAggregateRequiredForTenants のデフォルト値は **False** です。このパラメーターが **False** の場合、ホストアグリゲートに割り当てられていないプロジェクトは、任意のホストアグリゲートでインスタンスを作成することができます。

4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml \
```

6. ホストアグリゲートを作成します。
7. プロジェクト ID の一覧を取得します。

```
(overcloud)# openstack project list
```

8. **filter_tenant_id<suffix>** メタデータキーを使用して、プロジェクトをホストアグリゲートに割り当てます。

```
(overcloud)# openstack aggregate set \
--property filter_tenant_id<ID0>=<project_id0> \
--property filter_tenant_id<ID1>=<project_id1> \
...
--property filter_tenant_id<IDn>=<project_idn> \
<aggregate_name>
```

- **<ID0>**、**<ID1>**、および **<IDn>** までのすべての ID を、作成する各プロジェクトフィルターの一意の値に置き換えてください。
- **<project_id0>**、**<project_id1>**、および **<project_idn>** までのすべてのプロジェクト ID を、ホストアグリゲートに割り当てる各プロジェクトの ID に置き換えてください。
- **<aggregate_name>** をプロジェクト分離ホストアグリゲートの名前に置き換えてください。

たとえば、プロジェクト **78f1**、**9d3t**、および **aa29** をホストアグリゲート **project-isolated-aggregate** に割り当てるには、以下の構文を使用します。

```
(overcloud)# openstack aggregate set \
--property filter_tenant_id=78f1 \
--property filter_tenant_id1=9d3t \
--property filter_tenant_id2=aa29 \
project-isolated-aggregate
```

ヒント

filter_tenant_id メタデータキーの接尾辞を省略することで、単一の特定プロジェクトでのみ利用可能なホストアグリゲートを作成することができます。

```
(overcloud)# openstack aggregate set \  
--property filter_tenant_id=78f1 \  
single-project-isolated-aggregate
```

関連情報

- ホストアグリゲートの作成に関する詳細は、[Creating and managing host aggregates](#) を参照してください。

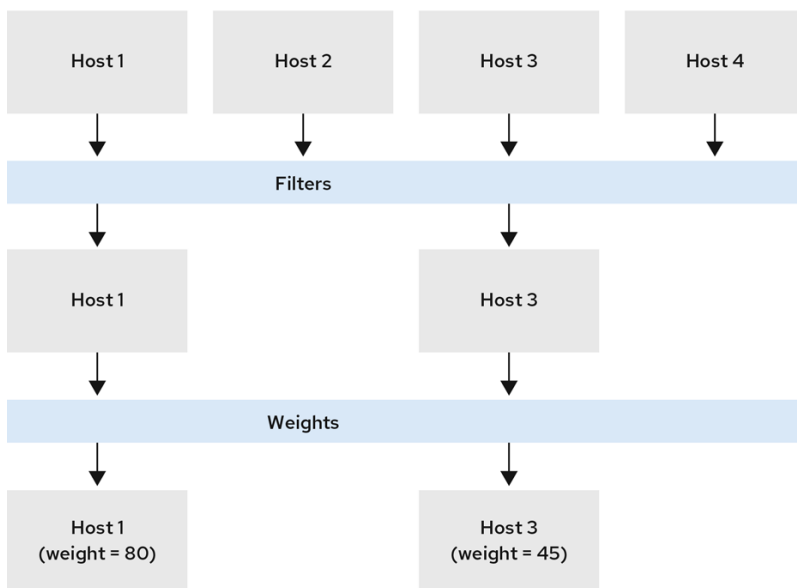
第7章 インスタンスのスケジューリングと配置の設定

Compute スケジューラーサービスは、インスタンスの配置先となるコンピュータードまたはホストアグリゲートを決定します。Compute (nova) サービスがインスタンスの起動または移動に関するリクエストを受け取ると、リクエスト、フレーバー、およびイメージで提供される仕様を使用して適切なホストを決定します。たとえば、フレーバーでは、ストレージディスクの種別や Intel CPU 拡張命令セットなど、インスタンスがホストに要求する特性を指定することができます。

Compute スケジューラーサービスは、以下の順序で以下のコンポーネントの設定を使用して、インスタンスを起動または移動するコンピュータードを決定します。

1. **Placement サービスのプレフィルタ:** Compute スケジューラーサービスは Placement サービスを使用して、特定の属性に基づいて候補のコンピュータードのセットを絞り込みます。たとえば、Placement サービスは無効な状態のコンピュータードを自動的に除外します。
2. **フィルタ:** Compute スケジューラーサービスは、これを使用してインスタンスを起動するコンピュータードの初期セットを決定します。
3. **重み:** Compute スケジューラーサービスは、重み付けシステムを使用して絞り込まれたコンピュータードの優先順位付けを行います。最も高い重みが最も優先されます。

下図では、絞り込み後、Host1 および 3 が条件を満たしています。Host1 の重みが最も高いため、スケジューリングで最も優先されます。



81_OpenStack_0520

7.1. PLACEMENT サービスを使用した事前絞り込み

Compute サービス (nova) は、Placement サービスと協調してインスタンスを作成および管理します。Placement サービスは、コンピュータード、共有ストレージプール、または IP 割り当てプールなど、リソースプロバイダーのインベントリおよび使用状況、ならびに利用可能な仮想 CPU 数などのリソースの量的情報を追跡します。リソースの選択および消費を管理する必要があるサービスは、すべて Placement サービスを使用することができます。

Placement サービスは、リソースプロバイダーのストレージディスク特性の種別など、リソースの機能的情報とリソースプロバイダー間のマッピングも追跡します。

Placement サービスは、Placement サービスリソースプロバイダーインベントリーおよび特性に基づいて、候補のコンピューターノードセットにプレフィルタを適用します。以下の尺度に基づいてプレフィルタを作成することができます。

- サポートされるイメージ種別
- 特性
- プロジェクトまたはテナント
- アベイラビリティーゾーン

7.1.1. 要求されたイメージ種別のサポートによる絞り込み

インスタンスの起動に使用するイメージのディスク形式をサポートしないコンピューターノードを除外することができます。これは、環境の一時バックエンドに QCOW2 イメージをサポートしない Red Hat Ceph Storage が使用される場合に有用です。この機能を有効にすると、スケジューラーは QCOW2 イメージを使用するインスタンスの起動要求を Red Hat Ceph Storage ベースのコンピューターノードに送信しないようになります。

手順

1. Compute 環境ファイルを開きます。
2. インスタンスの起動に使用するイメージのディスク形式をサポートしないコンピューターノードを除外するには、Compute 環境ファイルの **NovaSchedulerQueryImageType** パラメーターを **True** に設定します。
3. 更新内容を Compute 環境ファイルに保存します。
4. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

7.1.2. リソースプロバイダー特性による絞り込み

各リソースプロバイダーには特性のセットがあります。特性は、ストレージディスクの種別や Intel CPU 拡張命令セットなど、リソースプロバイダーの機能的な要素です。

コンピューターノードは、その機能を特性として Placement サービスに報告します。インスタンスは、要求する特性またはリソースプロバイダーにあってはいけない特性を指定することができます。Compute スケジューラーは、これらの特性を使用して、インスタンスをホストするのに適したコンピューターノードまたはホストアグリゲートを特定することができます。

クラウドユーザーが特定の特性を持つホストにインスタンスを作成できるようにするには、特定の特性を要求または禁止するフレーバーを定義して、その特性を要求または禁止するイメージを作成することができます。

利用可能な特性の一覧は、[os-traits ライブラリー](#) を参照してください。必要に応じて、カスタムの特性を作成することもできます。

7.1.2.1. リソースプロバイダー特性を要求または禁止するイメージの作成

クラウドユーザーが特定の特性を持つホストでインスタンスを起動するのに使用することのできるインスタンスイメージを作成することができます。

前提条件

- Placement サービスにクエリーを行うには、アンダークラウドに **python3-osc-placement** パッケージをインストールします。

手順

1. 新規イメージを作成します。

```
(overcloud)$ openstack image create ... trait-image
```

2. ホストまたはホストアグリゲートに必要な特性を識別します。既存の特性を選択するか、新たな特性を作成することができます。

- 既存の特性を使用するには、既存特性の一覧を表示して特性名を取得します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait list
```

- 新規特性を作成するには、以下のコマンドを入力します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait \
create CUSTOM_TRAIT_NAME
```

カスタムの特性は接頭辞 **CUSTOM_** で始まり、A から Z までの文字、0 から 9 までの数字、およびアンダースコア_だけを使用する必要があります。

3. 各ホストの既存のリソースプロバイダー特性を収集します。

```
(overcloud)$ existing_traits=$(openstack --os-placement-api-version 1.6 resource provider
trait list -f value <host_uuid> | sed 's/^/--trait /')
```

4. 既存のリソースプロバイダー特性に、ホストまたはホストアグリゲートに必要な特性があることを確認します。

```
(overcloud)$ echo $existing_traits
```

5. 必要な特性がまだリソースプロバイダーに追加されていない場合は、既存の特性と必要な特性を各ホストのリソースプロバイダーに追加してください。

```
(overcloud)$ openstack --os-placement-api-version 1.6 \
resource provider trait set $existing_traits \
--trait <TRAIT_NAME> \
<host_uuid>
```

<TRAIT_NAME> を、リソースプロバイダーに追加する特性の名前に置き換えます。必要に応じて、**--trait** オプションを複数回使用して、さらに特性を追加することができます。



注記

このコマンドは、リソースプロバイダーの特性をすべて置き換えます。したがって、ホスト上の既存のリソースプロバイダー特性の一覧を取得して、削除されないように再度設定する必要があります。

6. 要求された特性を持つホストまたはホストアグリゲートにインスタンスをスケジュールするには、イメージの追加スペックに特性を追加します。たとえば、AVX-512 をサポートするホストまたはホストアグリゲートにインスタンスをスケジュールするには、イメージの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack image set \
  --property trait:HW_CPU_X86_AVX512BW=required \
  trait-image
```

7. 禁止された特性を持つホストまたはホストアグリゲートを除外するには、イメージの追加スペックに特性を追加します。たとえば、ボリュームの複数接続をサポートするホストまたはホストアグリゲートにインスタンスがスケジュールされるのを防ぐには、イメージの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack image set \
  --property trait:COMPUTE_VOLUME_MULTI_ATTACH=forbidden \
  trait-image
```

7.1.2.2. リソースプロバイダー特性を要求または禁止するフレーバーの作成

クラウドユーザーが特定の特性を持つホストでインスタンスを起動するのに使用することのできるフレーバーを作成することができます。

前提条件

- Placement サービスにクエリーを行うには、アンダークラウドに **python3-osc-placement** パッケージをインストールします。

手順

1. フレーバーを作成します。

```
(overcloud)$ openstack flavor create --vcpus 1 --ram 512 \
  --disk 2 trait-flavor
```

2. ホストまたはホストアグリゲートに必要な特性を識別します。既存の特性を選択するか、新たな特性を作成することができます。

- 既存の特性を使用するには、既存特性の一覧を表示して特性名を取得します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait list
```

- 新規特性を作成するには、以下のコマンドを入力します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait \
  create CUSTOM_TRAIT_NAME
```

カスタムの特性は接頭辞 **CUSTOM_** で始まり、A から Z までの文字、0 から 9 までの数字、およびアンダースコア_だけを使用する必要があります。

3. 各ホストの既存のリソースプロバイダー特性を収集します。

```
(overcloud)$ existing_traits=$(openstack --os-placement-api-version 1.6 resource provider trait list -f value <host_uuid> | sed 's/^/--trait /')
```

4. 既存のリソースプロバイダー特性に、ホストまたはホストアグリゲートに必要な特性があることを確認します。

```
(overcloud)$ echo $existing_traits
```

5. 必要な特性がまだリソースプロバイダーに追加されていない場合は、既存の特性と必要な特性を各ホストのリソースプロバイダーに追加してください。

```
(overcloud)$ openstack --os-placement-api-version 1.6 \
resource provider trait set $existing_traits \
--trait <TRAIT_NAME> \
<host_uuid>
```

<TRAIT_NAME> を、リソースプロバイダーに追加する特性の名前に置き換えます。必要に応じて、**--trait** オプションを複数回使用して、さらに特性を追加することができます。



注記

このコマンドは、リソースプロバイダーの特性をすべて置き換えます。したがって、ホスト上の既存のリソースプロバイダー特性の一覧を取得して、削除されないように再度設定する必要があります。

6. 要求された特性を持つホストまたはホストアグリゲートにインスタンスをスケジュールするには、フレーバーの追加スペックに特性を追加します。たとえば、AVX-512 をサポートするホストまたはホストアグリゲートにインスタンスをスケジュールするには、フレーバーの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack flavor set \
--property trait:HW_CPU_X86_AVX512BW=required \
trait-flavor
```

7. 禁止された特性を持つホストまたはホストアグリゲートを除外するには、フレーバーの追加スペックに特性を追加します。たとえば、ボリュームの複数接続をサポートするホストまたはホストアグリゲートにインスタンスがスケジュールされるのを防ぐには、フレーバーの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack flavor set \
--property trait:COMPUTE_VOLUME_MULTI_ATTACH=forbidden \
trait-flavor
```

7.1.3. ホストアグリゲートの分離による絞り込み

ホストアグリゲートへのスケジューリングを、フレーバーおよびイメージの特性がホストアグリゲートのメタデータと一致するインスタンスだけに制限することができます。フレーバーとイメージのメタデータの組み合わせでは、そのホストアグリゲートに属するコンピュータードへのスケジューリング

を有効にするホストアグリゲート特性をすべて要求する必要があります。

前提条件

- Placement サービスにクエリーを行うには、アンダークラウドに **python3-osc-placement** パッケージをインストールします。

手順

1. Compute 環境ファイルを開きます。
2. ホストアグリゲートを分離してフレーバーおよびイメージの特性がアグリゲートのメタデータと一致するインスタンスだけをホストするには、Compute 環境ファイルの **NovaSchedulerEnableIsolatedAggregateFiltering** パラメーターを **True** に設定します。
3. 更新内容を Compute 環境ファイルに保存します。
4. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

5. ホストアグリゲートを分離する対象の特性を特定します。既存の特性を選択するか、新たな特性を作成することができます。

- 既存の特性を使用するには、既存特性の一覧を表示して特性名を取得します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait list
```

- 新規特性を作成するには、以下のコマンドを入力します。

```
(overcloud)$ openstack --os-placement-api-version 1.6 trait \
create CUSTOM_TRAIT_NAME
```

カスタムの特性は接頭辞 **CUSTOM_** で始まり、A から Z までの文字、0 から 9 までの数字、およびアンダースコア_だけを使用する必要があります。

6. 各コンピューターノードの既存のリソースプロバイダー特性を収集します。

```
(overcloud)$ existing_traits=$(openstack --os-placement-api-version 1.6 resource provider
trait list -f value <host_uuid> | sed 's/^/--trait /')
```

7. 既存のリソースプロバイダー特性で、ホストアグリゲートを分離する特性を確認します。

```
(overcloud)$ echo $existing_traits
```

8. 必要な特性がまだリソースプロバイダーに追加されていない場合は、既存の特性と必要な特性をホストアグリゲートの各コンピューターノードのリソースプロバイダーに追加してください。

```
(overcloud)$ openstack --os-placement-api-version 1.6 \
resource provider trait set $existing_traits \
--trait <TRAIT_NAME> \
```

```
<host_uuid>
```

<TRAIT_NAME> を、リソースプロバイダーに追加する特性の名前に置き換えます。必要に応じて、**--trait** オプションを複数回使用して、さらに特性を追加することができます。



注記

このコマンドは、リソースプロバイダーの特性をすべて置き換えます。したがって、ホスト上の既存のリソースプロバイダー特性の一覧を取得して、削除されないように再度設定する必要があります。

9. ホストアグリゲートに属する各コンピュータノードで、ステップ 6-8 を繰り返します。
10. 特性のメタデータ属性をホストアグリゲートに追加します。

```
(overcloud)$ openstack --os-compute-api-version 2.53 aggregate set \
--property trait:<TRAIT_NAME>=required <aggregate_name>
```

11. フレーバーまたはイメージに特性を追加します。

```
(overcloud)$ openstack flavor set \
--property trait:<TRAIT_NAME>=required <flavor>
(overcloud)$ openstack image set \
--property trait:<TRAIT_NAME>=required <image>
```

7.1.4. Placement サービスを使用したアベイラビリティゾーンによる絞り込み

Placement サービスを使用して、アベイラビリティゾーンの要求を適用することができます。Placement サービスを使用してアベイラビリティゾーンで絞り込むには、アベイラビリティゾーンホストアグリゲートのメンバーシップおよび UUID と一致する配置アグリゲートが存在する必要があります。

前提条件

- Placement サービスにクエリーを行うには、アンダークラウドに **python3-osc-placement** パッケージをインストールします。

手順

1. Compute 環境ファイルを開きます。
2. Placement サービスを使用してアベイラビリティゾーンで絞り込むには、Compute 環境ファイルの **NovaSchedulerQueryPlacementForAvailabilityZone** パラメーターを **True** に設定します。
3. **NovaSchedulerDefaultFilters** パラメーターから **AvailabilityZoneFilter** フィルターを削除します。
4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

関連情報

- アベイラビリティゾーンとして使用するホストアグリゲートの作成に関する詳細は、[Creating an availability zone](#) を参照してください。

7.2. COMPUTE スケジューラーサービス用フィルターおよび重みの設定

インスタンスを起動するコンピュータノードの初期セットを決定するには、Compute スケジューラーサービス用にフィルターおよび重みを設定する必要があります。

手順

1. Compute 環境ファイルを開きます。
2. スケジューラーが使用するフィルターを **NovaSchedulerDefaultFilters** パラメーターに追加します。以下に例を示します。

```
parameter_defaults:
  NovaSchedulerDefaultFilters:
    AggregateInstanceExtraSpecsFilter,ComputeFilter,ComputeCapabilitiesFilter,ImagePropertiesFilter
```

3. 各コンピュータノードの重みを計算するのに使用する属性を指定します。以下に例を示します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
      filter_scheduler/weight_classes:
        value: nova.scheduler.weights.all_weighters
```

利用可能な属性についての詳しい情報は、[Compute スケジューラーの重み](#)を参照してください。

4. オプション: 各重み付け関数に適用する重みの乗数を設定します。たとえば、コンピュータノードの利用可能な RAM が他のデフォルトの重み付け関数よりも高い重みを持つように指定し、Compute スケジューラーが、利用可能な RAM が少ないコンピュータノードよりも、利用可能な RAM が多いコンピュータノードを優先するには、以下の設定を使用します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
      filter_scheduler/weight_classes:
        value: nova.scheduler.weights.all_weighters
      filter_scheduler/ram_weight_multiplier:
        value: 2.0
```

ヒント

また、重みの乗数を負の値に設定することもできます。上記の例では、利用可能な RAM が多いコンピュートノードよりも利用可能な RAM が少ないノードを優先するには、**ram_weight_multiplier** を **-2.0** に設定します。

5. 更新内容を Compute 環境ファイルに保存します。
6. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

関連情報

- 利用可能な Compute スケジューラーサービスのフィルター一覧は、[Compute scheduler filters](#) を参照してください。
- 利用可能な重みの設定オプション一覧は、[Compute scheduler weights](#) を参照してください。

7.3. COMPUTE スケジューラーのフィルター

インスタンスをホストするのに適切なコンピュートノードを選択する際に Compute スケジューラーが適用しなければならないフィルターを指定するには、Compute 環境ファイルの **NovaSchedulerDefaultFilters** パラメーターを設定します。デフォルト設定では、以下のフィルターが適用されます。

- **AvailabilityZoneFilter**: コンピュートノードは要求されたアベイラビリティゾーンに属していない。
- **ComputeFilter**: コンピュートノードは要求に対応することができる。
- **ComputeCapabilitiesFilter**: コンピュートノードはフレーバーの追加スペックを満足する。
- **ImagePropertiesFilter**: コンピュートノードは要求されたイメージ属性を満足する。
- **ServerGroupAntiAffinityFilter**: コンピュートノードは、まだ指定されたグループに属するインスタンスをホストしていない。
- **ServerGroupAffinityFilter**: コンピュートノードは、すでに指定されたグループに属するインスタンスをホストしている。

フィルターを追加および削除することができます。利用可能なすべてのフィルターの詳細を以下の表に示します。

表7.1 Compute スケジューラーのフィルター

フィルター	説明
-------	----

フィルター	説明
AggregateImagePropertiesIsolation	<p>このフィルターを使用して、インスタンスのイメージメタデータとホストアグリゲートのメタデータを照合します。いずれかのホストアグリゲートのメタデータがイメージのメタデータと一致する場合は、そのホストアグリゲートに属するコンピュートノードはそのイメージからインスタンスを起動する候補となります。スケジューラーは、有効なイメージメタデータ属性のみを認識します。有効なイメージメタデータ属性に関する詳細は、Image metadata properties を参照してください。</p>
AggregateInstanceExtraSpecsFilter	<p>このフィルターを使用して、インスタンスのフレーバー追加スペックで定義された名前空間属性とホストアグリゲートのメタデータを照合します。</p> <p>フレーバー extra_specs キーのスコープは、aggregate_instance_extra_specs: namespace の前に付けて定義する必要があります。</p> <p>いずれかのホストアグリゲートのメタデータがフレーバー追加スペックのメタデータと一致する場合は、そのホストアグリゲートに属するコンピュートノードはそのイメージからインスタンスを起動する候補となります。</p>
AggregateIopsFilter	<p>このフィルターを使用して、ホストアグリゲートごとの filter_scheduler/max_io_ops_per_host 値を条件に I/O 操作でホストを絞り込みます。ホストアグリゲートごとの値が確認されない場合は、値はグローバル設定にフォールバックします。ホストが複数のアグリゲートに属し、複数の値が確認された場合、スケジューラーは最小の値を使用します。</p>
AggregateMultiTenancyIsolation	<p>このフィルターを使用して、プロジェクト分離ホストアグリゲートに属するコンピュートノードの可用性を、指定したプロジェクトのセットに制限します。filter_tenant_id メタデータキーを使用して指定したプロジェクトだけが、ホストアグリゲートに属するコンピュートノードでインスタンスを起動することができます。詳しくは、Creating a project-isolated host aggregate を参照してください。</p> <div>  <div> <p>注記</p> <p>プロジェクトが他のホストにインスタンスを配置することは可能です。これを制限するには、NovaSchedulerPlacementAggregateRequiredForTenants パラメーターを使用します。</p> </div> </div>
AggregateNumInstancesFilter	<p>このフィルターを使用して、アグリゲートに属する各コンピュートノードでホスト可能なインスタンスの数を制限します。filter_scheduler/max_instances_per_host パラメーターを使用して、ホストアグリゲートごとのインスタンスの最大数を設定することができます。ホストアグリゲートごとの値が確認されない場合は、値はグローバル設定にフォールバックします。コンピュートノードが複数のホストアグリゲートに属する場合、スケジューラーは最小の max_instances_per_host 値を使用します。</p>

フィルター	説明
AggregateTypeAffinityFilter	<p>フレーバーメタデータキーが設定されていない場合や、フレーバーアグリゲートメタデータの値に要求するフレーバーの名前が含まれる場合には、このフィルターを使用してホスト渡します。フレーバーメタデータエントリーの値は、単一のフレーバー名またはフレーバー名のコンマ区切りリストのいずれかを含む文字列です (例: m1.nano または m1.nano,m1.small)。</p>
AllHostsFilter	<p>このフィルターを使用して、利用可能なすべてのコンピュータノードをインスタンスのスケジューリング対象として考慮します。</p> <div>  <div> <p>注記</p> <p>このフィルターを使用しても、他のフィルターは無効になりません。</p> </div> </div>
AvailabilityZoneFilter	<p>このフィルターを使用して、インスタンスが指定するアベイラビリティーゾーンに属するコンピュータノードでインスタンスを起動します。</p>
ComputeCapabilitiesFilter	<p>このフィルターを使用して、インスタンスのフレーバー追加スペックで定義した名前空間属性とコンピュータノードのケイパビリティを照合します。フレーバーの追加スペックに capabilities: 名前空間の接頭辞を指定する必要があります。</p> <p>ComputeCapabilitiesFilter フィルターを使用するよりも効率的な方法は、フレーバーで CPU 特性を使用することです。これは、Placement サービスに報告されます。特性により、CPU 機能に一貫性のある名前が付けられます。詳細は、Filtering by using resource provider traits を参照してください。</p>
ComputeFilter	<p>このフィルターを使用して、稼働中で有効なすべてのコンピュータノードを渡します。このフィルターは常に設定されている必要があります。</p>
DifferentHostFilter	<p>このフィルターを使用して、特定のインスタンスセットとは異なるコンピュータノードへのインスタンスのスケジューリングを有効にします。インスタンスの起動時にこれらのインスタンスを指定するには、--hint 引数を使用して different_host およびインスタンスの UUID をキー/値のペアとして指定します。</p> <pre> \$ openstack server create --image cedef40a-ed67-4d10-800e-17455edce175 \ --flavor 1 --hint different_host=a0cf03a5-d921-4877-bb5c-86d26cf818e1 \ --hint different_host=8c19174f-4220-44f0-824a-cd1eeef10287 server-1 </pre>

フィルター	説明
ImagePropertiesFilter	<p>このフィルターを使用して、インスタンスイメージで定義された以下の属性に基づいてコンピュートノードを絞り込みます。</p> <ul style="list-style-type: none"> ● hw_architecture: ホストのアーキテクチャーを定義します (例: x86、ARM、および Power)。 ● img_hv_type: ハイパーバイザーの種別を定義します (例: KVM、QEMU、Xen、および LXC)。 ● img_hv_requested_version: Compute サービスが報告するハイパーバイザーのバージョンを定義します。 ● hw_vm_mode: ハイパーバイザーの種別を定義します (例: hvm、xen、uml、または exe)。 <p>インスタンスに含まれる指定のイメージ属性をサポートできるコンピュートノードが、スケジューラーに渡されます。イメージの属性に関する詳細は、Image metadata properties を参照してください。</p>
IsolatedHostsFilter	<p>このフィルターを使用して、分離したコンピュートノード上で分離したイメージだけを使用してインスタンスをスケジュールします。また、filter_scheduler/restrict_isolated_hosts_to_isolated_images を設定して、分離したコンピュートノード上でのインスタンスのビルドに分離していないイメージが使用されるのを防ぐこともできます。</p> <p>分離されたイメージとホストのセットを指定するには、filter_scheduler/isolated_hosts および filter_scheduler/isolated_images 設定オプションを使用します。以下に例を示します。</p> <pre>parameter_defaults: ComputeExtraConfig: nova::config::nova_config: filter_scheduler/isolated_hosts: value: server1, server2 filter_scheduler/isolated_images: value: 342b492c-128f-4a42-8d3a-c5088cf27d13, ebd267a6-ca86-4d6c-9a0e-bd132d6b7d09</pre>
IoOpsFilter	<p>このフィルターを使用して、(ホストで実行可能な高 I/O 負荷インスタンスの最大数を指定する) filter_scheduler/max_io_ops_per_host の設定値を超える同時 I/O 操作があるホストを除外します。</p>

フィルター	説明
MetricsFilter	<p>このフィルターを使用して、metrics/weight_setting を使用して設定されたメトリックを報告するコンピュータノードにスケジューリングを制限します。</p> <p>このフィルターを使用するには、Compute 環境ファイルに以下の設定を追加します。</p> <pre>parameter_defaults: ComputeExtraConfig: nova::config::nova_config: DEFAULT/compute_monitors: value: 'cpu.virt_driver'</pre> <p>デフォルトでは、Compute スケジューラーサービスは 60 秒ごとにメトリックを更新します。メトリックが最新の状態になるようにするには、update_resources_interval 設定オプションを使用してメトリックデータの更新頻度を上げることができます。たとえば、以下の設定を使用すると、2 秒ごとにメトリックデータが更新されます。</p> <pre>parameter_defaults: ComputeExtraConfig: nova::config::nova_config: DEFAULT/update_resources_interval: value: '2'</pre>
NUMATopologyFilter	<p>このフィルターを使用して、NUMA 対応コンピュータノードに NUMA トポロジーが設定されたインスタンスをスケジュールします。フレーバー extra_specs およびイメージ属性を使用して、インスタンスの NUMA トポロジーを指定します。このフィルターは、各ホストの NUMA セルのオーバーサブスクリプション限度を考慮して、インスタンスの NUMA トポロジーをコンピュータノードのトポロジーに一致させるを試みます。</p>
NumInstancesFilter	<p>このフィルターを使用して、max_instances_per_host オプションで指定した以上のインスタンスを実行中のコンピュータノードを除外します。</p>
PciPassthroughFilter	<p>このフィルターを使用して、フレーバー extra_specs を使用してインスタンスが要求するデバイスを持つコンピュータノードにインスタンスをスケジュールします。</p> <p>(通常高価で使用が制限される) PCI デバイスを要求するインスタンス用に、そのデバイスを持つノードを確保する場合に、このフィルターを使用します。</p>

フィルター	説明
SameHostFilter	<p>このフィルターを使用して、特定のインスタンスセットと同じコンピュータノードへのインスタンスのスケジューリングを有効にします。インスタンスの起動時にこれらのインスタンスを指定するには、--hint 引数を使用して same_host およびインスタンスの UUID をキー/値のペアとして指定します。</p> <pre>\$ openstack server create --image cedef40a-ed67-4d10-800e-17455edce175 \ --flavor 1 --hint same_host=a0cf03a5-d921-4877-bb5c-86d26cf818e1 \ --hint same_host=8c19174f-4220-44f0-824a-cd1eeef10287 server-1</pre>
ServerGroupAffinityFilter	<p>このフィルターを使用して、同じコンピュータノード上でアフィニティーサーバーグループに属するインスタンスをスケジュールします。サーバーグループを作成するには、以下のコマンドを入力します。</p> <pre>\$ openstack server group create --policy affinity <group_name></pre> <p>このグループに属するインスタンスを起動するには、--hint 引数を使用して group およびグループの UUID をキー/値のペアとして指定します。</p> <pre>\$ openstack server create --image <image> \ --flavor <flavor> \ --hint group=<group_uuid> <instance_name></pre>
ServerGroupAntiAffinityFilter	<p>このフィルターを使用して、異なるコンピュータノード上で非アフィニティーサーバーグループに属するインスタンスをスケジュールします。サーバーグループを作成するには、以下のコマンドを入力します。</p> <pre>\$ openstack server group create --policy anti-affinity <group_name></pre> <p>このグループに属するインスタンスを起動するには、--hint 引数を使用して group およびグループの UUID をキー/値のペアとして指定します。</p> <pre>\$ openstack server create --image <image> \ --flavor <flavor> \ --hint group=<group_uuid> <instance_name></pre>
SimpleCIDRAffinityFilter	<p>このフィルターを使用して、特定の IP サブネット範囲を持つコンピュータノードにインスタンスをスケジュールします。必要な範囲を指定するには、--hint 引数を使用してインスタンスの起動時にキー build_near_host_ip および cidr を渡します。</p> <pre>\$ openstack server create --image <image> \ --flavor <flavor> \ --hint build_near_host_ip=<ip_address> \ --hint cidr=<subnet_mask> <instance_name></pre>

7.4. COMPUTE スケジューラーの重み

それぞれのコンピュータノードは重みを持ち、スケジューラーはこれを使用してインスタンスのスケジューリングの優先度を決定します。フィルターを適用した後、Compute スケジューラーは残った候補のコンピュータノードから最大の重みを持つコンピュータノードを選択します。

Compute スケジューラーは、以下のタスクを実行することにより、各コンピュータノードの重みを決定します。

1. スケジューラーは、各重みを 0.0 から 1.0 までの値に正規化します。
2. スケジューラーは、正規化された重みを重み付け関数の乗数で乗算します。

Compute スケジューラーは、候補のコンピュータノード全体でリソースの可用性の低い値および高い値を使用して、各リソース種別の重みの正規化を計算します。

- 最も低いリソース可用性 (minval) を持つノードには、0 が割り当てられます。
- 最も高いリソース可用性 (maxval) を持つノードには1が割り当てられます。
- minval と maxval 内の範囲のリソース可用性を持つノードには、以下の式を使用して正規化された重みが割り当てられます。

$$(node_resource_availability - minval) / (maxval - minval)$$

すべてのコンピュータノードが同じリソース可用性を持つ場合、それらはすべて 0 に正規化されます。

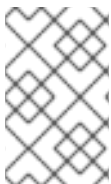
たとえば、スケジューラーは、利用可能な仮想 CPU の数がそれぞれ異なる 10 個のコンピュータノードに関して、利用可能な仮想 CPU の正規化された重みを以下のように計算します。

コンピュータノード	1	2	3	4	5	6	7	8	9	10
仮想 CPU の数	5	5	10	10	15	20	20	15	10	5
正規化された重み	0	0	0.33	0.33	0.67	1	1	0.67	0.33	0

Compute スケジューラーは、以下の式を使用してコンピュータノードの重みを計算します。

$$(w1_multiplier * norm(w1)) + (w2_multiplier * norm(w2)) + \dots$$

重みに使用することのできる設定オプションの詳細を以下の表に示します。



注記

以下の表で説明するオプションと同じ名前のアグリゲートメタデータのキーを使用して、ホストアグリゲートに重みを設定することができます。ホストアグリゲートに設定すると、ホストアグリゲートの値が優先されます。



表7.2 Compute スケジューラーの重み

設定オプション	型	説明
filter_scheduler/weight_classes	文字列	<p>このパラメーターを使用して、各コンピュートノードの重みを計算するのに以下の属性のどれを使用するかを設定します。</p> <ul style="list-style-type: none"> ● nova.scheduler.weights.ram.RAMWeigher: コンピュートノードで利用可能な RAM を重み付けします。 ● nova.scheduler.weights.cpu.CPUWeigher: コンピュートノードで利用可能な CPU の数を重み付けします。 ● nova.scheduler.weights.disk.DiskWeigher: コンピュートノードで利用可能なディスクを重み付けします。 ● nova.scheduler.weights.metrics.MetricsWeigher: コンピュートノードのメトリクスを重み付けします。 ● nova.scheduler.weights.affinity.ServerGroupSoftAffinityWeigher: 指定したインスタンスグループの他のノードとコンピュートノードの近接性を重み付けします。 ● nova.scheduler.weights.affinity.ServerGroupSoftAntiAffinityWeigher: 指定したインスタンスグループの他のノードとコンピュートノードの近接性を重み付けします。 ● nova.scheduler.weights.compute.BuildFailureWeigher: 直近ブート試行の失敗回数でコンピュートノードを重み付けします。 ● nova.scheduler.weights.io_ops.IoOpsWeigher: ワークロードでコンピュートノードを重み付けします。 ● nova.scheduler.weights.pci.PCIWeigher: PCI の可用性でコンピュートノードを重み付けします。 ● nova.scheduler.weights.cross_cell.CrossCellWeigher: 置かれているセルに基づいてコンピュートノードを重み付けします。インスタンスを移動する際に、移行元セル内にあるコンピュートノードを優先します。 ● nova.scheduler.weights.all_weighters: (デフォルト) 上記の重み付け関数をすべて使用します。

設定オプション	型	説明
filter_scheduler/ram_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、利用可能な RAM 容量に基づいてホストを重み付けするのに使用する重みの乗数を指定します。</p> <p>利用可能な RAM 容量がより大きいホストを優先するには、正の値に設定します。この場合、インスタンスは多くのホストに分散されます。</p> <p>利用可能な RAM 容量がより小さいホストを優先するには、負の値に設定します。この場合、可能な限り多くのインスタンスをホストに分担 (スタック) させた後に、使用率が低いホストをスケジューリングします。</p> <p>正または負の絶対値で、他の重み付け関数と比べて {b}RAM の重み付け関数をどれだけ優先するかを指定します。</p> <p>デフォルトは 1.0 で、スケジューラーはインスタンスをすべてのホストに均等に分散します。</p>
filter_scheduler/disk_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、利用可能なディスク容量に基づいてホストを重み付けするのに使用する重みの乗数を指定します。</p> <p>利用可能なディスク容量がより大きいホストを優先するには、正の値に設定します。この場合、インスタンスは多くのホストに分散されます。</p> <p>利用可能なディスク容量がより小さいホストを優先するには、負の値に設定します。この場合、可能な限り多くのインスタンスをホストに分担 (スタック) させた後に、使用率が低いホストをスケジューリングします。</p> <p>正または負の絶対値で、他の重み付け関数と比べてディスクの重み付け関数をどれだけ優先するかを指定します。</p> <p>デフォルトは 1.0 で、スケジューラーはインスタンスをすべてのホストに均等に分散します。</p>

設定オプション	型	説明
filter_scheduler/cpu_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、利用可能な仮想 CPU の数に基づいてホストを重み付けするのに使用する重みの乗数を指定します。</p> <p>利用可能な仮想 CPU の数がより多いホストを優先するには、正の値に設定します。この場合、インスタンスは多くのホストに分散されます。</p> <p>利用可能な仮想 CPU の数がより少ないホストを優先するには、負の値に設定します。この場合、可能な限り多くのインスタンスをホストに分担 (スタック) させた後に、使用率が低いホストをスケジューリングします。</p> <p>正または負の絶対値で、他の重み付け関数と比べて仮想 CPU の重み付け関数をどれだけ優先するかを指定します。</p> <p>デフォルトは 1.0 で、スケジューラーはインスタンスをすべてのホストに均等に分散します。</p>
filter_scheduler/io_ops_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、負荷に基づいてホストを重み付けするのに使用する重みの乗数を指定します。</p> <p>負荷がより軽いホストを優先するには、負の値に設定します。この場合、負荷はより多くのホストに分散されます。</p> <p>負荷がより重いホストを優先するには、正の値に設定します。この場合、インスタンスはすでにビジー状態にあるホストにスケジューリングされます。</p> <p>正または負の絶対値で、他の重み付け関数と比べて I/O 操作の重み付け関数をどれだけ優先するかを指定します。</p> <p>デフォルトは -1.0 で、スケジューラーは負荷をより多くのホストに分散します。</p>
filter_scheduler/build_failure_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、直近のビルド失敗回数に基づいてホストを重み付けするのに使用する重みの乗数を指定します。</p> <p>ホストから報告される直近のビルド失敗回数をより重要視するには、正の値に設定します。直近ビルドに失敗したホストは、選択されにくくなります。</p> <p>直近の失敗回数でコンピュートホストを重み付けするのを無効にするには、0 に設定します。</p> <p>デフォルト: 1000000.0</p>

設定オプション	型	説明
filter_scheduler/cross_cell_move_weight_multiplier	浮動小数点	<p>このパラメーターを使用して、セルを越えて移動する際にホストを重み付けするのに使用する重みの乗数を指定します。このオプションは、インスタンスを移動する際に、同じ移動元セル内にあるホストに加える重みを決定します。インスタンスを移行する場合、デフォルトではスケジューラーは同じ移行元セル内にあるホストを優先します。</p> <p>現在インスタンスを実行中のセル内にあるホストを優先するには、正の値に設定します。現在インスタンスを実行中のセルとは別のセルにあるホストを優先するには、負の値に設定します。</p> <p>デフォルト: 1000000.0</p>
filter_scheduler/pci_weight_multiplier	正の浮動小数点	<p>このパラメーターを使用して、ホスト上の PCI デバイスの数とインスタンスが要求する PCI デバイスの数に基づいてホストを重み付けするのに使用する重みの乗数を指定します。インスタンスが PCI デバイスを要求する場合、より多くの PCI デバイスを持つコンピュータードにより高い重みが割り当てられます。</p> <p>たとえば、ホストが 3 台利用可能で、PCI デバイスが 1 つのホストが 1 台、複数の PCI デバイスがあるホストが 1 台、PCI デバイスがないホストが 1 台の場合には、Compute スケジューラーはインスタンスの需要に基づいてこれらのホストの優先順位付けを行います。スケジューラーは、インスタンスが PCI デバイスを 1 つ要求している場合には 1 番目のホストを、複数の PCI デバイスを要求している場合には 2 番目のホストを、PCI デバイスを要求していない場合には 3 番目のホストを、それぞれ優先するはずでず。</p> <p>このオプションを設定して、PCI を要求しないインスタンスが PCI デバイスを持つホストのリソースを占有するのを防ぎます。</p> <p>デフォルト: 1.0</p>

設定オプション	型	説明
filter_scheduler/hot_subset_size	Integer	<p>このパラメーターを使用して、ホストを選択するサブセット (絞り込まれたホストのサブセット) のサイズを指定します。このオプションを1以上に設定する必要があります。値を1に指定した場合には、重み付け関数によって最初に返されるホストが選択されます。スケジューラーは1未満の値を無視し、代わりに1を使用します。</p> <p>類似の要求を処理する複数のスケジューラープロセスが同じホストを選択して競合状態が生じるのを防ぐには、1より大きい値に設定します。要求に最も適したN台のホストからホストを無作為に選択することで、競合の可能性が低減されます。ただし、この値を高く設定すると、選択されるホストが特定の要求に対して最適ではない可能性が高くなります。</p> <p>デフォルト:1</p>
filter_scheduler/soft_affinity_weight_multiplier	正の浮動小数点	<p>このパラメーターを使用して、グループのソフトアフィニティのホストを重み付けするのに使用する重みの乗数を指定します。</p> <div>  <div> <p>注記</p> <p>このポリシーでグループを作成する場合は、マイクロバージョンを指定する必要があります。</p> <pre>\$ openstack --os-compute-api-version 2.15 server group create --policy soft-affinity <group_name></pre> </div> </div> <p>デフォルト:1.0</p>
filter_scheduler/soft_anti_affinity_weight_multiplier	正の浮動小数点	<p>このパラメーターを使用して、グループのソフト非アフィニティのホストを重み付けするのに使用する重みの乗数を指定します。</p> <div>  <div> <p>注記</p> <p>このポリシーでグループを作成する場合は、マイクロバージョンを指定する必要があります。</p> <pre>\$ openstack --os-compute-api-version 2.15 server group create --policy soft-affinity <group_name></pre> </div> </div> <p>デフォルト:1.0</p>

設定オプション	型	説明
metrics/weight_multiplier	浮動小数点	<p>このパラメーターを使用して、メトリックの重み付けに使用する重みの乗数を指定します。デフォルトでは weight_multiplier=1.0 に設定されており、使用可能な全ホストの間でインスタンスを分散します。</p> <p>重み全体でメトリックの影響を増大させるには、1.0 を超える数値に設定します。</p> <p>重み全体でメトリックの影響を減少させるには、0.0 と 1.0 の間の数値に設定します。</p> <p>メトリックの値を無視して weight_of_unavailable オプションの値を返すには、0.0 に設定します。</p> <p>低いメトリックのホストを優先してインスタンスをホストにスタックするには、負の数値に設定します。</p> <p>デフォルト: 1.0</p>
metrics/weight_setting	metric=ratio ペアのコンマ区切りリスト	<p>このパラメーターを使用して、重み付けに使用するメトリック、および各メトリックの重みを計算するのに使用する比率を指定します。有効なメトリック名は以下のとおりです。</p> <ul style="list-style-type: none"> ● cpu.frequency: CPU の周波数 ● cpu.user.time: CPU のユーザーモード時間 ● cpu.kernel.time: CPU のカーネル時間 ● cpu.idle.time: CPU のアイドル時間 ● cpu.iowait.time: CPU の I/O 待機時間 ● cpu.user.percent: CPU のユーザーモード率 ● cpu.kernel.percent: CPU のカーネル率 ● cpu.idle.percent: CPU のアイドル率 ● cpu.iowait.percent: CPU の I/O 待機率 ● cpu.percent: 汎用 CPU の使用率 <p>例: weight_setting=cpu.user.time=1.0</p>

設定オプション	型	説明
metrics/required	Boolean	<p>このパラメーターを使用して、設定した metrics/weight_setting メトリックが使用できない場合の処理方法を指定します。</p> <ul style="list-style-type: none">● True: メトリックは必須です。メトリックが使用できない場合には、例外が発生します。この例外を回避するには、NovaSchedulerDefaultFilters の MetricsFilter フィルターを使用します。● False: 使用できないメトリックは、重み付け処理において負の係数として扱われます。weight_of_unavailable 設定オプションを使用して、戻り値を設定します。
metrics/weight_of_unavailable	浮動小数点	<p>このパラメーターを使用して、metrics/weight_setting メトリックが使用できず、かつ metrics/required=False の場合に用いる重みを指定します。</p> <p>デフォルト: -10000.0</p>

第8章 インスタンス起動用のフレーバーの作成

インスタンスのフレーバーは、インスタンス用の仮想ハードウェアプロファイルを指定するリソースのテンプレートです。クラウドユーザーは、インスタンスを起動する際にフレーバーを指定する必要があります。

フレーバーにより、Compute サービスがインスタンスに割り当てる必要のある以下のリソースの量を指定することができます。

- 仮想 CPU の数
- RAM (MB 単位)
- ルートディスク (GB 単位)
- セカンダリー一時ストレージおよびスワップディスクを含む仮想ストレージ

フレーバーを全プロジェクトに公開したり、特定のプロジェクトまたはドメインを対象にプライベートに設定したりすることで、フレーバーを使用できるユーザーを指定することができます。

フレーバーでは、メタデータ (追加スペックとも呼ばれる) を使用して、インスタンス用ハードウェアのサポートおよびクォータを指定することができます。フレーバーのメタデータは、インスタンスの配置、リソースの使用上限、およびパフォーマンスに影響を及ぼします。利用可能なメタデータ属性の完全なリストは、[Flavor metadata](#) を参照してください。

フレーバーメタデータキーを使用することで、ホストアグリゲートで設定した **extra_specs** メタデータを照合して、インスタンスをホストするのに適したホストアグリゲートを探すこともできます。ホストアグリゲートでインスタンスをスケジュールするには、**extra_specs** キーに **aggregate_instance_extra_specs:** 名前空間の接頭辞を指定して、フレーバーのメタデータのスコープを定義する必要があります。詳細は、[Creating and managing host aggregates](#) を参照してください。

Red Hat OpenStack Platform (RHOSP) のデプロイメントには、クラウドユーザーが使用可能な以下のデフォルトパブリックフレーバーのセットが含まれています。

表8.1 デフォルトのフレーバー

名前	仮想 CPU の数	メモリー	ルートディスクのサイズ
m1.nano	1	128 MB	1 GB
m1.micro	1	192 MB	1 GB



注記

フレーバー属性を使用して設定した動作は、イメージを使用して設定した動作よりも優先されます。クラウドユーザーがインスタンスを起動する際、指定したフレーバーの属性がイメージの属性よりも優先されます。

8.1. フレーバーの作成

特定の機能や動作のために特化したフレーバーを作成および管理することができます。以下に例を示します。

- 基になるハードウェアの要件に応じて、デフォルトのメモリーと容量を変更する

- インスタンスに特定の I/O レートを強制するためのメタデータ、またはホストアグリゲートと一致させるためのメタデータを追加する

手順

1. インスタンスが利用可能な基本的なリソースを指定するフレーバーを作成します。

```
(overcloud)$ openstack flavor create --ram <size_mb> \
--disk <size_gb> --vcpus <no_vcpus> \
[--private --project <project_id>] <flavor_name>
```

- **<size_mb>** を、このフレーバーで作成するインスタンスに割り当てる RAM の容量に置き換えます。
- **<size_gb>** を、このフレーバーで作成するインスタンスに割り当てるルートディスクのサイズに置き換えます。
- **<no_vcpus>** を、このフレーバーで作成するインスタンスに確保する仮想 CPU の数に置き換えます。
- (オプション) **--private** および **--project** オプションを指定して、特定のプロジェクトまたはユーザーグループだけがフレーバーにアクセスできるようにします。**<project_id>** を、このフレーバーを使用してインスタンスを作成できるプロジェクトの ID に置き換えます。アクセシビリティを指定しない場合は、フレーバーはデフォルトで public に設定されます。つまり、すべてのプロジェクトで使用可能になります。



注記

パブリックフレーバーの作成後は、そのフレーバーをプライベートにすることはできません。

- **<flavor_name>** を、一意のフレーバー名に置き換えます。
フレーバーの引数についての詳細は、[Flavor arguments](#) を参照してください。
2. (オプション) フレーバーのメタデータを指定するには、キー/値のペアを使用して必要な属性を設定します。

```
(overcloud)$ openstack flavor set \
--property <key=value> --property <key=value> ... <flavor_name>
```

- **<key>** を、このフレーバーで作成するインスタンスに割り当てる属性のメタデータキーに置き換えます。利用可能なメタデータキーの一覧は、[Flavor metadata](#) を参照してください。
- **<value>** を、このフレーバーで作成するインスタンスに割り当てるメタデータキーの値に置き換えます。
- **<flavor_name>** を、フレーバーの名前に置き換えます。
たとえば、以下のフレーバーを使用して起動されるインスタンスには 2 つの CPU ソケットがあり、それぞれに 2 つの CPU が割り当てられます。

```
(overcloud)$ openstack flavor set \
--property hw:cpu_sockets=2 \
--property hw:cpu_cores=2 processor_topology_flavor
```

8.2. フレーバーの引数

openstack flavor create コマンドには、新しいフレーバーの名前を指定する1つの位置引数 **<flavor_name>** を使用することができます。

以下の表には、新規フレーバーを作成する際に必要に応じて指定することのできる、オプションの引数の詳細をまとめています。

表8.2 オプションのフレーバー引数

オプションの引数	説明
--id	一意のフレーバー ID。デフォルト値は auto で、UUID4 値を生成します。この引数を使用して、整数値を手動で指定することや、UUID4 値を生成することができます。
--ram	(必須) インスタンスが利用可能なメモリーの容量 (MB 単位) デフォルト: 256 MB
--disk	<p>(必須) ルート (/) パーティションに使用するディスク容量 (GB 単位)。ルートディスクは、ベースイメージがコピーされる一時ディスクです。インスタンスが永続ボリュームからブートする場合、ルートディスクは使用されません。</p> <div>  <div> <p>注記</p> <p>--disk が 0 に設定されたフレーバーでインスタンスを作成する場合、インスタンスはボリュームからブートする必要があります。</p> </div> </div> <p>デフォルト: 0 GB</p>
--ephemeral	<p>一時ディスクに使用するディスク容量 (GB 単位) デフォルト値は 0 GB で、セカンダリー一時ディスクは作成されません。一時ディスクは、インスタンスのライフサイクルにリンクしたマシンのローカルディスクストレージを提供します。一時ディスクは、いずれのスナップショットにも含まれません。インスタンスが削除されると、このディスクは破棄されすべてのデータが失われます。</p> <p>デフォルト: 0 GB</p>
--swap	<p>スワップディスクのサイズ (MB 単位) Compute サービスのバックエンドストレージがローカルストレージでない場合は、フレーバーに swap を指定しないでください。</p> <p>デフォルト: 0 GB</p>
--vcpus	<p>(必須) インスタンス用の仮想 CPU の数</p> <p>デフォルト: 1</p>

オプションの引数	説明
--public	フレーバーは、すべてのプロジェクトで利用可能です。デフォルトでは、フレーバーはパブリックですべてのプロジェクトで利用することができます。
--private	フレーバーは、 --project オプションで指定したプロジェクトでのみ利用可能です。プライベートのフレーバーを作成し、フレーバーにプロジェクトを追加しない場合、クラウド管理者だけがそのフレーバーを利用することができます。
--property	以下の形式のキー/値のペアで指定されるメタデータまたは追加スペック --property <key=value> 複数の属性を設定するには、このオプションを繰り返します。
--project	プライベートのフレーバーを使用することができるプロジェクトを指定します。この引数は、 --private オプションと共に使用する必要があります。プロジェクトを指定しないと、フレーバーは admin ユーザーだけに表示されます。 複数のプロジェクトにアクセスを許可するには、このオプションを繰り返します。
--project-domain	プライベートのフレーバーを使用することができるプロジェクトドメインを指定します。この引数は、 --private オプションと共に使用する必要があります。 複数のプロジェクトドメインにアクセスを許可するには、このオプションを繰り返します。
--description	フレーバーの説明。長さは 65535 文字に制限されます。使用できるのは印刷可能文字だけです。

8.3. フレーバーのメタデータ

フレーバーを作成する際に、**--property** オプションを使用してフレーバーのメタデータを指定します。フレーバーのメタデータは **追加スペック** と呼ばれます。フレーバーのメタデータで指定するインスタンス用ハードウェアのサポートおよびクォータは、インスタンスの配置、インスタンスの制限、およびパフォーマンスに影響を及ぼします。

インスタンスによるリソースの使用

以下の表に示す属性キーを使用して、インスタンスによる CPU、メモリー、およびディスク I/O の使用に制限を設定します。

表8.3 リソースの使用を制限するためのフレーバーメタデータ

キー	説明
quota:cpu_shares	ドメイン内の CPU 時間の配分に使用する重みを指定します。デフォルトは OS が提供するデフォルト値です。Compute スケジューラーは、この属性の設定と同じドメイン内にある他のインスタンスの設定を比較して、相対的な重み付けを行います。たとえば、設定が quota:cpu_shares=2048 のインスタンスには、設定が quota:cpu_shares=1024 のインスタンスの 2 倍の CPU 時間が割り当てられます。
quota:cpu_period	cpu_quota を適用する期間を指定します (マイクロ秒単位)。この cpu_period の期間、各仮想 CPU は cpu_quota を超えるランタイムを使用することはできません。1000 - 10000000 の範囲で値を設定します。無効にするには 0 に設定します。
quota:cpu_quota	<p>各 cpu_period における仮想 CPU の最大許容帯域幅を指定します (マイクロ秒単位)。</p> <ul style="list-style-type: none"> 1000 - 18446744073709551 の範囲で値を設定します。 無効にするには 0 に設定します。 帯域幅に制限を設けない場合は、負の値に設定します。 <p>cpu_quota および cpu_period を使用して、全仮想 CPU が同じ速度で実行されるようにすることができます。たとえば、以下のフレーバーを使用して、物理 CPU の処理能力の最大 50% しか消費できないインスタンスを起動することができます。</p> <pre>\$ openstack flavor set cpu_limits_flavor \ --property quota:cpu_quota=10000 \ --property quota:cpu_period=20000</pre>

インスタンスディスクのチューニング

以下の表に示す属性キーを使用して、インスタンスのディスクのパフォーマンスをチューニングします。



注記

Compute サービスは、以下の QoS 設定を、Compute サービスがプロビジョニングしたストレージ (一時ストレージなど) に適用します。Block Storage (cinder) ボリュームのパフォーマンスを調整するには、ボリューム種別の Quality-of-Service (QOS) の値も設定する必要があります。詳細は、**Storage Guide**の[Use Quality-of-Service Specifications](#)を参照してください。

表8.4 ディスクチューニング用のフレーバーメタデータ

キー	説明
----	----

キー	説明
quota:disk_read_bytes_sec	インスタンスが利用可能な最大ディスク読み取りを指定します (バイト毎秒単位)。
quota:disk_read_iops_sec	インスタンスが利用可能な最大ディスク読み取りを指定します (IOPS 単位)。
quota:disk_write_bytes_sec	インスタンスが利用可能な最大ディスク書き込みを指定します (バイト毎秒単位)。
quota:disk_write_iops_sec	インスタンスが利用可能な最大ディスク書き込みを指定します (IOPS 単位)。
quota:disk_total_bytes_sec	インスタンスが利用可能な最大 I/O 操作を指定します (バイト毎秒単位)。
quota:disk_total_iops_sec	インスタンスが利用可能な最大 I/O 操作を指定します (IOPS 単位)。

インスタンスのネットワークトラフィックの帯域幅

以下の表に示す属性キーを使用して、VIF I/O オプションの設定により、インスタンスのネットワークトラフィックの帯域幅上限を設定します。



注記

quota:vif_* 属性は非推奨になりました。この属性の代わりに、Networking (neutron) サービスの Quality of Service (QoS) ポリシーを使用する必要があります。QoS ポリシーについての詳細は、[Networking Guide](#) の [Configuring Quality of Service \(QoS\) policies](#) を参照してください。**quota:vif_*** プロパティは、**NeutronOVSFirewallDriver** が **iptables_hybrid** に設定されている ML2/OVS メカニズムドライバーを使用する場合にのみサポートされます。

表8.5 帯域幅を制限するためのフレーバーメタデータ

キー	説明
quota:vif_inbound_average	(非推奨) インスタンスに送付されるトラフィックに要求される平均ビットレートを指定します (kbps 単位)。
quota:vif_inbound_burst	(非推奨) ピークの速度でバースト処理することができる受信トラフィックの最大量を指定します (KB 単位)。
quota:vif_inbound_peak	(非推奨) インスタンスが受信することのできるトラフィックの最大レートを指定します (kbps 単位)。

キー	説明
quota:vif_outbound_average	(非推奨) インスタンスから送信されるトラフィックに要求される平均ビットレートを指定します (kbps 単位)。
quota:vif_outbound_burst	(非推奨) ピークの速度でバースト処理することができる送信トラフィックの最大量を指定します (KB 単位)。
quota:vif_outbound_peak	(非推奨) インスタンスが送信することのできるトラフィックの最大レートを指定します (kbps 単位)。

ハードウェアビデオ RAM

以下の表に示す属性キーを使用して、ビデオデバイスに使用するインスタンス RAM の上限を設定します。

表8.6 ビデオデバイス用のフレーバーメタデータ

キー	説明
hw_video:ram_max_mb	ビデオデバイスに使用する最大 RAM を指定します (MB 単位)。 hw_video_ram イメージ属性で使用します。 hw_video_ram は hw_video:ram_max_mb 以下でなければなりません。

ウォッチドッグの動作

以下の表に示す属性キーを使用して、インスタンスで仮想ハードウェアのウォッチドッグデバイスを有効にします。

表8.7 ウォッチドッグの動作を設定するためのフレーバーメタデータ

キー	説明
----	----

キー	説明
hw:watchdog_action	<p>仮想ハードウェアのウォッチドッグデバイスを有効にするかどうかを指定し、その動作を設定します。インスタンスがハングアップまたはエラーが発生した場合に、ウォッチドッグデバイスは設定されたアクションを実行します。ウォッチドッグは i6300esb デバイスを使用し、PCI Intel 6300ESB をエミュレートします。hw:watchdog_action が指定されていない場合には、ウォッチドッグは無効になります。</p> <p>以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● disabled: (デフォルト) デバイスは接続されません。 ● reset: インスタンスを強制的にリセットします。 ● poweroff: インスタンスを強制的にシャットダウンします。 ● pause: インスタンスを一時停止します。 ● none: ウォッチドッグを有効にしますが、インスタンスがハングアップまたはエラーが発生しても何も実行しません。 <div>  <div> <p>注記</p> <p>特定のイメージの属性を使用して設定するウォッチドッグの動作は、フレーバーを使用して設定する動作よりも優先されます。</p> </div> </div>

乱数ジェネレーター

以下の表に示す属性キーを使用して、インスタンスで乱数ジェネレーターデバイスを有効にします。

表8.8 乱数ジェネレーターを設定するためのフレーバーメタデータ

キー	説明
hw_rng:allowed	イメージ属性によりインスタンスに追加された乱数ジェネレーターデバイスを有効にするには、 True に設定します。
hw_rng:rate_bytes	期間中、インスタンスがホストのエントロピーから読み取ることのできる最大バイト数を指定します。
hw_rng:rate_period	読み取り期間を指定します (ミリ秒単位)。

仮想パフォーマンス監視ユニット (vPMU)

以下の表に示す属性キーを使用して、インスタンスの仮想 PMU を有効にします。

表8.9 仮想 PMU 用のフレーバーメタデータ

キー	説明
hw:pmu	<p>インスタンスの仮想 PMU を有効にするには、True に設定します。</p> <p>perf 等のツールは、インスタンスの仮想 PMU を使用して、インスタンスのパフォーマンスをプロファイリングおよびモニターリングするためのより正確な情報を提供します。リアルタイム負荷のケースでは、仮想 PMU のエミュレーションにより、望ましくないレイテンシーが付加される場合があります。テレメトリーの提供が不要な場合は、hw:pmu=False に設定します。</p>

インスタンスの CPU トポロジー

以下の表に示す属性キーを使用して、インスタンス内のプロセッサのトポロジーを定義します。

表8.10 CPU トポロジー用のフレーバーメタデータ

キー	説明
hw:cpu_sockets	<p>インスタンスの推奨ソケット数を指定します。</p> <p>デフォルト: 要求される仮想 CPU の数</p>
hw:cpu_cores	<p>インスタンスの1ソケットあたりの推奨コア数を指定します。</p> <p>デフォルト: 1</p>
hw:cpu_threads	<p>インスタンスの1コアあたりの推奨スレッド数を指定します。</p> <p>デフォルト: 1</p>
hw:cpu_max_sockets	<p>イメージ属性を使用してユーザーがインスタンスに選択できる最大ソケット数を指定します。</p> <p>例: hw:cpu_max_sockets=2</p>
hw:cpu_max_cores	<p>イメージ属性を使用してユーザーがインスタンスに選択できる、1ソケットあたりの最大コア数を指定します。</p>
hw:cpu_max_threads	<p>イメージ属性を使用してユーザーがインスタンスに選択できる、1コアあたりの最大スレッド数を指定します。</p>

シリアルポート

以下の表に示す属性キーを使用して、1インスタンスあたりのシリアルポートの数を設定します。

表8.11 シリアルポート用のフレーバーメタデータ

キー	説明
<code>hw:serial_port_count</code>	1 インスタンスあたりの最大シリアルポート数

CPU ピニングポリシー

デフォルトでは、インスタンスの仮想 CPU (vCPU) は 1 コア 1 スレッドのソケットです。属性を使用して、インスタンスの仮想 CPU をホストの物理 CPU コア (pCPU) に固定するフレーバーを作成することができます。1 つまたは複数のコアがスレッドシブリングを持つ同時マルチスレッド (SMT) アーキテクチャーで、ハードウェア CPU スレッドの動作を設定することもできます。

以下の表に示す属性キーを使用して、インスタンスの CPU ピニングポリシーを定義します。

表 8.12 CPU ピニング用のフレーバーメタデータ

キー	説明
<code>hw:cpu_policy</code>	<p>使用する CPU ポリシーを指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● shared: (デフォルト) インスタンスの仮想 CPU は、ホストの物理 CPU 全体で共有されます。 ● dedicated: インスタンスの仮想 CPU をホストの物理 CPU のセットに固定します。これにより、インスタンスが固定される CPU のトポロジに一致するインスタンス CPU のトポロジが作成されます。このオプションの場合、必然的にオーバーコミット比は 1.0 になります。

キー	説明
hw:cpu_thread_policy	<p>設定が hw:cpu_policy=dedicated の場合に使用する CPU スレッドポリシーを指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● prefer: (デフォルト) ホストは SMT アーキテクチャーを持つ場合と持たない場合があります。SMT アーキテクチャーが存在する場合、Compute スケジューラーはスレッドシブリングを優先します。 ● isolate: ホストは SMT アーキテクチャーを持たないか、あるいは SMT 以外のアーキテクチャーをエミュレートする必要があります。このポリシーにより、Compute スケジューラーは HW_CPU_HYPERTHREADING 特性が設定されていないホストを要求して、SMT を持たないホストにインスタンスを配置するようになります。以下の属性を使用して、この特性を明示的に要求することもできます。 <pre>--property trait:HW_CPU_HYPERTHREADING=forbidden</pre> <p>ホストが SMT アーキテクチャーを持たない場合、Compute サービスはそれぞれの仮想 CPU を想定どおりに異なるコアに配置します。ホストが SMT アーキテクチャーを持つ場合は、動作は [workarounds]/disable_fallback_pcpu_query パラメーターの設定により決定されます。</p> <ul style="list-style-type: none"> ○ True: SMT アーキテクチャーを持つホストは使用されず、スケジューリングに失敗します。 ○ False: Compute サービスはそれぞれの仮想 CPU を異なる物理コアに配置します。Compute サービスは、別のインスタンスからの仮想 CPU を同じコア上に配置しません。したがって、使用される各コアのスレッドシブリングは、1つを除きすべて使用できなくなります。 ● require: ホストは SMT アーキテクチャーを持つ必要があります。このポリシーにより、Compute スケジューラーは HW_CPU_HYPERTHREADING 特性が設定されたホストを要求して、SMT を持つホストにインスタンスを配置するようになります。以下の属性を使用して、この特性を明示的に要求することもできます。 <pre>--property trait:HW_CPU_HYPERTHREADING=required</pre> <p>Compute サービスは、それぞれの仮想 CPU をスレッドシブリングに割り当てます。ホストが SMT アーキテクチャーを持たない場合、そのホストは使用されません。ホストは SMT アーキテクチャーを持つが、スレッドシブリングが使用されていないコアが十分でない場合、スケジューリングに失敗します。</p>

インスタンス PCI NUMA アフィニティーポリシー

以下の表に示す属性キーを使用して、PCI パススルーデバイスおよび SR-IOV インターフェイスの NUMA アフィニティーポリシーを指定するフレーバーを作成します。

表8.13 PCI NUMA アフィニティーポリシー用のフレーバーメタデータ

キー	説明
hw:pci_numa_affinity_policy	<p>PCI パススルーデバイスおよび SR-IOV インターフェイスの NUMA アフィニティーポリシーを指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● required: インスタンスの NUMA ノードの少なくとも 1 つが PCI デバイスとのアフィニティーを持つ場合に限り、Compute サービスは PCI デバイスを要求するインスタンスを作成します。このオプションは、最高のパフォーマンスを提供します。 ● preferred: Compute サービスは、NUMA アフィニティーに基づきベストエフォートで PCI デバイスの選択を試みます。これができない場合には、Compute サービスは PCI デバイスとのアフィニティーを持たない NUMA ノード上でインスタンスをスケジュールします。 ● legacy: (デフォルト) 以下のどちらかのケースで、Compute サービスは PCI デバイスを要求するインスタンスを作成します。 <ul style="list-style-type: none"> ○ PCI デバイスが少なくとも 1 つの NUMA ノードとのアフィニティーを持つ。 ○ PCI デバイスが NUMA アフィニティーに関する情報を提供しない。

インスタンスの NUMA トポロジー

属性を使用して、インスタンスの仮想 CPU スレッドのホスト NUMA 配置、ならびにホスト NUMA ノードからのインスタンスの仮想 CPU およびメモリーの割り当てを定義するフレーバーを作成することができます。

メモリーおよび仮想 CPU の割り当てがコンピュータホスト内の NUMA ノードのサイズよりも大きいフレーバーの場合、インスタンスの NUMA トポロジーを定義するとインスタンス OS のパフォーマンスが向上します。

Compute スケジューラーは、これらの属性を使用してインスタンスに適したホストを決定します。たとえば、クラウドユーザーは以下のフレーバーを使用してインスタンスを起動します。

```
$ openstack flavor set numa_top_flavor \
--property hw:numa_nodes=2 \
--property hw:numa_cpus.0=0,1,2,3,4,5 \
--property hw:numa_cpus.1=6,7 \
--property hw:numa_mem.0=3072 \
--property hw:numa_mem.1=1024
```

Compute スケジューラーは、2 つの NUMA ノード (1 つは 3 GB の RAM を持ち 6 つの CPU を実行できるノード、もう 1 つは 1 GB の RAM を持ち 2 つの CPU を実行できるノード) を持つホストを探しま

す。4 GB の RAM を持ち 8 つの CPU を実行できる 1 つの NUMA ノードを持つホストの場合、Compute スケジューラーは有効な一致とは見なしません。



注記

フレーバーで定義された NUMA トポロジは、イメージで定義された NUMA トポロジでオーバーライドされることはありません。イメージの NUMA トポロジがフレーバーの NUMA トポロジと競合する場合、Compute サービスは **ImageNUMATopologyForbidden** エラーを報告します。

注意

この機能を使用して、インスタンスを特定のホスト CPU または NUMA ノードに制限することはできません。包括的なテストおよびパフォーマンス計測が完了した後にのみ、この機能を使用してください。代わりに **hw:pci_numa_affinity_policy** プロパティを使用することができます。

以下の表に示す属性キーを使用して、インスタンスの NUMA トポロジを定義します。

表8.14 NUMA トポロジ用のフレーバーメタデータ

キー	説明
hw:numa_nodes	<p>インスタンスの仮想 CPU スレッドの実行先として制限するホスト NUMA ノードの数を指定します。指定しない場合は、利用可能な任意の数のホスト NUMA ノード上で仮想 CPU スレッドを実行することができます。</p>
hw:numa_cpus.N	<p>インスタンスの NUMA ノード N にマッピングするインスタンスの仮想 CPU のコンマ区切りリスト。このキーを指定しない場合、仮想 CPU は利用可能な NUMA ノード間で均等に配分されます。</p> <p>N は 0 から始まります。*.N 値を使用する場合には注意が必要です。NUMA ノードが少なくとも 2 つある場合に限り使用してください。</p> <p>この属性は hw:numa_nodes を設定している場合にのみ有効で、インスタンスの NUMA ノードに CPU および RAM が対称的に割り当てられていない場合 (一部の NFV 負荷で重要) にのみ必要です。</p>
hw:numa_mem.N	<p>インスタンスの NUMA ノード N にマッピングするインスタンスのメモリー容量 (MB 単位)。このキーを指定しない場合、メモリーは利用可能な NUMA ノード間で均等に配分されます。</p> <p>N は 0 から始まります。*.N 値を使用する場合には注意が必要です。NUMA ノードが少なくとも 2 つある場合に限り使用してください。</p> <p>この属性は hw:numa_nodes を設定している場合にのみ有効で、インスタンスの NUMA ノードに CPU および RAM が対称的に割り当てられていない場合 (一部の NFV 負荷で重要) にのみ必要です。</p>

**警告**

hw:numa_cpus.N で指定する仮想 CPU の総数または **hw:numa_mem.N** で指定するメモリー容量が、それぞれ利用可能な CPU の数またはメモリー容量よりも大きい場合、Compute サービスは例外を発生させます。

インスタンスメモリーの暗号化

以下の表に示す属性キーを使用して、インスタンスのメモリーの暗号化を有効にします。

表8.15 メモリー暗号化用のフレーバーメタデータ

キー	説明
hw:mem_encryption	インスタンスのメモリーの暗号化を要求するには、 True に設定します。詳しくは、 Configuring AMD SEV Compute nodes to provide memory encryption for instances を参照してください。

CPU リアルタイムポリシー


以下の表に示す属性キーを使用して、インスタンス内のプロセッサのリアルタイムポリシーを定義します。

**注記**

- インスタンスのほとんどの仮想 CPU は、リアルタイムポリシーを設定して実行することができますが、非リアルタイムのゲストプロセスとエミュレーターのオーバーヘッドプロセスの両方に使用するために、少なくとも1つの仮想 CPU を非リアルタイムと識別する必要があります。
- この追加スペックを使用するには、ピンングされた CPU を有効にする必要があります。

表8.16 CPU リアルタイムポリシー用のフレーバーメタデータ

キー	説明
hw:cpu_realtime	リアルタイムポリシーをインスタンスの仮想 CPU に割り当てるフレーバーを作成するには、 yes に設定します。 デフォルト: no

キー	説明
hw:cpu_realtime_mask	<p>リアルタイムポリシーを割り当てない仮想 CPU を指定します。マスクする値の前にキャレット記号 (^) を追加する必要があります。仮想 CPU 0 および 1 を除くすべての仮想 CPU にリアルタイムポリシーを設定する場合の例を以下に示します。</p> <pre>\$ openstack flavor set <flavor> \ --property hw:cpu_realtime="yes" \ --property hw:cpu_realtime_mask="^0-1"</pre> <p> 注記</p> <p>hw_cpu_realtime_mask 属性がイメージで設定されている場合、フレーバーで設定した hw:cpu_realtime_mask 属性よりも優先されます。</p>

エミュレータースレッドポリシー

物理 CPU をインスタンスに割り当てて、エミュレータースレッドに使用することができます。エミュレータースレッドとは、インスタンスと直接関係しないエミュレータープロセスを指します。リアルタイム負荷には、専用のエミュレータースレッド用物理 CPU が必要です。エミュレータースレッドポリシーを使用するには、以下の属性を設定してピンングされた CPU を有効にする必要があります。

```
--property hw:cpu_policy=dedicated
```

以下の表に示す属性キーを使用して、インスタンスのエミュレータースレッドポリシーを定義します。

表8.17 エミュレータースレッドポリシー用のフレーバーメタデータ

キー	説明
hw:emulator_threads_policy	<p>インスタンスに使用するエミュレータースレッドポリシーを指定します。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● share: エミュレータースレッドは、NovaComputeCpuSharedSet heat パラメーターで定義される物理 CPU 全体で共有されます。NovaComputeCpuSharedSet が設定されていない場合、エミュレータースレッドはインスタンスに関連付けられたピンングされた CPU 全体で共有されます。 ● isolate: エミュレータースレッド用に、インスタンスごとに専用の物理 CPU をさらに確保します。このポリシーは過度にリソースを消費するので、使用には注意が必要です。 ● unset: (デフォルト) エミュレータースレッドポリシーは有効ではなく、エミュレータースレッドはインスタンスに関連付けられたピンングされた CPU 全体で共有されます。

インスタンスのメモリーページサイズ

以下の表に示す属性キーを使用して、明示的なメモリーページサイズでインスタンスを作成します。

表8.18 メモリーページサイズ用のフレーバーメタデータ

キー	説明
hw:mem_page_size	<p>インスタンスをサポートするのに使用するラージページのサイズを指定します。このオプションを使用すると、hw:numa_nodes で特に指定しない限り、1 NUMA ノードの暗黙的な NUMA トポロジが作成されます。以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> ● large: ホストでサポートされる最小のページサイズより大きいページサイズを選択します。x86_64 システムでは 2 MB または 1GB です。 ● small: ホストでサポートされる最小のページサイズを選択します。X86_64 システムでは、4 kB (通常のページ) です。 ● any: libvirt ドライバーで決定される、利用可能な最大のヒュージページサイズを選択します。 ● <pagesize>: (文字列) ワークロードに具体的な要件がある場合、ページサイズを明示的に設定します。ページサイズには整数値を使用し、kB またはその他の標準単位で指定します。(例: 4KB、2MB、2048、1GB)。 ● unset: (デフォルト) インスタンスをサポートするのにラージページは使用されず、暗黙的な NUMA トポロジは生成されません。

PCI パススルー

以下の表に示す属性キーを使用して、グラフィックカードやネットワークデバイス等の物理 PCI デバイスをインスタンスにアタッチします。PCI パススルーの使用に関する詳細は、[Configuring PCI passthrough](#) を参照してください。

表8.19 PCI パススルー用のフレーバーメタデータ

キー	説明
pci_passthrough:alias	<p>以下の形式を使用して、インスタンスに割り当てる PCI デバイスを指定します。</p> <pre><alias>:<count></pre> <ul style="list-style-type: none"> ● <alias> を、特定の PCI デバイスクラスに対応するエイリアスに置き換えます。 ● <count> を、インスタンスに割り当てる種別 <alias> の PCI デバイスの数に置き換えます。

ハイパーバイザーの署名

以下の表に示す属性キーを使用して、ハイパーバイザーの署名をインスタンスからは非表示にします。

表8.20 ハイパーバイザーの署名を非表示にするためのフレーバーメタデータ

キー	説明
<code>hide_hypervisor_id</code>	ハイパーバイザーの署名をインスタンスからは非表示にするには、 True に設定します。これにより、すべてのドライバーがインスタンスで読み込みおよび操作を行うことができます。

インスタンスのリソース特性

各リソースプロバイダーには特性のセットがあります。特性は、ストレージディスクの種別や Intel CPU 拡張命令セットなど、リソースプロバイダーの機能的な要素です。インスタンスは、これらの中から要求する特性を指定することができます。

指定することのできる特性は **os-traits** ライブラリーで定義されます。特性の例を以下に示します。

- **COMPUTE_TRUSTED_CERTS**
- **COMPUTE_NET_ATTACH_INTERFACE_WITH_TAG**
- **COMPUTE_IMAGE_TYPE_RAW**
- **HW_CPU_X86_AVX**
- **HW_CPU_X86_AVX512VL**
- **HW_CPU_X86_AVX512CD**

os-traits ライブラリーの使用方法の詳細は、[Usage](#) を参照してください。

以下の表に示す属性キーを使用して、インスタンスのリソース特性を定義します。

表8.21 リソース特性用のフレーバーメタデータ

キー	説明
<code>trait:<trait_name></code>	<p>コンピュートノードの特性を指定します。特性を、以下の有効な値のいずれかに設定します。</p> <ul style="list-style-type: none"> • required: インスタンスをホストするために選択したコンピュートノードになければいけない特性 • forbidden: インスタンスをホストするために選択したコンピュートノードにあってはいけない特性 <p>例:</p> <pre>\$ openstack flavor set --property trait:HW_CPU_X86_AVX512BW=required avx512- flavor</pre>

インスタンスのベアメタルリソースクラス

以下の表に示す属性キーを使用して、インスタンスのベアメタルリソースクラスを要求します。

表8.22 ベアメタルリソースクラス用のフレーバーメタデータ

キー	説明
resources:<resource_class_name>	<p>この属性を使用して、値をオーバーライドする標準のベアメタルリソースクラスを指定するか、インスタンスが要求するカスタムのベアメタルリソースクラスを指定します。</p> <p>オーバーライドすることができる標準のリソースクラスは VCPU、MEMORY_MB、および DISK_GB です。Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぐには、標準のリソースクラスの値を 0 に設定します。</p> <p>カスタムリソースクラスの名前は、CUSTOM_ で始まる必要があります。Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、すべての句読点をアンダースコアに置き換え、CUSTOM_ の接頭辞を追加します。</p> <p>たとえば、--resource-class baremetal.SMALL のノードにインスタンスをスケジュールするには、以下のフレーバーを作成します。</p> <pre>\$ openstack flavor set \ --property \ resources:CUSTOM_BAREMETAL_SMALL=1 \ --property resources:VCPU=0 --property \ resources:MEMORY_MB=0 \ --property resources:DISK_GB=0 compute-small</pre>

第9章 インスタンスへのメタデータの追加

Compute (nova) サービスは、メタデータを使用してインスタンスの起動時に設定情報を渡します。インスタンスは、コンフィグドライブまたはメタデータサービスを使用してメタデータにアクセスすることができます。

コンフィグドライブ

コンフィグドライブは、インスタンスのブート時にアタッチすることのできる特別なドライブです。コンフィグドライブは読み取り専用ドライブとしてインスタンスに提示されます。インスタンスはこのドライブをマウントしてそこからファイルを読み取り、通常メタデータサービスから利用する情報を取得することができます。

メタデータサービス

Compute サービスは、REST API としてメタデータサービスを提供します。これを使用して、インスタンス固有のデータを取得することができます。インスタンスは、**169.254.169.254** または **fe80::a9fe:a9fe** からこのサービスにアクセスします。

9.1. インスタンスメタデータの種別

クラウドユーザー、クラウド管理者、および Compute サービスは、メタデータをインスタンスに渡すことができます。

クラウドユーザーが提供するデータ

クラウドユーザーは、インスタンスがブート時に実行するシェルスクリプトなど、インスタンスを起動する際に使用する追加データを指定することができます。クラウドユーザーは、インスタンスを作成または更新する際に、ユーザーデータ機能を使用し、キー/値のペアを必要な属性として渡すことで、データをインスタンスに渡すことができます。

クラウド管理者が提供するデータ

RHOSP 管理者は、ベンダーデータ機能を使用してデータをインスタンスに渡します。Compute サービスの提供するベンダーデータモジュール **StaticJSON** および **DynamicJSON** により、管理者はメタデータをインスタンスに渡すことができます。

- **StaticJSON**: (デフォルト) 全インスタンスで共通のメタデータに使用します。
- **DynamicJSON**: 各インスタンスで異なるメタデータに使用します。このモジュールは外部の REST サービスにリクエストを行い、インスタンスに追加するメタデータを決定します。

ベンダーデータの設定は、インスタンスの以下の読み取り専用ファイルのいずれかにあります。

- `/openstack/{version}/vendor_data.json`
- `/openstack/{version}/vendor_data2.json`

Compute サービスが提供するデータ

Compute サービスはメタデータサービスの内部実装を使用して、要求されたインスタンスのホスト名やインスタンスが属するアベイラビリティゾーン等の情報をインスタンスに渡します。この操作はデフォルトで実施され、クラウドユーザーまたは管理者が設定を行う必要はありません。

9.2. 全インスタンスへのコンフィグドライブの追加

管理者は Compute サービスを設定し、常にインスタンス用のコンフィグドライブを作成し、コンフィグドライブにデプロイメント固有のメタデータを設定することができます。たとえば、以下の理由によりコンフィグドライブを使用する場合があります。

- デプロイメントにおいて、インスタンスへの IP アドレスの割り当てに DHCP を使用しない場合に、ネットワーク設定を渡すため。インスタンスのネットワーク設定を行う前に、コンフィグドライブを通じてインスタンスの IP アドレス設定を渡すことができます。インスタンスは、コンフィグドライブをマウントして設定にアクセスすることができます。
- Active Directory ポストブートにインスタンスを登録するのに使用される暗号化トークン等、インスタンスを起動するユーザーがアクセスできないデータをインスタンスに渡すため。
- ローカルにキャッシュされたディスク読み取りを作成し、インスタンスのリクエストの負荷を管理するため。これにより、ファクトのチェックインおよびビルドのために定期的にメタデータサーバーにアクセスするインスタンスの影響が軽減されます。

ISO 9660 または VFAT ファイルシステムをマウントできるインスタンスのオペレーティングシステムは、すべてコンフィグドライブを使用することができます。

手順

1. Compute 環境ファイルを開きます。
2. インスタンスの起動時に常にコンフィグドライブをアタッチするには、以下のパラメーターを **True** に設定します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::compute::force_config_drive: 'true'
```

3. (オプション) コンフィグドライブの形式をデフォルト値の **iso9660** から **vfat** に変更するには、設定に **config_drive_format** パラメーターを追加します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::compute::force_config_drive: 'true'
    nova::compute::config_drive_format: vfat
```

4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml \
```

検証

1. インスタンスを作成します。

```
(overcloud)$ openstack server create --flavor m1.tiny \
--image cirros test-config-drive-instance
```

2. インスタンスにログインします。
3. コンフィグドライブをマウントします。

- インスタンスの OS が **udev** を使用する場合:

```
# mkdir -p /mnt/config
# mount /dev/disk/by-label/config-2 /mnt/config
```

- インスタンスの OS が **udev** を使用しない場合は、まずコンフィグドライブに対応するブロックデバイスを特定する必要があります。

```
# blkid -t LABEL="config-2" -o device
/dev/vdb
# mkdir -p /mnt/config
# mount /dev/vdb /mnt/config
```

4. マウントされたコンフィグドライブディレクトリー **mnt/config/openstack/{version}/** で、メタデータのファイルを検査します。

9.3. インスタンスへの動的メタデータの追加

デプロイメントを設定してインスタンス固有のメタデータを作成し、そのインスタンスが JSON ファイルを使用してメタデータを利用できるようにすることができます。

ヒント

アンダークラウド上で動的メタデータを使用して、director を Red Hat Identity Management (IdM) サーバーと統合することができます。IdM サーバーは認証局として使用することができます、オーバークラウドで SSL/TLS が有効な場合にオーバークラウドの証明書を管理することができます。詳細は、[IdM へのアンダークラウドの追加](#) を参照してください。

手順

1. Compute 環境ファイルを開きます。
2. ベンダーデータプロバイダーモジュールに **DynamicJSON** を追加します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
    ...
    api/vendordata_providers:
      value: StaticJSON,DynamicJSON
```

3. メタデータを生成するためにアクセスする REST サービスを指定します。必要な数だけ目的の REST サービスを指定することができます。以下に例を示します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
    ...
    api/vendordata_providers:
      value: StaticJSON,DynamicJSON
    api/vendordata_dynamic_targets:
      value: target1@http://127.0.0.1:125
    api/vendordata_dynamic_targets:
      value: target2@http://127.0.0.1:126
```

Compute サービスは設定されたターゲットサービスから取得したメタデータが含まれる JSON ファイル **vendordata2.json** を生成し、それをコンフィグドライブディレクトリーに保存します。



注記

ターゲットサービスに同じ名前を複数回使用しないでください。

4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \  
-e [your environment files] \  
-e /home/stack/templates/<compute_environment_file>.yaml
```

第10章 インスタンスのメモリーを暗号化するための AMD SEV コンピュートノードの設定



重要

この機能は、本リリースでは **テクノロジープレビュー** として提供しているため、Red Hat では全面的にはサポートしていません。これは、テスト用途にのみご利用いただく機能で、実稼働環境にデプロイすべきではありません。テクノロジープレビュー機能についての詳しい情報は、[対象範囲の詳細](#) を参照してください。

クラウドユーザーは、メモリーの暗号化が有効な SEV 対応コンピュートノード上で動作するインスタンスを作成することができます。

この機能は、2nd Gen AMD EPYC™ 7002 Series (Rome) から利用できます。

クラウドユーザーがメモリーの暗号化を使用するインスタンスを作成できるようにするには、以下のタスクを実施する必要があります。

1. メモリーの暗号化用に AMD SEV コンピュートノードを指定する。
2. メモリーの暗号化用にコンピュートノードを設定する。
3. オークラウドをデプロイする。
4. メモリーを暗号化してインスタンスを起動するためのフレーバーまたはイメージを作成する。

ヒント

AMD SEV ハードウェアが制限されている場合は、ホストアグリゲートを設定して AMD SEV コンピュートノードでのスケジューリングを最適化することもできます。メモリーの暗号化を要求するインスタンスのみを AMD SEV コンピュートノードにスケジュールするには、AMD SEV ハードウェアを持つコンピュートノードのホストアグリゲートを作成し、Compute スケジューラーがメモリーの暗号化を要求するインスタンスのみをホストアグリゲートに配置するように設定します。詳細は、[Creating and managing host aggregates](#) および [Filtering by isolating host aggregates](#) を参照してください。

10.1. SECURE ENCRYPTED VIRTUALIZATION (SEV)

AMD が提供する Secure Encrypted Virtualization (SEV) は、動作中の仮想マシンインスタンスが使用している DRAM のデータを保護します。SEV は、各インスタンスのメモリーを一意的な鍵で暗号化します。

SEV は、不揮発性メモリーテクノロジー (NVDIMM) を使用する際にセキュリティを強化します。ハードドライブと同様に、NVDIMM チップはデータが保存されたままシステムから物理的に取り外すことができるためです。暗号化しないと、機密データ、パスワード、またはシークレットキー等の保存された情報が危険にさらされる可能性があります。

詳細は、[AMD Secure Encrypted Virtualization \(SEV\)](#) のドキュメントを参照してください。

メモリー暗号化を使用する場合のインスタンスの制限

- メモリー暗号化を使用するインスタンスのライブマイグレーションや、インスタンスを一時停止および再開することはできません。

- PCI パススルーを使用して、メモリーの暗号化を使用するインスタンス上のデバイスに直接アクセスすることはできません。
- kernel-4.18.0-115.el8 (RHEL-8.1.0) 以前の Red Hat Enterprise Linux (RHEL) カーネルでメモリー暗号化を使用するインスタンスのブートディスクとして **virtio-blk** を使用することはできません。



注記

virtio-scsi または **SATA** をブートディスクとして使用することができます。また、ブートディスク以外の用途に **virtio-blk** を使用することができます。

- 暗号化されたインスタンスで実行されているオペレーティングシステムは、SEV をサポートしている必要があります。詳細は、Red Hat ナレッジベースのソリューション [Enabling AMD Secure Encrypted Virtualization in RHEL 8](#) を参照してください。
- SEV をサポートするマシンでは、暗号鍵を格納するためのメモリーコントローラーのスロット数に制限があります。動作中のメモリーが暗号化された各インスタンスは、これらのスロットの1つを使用します。したがって、同時に実行できるメモリー暗号化インスタンスの数は、メモリーコントローラーのスロット数に制限されます。たとえば、1st Gen AMD EPYC™ 7001 Series (Naples) の場合、制限は 16 で、2nd Gen AMD EPYC™ 7002 Series (Rome) では上限は 255 です。
- メモリー暗号化を使用するインスタンスの RAM ページの固定 Compute サービスはこれらのページをスワップすることができないため、メモリーが暗号化されたインスタンスをホストするコンピュータノードでメモリーをオーバーコミットすることはできません。
- NUMA ノードが複数あるインスタンスでは、メモリーの暗号化を使用することはできません。

10.2. メモリー暗号化用 AMD SEV コンピュートノードの指定

メモリーの暗号化を使用するインスタンス用に AMD SEV コンピュートノードを指定するには、AMD SEV ロールを設定するための新規ロールファイルを作成し、メモリーの暗号化のためにコンピュータノードをタグ付けするための新規オーバークラウドフレーバーおよび AMD SEV リソースクラスを設定する必要があります。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. オーバークラウドに必要なその他のロールに加えて **ComputeAMDSEV** ロールが含まれる新しいロールデータファイルを生成します。以下の例では、ロールデータファイル **roles_data_amd_sev.yaml** を生成します。これには、**Controller** および **ComputeAMDSEV** ロールが含まれます。

```
(undercloud)$ openstack overcloud roles \
generate -o /home/stack/templates/roles_data_amd_sev.yaml \
Compute:ComputeAMDSEV Controller
```

4. **roles_data_amd_sev.yaml** を開き、以下のパラメーターおよびセクションを編集または追加します。

セクション/パラメーター	現在の値	新しい値
ロールのコメント	Role: Compute	Role: ComputeAMDSEV
ロール名	Compute	name: ComputeAMDSEV
description	Basic Compute Node role	AMD SEV Compute Node role
HostnameFormatDefault	%stackname%-novacompute-%index%	%stackname%-novacomputeamdsev-%index%
deprecated_nic_config_name	compute.yaml	compute-amd-sev.yaml

5. オーバークラウド用の AMD SEV コンピュートノードをノード定義のテンプレート **node.json** または **node.yaml** に追加して、そのノードを登録します。詳しい情報は、**Director Installation and Usage** の [Registering nodes for the overcloud](#) を参照してください。
6. ノードのハードウェアを検査します。

```
(undercloud)$ openstack overcloud node introspect \
--all-manageable --provide
```

詳細は、**Director Installation and Usage** ガイドの [Creating an inventory of the bare-metal node hardware](#) を参照してください。

7. AMD SEV コンピュートノード用の **compute-amd-sev** オーバークラウドフレーバーを作成します。

```
(undercloud)$ openstack flavor create --id auto \
--ram <ram_size_mb> --disk <disk_size_gb> \
--vcpus <no_vcpus> compute-amd-sev
```

- **<ram_size_mb>** をベアメタルノードの RAM (MB 単位) に置き換えます。
- **<disk_size_gb>** をベアメタルノード上のディスク容量 (GB 単位) に置き換えます。
- **<no_vcpus>** をベアメタルノードの CPU 数に置き換えます。



注記

これらの属性は、インスタンスのスケジューリングには使用されません。ただし Compute スケジューラーは、ディスク容量を使用してルートパーティションのサイズを決定します。

8. ノード一覧を取得して UUID を把握します。

```
(undercloud)$ openstack baremetal node list
```

9. メモリーの暗号化用に指定する各ベアメタルノードに、カスタムの AMD SEV リソースクラスをタグ付けします。

```
(undercloud)$ openstack baremetal node set \
--resource-class baremetal.AMD-SEV <node>
```

<node> をベアメタルノードの ID に置き換えてください。

10. **compute-amd-sev** フレーバーをカスタム AMD SEV リソースクラスに関連付けます。

```
(undercloud)$ openstack flavor set \
--property resources:CUSTOM_BAREMETAL_AMD_SEV=1 \
compute-amd-sev
```

Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、それぞれの句読点をアンダースコアに置き換え、**CUSTOM_** の接頭辞を追加します。



注記

フレーバーが要求できるのは、ベアメタルリソースクラスの1つのインスタンスだけです。

11. 以下のフレーバー属性を設定して、Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぎます。

```
(undercloud)$ openstack flavor set \
--property resources:VCPU=0 --property resources:MEMORY_MB=0 \
--property resources:DISK_GB=0 compute-amd-sev
```

12. (オプション) **ComputeAMDSEV** ロールのネットワークポロジが **Compute** ロールのネットワークポロジと異なる場合は、カスタムネットワークインターフェーステンプレートを作成します。詳しくは、**Advanced Overcloud Customization** の [Custom network interface templates](#) を参照してください。

ComputeAMDSEV ロールのネットワークポロジが **Compute** ロールと同じ場合は、**compute.yaml** で定義されるデフォルトのネットワークポロジを使用することができます。

13. **ComputeAMDSEV** ロールの **Net::SoftwareConfig** を **network-environment.yaml** ファイルに登録します。

```
resource_registry:
  OS::TripleO::Compute::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/compute.yaml
  OS::TripleO::ComputeCPUPinning::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/<amd_sev_net_top>.yaml
  OS::TripleO::Controller::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/controller.yaml
```

<amd_sev_net_top> を **ComputeAMDSEV** ロールのネットワークポロジが含まれるファイルの名前に置き換えます。たとえば、デフォルトのネットワークポロジを使用する場合は **compute.yaml** です。

14. 以下のパラメーターを **node-info.yaml** ファイルに追加して、AMD SEV コンピュートノードの数および AMD SEV 対応コンピュートノード用に使用するフレーバーを指定します。

```
parameter_defaults:
  OvercloudComputeAMDSEVFlavor: compute-amd-sev
  ComputeAMDSEVCount: 3
```

15. ロールが作成されたことを確認するには、以下のコマンドを入力します。

```
(undercloud)$ openstack baremetal node list --long -c "UUID" \
-c "Instance UUID" -c "Resource Class" -c "Provisioning State" \
-c "Power State" -c "Last Error" -c "Fault" -c "Name" -f json
```

出力例:

```
[
  {
    "Fault": null,
    "Instance UUID": "e8e60d37-d7c7-4210-acf7-f04b245582ea",
    "Last Error": null,
    "Name": "compute-0",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "baremetal.AMD-SEV",
    "UUID": "b5a9ac58-63a7-49ba-b4ad-33d84000ccb4"
  },
  {
    "Fault": null,
    "Instance UUID": "3ec34c0b-c4f5-4535-9bd3-8a1649d2e1bd",
    "Last Error": null,
    "Name": "compute-1",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "compute",
    "UUID": "432e7f86-8da2-44a6-9b14-dfacdf611366"
  },
  {
    "Fault": null,
    "Instance UUID": "4992c2da-adde-41b3-bef1-3a5b8e356fc0",
    "Last Error": null,
    "Name": "controller-0",
    "Power State": "power on",
    "Provisioning State": "active",
    "Resource Class": "controller",
    "UUID": "474c2fc8-b884-4377-b6d7-781082a3a9c0"
  }
]
```

10.3. メモリー暗号化用 AMD SEV コンピュートノードの設定

クラウドユーザーがメモリーの暗号化を使用するインスタンスを作成できるようにするには、AMD SEV ハードウェアを持つコンピュートノードを設定する必要があります。

前提条件

- デプロイメントには、SEV に対応する AMD ハードウェア (AMD EPYC CPU 等) 上で実行されるコンピュートノードが含まれている必要があります。以下のコマンドを使用して、デプロイメントが SEV に対応しているかどうか判断することができます。

```
$ lscpu | grep sev
```

手順

1. Compute 環境ファイルを開きます。
2. オプション: AMD SEV コンピュートノードが同時にホストできるメモリーが暗号化されたインスタンス数の最大値を指定するには、以下の設定を Compute 環境ファイルに追加します。

```
parameter_defaults:
  ComputeAMDSEVExtraConfig:
    nova::config::nova_config:
      libvirt/num_memory_encrypted_guests:
        value: 15
```



注記

libvirt/num_memory_encrypted_guests パラメーターのデフォルト値は **none** です。カスタム値を設定しない場合、AMD SEV コンピュートノードは、ノードが同時にホストできるメモリーが暗号化されたインスタンスの数に制限を設けません。代わりに、ハードウェアが、AMD SEV コンピュートノードが同時にホストできるメモリーが暗号化されたインスタンス数の最大値を決定します。この場合、メモリーが暗号化されたインスタンスの一部が起動に失敗する可能性があります。

3. (オプション) デフォルトでは、すべての x86_64 イメージが q35 マシン種別を使用するように指定するには、以下の設定を Compute 環境ファイルに追加します。

```
parameter_defaults:
  ComputeAMDSEVParameters:
    NovaHWMachineType: x86_64=q35
```

このパラメーターの値を指定する場合、すべての AMD SEV インスタンスイメージで **hw_machine_type** 属性を **q35** に設定する必要はありません。

4. AMD SEV コンピュートノードがホストレベルのサービスが機能するのに十分なメモリーが確保するようにするには、考え得る AMD SEV インスタンスごとに 16 MB を追加します。

```
parameter_defaults:
  ComputeAMDSEVParameters:
    ...
    NovaReservedHostMemory: <libvirt/num_memory_encrypted_guests * 16>
```

5. AMD SEV コンピュートノード用のカーネルパラメーターを設定します。

```
parameter_defaults:
  ComputeAMDSEVParameters:
    ...
```

```
KernelArgs: "hugepagesz=1GB hugepages=32 default_hugepagesz=1GB
mem_encrypt=on kvm_amd.sev=1"
```

6. 更新内容を Compute 環境ファイルに保存します。
7. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

10.4. メモリー暗号化用のイメージの作成

オーバークラウドに AMD SEV コンピュートノードが含まれる場合、AMD SEV インスタンスイメージを作成することができます。クラウドユーザーはこのイメージを使用して、メモリーが暗号化されたインスタンスを起動することができます。

手順

1. メモリー暗号化用の新規イメージの作成

```
(overcloud)$ openstack image create ... \
--property hw_firmware_type=uefi amd-sev-image
```



注記

既存のイメージを使用する場合、イメージの **hw_firmware_type** 属性が **uefi** に設定されている必要があります。

2. (オプション) イメージに属性 **hw_mem_encryption=True** を追加して、イメージで AMD SEV のメモリー暗号化を有効にします。

```
(overcloud)$ openstack image set \
--property hw_mem_encryption=True amd-sev-image
```

ヒント

フレーバーでメモリー暗号化を有効にすることができます。詳細は、[Creating a flavor for memory encryption](#) を参照してください。

3. オプション: コンピュートノード設定のマシン種別がまだ **q35** に設定されていない場合には、そのように設定します。

```
(overcloud)$ openstack image set \
--property hw_machine_type=q35 amd-sev-image
```

4. オプション: SEV 対応ホストアグリゲートでメモリーが暗号化されたインスタンスをスケジューリングするには、イメージの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack image set \
--property trait:HW_CPU_X86_AMD_SEV=required amd-sev-image
```

■

ヒント

フレーバーでこの特性を指定することもできます。詳細は、[Creating a flavor for memory encryption](#) を参照してください。

10.5. メモリー暗号化用のフレーバーの作成

オーバークラウドに AMD SEV コンピュートノードが含まれる場合、1つまたは複数の AMD SEV フレーバーを作成することができます。クラウドユーザーはこのイメージを使用して、メモリーが暗号化されたインスタンスを起動することができます。



注記

AMD SEV フレーバーは、**hw_mem_encryption** 属性がイメージで設定されていない場合にのみ必要です。

手順

1. メモリー暗号化用のフレーバーを作成します。

```
(overcloud)$ openstack flavor create --vcpus 1 --ram 512 --disk 2 \
--property hw:mem_encryption=True m1.small-amd-sev
```

2. SEV 対応ホストアグリゲートでメモリーが暗号化されたインスタンスをスケジュールするには、フレーバーの追加スペックに以下の特性を追加します。

```
(overcloud)$ openstack flavor set \
--property trait:HW_CPU_X86_AMD_SEV=required m1.small-amd-sev
```

10.6. メモリーが暗号化されたインスタンスの起動

メモリーの暗号化を有効にして AMD SEV コンピュートノードでインスタンスを起動できることを確認するには、メモリー暗号化フレーバーまたはイメージを使用してインスタンスを作成します。

手順

1. AMD SEV フレーバーまたはイメージを使用してインスタンスを作成します。以下の例では、[メモリー暗号化用のフレーバーの作成](#) で作成したフレーバーおよび [メモリー暗号化用のイメージの作成](#) で作成したイメージを使用してインスタンスを作成します。

```
(overcloud)$ openstack server create --flavor m1.small-amd-sev \
--image amd-sev-image amd-sev-instance
```

2. クラウドユーザーとしてインスタンスにログインします。
3. インスタンスがメモリーの暗号化を使用していることを確認するには、インスタンスから以下のコマンドを入力します。

```
$ dmesg | grep -i sev
AMD Secure Encrypted Virtualization (SEV) active
```

第11章 インスタンスに永続メモリーを提供する NVDIMM コンピュートノードの設定

不揮発性デュアルインラインメモリーモジュール (NVDIMM) は、メモリーが永続的 (PMEM) な DRAM を提供するテクノロジーです。標準的なコンピューターのメモリーは、電源が切れるとそのデータを喪失します。NVDIMM は、電源が切れてもそのデータを維持します。PMEM を使用するインスタンスでは、アプリケーションは大量の連続したメモリーセグメントを読み込むことができ、このセグメントは電源の切断後もアプリケーションデータを維持します。この機能は、大量のメモリーを要求する高パフォーマンスコンピューティング (HPC) に役立ちます。

NVDIMM ハードウェアを持つコンピュートノードで PMEM 名前空間を作成および設定することで、クラウド管理者はインスタンスが仮想 PMEM (vPMEM) として PMEM を利用できるようにすることができます。これにより、インスタンスのシャットダウン後にコンテンツを維持する必要がある場合、クラウドユーザーは vPMEM を要求するインスタンスを作成することができます。

クラウドユーザーが PMEM を使用するインスタンスを作成できるようにするには、以下の手順を完了する必要があります。

1. PMEM 用のコンピュートノードを指定する。
2. NVDIMM ハードウェアを持つ PMEM 用のコンピュートノードを設定する。
3. オーバークラウドをデプロイする。
4. 仮想 PMEM が設定されたインスタンスを起動するための PMEM フレーバーを作成する。

ヒント

NVDIMM ハードウェアが制限されている場合は、ホストアグリゲートを設定して PMEM コンピュートノードでのスケジューリングを最適化することもできます。vPMEM を要求するインスタンスのみを PMEM コンピュートノードにスケジュールするには、NVDIMM ハードウェアを持つコンピュートノードのホストアグリゲートを作成し、Compute スケジューラーが PMEM インスタンスのみをホストアグリゲートに配置するように設定します。詳細は、[Creating and managing host aggregates](#) および [Filtering by isolating host aggregates](#) を参照してください。

前提条件

- コンピュートノードに、Intel® Optane™ DC Persistent Memory 等の永続メモリーハードウェアが使用されている。
- PMEM ハードウェアデバイスにバックエンド NVDIMM リージョンを設定し、PMEM 名前空間を作成している。Intel が提供する [ipmctl](#) ツールを使用して、PMEM ハードウェアを設定することができます。

PMEM デバイスを使用する際の制限

- 仮想 PMEM を使用するインスタンスについて、コールドマイグレーション、ライブマイグレーション、サイズ変更、または休止および再開を行うことはできません。
- 仮想 PMEM を使用することができるのは、RHEL8 を実行しているインスタンスだけです。
- 仮想 PMEM インスタンスを再ビルドする場合、永続メモリー名前空間は削除され、インスタンスの初期状態に戻ります。

- 新しいフレーバーを使用してインスタンスのサイズを変更する場合、元の仮想永続メモリーの内容は新しい仮想永続メモリーにはコピーされません。
- 仮想永続メモリーのホットプラグはサポートされていません。
- 仮想 PMEM インスタンスのスナップショットを作成する場合、仮想永続イメージは含まれません。

11.1. PMEM 用コンピュートノードの指定

PMEM 負荷用のコンピュートノードを指定するには、PMEM ロールを設定するための新規ロールファイルを作成し、NVDIMM コンピュートノードをタグ付けするための PMEM 用新規オーバークラウドフレーバーおよびリソースクラスを設定する必要があります。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. **roles_data_pmem.yaml** という名前で、**Controller**、**Compute**、および **ComputePMEM** ロールが含まれる新しいロールデータファイルを生成します。

```
(undercloud)$ openstack overcloud roles \
generate -o /home/stack/templates/roles_data_pmem.yaml \
Compute:ComputePMEM Compute Controller
```

4. **roles_data_pmem.yaml** を開き、以下のパラメーターおよびセクションを編集または追加します。

セクション/パラメーター	現在の値	新しい値
ロールのコメント	Role: Compute	Role: ComputePMEM
ロール名	Compute	name: ComputePMEM
description	Basic Compute Node role	PMEM Compute Node role
HostnameFormatDefault	%stackname%-novacompute-%index%	%stackname%-novacomputepmem-%index%
deprecated_nic_config_name	compute.yaml	compute-pmem.yaml

5. オーバークラウド用の NVDIMM コンピュートノードをノード定義のテンプレート **node.json** または **node.yaml** に追加して、そのノードを登録します。詳細は、**director** のインストールと使用方法の [オーバークラウドノードの登録](#) を参照してください。
6. ノードのハードウェアを検査します。

```
(undercloud)$ openstack overcloud node introspect --all-manageable --provide
```

詳細は、**Director Installation and Usage** ガイドの [Creating an inventory of the bare-metal node hardware](#) を参照してください。

- PMEM コンピュートノード用の **compute-pmem** オーバークラウドフレーバーを作成します。

```
(undercloud)$ openstack flavor create --id auto \
--ram <ram_size_mb> --disk <disk_size_gb> \
--vcpus <no_vcpus> compute-pmem
```

- **<ram_size_mb>** をベアメタルノードの RAM (MB 単位) に置き換えます。
- **<disk_size_gb>** をベアメタルノード上のディスク容量 (GB 単位) に置き換えます。
- **<no_vcpus>** をベアメタルノードの CPU 数に置き換えます。



注記

これらの属性は、インスタンスのスケジューリングには使用されません。ただし Compute スケジューラーは、ディスク容量を使用してルートパーティションのサイズを決定します。

- ノード一覧を取得して UUID を把握します。

```
(undercloud)$ openstack baremetal node list
```

- PMEM 負荷用に指定する各ベアメタルノードに、カスタムの PMEM リソースクラスをタグ付けします。

```
(undercloud)$ openstack baremetal node set \
--resource-class baremetal.PMEM <node>
```

<node> をベアメタルノードの ID に置き換えてください。

- compute-pmem** フレーバーをカスタムの PMEM リソースクラスに関連付けます。

```
(undercloud)$ openstack flavor set \
--property resources:CUSTOM_BAREMETAL_PMEM=1 \
compute-pmem
```

Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、すべての句読点をアンダースコアに置き換え、**CUSTOM_** の接頭辞を追加します。



注記

フレーバーが要求できるのは、ベアメタルリソースクラスの1つのインスタンスだけです。

- 以下のフレーバー属性を設定して、Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぎます。

```
(undercloud)$ openstack flavor set \
--property resources:VCPU=0 --property resources:MEMORY_MB=0 \
--property resources:DISK_GB=0 compute-pmem
```

- 以下のパラメーターを **node-info.yaml** ファイルに追加して、PMEM コンピュートノードの数および PMEM 対応コンピュートノード用に使用するフレーバーを指定します。

```
parameter_defaults:
  OvercloudComputePMEMFlavor: compute-pmem
  ComputePMEMCount: 3 #set to the no of NVDIMM devices you have
```

- ロールが作成されたことを確認するには、以下のコマンドを入力します。

```
(undercloud)$ openstack overcloud profiles list
```

11.2. PMEM 用コンピュートノードの設定

クラウドユーザーが仮想 PMEM を使用するインスタンスを作成できるようにするには、NVDIMM ハードウェアを持つコンピュートノードを設定する必要があります。

手順

- NVDIMM コンピュートノードを設定するための新規 Compute 環境ファイルを作成します (例: **env_pmem.yaml**)。
- NVDIMM リージョンをインスタンスが使用できる PMEM 名前空間に分割するには、Compute 環境ファイルの PMEM ロールに **NovaPMEMNamespaces** ロール固有パラメーターを追加し、以下の形式で値を設定します。

```
<size>:<namespace_name>[,<size>:<namespace_name>]
```

サイズを表すには、以下の接尾辞を使用します。

- KiB:k または K
 - MiB:m または M
 - GiB:g または G
 - TiB:t または T
- たとえば、以下の設定では、4 つの名前空間 (サイズが 6 GiB の 3 つの名前空間およびサイズが 100 GiB の 1 つの名前空間) が作成されます。

```
parameter_defaults:
  ComputePMEMParameters:
    NovaPMEMNamespaces: "6G:ns0,6G:ns1,6G:ns2,100G:ns3"
```

- PMEM 名前空間をフレーバーで使用できるラベルにマッピングするには、Compute 環境ファイルの PMEM ロールに **NovaPMEMMappings** ロール固有パラメーターを追加し、以下の形式で値を設定します。

```
<label>:<namespace_name>[,<namespace_name>][,<label>:<namespace_name>[<namespace_name>]].
```

たとえば、以下の設定では、3 つの 6 GiB 名前空間がラベル 6GB にマッピングされ、100 GiB 名前空間がラベル LARGE にマッピングされます。

```
parameter_defaults:
  ComputePMEMParameters:
    NovaPMEMNamespaces: "6G:ns0,6G:ns1,6G:ns2,100G:ns3"
    NovaPMEMMappings: "6GB:ns0|ns1|ns2,LARGE:ns3"
```

4. 更新内容を Compute 環境ファイルに保存します。
5. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-r /home/stack/templates/roles_data_pmem.yaml \
-e /home/stack/templates/node-info.yaml \
-e [your environment files] \
-e /home/stack/templates/env_pmem.yaml
```

6. クラウドユーザーが仮想 PMEM の設定されたインスタンスを起動するのに使用できるフレーバーを作成および設定します。以下の例では、ステップ 3 でマッピングした小さい PMEM デバイス (6 GB) を要求するフレーバーが作成されます。

```
(overcloud)$ openstack flavor create --vcpus 1 --ram 512 --disk 2 \
--property hw:pmem='6GB' small_pmem_flavor
```

検証

1. PMEM フレーバーのいずれかを使用して、インスタンスを作成します。

```
(overcloud)$ openstack flavor list
(overcloud)$ openstack server create --flavor small_pmem_flavor \
--image rhel8 pmem_instance
```

2. クラウドユーザーとしてインスタンスにログインします。詳細は、[Connecting to an instance](#) を参照してください。
3. インスタンスに接続されたすべてのディスクデバイスを一覧表示します。

```
$ sudo fdisk -l /dev/pmem0
```

一覧表示されるデバイスのいずれかが NVDIMM であれば、インスタンスに仮想 PMEM が設定されています。

第12章 インスタンス用の仮想 GPU の設定

インスタンスで GPU ベースのレンダリングをサポートするには、利用可能な物理 GPU デバイスおよびハイパーバイザーの種別に応じて、仮想 GPU (vGPU) リソースを定義し、管理できます。この設定を使用して、レンダリングの負荷をすべての物理 GPU デバイス間でより効果的に分割し、仮想 GPU 対応のインスタンスをスケジューリングする際の制御性を向上させることができます。

Compute (nova) サービスで仮想 GPU を有効にするには、クラウドユーザーが仮想 GPU デバイスの設定された Red Hat Enterprise Linux (RHEL) インスタンスを作成するのに使用できるフレーバーを作成します。これにより、各インスタンスは物理 GPU デバイスに対応する仮想 GPU デバイスで GPU 負荷に対応することができます。

Compute サービスは、各ホストに定義する GPU プロファイルで利用可能な仮想 GPU デバイスの数を追跡します。Compute サービスはフレーバーに基づいてこれらのホストにインスタンスをスケジューリングし、デバイスをアタッチし、使用状況を継続的に監視します。インスタンスが削除されると、Compute サービスは仮想 GPU デバイスを利用可能なプールに戻します。

重要

Red Hat では、サポート例外を要求せずに RHOSP での NVIDIA 仮想 GPU の使用を有効にしています。ただし、Red Hat は、NVIDIA 仮想 GPU ドライバーのテクニカルサポートを提供しません。NVIDIA 仮想 GPU ドライバーは、NVIDIA により提供され、サポートされます。NVIDIA 仮想 GPU ソフトウェアの NVIDIA Enterprise サポートを取得するには、NVIDIA 認定サポートサービスサブスクリプションが必要です。サポートされるコンポーネントで問題を再現できない NVIDIA 仮想 GPU の使用から生じる問題については、以下のサポートポリシーが適用されます。

- サードパーティーコンポーネントが問題に関与していないと Red Hat が考える場合は、通常の [サポート対象範囲](#) および [Red Hat SLA](#) が適用されます。
- サードパーティーコンポーネントが問題に関与していると Red Hat が考える場合は、お客様は Red Hat の [サードパーティーサポートおよび認定ポリシー](#) に従って NVIDIA に問い合わせを依頼されます。詳細は、ナレッジベースの記事 [Obtaining Support from NVIDIA](#) を参照してください。

12.1. サポートされる設定および制限

サポートされる GPU カード

サポートされる NVIDIA GPU カードの一覧については、NVIDIA の Web サイトで [Virtual GPU Software Supported Products](#) を参照してください。

仮想 GPU デバイスを使用する際の制限

- 各コンピュータノードで有効にできる仮想 GPU の種別は1つだけです。
- 各インスタンスが使用できる仮想 GPU のリソースは1つだけです。
- ホスト間の vGPU インスタンスのライブマイグレーションはサポートされていません。
- vGPU インスタンスの退避はサポートされていません。
- 仮想 GPU インスタンスをホストするコンピュータノードをリブートする必要がある場合、仮想 GPU は自動的に再作成されたインスタンスに再割り当てされません。コンピュータノードをリブートする前にインスタンスのコールドマイグレーションを行うか、リブート後に各仮想 GPU

を正しいインスタンスに手動で割り当てる必要があります。各仮想 GPU を手動で割り当てるには、リブートする前にコンピュータノードで実行される各仮想 GPU インスタンスのインスタンス XML から **mdev** UUID を取得する必要があります。以下のコマンドを使用して、各インスタンスの **mdev** UUID を検出することができます。

```
# virsh dumpxml <instance_name> | grep mdev
```

<instance_name> を、Compute API への **/servers** リクエストで返される libvirt インスタンス名 (**OS-EXT-SRV-ATTR:instance_name**) に置き換えます。

- libvirt の制限により、仮想 GPU 対応インスタンスでの休止操作はサポートされていません。代わりに、インスタンスのスナップショット作成またはシェルフ処理が可能です。
- 仮想 GPU フレーバーが設定されたインスタンスでサイズ変更およびコールドマイグレーション操作を行っても、仮想 GPU のリソースはインスタンスに自動的に再割り当てされません。インスタンスのサイズ変更または移行後に、インスタンスを手動で再ビルドし、仮想 GPU のリソースを再割り当てする必要があります。
- デフォルトでは、コンピュータホストの仮想 GPU の種別は API ユーザーに公開されません。アクセス権限を付与するには、ホストをホストアグリゲートに追加します。詳細は、[Creating and managing host aggregates](#) を参照してください。
- NVIDIA アクセラレーターハードウェアを使用する場合は、NVIDIA ライセンス要件に従う必要があります。たとえば、NVIDIA vGPU GRID にはライセンスサーバーが必要です。NVIDIA のライセンス要件の詳細は、NVIDIA の Web サイトで [Virtual GPU License Server Release Notes](#) を参照してください。

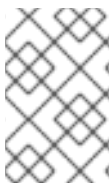
12.2. コンピュートノードでの仮想 GPU の設定

クラウドユーザーが仮想 GPU (vGPU) を使用するインスタンスを作成できるようにするには、物理 GPU を持つコンピュータノードを設定する必要があります。

1. vGPU 用のコンピュータノードを指定する。
2. 仮想 GPU 用のコンピュータノードを設定する。
3. オーバークラウドをデプロイする。
4. vGPU を持つインスタンスを起動するための vGPU フレーバーを作成します。

ヒント

GPU ハードウェアが制限されている場合は、ホストアグリゲートを設定して vGPU コンピュートノードでのスケジューリングを最適化することもできます。仮想 GPU を要求するインスタンスのみを仮想 GPU コンピュートノードにスケジュールするには、仮想 GPU が設定されたコンピュータノードのホストアグリゲートを作成し、Compute スケジューラーが仮想 GPU インスタンスのみをホストアグリゲートに配置するように設定します。詳細は、[Creating and managing host aggregates](#) および [Filtering by isolating host aggregates](#) を参照してください。



注記

NVIDIA GRID vGPU を使用するには、NVIDIA GRID ライセンス要件に従う共に、セルフホストライセンスサーバーの URL が必要です。詳細は、[Virtual GPU License Server Release Notes](#) の Web ページを参照してください。

12.2.1. 前提条件

- NVIDIA の Web サイトから、GPU デバイスに対応する NVIDIA GRID ホストドライバー RPM パッケージをダウンロードしている。必要なドライバーを確認するには、[NVIDIA ドライバーダウンロードポータル](#)を参照してください。ポータルからドライバーをダウンロードするには、NVIDIA カスタマーとして登録されている必要があります。
- NVIDIA GRID ホストドライバーがインストールされているカスタムオーバークラウドイメージをビルドしている。

12.2.2. 仮想 GPU 用コンピュートノードの指定

仮想 GPU 負荷用のコンピュートノードを指定するには、仮想 GPU ロールを設定するための新規ロールファイルを作成し、GPU 対応のコンピュートノードをタグ付けするための新規オーバークラウドフレーバーおよびリソースクラスを設定する必要があります。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. **stackrc** ファイルを取得します。

```
[stack@director ~]$ source ~/stackrc
```

3. **roles_data_gpu.yaml** という名前で、**Controller**、**Compute**、および **ComputeGpu** ロールが含まれる新しいロールデータファイルを生成します。

```
(undercloud)$ openstack overcloud roles \
generate -o /home/stack/templates/roles_data_gpu.yaml \
Compute:ComputeGpu Compute Controller
```

4. **roles_data_gpu.yaml** を開き、以下のパラメーターおよびセクションを編集または追加します。

セクション/パラメーター	現在の値	新しい値
ロールのコメント	Role: Compute	Role: ComputeGpu
name (ロール名)	Compute	ComputeGpu
description	Basic Compute Node role	GPU Compute Node role
ImageDefault	該当なし	overcloud-full-gpu
HostnameFormatDefault	-compute-	-computegpu-
deprecated_nic_config_name	compute.yaml	compute-gpu.yaml

5. オーバークラウド用の GPU 対応コンピュートノードをノード定義のテンプレート **node.json** または **node.yaml** に追加して、そのノードを登録します。詳細は、**director** のインストールと使用方法の [オーバークラウドノードの登録](#) を参照してください。

6. ノードのハードウェアを検査します。

```
(undercloud)$ openstack overcloud node introspect --all-manageable \
--provide
```

詳細は、**Director Installation and Usage** ガイドの [Creating an inventory of the bare-metal node hardware](#) を参照してください。

7. 仮想 GPU コンピュートノード用の **compute-vgpu-nvidia** オーバークラウドフレーバーを作成します。

```
(undercloud)$ openstack flavor create --id auto \
--ram <ram_size_mb> --disk <disk_size_gb> \
--vcpus <no_vcpus> compute-vgpu-nvidia
```

- **<ram_size_mb>** をベアメタルノードの RAM (MB 単位) に置き換えます。
- **<disk_size_gb>** をベアメタルノード上のディスク容量 (GB 単位) に置き換えます。
- **<no_vcpus>** をベアメタルノードの CPU 数に置き換えます。



注記

これらの属性は、インスタンスのスケジューリングには使用されません。ただし Compute スケジューラーは、ディスク容量を使用してルートパーティションのサイズを決定します。

8. ノード一覧を取得して UUID を把握します。

```
(undercloud)$ openstack baremetal node list
```

9. GPU 負荷用に指定する各ベアメタルノードに、カスタムの GPU リソースクラスをタグ付けします。

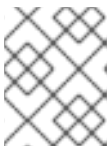
```
(undercloud)$ openstack baremetal node set \
--resource-class baremetal.GPU <node>
```

<node> をベアメタルノードの ID に置き換えてください。

10. **compute-vgpu-nvidia** フレーバーをカスタムの GPU リソースクラスに関連付けます。

```
(undercloud)$ openstack flavor set \
--property resources:CUSTOM_BAREMETAL_GPU=1 \
compute-vgpu-nvidia
```

Bare Metal サービスノードのリソースクラスに対応するカスタムリソースクラスの名前を指定するには、リソースクラスを大文字に変換し、すべての句読点をアンダースコアに置き換え、**CUSTOM_** の接頭辞を追加します。



注記

フレーバーが要求できるのは、ベアメタルリソースクラスの1つのインスタンスだけです。

- 以下のフレーバー属性を設定して、Compute スケジューラーがインスタンスのスケジューリングにベアメタルフレーバー属性を使用するのを防ぎます。

```
(undercloud)$ openstack flavor set \
  --property resources:VCPU=0 --property resources:MEMORY_MB=0 \
  --property resources:DISK_GB=0 compute-vgpu-nvidia
```

- ロールが作成されたことを確認するには、以下のコマンドを入力します。

```
(undercloud)$ openstack overcloud profiles list
```

12.2.3. 仮想 GPU 用コンピュートノードの設定およびオーバークラウドのデプロイ

環境内の物理 GPU デバイスに対応する仮想 GPU の種別を取得して割り当て、仮想 GPU 用コンピュートノードを設定するための環境ファイルを準備する必要があります。

手順

- Red Hat Enterprise Linux と NVIDIA GRID ドライバーを一時コンピュートノードにインストールし、そのノードを起動します。
- コンピュートノードで、有効にする物理 GPU デバイスに対応する仮想 GPU の種別を確認します。libvirt の場合、仮想 GPU は仲介デバイスまたは **mdev** 種別のデバイスです。サポートされている **mdev** デバイスを検出するには、以下のコマンドを入力します。

```
[root@overcloud-computegpu-0 ~]# ls
/sys/class/mdev_bus/0000\:06\:00.0/mdev_supported_types/
nvidia-11 nvidia-12 nvidia-13 nvidia-14 nvidia-15 nvidia-16 nvidia-17 nvidia-18 nvidia-19
nvidia-20 nvidia-21 nvidia-210 nvidia-22

[root@overcloud-computegpu-0 ~]# cat
/sys/class/mdev_bus/0000\:06\:00.0/mdev_supported_types/nvidia-18/description
num_heads=4, frl_config=60, framebuffer=2048M, max_resolution=4096x2160,
max_instance=4
```

- ComputeGpu** ロールの **Net::SoftwareConfig** を **network-environment.yaml** ファイルに登録します。

```
resource_registry:
  OS::TripleO::Compute::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/compute.yaml
  OS::TripleO::ComputeGpu::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/compute-gpu.yaml
  OS::TripleO::Controller::Net::SoftwareConfig: /home/stack/templates/nic-
  configs/controller.yaml
```

- 以下のパラメーターを **node-info.yaml** ファイルに追加して、GPU コンピュートノードの数および GPU が指定されたコンピュートノード用に使用するフレーバーを指定します。

```
parameter_defaults:
  OvercloudControllerFlavor: control
  OvercloudComputeFlavor: compute
  OvercloudComputeGpuFlavor: compute-vgpu-nvidia
```

```
ControllerCount: 1
ComputeCount: 0
ComputeGpuCount: 1
```

5. **gpu.yaml** ファイルを作成し、GPU デバイスの仮想 GPU 種別を指定します。

```
parameter_defaults:
  ComputeGpuExtraConfig:
    nova::compute::vgpu::enabled_vgpu_types:
      - nvidia-18
```



注記

各物理 GPU がサポートするのは1つの仮想 GPU 種別だけです。この属性で複数の仮想 GPU 種別を指定した場合は、最初の種別だけが使用されます。

6. 更新内容を Compute 環境ファイルに保存します。
7. その他の環境ファイルと共に新しいロールファイルおよび環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-r /home/stack/templates/roles_data_gpu.yaml \
-e /home/stack/templates/network-environment.yaml \
-e /home/stack/templates/gpu.yaml \
-e /home/stack/templates/node-info.yaml
```

12.3. カスタム GPU インスタンスイメージの作成

クラウドユーザーが仮想 GPU (vGPU) を使用するインスタンスを作成できるようにするには、インスタンス起動用のカスタムの仮想 GPU 対応イメージを作成します。NVIDIA GRID ゲストドライバーおよびライセンスファイルを使用してカスタムの仮想 GPU 対応インスタンスイメージを作成するには、以下の手順を使用します。

前提条件

- GPU 対応のコンピュータードと共にオーバークラウドを設定およびデプロイしている。

手順

1. アンダークラウドに **stack** ユーザーとしてログインします。
2. source コマンドで **overcloudrc** 認証情報ファイルを読み込みます。

```
$ source ~/overcloudrc
```

3. 仮想 GPU インスタンスが必要とするハードウェアおよびソフトウェアプロファイルでインスタンスを作成します。

```
(overcloud)$ openstack server create --flavor <flavor> \
--image <image> temp_vgpu_instance
```

- **<flavor>** を、仮想 GPU インスタンスが必要とするハードウェアプロファイルを持つフレーバーの名前または ID に置き換えてください。仮想 GPU フレーバー作成に関する詳細は、[Creating a vGPU flavor for instances](#) を参照してください。
 - **<image>** を、仮想 GPU インスタンスが必要とするソフトウェアプロファイルを持つイメージの名前または ID に置き換えてください。RHEL クラウドイメージのダウンロードに関する情報は、[Image service](#) を参照してください。
4. クラウドユーザーとしてインスタンスにログインします。詳細は、[Connecting to an instance](#) を参照してください。
 5. NVIDIA のガイドンス ([Licensing an NVIDIA vGPU on Linux by Using a Configuration File](#)) に従って、インスタンス上に **gridd.conf** NVIDIA GRID ライセンスファイルを作成します。
 6. インスタンスに GPU ドライバーをインストールします。NVIDIA ドライバーのインストールについての詳細は、[Installing the NVIDIA vGPU Software Graphics Driver on Linux](#) を参照してください。



注記

hw_video_model イメージ属性を使用して GPU ドライバーの種別を定義します。仮想 GPU インスタンスのエミュレートされた GPU を無効にする場合は、**none** を選択します。サポートされているドライバーについての詳しい情報は、[Image metadata](#) を参照してください。

7. インスタンスのイメージスナップショットを作成します。

```
(overcloud)$ openstack server image create \
--name vgpu_image temp_vgpu_instance
```

8. オプション: インスタンスを削除します。

12.4. インスタンス用の仮想 GPU フレーバーの作成

クラウドユーザーが GPU 負荷用のインスタンスを作成できるようにするには、仮想 GPU インスタンスを起動するための GPU フレーバーを作成し、仮想 GPU のリソースをそのフレーバーに割り当てます。

前提条件

- GPU 対応コンピュートノードと共にオーバークラウドを設定およびデプロイしている。

手順

1. NVIDIA GPU フレーバーを作成します。以下に例を示します。

```
(overcloud)$ openstack flavor create --vcpus 6 \
--ram 8192 --disk 100 m1.small-gpu
+-----+-----+
| Field                | Value                |
+-----+-----+
| OS-FLV-DISABLED:disabled | False                |
| OS-FLV-EXT-DATA:ephemeral | 0                    |
| disk                  | 100                  |
| id                    | a27b14dd-c42d-4084-9b6a-225555876f68 |
```

```
| name          | m1.small-gpu          |
| os-flavor-access:is_public | True                  |
| properties    |                        |
| ram           | 8192                   |
| rxtx_factor   | 1.0                    |
| swap          |                        |
| vcpus         | 6                      |
+-----+-----+
```

- 作成したフレーバーに仮想 GPU のリソースを割り当てます。各インスタンスに割り当てられる仮想 GPU は1つだけです。

```
(overcloud)$ openstack flavor set m1.small-gpu \
--property "resources:VGPU=1"

(overcloud)$ openstack flavor show m1.small-gpu
+-----+-----+
| Field          | Value                  |
+-----+-----+
| OS-FLV-DISABLED:disabled | False                  |
| OS-FLV-EXT-DATA:ephemeral | 0                      |
| access_project_ids      | None                   |
| disk                   | 100                    |
| id                     | a27b14dd-c42d-4084-9b6a-225555876f68 |
| name                  | m1.small-gpu          |
| os-flavor-access:is_public | True                  |
| properties             | resources:VGPU='1'    |
| ram                    | 8192                   |
| rxtx_factor            | 1.0                    |
| swap                   |                        |
| vcpus                  | 6                      |
+-----+-----+
```

12.5. 仮想 GPU インスタンスの起動

GPU 負荷用の GPU 対応インスタンスを作成することができます。

手順

- GPU フレーバーおよびイメージを使用して、インスタンスを作成します。以下に例を示します。

```
(overcloud)$ openstack server create --flavor m1.small-gpu \
--image vgpu_image --security-group web --nic net-id=internal0 \
--key-name lambda vgpu-instance
```

- クラウドユーザーとしてインスタンスにログインします。詳細は、[Connecting to an instance](#) を参照してください。
- インスタンスが GPU にアクセスできることを確認するには、インスタンスから以下のコマンドを入力します。

```
$ lspci -nn | grep <gpu_name>
```

12.6. GPU デバイスの PCI パススルーの有効化

PCI パススルーを使用して、グラフィックカード等の物理 PCI デバイスをインスタンスに接続することができます。デバイスに PCI パススルーを使用する場合、インスタンスはタスクを実行するためにデバイスへの排他的アクセスを確保し、ホストはデバイスを利用することができません。

前提条件

- **pciutils** パッケージが PCI カードを持つ物理サーバーにインストールされている。
- GPU デバイスのドライバーが、デバイスをパススルーするインスタンスにインストールされている必要があります。したがって、必要な GPU ドライバーがインストールされたカスタムのインスタンスイメージを作成している必要があります。GPU ドライバーがインストールされたカスタムのインスタンスイメージを作成する方法についての詳細は、[Creating a custom GPU instance image](#) を参照してください。

手順

1. 各パススルーデバイス種別のベンダー ID および製品 ID を確認するには、PCI カードを持つ物理サーバーで以下のコマンドを入力します。

```
# lspci -nn | grep -i <gpu_name>
```

たとえば、NVIDIA GPU のベンダーおよび製品 ID を確認するには、以下のコマンドを入力します。

```
# lspci -nn | grep -i nvidia
3b:00.0 3D controller [0302]: NVIDIA Corporation TU104GL [Tesla T4] [10de:1eb8] (rev a1)
d8:00.0 3D controller [0302]: NVIDIA Corporation TU104GL [Tesla T4] [10de:1db4] (rev a1)
```

2. 各 PCI デバイスに Single Root I/O Virtualization (SR-IOV) 機能があるかどうかを確認するには、PCI カードを持つ物理サーバーで以下のコマンドを入力します。

```
# lspci -v -s 3b:00.0
3b:00.0 3D controller: NVIDIA Corporation TU104GL [Tesla T4] (rev a1)
...
Capabilities: [bcc] Single Root I/O Virtualization (SR-IOV)
...
```

3. PCI パススルー用にオーバークラウド上のコントローラーノードを設定するには、環境ファイル (例: **pci_passthru_controller.yaml**) を作成します。
4. **pci_passthru_controller.yaml** の **NovaSchedulerDefaultFilters** パラメーターに **PciPassthroughFilter** を追加します。

```
parameter_defaults:
  NovaSchedulerDefaultFilters:
    ['AvailabilityZoneFilter','ComputeFilter','ComputeCapabilitiesFilter','ImagePropertiesFilter','ServerGroupAntiAffinityFilter','ServerGroupAffinityFilter','PciPassthroughFilter','NUMATopologyFilter']
```

5. コントローラーノード上のデバイスの PCI エイリアスを指定するには、以下の設定を **pci_passthru_controller.yaml** に追加します。
 - PCI デバイスに SR-IOV 機能がある場合:

```
ControllerExtraConfig:
nova::pci::aliases:
- name: "t4"
  product_id: "1eb8"
  vendor_id: "10de"
  device_type: "type-PF"
- name: "v100"
  product_id: "1db4"
  vendor_id: "10de"
  device_type: "type-PF"
```

- PCI デバイスに SR-IOV 機能がない場合:

```
ControllerExtraConfig:
nova::pci::aliases:
- name: "t4"
  product_id: "1eb8"
  vendor_id: "10de"
- name: "v100"
  product_id: "1db4"
  vendor_id: "10de"
```

device_type フィールドの設定に関する詳細は、[PCI passthrough device type field](#) を参照してください。



注記

nova-api サービスが Controller 以外のロールで実行されている場合は、**ControllerExtraConfig** を **<Role>ExtraConfig** の形式でユーザーロールに置き換えます。

6. PCI パススルー用にオーバークラウド上のコンピュートノードを設定するには、環境ファイル (例: **pci_passthru_compute.yaml**) を作成します。
7. コンピュートノード上のデバイスで利用可能な PCI を指定するには、以下の設定を **pci_passthru_compute.yaml** に追加します。

```
parameter_defaults:
NovaPCIPassthrough:
- vendor_id: "10de"
  product_id: "1eb8"
```

8. インスタンスの移行およびサイズ変更の操作を行うために、コンピュートノードの PCI エイリアスのコピーを作成する必要があります。コンピュートノード上のデバイスの PCI エイリアスを指定するには、以下の設定を **pci_passthru_compute.yaml** に追加します。
- PCI デバイスに SR-IOV 機能がある場合:

```
ComputeExtraConfig:
nova::pci::aliases:
- name: "t4"
  product_id: "1eb8"
  vendor_id: "10de"
  device_type: "type-PF"
```

```
- name: "v100"
  product_id: "1db4"
  vendor_id: "10de"
  device_type: "type-PF"
```

- PCI デバイスに SR-IOV 機能がない場合:

```
ComputeExtraConfig:
nova::pci::aliases:
  - name: "t4"
    product_id: "1eb8"
    vendor_id: "10de"
  - name: "v100"
    product_id: "1db4"
    vendor_id: "10de"
```



注記

コンピュートノードのエイリアスは、コントローラーノードのエイリアスと同じでなければなりません。

9. PCI パススルーをサポートするためにコンピュートノードのサーバー BIOS で IOMMU を有効にするには、**pci_passthru_compute.yaml** に **KernelArgs** パラメーターを追加します。

```
parameter_defaults:
...
ComputeParameters:
  KernelArgs: "intel_iommu=on iommu=pt"
```

10. その他の環境ファイルと共にこれらのカスタム環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/pci_passthru_controller.yaml \
-e /home/stack/templates/pci_passthru_compute.yaml
```

11. PCI デバイスを要求するためのフレーバーを設定します。以下の例では、それぞれベンダー ID および製品 ID が **10de** および **13f2** の 2 つのデバイスをリクエストします。

```
# openstack flavor set m1.large \
--property "pci_passthrough:alias"="t4:2"
```

検証

1. PCI パススルーデバイスを設定してインスタンスを作成します。

```
# openstack server create --flavor m1.large \
--image <custom_gpu> --wait test-pci
```

<custom_gpu> を、必要な GPU ドライバーがインストールされたカスタムインスタンスイメージの名前に置き換えます。

2. クラウドユーザーとしてインスタンスにログインします。
3. インスタンスが GPU にアクセスできることを確認するには、インスタンスから以下のコマンドを入力します。

```
$ lspci -nn | grep <gpu_name>
```

4. NVIDIA System Management Interface のステータスを確認するには、インスタンスから以下のコマンドを入力します。

```
$ nvidia-smi
```

出力例:

```
-----
| NVIDIA-SMI 440.33.01    Driver Version: 440.33.01    CUDA Version: 10.2    |
|-----+-----|
| GPU Name      Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|=====|
| 0  Tesla T4           Off | 00000000:01:00.0 Off |                    0 |
| N/A   43C    P0   20W / 70W |    0MiB / 15109MiB |      0%      Default |
|-----|
|
| Processes:
| GPU      PID   Type   Process name                      Usage      |
|=====|
| No running processes found                      |
|-----|
```

第13章 リアルタイムコンピュートの設定

クラウド管理者は、コンピュートノードに、低レイテンシーのポリシーを順守しリアルタイム処理を実行するためのインスタンスが必要となる場合があります。リアルタイムコンピュートノードには、リアルタイム対応のカーネル、特定の仮想化モジュール、および最適化されたデプロイメントパラメーターが設定され、リアルタイム処理の要求に対応してレイテンシーを最小限に抑えます。

リアルタイムコンピュートを有効にするプロセスは、以下のステップで設定されます。

- コンピュートノードの BIOS 設定の定義
- real-time カーネルおよび Real-Time KVM (RT-KVM) カーネルモジュールを持つ real-time のイメージのビルド
- コンピュートノードへの **ComputeRealTime** ロールの割り当て

NFV 負荷用リアルタイムコンピュートのデプロイメントのユースケース例については、[ネットワーク機能仮想化 \(NFV\) のプランニングおよび設定ガイドの 例: OVS-DPDK および SR-IOV ならびに VXLAN トンネリングの設定](#) セクションを参照してください。



注記

リアルタイムコンピュートノードは、Red Hat Enterprise Linux バージョン 7.5 以降でのみサポートされます。

13.1. リアルタイム処理用コンピュートノードの準備

オーバークラウドにリアルタイムコンピュートをデプロイするには、Red Hat Enterprise Linux Real-Time KVM (RT-KVM) を有効にし、リアルタイムをサポートするように BIOS を設定し、リアルタイムのオーバークラウドイメージをビルドする必要があります。

前提条件

- RT-KVM コンピュートノードには、Red Hat 認定済みサーバーを使用する必要があります。詳しくは、[Red Hat Enterprise Linux for Real Time 7 用認定サーバー](#) を参照してください。
- **rhel-8-for-x86_64-nfv-rpms** リポジトリにアクセスするには、別途 **Red Hat OpenStack Platform for Real Time** に対するサブスクリプションが必要です。アンダークラウド用のリポジトリおよびサブスクリプションの管理に関する詳細は、**Director Installation and Usage** の [Registering the undercloud and attaching subscriptions](#) を参照してください。

手順

1. real-time のオーバークラウドイメージをビルドするには、RT-KVM 用の **rhel-8-for-x86_64-nfv-rpms** リポジトリを有効にする必要があります。リポジトリからインストールされるパッケージを確認するには、以下のコマンドを入力します。

```
$ dnf repo-pkgs rhel-8-for-x86_64-nfv-rpms list
Loaded plugins: product-id, search-disabled-repos, subscription-manager
Available Packages
kernel-rt.x86_64          4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-rpms
kernel-rt-debug.x86_64   4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-
rpms
kernel-rt-debug-devel.x86_64 4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-
rpms
```

```
kernel-rt-debug-kvm.x86_64      4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-
rpms
kernel-rt-devel.x86_64         4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-
rpms
kernel-rt-doc.noarch           4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-rpms
kernel-rt-kvm.x86_64           4.18.0-80.7.1.rt9.153.el8_0      rhel-8-for-x86_64-nfv-
rpms
[ output omitted...]
```

- リアルタイムコンピュートノード用のオーバークラウドイメージをビルドするには、アンダークラウドに **libguestfs-tools** パッケージをインストールして、**virt-customize** ツールを取得します。

```
(undercloud)$ sudo dnf install libguestfs-tools
```



重要

アンダークラウドに **libguestfs-tools** パッケージをインストールする場合は、アンダークラウドの **tripleo_iscsid** サービスとのポートの競合を避けるために **iscsid.socket** を無効にします。

```
$ sudo systemctl disable --now iscsid.socket
```

- イメージを抽出します。

```
(undercloud)$ tar -xf /usr/share/rhosp-director-images/overcloud-full.tar
(undercloud)$ tar -xf /usr/share/rhosp-director-images/ironic-python-agent.tar
```

- デフォルトのイメージをコピーします。

```
(undercloud)$ cp overcloud-full.qcow2 overcloud-realtime-compute.qcow2
```

- イメージを登録して、必要なサブスクリプションを設定します。

```
(undercloud)$ virt-customize -a overcloud-realtime-compute.qcow2 --run-command
'subscription-manager register --username=<username> --password=<password>'
[ 0.0] Examining the guest ...
[ 10.0] Setting a random seed
[ 10.0] Running: subscription-manager register --username=<username> --password=
<password>
[ 24.0] Finishing off
```

username および **password** の値を、ご自分の Red Hat カスタマーアカウント情報に置き換えてください。

リアルタイムのオーバークラウドイメージのビルドに関する一般的な情報は、ナレッジベースのアーティクル [Modifying the Red Hat Enterprise Linux OpenStack Platform Overcloud Image with virt-customize](#) を参照してください。

- 以下の例に示すように、**Red Hat OpenStack Platform for Real Time** サブスクリプションの SKU を探します。SKU は、同じアカウントおよび認証情報を使用してすでに Red Hat サブスクリプションマネージャーに登録済みのシステムに置かれている場合があります。

```
$ sudo subscription-manager list
```

7. Red Hat OpenStack Platform for Real Timeサブスクリプションをイメージにアタッチします。

```
(undercloud)$ virt-customize -a overcloud-realtime-compute.qcow2 --run-command  
'subscription-manager attach --pool [subscription-pool]'
```

8. イメージ上で **rt** を設定するためのスクリプトを作成します。

```
(undercloud)$ cat rt.sh  
#!/bin/bash  
  
set -eux  
  
subscription-manager repos --enable=[REPO_ID]  
dnf -v -y --setopt=protected_packages= erase kernel.$(uname -m)  
dnf -v -y install kernel-rt kernel-rt-kvm tuned-profiles-nfv-host  
  
# END OF SCRIPT
```

9. リアルタイムイメージを設定するスクリプトを実行します。

```
(undercloud)$ virt-customize -a overcloud-realtime-compute.qcow2 -v --run rt.sh 2>&1 | tee  
virt-customize.log
```

10. SELinux の再ラベル付けをします。

```
(undercloud)$ virt-customize -a overcloud-realtime-compute.qcow2 --selinux-relabel
```

11. **vmlinux** および **initrd** を抽出します。以下に例を示します。

```
(undercloud)$ mkdir image  
(undercloud)$ guestmount -a overcloud-realtime-compute.qcow2 -i --ro image  
(undercloud)$ cp image/boot/vmlinux-4.18.0-80.7.1.rt9.153.el8_0.x86_64 ./overcloud-  
realtime-compute.vmlinux  
(undercloud)$ cp image/boot/initramfs-4.18.0-80.7.1.rt9.153.el8_0.x86_64.img ./overcloud-  
realtime-compute.initrd  
(undercloud)$ guestunmount image
```



注記

vmlinux および **initramfs** のファイル名に含まれるソフトウェアバージョンは、カーネルバージョンによって異なります。

12. イメージをアップロードします。

```
(undercloud)$ openstack overcloud image upload \  
--update-existing --os-image-name  
overcloud-realtime-compute.qcow2
```

これで、選択したコンピュートノード上の **ComputeRealTime** コンポーザブルロールで使用する
ことのできる real-time イメージの準備ができました。

- リアルタイムコンピュートノードのレイテンシーを短縮するには、コンピュートノードの BIOS 設定を変更する必要があります。コンピュートノードの BIOS 設定で、以下のコンポーネントの全オプションを無効にする必要があります。

- 電源管理
- ハイパースレッディング
- CPU のスリープ状態
- 論理プロセッサ

これらの設定に関する説明と、無効化の影響については、Red Hat Enterprise Linux for Real Time チューニングガイドの [BIOS パラメーターの設定](#) を参照してください。BIOS 設定の変更方法に関する詳しい情報は、ハードウェアの製造会社のドキュメントを参照してください。

13.2. リアルタイム COMPUTE ロールのデプロイメント

Red Hat OpenStack Platform (RHOSP) director では、**ComputeRealTime** ロールのテンプレートが利用可能です。これを使用して、real-time コンピュートノードをデプロイすることができます。コンピュートノードをリアルタイム用に指定するためには、追加のステップを実施する必要があります。

手順

- `/usr/share/openstack-tripleo-heat-templates/environments/compute-real-time-example.yaml` ファイルをベースに、**ComputeRealTime** ロールのパラメーターを設定する `compute-real-time.yaml` 環境ファイルを作成します。

```
cp /usr/share/openstack-tripleo-heat-templates/environments/compute-real-time-example.yaml /home/stack/templates/compute-real-time.yaml
```

ファイルには、以下のパラメーター値を含める必要があります。

- **IsolCpusList** および **NovaComputeCpuDedicatedSet**: リアルタイム負荷のために確保する分離 CPU コアおよび仮想 CPU のピンギングの一覧。この値は、real-time コンピュートノードの CPU ハードウェアにより異なります。
 - **NovaComputeCpuSharedSet**: エミュレータスレッド用に確保するホスト CPU の一覧
 - **KernelArgs**: Real-time コンピュートノードのカーネルに渡す引数。たとえば、`default_hugepagesz=1G hugepagesz=1G hugepages=<number_of_1G_pages_to_reserve> hugepagesz=2M hugepages=<number_of_2M_pages>` を使用して、複数のサイズのヒュージページを持つゲストのメモリー要求を定義することができます。この例では、デフォルトのサイズは 1 GB ですが、2 MB のヒュージページを確保することもできます。
 - **NovaComputeDisableIrqBalance**: リアルタイムデプロイメントでは `irqbalance` サービスではなく `tuned` サービスが IRQ 分散を管理するので、リアルタイムコンピュートノードではこのパラメーターを `true` に設定するようにしてください。
- ComputeRealTime** ロールをロールデータのファイルに追加し、ファイルを再生成します。以下に例を示します。

```
$ openstack overcloud roles generate -o /home/stack/templates/rt_roles_data.yaml Controller
Compute ComputeRealTime
```

このコマンドにより、以下の例のような内容で **ComputeRealTime** ロールが生成され、**ImageDefault** オプションに **overcloud-realtime-compute** が設定されます。

```
- name: ComputeRealTime
  description: |
    Compute role that is optimized for real-time behaviour. When using this role
    it is mandatory that an overcloud-realtime-compute image is available and
    the role specific parameters IsolCpusList, NovaComputeCpuDedicatedSet and
    NovaComputeCpuSharedSet are set accordingly to the hardware of the real-time compute
    nodes.
  CountDefault: 1
  networks:
    InternalApi:
      subnet: internal_api_subnet
  Tenant:
    subnet: tenant_subnet
  Storage:
    subnet: storage_subnet
  HostnameFormatDefault: '%stackname%-computerealtime-%index%'
  ImageDefault: overcloud-realtime-compute
  RoleParametersDefault:
    TunedProfileName: "realtime-virtual-host"
    KernelArgs: "" # these must be set in an environment file
    IsolatedCpusList: "" # or similar according to the hardware
    NovaComputeCpuDedicatedSet: "" # of real-time nodes
    NovaComputeCpuSharedSet: "" #
    NovaLibvirtMemStatsPeriodSeconds: 0
  ServicesDefault:
    - OS::TripleO::Services::Aide
    - OS::TripleO::Services::AuditD
    - OS::TripleO::Services::BootParams
    - OS::TripleO::Services::CACerts
    - OS::TripleO::Services::CephClient
    - OS::TripleO::Services::CephExternal
    - OS::TripleO::Services::CertmongerUser
    - OS::TripleO::Services::Collectd
    - OS::TripleO::Services::ComputeCeilometerAgent
    - OS::TripleO::Services::ComputeNeutronCorePlugin
    - OS::TripleO::Services::ComputeNeutronL3Agent
    - OS::TripleO::Services::ComputeNeutronMetadataAgent
    - OS::TripleO::Services::ComputeNeutronOvsAgent
    - OS::TripleO::Services::Docker
    - OS::TripleO::Services::Fluentd
    - OS::TripleO::Services::IpaClient
    - OS::TripleO::Services::Ipsec
    - OS::TripleO::Services::Iscsid
    - OS::TripleO::Services::Kernel
    - OS::TripleO::Services::LoginDefs
    - OS::TripleO::Services::MetricsQdr
    - OS::TripleO::Services::MySQLClient
    - OS::TripleO::Services::NeutronBgpVpnBagpipe
    - OS::TripleO::Services::NeutronLinuxbridgeAgent
```

```
- OS::TripleO::Services::NeutronVppAgent
- OS::TripleO::Services::NovaCompute
- OS::TripleO::Services::NovaLibvirt
- OS::TripleO::Services::NovaLibvirtGuests
- OS::TripleO::Services::NovaMigrationTarget
- OS::TripleO::Services::ContainersLogrotateCron
- OS::TripleO::Services::OpenDaylightOvs
- OS::TripleO::Services::Podman
- OS::TripleO::Services::Rhsm
- OS::TripleO::Services::RsyslogSidecar
- OS::TripleO::Services::Securetty
- OS::TripleO::Services::SensuClient
- OS::TripleO::Services::SkydiveAgent
- OS::TripleO::Services::Snmp
- OS::TripleO::Services::Sshd
- OS::TripleO::Services::Timesync
- OS::TripleO::Services::Timezone
- OS::TripleO::Services::TripleoFirewall
- OS::TripleO::Services::TripleoPackages
- OS::TripleO::Services::Vpp
- OS::TripleO::Services::OVNController
- OS::TripleO::Services::OVNMetadataAgent
```

カスタムロールおよび **roles-data.yaml** に関する一般的な情報は、[Roles](#) セクションを参照してください。

- リアルタイム負荷用に指定するノードをタグ付けするために、**compute-realtime** フレーバーを作成します。以下に例を示します。

```
$ source ~/stackrc
$ openstack flavor create --id auto --ram 6144 --disk 40 --vcpus 4 compute-realtime
$ openstack flavor set --property "cpu_arch"="x86_64" --property
"capabilities:boot_option"="local" --property "capabilities:profile"="compute-realtime"
compute-realtime
```

- リアルタイム負荷用に指定するそれぞれのノードを、**compute-realtime** プロファイルでタグ付けします。

```
$ openstack baremetal node set --property capabilities='profile:compute-
realtime,boot_option:local' <node_uuid>
```

- 以下の内容の環境ファイルを作成して、**ComputeRealTime** ロールを **compute-realtime** フレーバーにマッピングします。

```
parameter_defaults:
  OvercloudComputeRealTimeFlavor: compute-realtime
```

- その他の環境ファイルと共に、これらの環境ファイルおよび新しいロールファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-r /home/stack/templates/rt~/my_roles_data.yaml \
-e home/stack/templates/compute-real-time.yaml
```

13.3. デプロイメントおよびテストシナリオの例

以下の手順の例では、単純な単一ノードのデプロイメントを使用して、環境変数およびその他の補助設定が正しく定義されていることをテストします。実際の実行結果は、クラウドにデプロイするノードおよびインスタンスの数により異なります。

手順

1. 以下のパラメーターで **compute-real-time.yaml** ファイルを作成します。

```
parameter_defaults:
  ComputeRealTimeParameters:
    IsolCpusList: "1"
    NovaComputeCpuDedicatedSet: "1"
    NovaComputeCpuSharedSet: "0"
    KernelArgs: "default_hugepagesz=1G hugepagesz=1G hugepages=16"
    NovaComputeDisableIrqBalance: true
```

2. **ComputeRealTime** ロールを設定して新たな **rt_roles_data.yaml** ファイルを作成します。

```
$ openstack overcloud roles generate \
-o ~/rt_roles_data.yaml Controller ComputeRealTime
```

3. その他の環境ファイルと共にこの **compute-real-time.yaml** および **rt_roles_data.yaml** をスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-r /home/stack/rt_roles_data.yaml \
-e [your environment files] \
-e /home/stack/templates/compute-real-time.yaml
```

このコマンドにより、コントローラーノードおよびリアルタイムコンピュートノードが1台ずつデプロイされます。

4. リアルタイムコンピュートノードにログインし、以下のパラメーターを確認します。

```
[root@overcloud-computerealttime-0 ~]# uname -a
Linux overcloud-computerealttime-0 4.18.0-80.7.1.rt9.153.el8_0.x86_64 #1 SMP PREEMPT
RT Wed Dec 13 13:37:53 UTC 2017 x86_64 x86_64 x86_64 GNU/Linux
[root@overcloud-computerealttime-0 ~]# cat /proc/cmdline
BOOT_IMAGE=/boot/vmlinuz-4.18.0-80.7.1.rt9.153.el8_0.x86_64 root=UUID=45ae42d0-
58e7-44fe-b5b1-993fe97b760f ro console=tty0 crashkernel=auto console=ttyS0,115200
default_hugepagesz=1G hugepagesz=1G hugepages=16
[root@overcloud-computerealttime-0 ~]# tuned-adm active
Current active profile: realtime-virtual-host
[root@overcloud-computerealttime-0 ~]# grep ^isolated_cores /etc/tuned/realtime-virtual-host-
variables.conf
isolated_cores=1
[root@overcloud-computerealttime-0 ~]# cat /usr/lib/tuned/realtime-virtual-
host/lapic_timer_adv_ns
4000 # The returned value must not be 0
[root@overcloud-computerealttime-0 ~]# cat
/sys/module/kvm/parameters/lapic_timer_advance_ns
4000 # The returned value must not be 0
# To validate hugepages at a host level:
```

```
[root@overcloud-computerealttime-0 ~]# cat /proc/meminfo | grep -E
HugePages_Total|Hugepagesize
HugePages_Total:    64
Hugepagesize:    1048576 kB
# To validate hugepages on a per NUMA level (below example is a two NUMA compute
host):
[root@overcloud-computerealttime-0 ~]# cat
/sys/devices/system/node/node0/hugepages/hugepages-1048576kB/nr_hugepages
32
[root@overcloud-computerealttime-0 ~]# cat
/sys/devices/system/node/node1/hugepages/hugepages-1048576kB/nr_hugepages
32
[root@overcloud-computerealttime-0 ~]# crudini --get /var/lib/config-data/puppet-
generated/nova_libvirt/etc/nova/nova.conf compute cpu_dedicated_set
1
[root@overcloud-computerealttime-0 ~]# crudini --get /var/lib/config-data/puppet-
generated/nova_libvirt/etc/nova/nova.conf compute cpu_shared_set
0
[root@overcloud-computerealttime-0 ~]# systemctl status irqbalance
● irqbalance.service - irqbalance daemon
   Loaded: loaded (/usr/lib/systemd/system/irqbalance.service; enabled; vendor preset:
   enabled)
   Active: inactive (dead) since Tue 2021-03-30 13:36:31 UTC; 2s ago
```

13.4. リアルタイムインスタンスの起動およびチューニング

リアルタイムコンピュートノードをデプロイして設定したら、それらのノードでリアルタイムインスタンスを起動することができます。CPU ピニング、NUMA トポロジーフィルター、およびヒュージページを使用して、これらのリアルタイムインスタンスをさらに設定することができます。

前提条件

- [Deploying the Real-time Compute role](#) で説明したとおり、オーバークラウド上に **compute-realtime** フレーバーが存在する。

手順

1. リアルタイムインスタンスを起動します。

```
# openstack server create --image <rhel> \
--flavor r1.small --nic net-id=<dpdk_net> test-rt
```

2. (オプション) インスタンスが割り当てられたエミュレータスレッドを使用していることを確認します。

```
# virsh dumpxml <instance_id> | grep vcpu -A1
<vcpu placement='static'>4</vcpu>
<cpupin>
  <vcpupin vcpu='0' cpuset='1'>
  <vcpupin vcpu='1' cpuset='3'>
  <vcpupin vcpu='2' cpuset='5'>
  <vcpupin vcpu='3' cpuset='7'>
  <emulatorpin cpuset='0-1'>
```

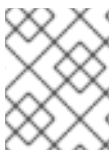
```
<vcpusched vcpus='2-3' scheduler='fifo'
priority='1'/>
</cputune>
```

CPU のピンニングおよびエミュレータースレッドポリシーの設定

リアルタイム負荷用に各リアルタイムコンピュートノードの CPU を十分に確保するためには、インスタンス用仮想 CPU (vCPU) の少なくとも 1 つをホストの物理 CPU (pCPU) にピンニングする必要があります。その結果、その仮想 CPU のエミュレータースレッドは、ピンニングした物理 CPU 専用として維持されます。

専用 CPU のポリシーを使用するようにフレーバーを設定します。そのためには、フレーバーで **hw:cpu_policy** パラメーターを **dedicated** に設定します。以下に例を示します。

```
# openstack flavor set --property hw:cpu_policy=dedicated 99
```



注記

リソースクォータに、Real-time コンピュートノードが消費するのに十分な pCPU があることを確認してください。

ネットワーク設定の最適化

デプロイメントのニーズによっては、特定のリアルタイム負荷に合わせてネットワークをチューニングするために、**network-environment.yaml** ファイルのパラメーターを設定しなければならない場合があります。

OVS-DPDK 用に最適化した設定の例を確認するには、**ネットワーク機能仮想化 (NFV) のプランニングおよび設定ガイド**の [OVS-DPDK パラメーターの設定](#) セクションを参照してください。

ヒュージページの設定

デフォルトのヒュージページサイズを 1GB に設定することを推奨します。このように設定しないと、TLB のフラッシュにより仮想 CPU の実行にジッターが生じます。ヒュージページの使用に関する一般的な情報については、DPDK Getting Started Guide for Linux の [Running DPDK applications](#) を参照してください。

Performance Monitoring Unit (PMU) エミュレーションの無効化

イメージまたはフレーバーに仮想 PMU を設定すると、インスタンスは PMU メトリックを提供することができます。PMU メトリックを提供することにより、レイテンシーが生じます。



注記

NovaLibvirtCPUMode が **host-passthrough** に設定されている場合、仮想 PMU はデフォルトで有効です。

PMU メトリックが必要ない場合は、インスタンスの作成に使用するイメージまたはフレーバーで PMU 属性を **False** に設定し、仮想 PMU を無効にしてレイテンシーを軽減します。

- イメージ: **hw_pmu=False**
- フレーバー: **hw:pmu=False**

第14章 インスタンスの管理

クラウド管理者は、クラウド上で実行されているインスタンスを監視および管理することができます。

14.1. データベースのクリーニング

Compute サービスには管理ツール **nova-manage** が含まれています。このツールを使用して、データベーススキーマの適用、アップグレード中のオンラインデータ移行の実行、データベースの管理およびクリーンアップ等の、デプロイメント、アップグレード、クリーンアップ、およびメンテナンス関連のタスクを実行することができます。

director は、cron を使用してオーバークラウドでの以下のデータベース管理タスクを自動化します。

- 削除された行を実稼働テーブルからシャドウテーブルに移動して、削除されたインスタンスレコードをアーカイブする。
- アーカイブ処理が完了した後に、シャドウテーブルから削除された行をパージする。

14.1.1. データベース管理の設定

cron ジョブは、デフォルト設定を使用してデータベース管理タスクを実行します。デフォルトでは、データベースをアーカイブする cron ジョブは毎日 00:01 に実行され、データベースをパージする cron ジョブは毎日 05:00 に実行されます。共にジッターは 0 秒から 3600 秒の間です。必要に応じて、これらの設定は heat パラメーターを使用して変更することができます。

手順

1. Compute 環境ファイルを開きます。
2. 追加または変更する cron ジョブを制御する heat パラメーターを追加します。たとえば、シャドウテーブルをアーカイブ直後にパージするには、次のパラメーターを True に設定します。

```
parameter_defaults:
...
NovaCronArchiveDeleteRowsPurge: True
```

データベースの cron ジョブを管理する heat パラメーターの完全一覧は、[Configuration options for the Compute service automated database management](#) を参照してください。

3. 更新内容を Compute 環境ファイルに保存します。
4. その他の環境ファイルと共に Compute 環境ファイルをスタックに追加して、オーバークラウドをデプロイします。

```
(undercloud)$ openstack overcloud deploy --templates \
-e [your environment files] \
-e /home/stack/templates/<compute_environment_file>.yaml
```

14.1.2. Compute サービスのデータベース自動管理用設定オプション

以下の heat パラメーターを使用して、データベースを管理する自動 cron ジョブを有効化および変更します。

表14.1 Compute (nova) サービスの cron パラメーター

パラメーター	説明
NovaCronArchiveDeleteAllCells	<p>すべてのセルから削除されたインスタンスレコードをアーカイブするには、このパラメーターを True に設定します。</p> <p>デフォルト: True</p>
NovaCronArchiveDeleteRowsAge	<p>このパラメーターを使用して、削除されたインスタンスレコードを経過時間 (日数単位) に基づいてアーカイブします。</p> <p>シャドウテーブル内の本日以前のデータをアーカイブするには、0 に設定します。</p> <p>デフォルト: 90</p>
NovaCronArchiveDeleteRowsDestination	<p>このパラメーターを使用して、削除されたインスタンスレコードを記録するファイルを設定します。</p> <p>デフォルト: /var/log/nova/nova-rowsflush.log</p>
NovaCronArchiveDeleteRowsHour	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動する cron コマンドを実行する時刻の時間部分を設定します。</p> <p>デフォルト: 0</p>
NovaCronArchiveDeleteRowsMaxDelay	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動するまでの最大遅延時間 (秒単位) を設定します。</p> <p>デフォルト: 3600</p>
NovaCronArchiveDeleteRowsMaxRows	<p>このパラメーターを使用して、別のテーブルに移動することのできる削除インスタンスレコード数の最大値を設定します。</p> <p>デフォルト: 1000</p>
NovaCronArchiveDeleteRowsMinute	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動する cron コマンドを実行する時刻の分部分を設定します。</p> <p>デフォルト: 1</p>
NovaCronArchiveDeleteRowsMonthday	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動する cron コマンドを実行する日にちを設定します。</p> <p>デフォルト: *(毎日)</p>

パラメーター	説明
NovaCronArchiveDeleteRowsMonth	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動する cron コマンドを実行する月を設定します。</p> <p>デフォルト: <code>*</code> (毎月)</p>
NovaCronArchiveDeleteRowsPurge	<p>スケジュールされたアーカイブ処理の直後にシャドウテーブルをパージするには、このパラメーターを <code>True</code> に設定します。</p> <p>デフォルト: False</p>
NovaCronArchiveDeleteRowsUntilComplete	<p>すべてのレコードを移動するまで削除されたインスタンスレコードを別のテーブルに移動し続けるには、このパラメーターを <code>True</code> に設定します。</p> <p>デフォルト: True</p>
NovaCronArchiveDeleteRowsUser	<p>このパラメーターを使用して、削除されたインスタンスレコードをアーカイブする crontab の所有権を持ち、crontab が使用するログファイルにアクセスできるユーザーを設定します。</p> <p>デフォルト: nova</p>
NovaCronArchiveDeleteRowsWeekday	<p>このパラメーターを使用して、削除されたインスタンスレコードを別のテーブルに移動する cron コマンドを実行する曜日を設定します。</p> <p>デフォルト: <code>*</code> (毎日)</p>
NovaCronPurgeShadowTablesAge	<p>このパラメーターを使用して、シャドウテーブルを経過時間 (日数単位) に基づいてパージします。</p> <p>本日以前のシャドウテーブルをパージするには、0 に設定します。</p> <p>デフォルト: 14</p>
NovaCronPurgeShadowTablesAllCells	<p>すべてのセルからシャドウテーブルをパージするには、このパラメーターを <code>True</code> に設定します。</p> <p>デフォルト: True</p>
NovaCronPurgeShadowTablesDestination	<p>このパラメーターを使用して、パージされたシャドウテーブルを記録するファイルを設定します。</p> <p>デフォルト: <code>/var/log/nova/nova-rowspurge.log</code></p>

パラメーター	説明
NovaCronPurgeShadowTablesHour	<p>このパラメーターを使用して、シャドウテーブルをパージする cron コマンドを実行する時刻の時間部分を設定します。</p> <p>デフォルト: 5</p>
NovaCronPurgeShadowTablesMaxDelay	<p>このパラメーターを使用して、シャドウテーブルをパージするまでの最大遅延時間 (秒単位) を設定します。</p> <p>デフォルト: 3600</p>
NovaCronPurgeShadowTablesMinute	<p>このパラメーターを使用して、シャドウテーブルをパージする cron コマンドを実行する時刻の分部分を設定します。</p> <p>デフォルト: 0</p>
NovaCronPurgeShadowTablesMonth	<p>このパラメーターを使用して、シャドウテーブルをパージする cron コマンドを実行する月を設定します。</p> <p>デフォルト: * (毎月)</p>
NovaCronPurgeShadowTablesMonthday	<p>このパラメーターを使用して、シャドウテーブルをパージする cron コマンドを実行する日にちを設定します。</p> <p>デフォルト: * (毎日)</p>
NovaCronPurgeShadowTablesUser	<p>このパラメーターを使用して、シャドウテーブルをパージする crontab の所有権を持ち、crontab が使用するログファイルにアクセスできるユーザーを設定します。</p> <p>デフォルト: nova</p>
NovaCronPurgeShadowTablesVerbose	<p>このパラメーターを使用して、パージされたシャドウテーブルに関するログファイルの詳細ロギングを有効にします。</p> <p>デフォルト: False</p>
NovaCronPurgeShadowTablesWeekday	<p>このパラメーターを使用して、シャドウテーブルをパージする cron コマンドを実行する曜日を設定します。</p> <p>デフォルト: * (毎日)</p>

14.2. コンピュートノード間の仮想マシンインスタンスの移行

メンテナンスを実行する場合やワークロードのリバランスを行う場合、あるいは障害が発生した/障害が発生しつつあるノードを置き換える場合に、あるコンピュートノードからオーバークラウド内の別のコンピュートノードにインスタンスを移行しなければならない場合があります。

コンピュートノードのメンテナンス

ハードウェアのメンテナンスや修理、カーネルのアップグレードおよびソフトウェアの更新を行うなどの理由により、コンピュートノードを一時的に停止する必要がある場合、コンピュートノード上で実行中のインスタンスを別のコンピュートノードに移行することができます。

障害が発生しつつあるコンピュートノード

コンピュートノードで障害が発生する可能性があり、ノードのサービスまたは置き換えが必要な場合、障害が発生しつつあるコンピュートノードから正常なコンピュートノードにインスタンスを移行することができます。

障害が発生したコンピュートノード

コンピュートノードですでに障害が発生している場合には、インスタンスを退避させることができます。同じ名前、UUID、ネットワークアドレス、およびコンピュートノードに障害が発生する前にインスタンスに割り当てられていたその他すべてのリソースを使用して、元のイメージから別のコンピュートノードにインスタンスを再ビルドすることができます。

ワークロードのリバランス

ワークロードをリバランスするために、1つまたは複数のインスタンスを別のコンピュートノードに移行することができます。たとえば、コンピュートノード上のインスタンスを1つにまとめて電力を節約する、他のネットワークリソースに物理的に近いコンピュートノードにインスタンスを移行してレイテンシーを低減する、インスタンスを全コンピュートノードに分散してホットスポットをなくし復元力を向上させる、等が可能です。

director は、すべてのコンピュートノードがセキュアな移行を提供するように設定します。すべてのコンピュートノードには、移行プロセス中それぞれのホストのユーザーが他のコンピュートノードにアクセスできるように、共有 SSH キーも必要です。director は、**OS::TripleO::Services::NovaCompute** コンポーザブルサービスを使用してこのキーを作成します。このコンポーザブルサービスは、すべての Compute ロールにデフォルトで含まれているメインのサービスの1つです。詳細は、**オーバークラウドの高度なカスタマイズ** ガイドの [コンポーザブルサービスとカスタムロール](#) を参照してください。



注記

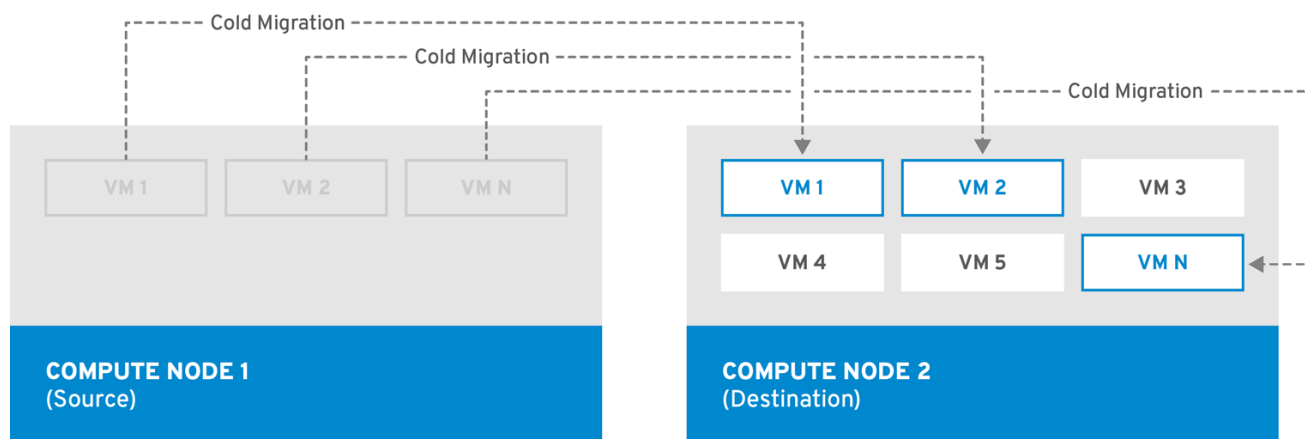
バックアップの目的で正常に機能しているコンピュートノードインスタンスのコピーを作成する場合や、インスタンスを別の環境にコピーする場合は、**director** のインストールと使用方法の [オーバークラウドへの仮想マシンのインポート](#) の手順に従います。

14.2.1. 移行の種別

Red Hat OpenStack Platform (RHOSP) では、以下の移行種別がサポートされています。

コールドマイグレーション

コールドマイグレーション (あるいは、非ライブマイグレーション) では、動作中のインスタンスをシャットダウンしてから、移行元コンピュートノードから移行先コンピュートノードに移行します。



OPENSTACK_11_0419

コールドマイグレーションでは、インスタンスに多少のダウンタイムが発生します。移行したインスタンスは、引き続き同じボリュームおよび IP アドレスにアクセスすることができます。

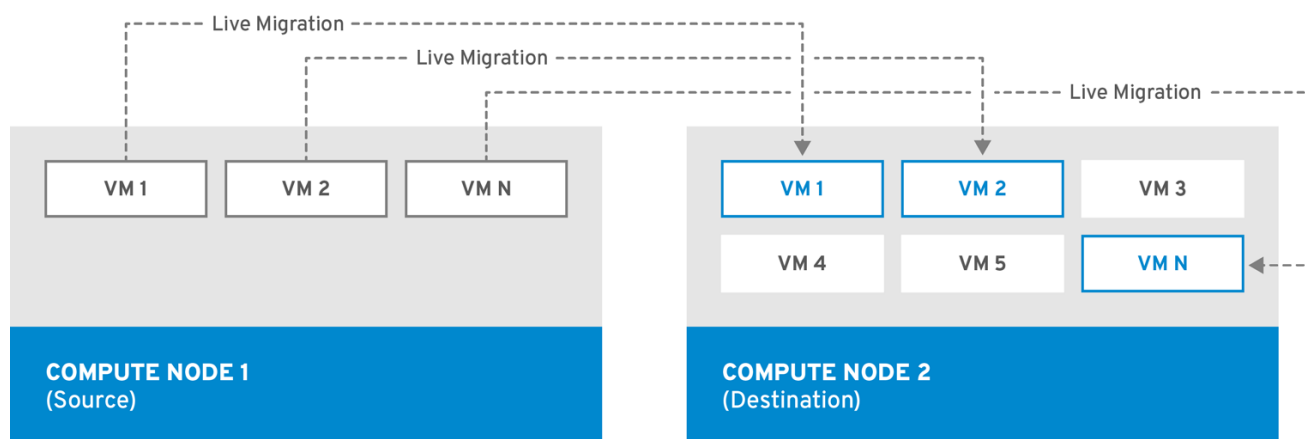


注記

コールドマイグレーションを行うためには、移行元および移行先両方のコンピュートノードが動作状態になければなりません。

ライブマイグレーション

ライブマイグレーションでは、インスタンスをシャットダウンせずに、動作状態を維持しながら移行元コンピュートノードから移行先コンピュートノードに移行します。



OPENSTACK_11_0419

インスタンスのライブマイグレーションを行う場合、ダウンタイムはほとんど発生しません。ただし、ライブマイグレーションは、移行操作中のパフォーマンスに影響を及ぼします。したがって、移行中のインスタンスは重要なパスから除外する必要があります。



重要

ライブマイグレーションは、移動されるワークロードのパフォーマンスに影響を与えます。Red Hat は、ライブマイグレーション中のパケット損失、ネットワーク遅延、メモリー遅延の増加、またはネットワーク帯域幅、メモリー帯域幅、ストレージ IO、または CPU パフォーマンスの低下をサポートしません。



注記

ライブマイグレーションを行うためには、移行元および移行先両方のコンピュートノードが動作状態になければなりません。

状況によっては、インスタンスのライブマイグレーションを行うことができない場合があります。詳細は、**移行の制約** を参照してください。

退避

コンピュートノードですでに障害が発生しているためインスタンスを移行する必要がある場合、インスタンスを退避させることができます。

14.2.2. 移行の制約

移行の制約は通常、ブロックマイグレーション、設定ディスク、またはいずれかのインスタンスがコンピュートノード上の物理ハードウェアにアクセスする場合に生じます。

CPU に関する制約

移行元および移行先コンピュートノードの CPU アーキテクチャーは、同一であることが必須です。たとえば、Red Hat では、**x86_64** CPU から **ppc64le** CPU へのインスタンスの移行をサポートしません。

異なる CPU モデル間の移行はサポートされていません。CPU ホストパススルーを使用するインスタンス等の場合には、移行元および移行先コンピュートノードの CPU は、完全に同一でなければなりません。すべてのケースで、移行先ノードの CPU 機能は、移行元ノードの CPU 機能の上位セットであることが必須です。

メモリーに関する制約

移行先コンピュートノードでは、十分な RAM が利用可能でなければなりません。メモリーのオーバーサブスクリプションが、移行失敗の原因となる可能性があります。

ブロックマイグレーションに関する制約

インスタンスの使用するディスクがコンピュートノード上にローカルに格納されている場合、その移行には、共有ストレージ (Red Hat Ceph Storage 等) を使用するボリュームベースのインスタンスよりもはるかに長い時間がかかります。このレイテンシーは、OpenStack Compute (nova) がコンピュートノード間でローカルディスクをブロックごとに移行するために発生します。この処理は、デフォルトではコントロールプレーンネットワークを通じて行われます。これとは対照的に、Red Hat Ceph Storage 等の共有ストレージを使用するボリュームベースのインスタンスでは、ボリュームを移行する必要がありません。それぞれのコンピュートノードが、共有ストレージにアクセスできるためです。



注記

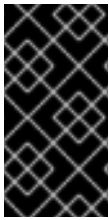
大量の RAM を消費するローカルディスクまたはインスタンスの移行により、コントロールプレーンネットワークに輻輳が生じ、コントロールプレーンネットワークを使用する他のシステム (RabbitMQ 等) のパフォーマンスに悪影響を及ぼす場合があります。

読み取り専用ドライブの移行に関する制約

ドライブの移行は、ドライブに読み取りおよび書き込み両方の機能がある場合に限りサポートされます。たとえば、OpenStack Compute (nova) は CD-ROM ドライブまたは読み取り専用のコンフィグドライブを移行することはできません。ただし、OpenStack Compute (nova) は、**vfat** 等のドライブ形式を持つコンフィグドライブなど、読み取りおよび書き込み両方の機能を持つドライブを移行することができます。

ライブマイグレーションに関する制約

インスタンスのライブマイグレーションでは、さらに制約が生じる場合があります。



重要

ライブマイグレーションは、移動されるワークロードのパフォーマンスに影響を与えます。Red Hat は、ライブマイグレーション中のパケット損失、ネットワーク遅延、メモリー遅延の増加、またはネットワーク帯域幅、メモリー帯域幅、ストレージ IO、または CPU パフォーマンスの低下をサポートしません。

移行中に新しい操作を行うことができない

移行元および移行先ノードのインスタンスのコピー間で状態の整合性を確保するために、RHOSP ではライブマイグレーション中の新規操作を拒否する必要があります。拒否しないと、ライブマイグレーションでメモリーの状態を複製する前にメモリーへの書き込みが行われた場合、ライブマイグレーションに長い時間がかかる状況や、永久に完了しない状況が発生する可能性があります。

NUMA を使用した CPU ピニング

Compute 設定の **NovaSchedulerDefaultFilters** パラメーターには、**AggregateInstanceExtraSpecsFilter** および **NUMATopologyFilter** の値が含まれている必要があります。

マルチセルクラウド

マルチセルクラウドの場合、同じセル内の異なるホストにインスタンスのライブマイグレーションを行うことはできますが、セルをまたがる移行を行うことはできません。

フローティングインスタンス

フローティングインスタンスのライブマイグレーションを行う場合、移行先コンピュートノードの **NovaComputeCpuSharedSet** の設定と移行元コンピュートノードの **NovaComputeCpuSharedSet** の設定が異なると、移行先コンピュートノードでは、インスタンスは共有の (ピンングされていない) インスタンス用に設定された CPU には割り当てられません。したがって、フローティングインスタンスのライブマイグレーションを行う必要がある場合は、専用の (ピンングされた) インスタンスおよび共有の (ピンングされていない) インスタンスに関して、すべてのコンピュートノードに同じ CPU マッピングを設定する必要があります。あるいは、共有のインスタンスにホストアグリゲートを使用します。

移行先コンピュートノードの容量

移行先コンピュートノードには、移行するインスタンスをホストするのに十分な空き容量が必要です。

SR-IOV ライブマイグレーション

SR-IOV ベースのネットワークインターフェイスを使用するインスタンスは、ライブマイグレーションが可能です。ダイレクトモードの SR-IOV ネットワークインターフェイスを持つインスタンスのライブマイグレーションでは、ネットワークのダウンタイムが発生します。これは、移行時に、ダイレクトモードのインターフェイスの接続を解除し再び接続する必要があるためです。

ML2/OVN デプロイでのパケット損失

ML2/OVN は、パケット損失のないライブマイグレーションをサポートしていません。これは、OVN が複数のポートバインディングを処理できないため、ポートがいつ移行されるかがわからないためです。

ライブマイグレーション中のパケット損失を最小限に抑えるには、移行が完了したら、宛先ホストでインスタンスをアナウンスするように、ML2/OVN デプロイメントを設定します。

```
parameter_defaults:
  ComputeExtraConfig:
    nova::config::nova_config:
```

```
workarounds/enable_qemu_monitor_announce_self:
  value: 'True'
```

ML2/OVS デプロイメントでのライブマイグレーション

ML2/OVS デプロイメントでインスタンスをライブマイグレーションする場合のパケット損失を最小限に抑えるには、ネットワーキングサービス (neutron) ライブマイグレーションイベントを有効にするように、ML2/OVS デプロイメントを設定し、移行が完了したら、宛先ホストでインスタンスをアナウンスします。

```
parameter_defaults:
  NetworkExtraConfig:
    neutron::config::neutron_config:
      nova/live_migration_events:
        value: 'True'
  ComputeExtraConfig:
    nova::config::nova_config:
      workarounds/enable_qemu_monitor_announce_self:
        value: 'True'
```

ライブマイグレーションの妨げとなる制約

以下の機能を使用するインスタンスのライブマイグレーションを行うことはできません。

PCI パススルー

QEMU/KVM ハイパーバイザーでは、コンピュータノード上の PCI デバイスをインスタンスにアタッチすることができます。PCI パススルーを使用すると、インスタンスは PCI デバイスに排他的にアクセスすることができ、これらのデバイスがインスタンスのオペレーティングシステムに物理的に接続されているかのように表示され、動作します。ただし、PCI パススルーには物理デバイスへの直接アクセスが必要なため、QEMU/KVM は PCI パススルーを使用するインスタンスのライブマイグレーションをサポートしません。

ポートリソースの要求

最小帯域幅を確保する QoS ポリシーなど、リソース要求が設定されたポートを使用するインスタンスのライブマイグレーションを行うことはできません。ポートにリソース要求があるかどうかを確認するには、以下のコマンドを使用します。

```
$ openstack port show <port_name/port_id>
```

14.2.3. 移行の準備

1つまたは複数のインスタンスを移行する前に、コンピュータノード名および移行するインスタンスの ID を把握する必要があります。

手順

1. 移行元コンピュータノードのホスト名および移行先コンピュータノードのホスト名を特定します。

```
(undercloud)$ source ~/overcloudrc
(overcloud)$ openstack compute service list
```

2. 移行元コンピュートノード上のインスタンスを一覧表示し、移行するインスタンスの ID を特定します。

```
(overcloud)$ openstack server list --host <source> --all-projects
```

<source> を移行元コンピュートノードの名前または ID に置き換えてください。

3. オプション: ノードのメンテナンส์を行うためにインスタンスを移行元コンピュートノードから移行する場合、ノードを無効にして、メンテナンส์中にスケジューラーがノードに新規インスタンスを割り当てるのを防ぐ必要があります。

```
(overcloud)$ source ~/stackrc
(undercloud)$ openstack compute service set <source> nova-compute --disable
```

<source> を移行元コンピュートノードの名前または ID に置き換えてください。

これで移行を行う準備が整いました。[Cold migrating an instance](#) または [Live migrating an instance](#) で詳しく説明されている必須手順に従います。

14.2.4. インスタンスのコールドマイグレーション

インスタンスのコールドマイグレーションでは、インスタンスを停止して別のコンピュートノードに移動します。コールドマイグレーションは、PCI パススルーを使用するインスタンスの移行など、ライブマイグレーションでは対応することのできない移行シナリオに対応します。移行先コンピュートノードは、スケジューラーにより自動的に選択されます。詳細は、[移行の制約](#) を参照してください。

手順

1. インスタンスのコールドマイグレーションを行うには、以下のコマンドを入力してインスタンスの電源をオフにして移動します。

```
(overcloud)$ openstack server migrate <instance> --wait
```

- **<instance>** を移行するインスタンスの名前または ID に置き換えてください。
 - ローカルに確保されたボリュームを移行する場合には、**--block-migration** フラグを指定します。
2. 移行が完了するまで待ちます。インスタンスの移行が完了するのを待つ間、移行のステータスを確認することができます。詳細は、[Checking migration status](#) を参照してください。
 3. インスタンスのステータスを確認します。

```
(overcloud)$ openstack server list --all-projects
```

ステータスが VERIFY_RESIZE と表示される場合は、移行を確認する、または元に戻す必要があることを示しています。

- 予想どおりに機能している場合は、移行を確認します。

```
(overcloud)$ openstack server resize --confirm <instance>
```

<instance> を移行するインスタンスの名前または ID に置き換えてください。ステータスが ACTIVE と表示される場合は、インスタンスを使用する準備が整っていることを示しています。

- 予想どおりに機能していない場合は、移行を元に戻します。

```
(overcloud)$ openstack server resize --revert <instance>
```

<instance> をインスタンスの名前または ID に置き換えてください。

4. インスタンスを再起動します。

```
(overcloud)$ openstack server start <instance>
```

<instance> をインスタンスの名前または ID に置き換えてください。

5. オプション: メンテナンスのために移行元コンピュートノードを無効にした場合は、新規インスタンスがノードに割り当てられるようにノードを再度有効にする必要があります。

```
(overcloud)$ source ~/stackrc
(undercloud)$ openstack compute service set <source> nova-compute --enable
```

<source> を移行元コンピュートノードのホスト名に置き換えてください。

14.2.5. インスタンスのライブマイグレーション

ライブマイグレーションでは、ダウンタイムを最小限に抑えて、インスタンスを移行元コンピュートノードから移行先コンピュートノードに移動します。ライブマイグレーションがすべてのインスタンスに適しているとは限りません。詳細は、[移行の制約](#) を参照してください。

手順

1. インスタンスのライブマイグレーションを行うには、インスタンスおよび移行先コンピュートノードを指定します。

```
(overcloud)$ openstack server migrate <instance> --live-migration [--host <dest>] --wait
```

- **<instance>** をインスタンスの名前または ID に置き換えてください。
- **<dest>** を移行先コンピュートノードの名前または ID に置き換えてください。



注記

openstack server migrate コマンドは、共有ストレージを持つインスタンスの移行が対象です。これがデフォルトの設定です。ローカルに確保されたボリュームを移行するには、**--block-migration** フラグを指定します。

```
(overcloud)$ openstack server migrate <instance> --live-migration [--host <dest>] --wait --block-migration
```

2. インスタンスが移行されていることを確認します。

```
(overcloud)$ openstack server show <instance>
```

```
+-----+-----+
| Field           | Value                               |
+-----+-----+
```

```
| ...      | ...      |
| status   | MIGRATING |
| ...      | ...      |
+-----+
```

3. 移行が完了するまで待ちます。インスタンスの移行が完了するのを待つ間、移行のステータスを確認することができます。詳細は、[Checking migration status](#) を参照してください。
4. インスタンスのステータスをチェックして、移行が成功したかどうかを確認します。

```
(overcloud)$ openstack server list --host <dest> --all-projects
```

<dest> を移行先コンピュートノードの名前または ID に置き換えてください。

5. オプション: メンテナンスのために移行元コンピュートノードを無効にした場合は、新規インスタンスがノードに割り当てられるようにノードを再度有効にする必要があります。

```
(overcloud)$ source ~/stackrc
(undercloud)$ openstack compute service set <source> nova-compute --enable
```

<source> を移行元コンピュートノードのホスト名に置き換えてください。

14.2.6. 移行ステータスの確認

移行が完了するまでに、さまざまな移行状態を遷移します。正常な移行では、通常、移行状態は以下のようになります。

1. **Queued:** Compute サービスはインスタンス移行の要求を受け入れ、移行は保留中です。
2. **Preparing:** Compute サービスはインスタンス移行の準備中です。
3. **Running:** Compute サービスはインスタンスを移行中です。
4. **Post-migrating:** Compute サービスはインスタンスを移行先コンピュートノードにビルドし、移行元コンピュートノードのリソースを解放しています。
5. **Completed:** Compute サービスはインスタンスの移行を完了し、移行元コンピュートノードのリソース解放を終了しています。

手順

1. インスタンスの移行 ID の一覧を取得します。

```
$ nova server-migration-list <instance>

+----+-----+-----+ (...)
| Id | Source Node | Dest Node | (...)
+----+-----+-----+ (...)
| 2 | -      | -      | (...)
+----+-----+-----+ (...)
```

<instance> をインスタンスの名前または ID に置き換えてください。

2. 移行のステータスを表示します。

```
$ nova server-migration-show <instance> <migration_id>
```

- **<instance>** をインスタンスの名前または ID に置き換えてください。
- **<migration_id>** を移行の ID に置き換えてください。
nova server-migration-show コマンドを実行すると、以下の例に示すような出力が返されます。

```
+-----+-----+
| Property      | Value                                |
+-----+-----+
| created_at     | 2017-03-08T02:53:06.000000         |
| dest_compute   | controller                          |
| dest_host      | -                                   |
| dest_node      | -                                   |
| disk_processed_bytes | 0                                  |
| disk_remaining_bytes | 0                                  |
| disk_total_bytes  | 0                                  |
| id             | 2                                   |
| memory_processed_bytes | 65502513                          |
| memory_remaining_bytes | 786427904                          |
| memory_total_bytes  | 1091379200                          |
| server_uuid     | d1df1b5a-70c4-4fed-98b7-423362f2c47c |
| source_compute  | compute2                            |
| source_node     | -                                   |
| status         | running                             |
| updated_at      | 2017-03-08T02:53:47.000000         |
+-----+-----+
```

ヒント

OpenStack Compute サービスは、コピーする残りのメモリーのバイト数によって移行の進捗を測定します。時間が経過してもこの数字が減少しない場合、移行を完了することができず、Compute サービスは移行を中止する可能性があります。

インスタンスの移行に長い時間がかったり、エラーが発生したりする場合があります。詳細は、[Troubleshooting migration](#) を参照してください。

14.2.7. インスタンスの退避

インスタンスを障害の発生したコンピュートノードまたはシャットダウンしたコンピュートノードから同じ環境内の新しいホストに移動する場合、インスタンスを退避させることができます。

退避プロセスにより元のインスタンスは破棄され、元のイメージ、インスタンス名、UUID、ネットワークアドレス、およびインスタンスに割り当てられていたその他すべてのリソースを使用して、別のコンピュートノードに元のインスタンスが再ビルドされます。

インスタンスが共有ストレージを使用する場合、インスタンスのルートディスクは退避プロセス中に再ビルドされません。移行先コンピュートノードが引き続きこのディスクにアクセス可能なためです。インスタンスが共有ストレージを使用しない場合は、インスタンスのルートディスクも移行先コンピュートノードに再ビルドされます。



注記

- コンピュートノードがフェンシングされ、API が報告するコンピュートノードの状態が down または forced-down である場合に限り、退避を行うことができます。コンピュートノードが down または forced-down と報告されない場合、**evacuate** コマンドは失敗します。
- クラウド管理者でなければ、退避を行うことはできません。

14.2.7.1. 単一のインスタンスの退避

インスタンスを一度に1つずつ退避させることができます。

手順

1. 障害の発生したコンピュートノードに管理者としてログインします。
2. コンピュートノードを無効にします。

```
(overcloud)[stack@director ~]$ openstack compute service set \
  <host> <service> --disable
```

- **<host>** をインスタンスの退避元となるコンピュートノードの名前に置き換えてください。
- **<service>** を無効にするサービスの名前 (例: **nova-compute**) に置き換えてください。

3. インスタンスを退避させるには、以下のコマンドを入力します。

```
(overcloud)[stack@director ~]$ nova evacuate [--password <pass>] <instance> [<dest>]
```

- **<pass>** を退避するインスタンスに設定する admin パスワードに置き換えてください。パスワードを指定しなかった場合には、無作為に生成され、退避の完了時に出力されます。
- **<instance>** を退避させるインスタンスの名前または ID に置き換えてください。
- **<dest>** をインスタンスの退避先となるコンピュートノードの名前に置き換えてください。退避先コンピュートノードを指定しなかった場合には、Compute のスケジューラーがノードを選択します。退避先に指定可能なコンピュートノードを確認するには、以下のコマンドを使用します。

```
(overcloud)[stack@director ~]$ openstack hypervisor list
```

14.2.7.2. ホスト上の全インスタンスの退避

指定したコンピュートノード上の全インスタンスを退避させることができます。

手順

1. 障害の発生したコンピュートノードに管理者としてログインします。
2. コンピュートノードを無効にします。

```
(overcloud)[stack@director ~]$ openstack compute service set \
  <host> <service> --disable
```

- **<host>** をインスタンスの退避元となるコンピュートノードの名前に置き換えてください。
- **<service>** を無効にするサービスの名前 (例: **nova-compute**) に置き換えてください。

3. 指定したコンピュートノード上の全インスタンスを退避させます。

```
(overcloud)[stack@director ~]$ nova host-evacuate [--target_host <dest>] <host>
```

- **<dest>** をインスタンスの退避先となるコンピュートノードの名前に置き換えてください。退避先を指定しなかった場合には、Compute のスケジューラーがノードを選択します。退避先に指定可能なコンピュートノードを確認するには、以下のコマンドを使用します。

```
(overcloud)[stack@director ~]$ openstack hypervisor list
```

- **<host>** をインスタンスの退避元となるコンピュートノードの名前に置き換えてください。

14.2.8. 移行に関するトラブルシューティング

インスタンスの移行時に、以下の問題が発生する可能性があります。

- 移行プロセスでエラーが生じる。
- 移行プロセスが終了しない。
- 移行後にインスタンスのパフォーマンスが低下する。

14.2.8.1. 移行中のエラー

以下の問題が発生すると、移行操作が **error** 状態に遷移します。

- 実行しているクラスターに異なるバージョンの Red Hat OpenStack Platform (RHOSP) が存在する。
- 指定したインスタンス ID が見つからない。
- 移行を試みているインスタンスが **error** 状態にある。
- Compute サービスが停止している。
- 競合状態が発生する。
- ライブマイグレーションが **failed** 状態に移行する。

ライブマイグレーションが **failed** 状態に移行すると、通常は **error** 状態になります。**failed** 状態の原因となる可能性のある典型的な問題を以下に示します。

- 移行先コンピュートホストが利用可能な状態にない。
- スケジューラーの例外が発生する。
- コンピュティングリソースが不十分なため、再ビルドプロセスに失敗する。
- サーバグループの確認に失敗する。
- 移行先コンピュートノードへの移行が完了する前に、移行元コンピュートノードのインスタンスが削除される。

14.2.8.2. ライブマイグレーションのスタック

ライブマイグレーションが完了せず、永久に **running** 状態のままになる可能性があります。ライブマイグレーションが永久に完了しない一般的な理由は、Compute サービスがインスタンスの変更を移行先コンピュートノードに複製するより早く、クライアントのリクエストにより移行元コンピュートノード上で実行中のインスタンスに変更が生じることです。

この状況に対処するには、以下のいずれかの方法を使用します。

- ライブマイグレーションを中止する。
- ライブマイグレーションを強制的に完了させる。

ライブマイグレーションの中止

移行プロセスがインスタンスの状態の変化を移行先ノードにコピーするより早くインスタンスの状態が変化する状況で、インスタンスの動作を一時的に中断したくない場合には、ライブマイグレーションを中止することができます。

手順

1. インスタンスの移行の一覧を取得します。

```
$ nova server-migration-list <instance>
```

<instance> をインスタンスの名前または ID に置き換えてください。

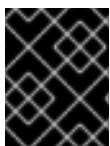
2. ライブマイグレーションを中止します。

```
$ nova live-migration-abort <instance> <migration_id>
```

- **<instance>** をインスタンスの名前または ID に置き換えてください。
- **<migration_id>** を移行の ID に置き換えてください。

ライブマイグレーション完了の強制

移行プロセスがインスタンスの状態の変化を移行先ノードにコピーするより早くインスタンスの状態が変化する状況で、インスタンスの動作を一時的に中断して移行を強制的に完了させたい場合には、ライブマイグレーションの手順を強制的に完了させることができます。



重要

ライブマイグレーションを強制的に完了させると、かなりのダウンタイムが発生する可能性があります。

手順

1. インスタンスの移行の一覧を取得します。

```
$ nova server-migration-list <instance>
```

<instance> をインスタンスの名前または ID に置き換えてください。

2. ライブマイグレーションを強制的に完了させます。

```
$ nova live-migration-force-complete <instance> <migration_id>
```

- **<instance>** をインスタンスの名前または ID に置き換えてください。
- **<migration_id>** を移行の ID に置き換えてください。

14.2.8.3. 移行後のインスタンスパフォーマンスの低下

NUMA トポロジーを使用するインスタンスの場合、移行元および移行先コンピュータノードの NUMA トポロジーおよび設定は同一でなければなりません。移行先コンピュータノードの NUMA トポロジーでは、十分なリソースが利用可能でなければなりません。移行元および移行先コンピュータノード間で NUMA 設定が同一でない場合、ライブマイグレーションは成功するがインスタンスのパフォーマンスが低下する可能性があります。たとえば、移行元コンピュータノードは NIC 1 を NUMA ノード 0 にマッピングするが、移行先コンピュータノードは NIC 1 を NUMA ノード 5 にマッピングする場合、移行後にインスタンスはバス内の最初の CPU からのネットワークトラフィックを NUMA ノード 5 の別の CPU にルーティングし、トラフィックを NIC 1 にルーティングする可能性があります。その結果、予想されたとおりに動作はしますが、パフォーマンスが低下します。同様に、移行元コンピュータノードの NUMA ノード 0 では十分な CPU および RAM が利用可能だが、移行先コンピュータノードの NUMA ノード 0 にリソースの一部を使用するインスタンスがすでに存在する場合、インスタンスは正しく動作するがパフォーマンスが低下する可能性があります。詳細は、[移行の制約](#) を参照してください。