Red Hat OpenStack Platform 13

大規模デプロイメントにおける推奨事項

大規模な OpenStack Platform をデプロイする際のハードウェア要件および設定

Last Updated: 2023-03-15

Red Hat OpenStack Platform 13 大規模デプロイメントにおける推奨事項

大規模な OpenStack Platform をデプロイする際のハードウェア要件および設定

Enter your first name here. Enter your surname here.
Enter your organisation's name here. Enter your organisational division here.
Enter your email address here.

法律上の通知

概要

本ガイドでは、大規模な Red Hat OpenStack Platform をデプロイする際のさまざまな推奨事項を記載します。これらの推奨事項には、ハードウェアの推奨事項、アンダークラウドのチューニング、およびオーバークラウドの設定が含まれます。

目次

第1章 はじめに

本書には、大規模な Red Hat OpenStack Platform 環境をデプロイするための推奨アンダークラウドおよびオーバークラウドの仕様および設定について説明します。50 を超えるオーバークラウドノードのデプロイメントは、大規模な環境として対応します。

第2章 推奨される仕様

2.1. アンダークラウド

最適なパフォーマンスを得るには、物理サーバーにアンダークラウドノードをインストールします。ただし、仮想アンダークラウドノードを使用する場合、仮想マシンには、以下の表で説明されている物理マシンと同様の十分なリソースを確保するようにしてください。

表2.1 推奨されるアンダークラウドノードの仕様

ノード数	1
CPU の数	32 コア、64 スレッド
ディスク	500 GB のルートディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1)) Swift 用 500 GB のディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1))
メモリー	64 GB
ネットワーク	10 Gbps ネットワークインターフェイス

2.2. オーバークラウドコントローラーノード

すべてのコントロールプレーンサービスは、3 つのノードで稼働する必要があります。通常、すべてのコントロールプレーンサービスは 3 つのコントローラーノードに分散してデプロイされます。

コントローラーサービスのスケーリング

コントローラーサービスで利用可能なリソースを増やすには、これらのサービスを追加のノードにスケーリングできます。たとえば、**db** または **messaging** コントローラーサービスを専用のノードにデプロイして、コントローラーノードの負荷を軽減することができます。

コントローラーサービスをスケーリングするには、コンポーザブルロールを使用してスケーリングするサービスのセットを定義します。コンポーザブルロールを使用する場合には、各サービスは 3 つの追加の専用ノードで実行される必要があり、Pacemaker のクォーラムを維持するために、コントロールプレーン内のノードの合計数を追加する必要があります。

この例のコントロールプレーンは、以下の 9 ノードで設定されます。

- コントローラーノード 3 台
- データベースノード 3 台
- メッセージングノード 3 台

詳しい情報は、**Advanced Overcloud Customization** の Composable services and custom roles を参照してください。

コンポーザブルロールを使用したコントローラーサービスのスケーリングに関する質問は、Red Hat Global Professional Services にお問い合わせください。

ストレージに関する考慮事項

オーバークラウドデプロイメントのコントローラーノードを計画する場合には、十分なストレージを追加します。OpenStack Telemetry Metrics (gnocchi) および OpenStack Image (glance) は、I/O 負荷の高いサービスです。オーバークラウドは I/O の負荷を Ceph OSD サーバーに移すので、Image サービスおよびテレメトリー用に Ceph Storage を使用します。

デプロイメントに Ceph ストレージが含まれていない場合には、Telemetry Metrics (gnocchi) および Image (glance) サービスが利用できる Object Storage (swift) 用に専用のディスクまたはノードを使用します。コントローラーノードで Object Storage を使用する場合は、ルートディスクとは別に NVMe デバイスを使用し、オブジェクトデータ保存時のディスク使用率を削減します。

表2.2 Ceph Storage ノードを使用する場合に推奨されるコントローラーノードの仕様

ノード数	Controller ロールに含まれるコントローラーサービスを持つ 3 台のコントローラーノード オプションとして、専用ノードでコントローラーサービスをスケーリングするには、コンポーザブルサービスを使用します。詳しい情報は、**Advanced Overcloud Customization** の Composable services and custom roles を参照してください。
CPU の数	2 ソケット (それぞれ 12 コア、24 スレッド)
ディスク	500 GB のルートディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1))
メモリー	128 GB
ネットワーク	25 Gbps のネットワークインターフェイスまたは 10 Gbps のネットワークインターフェイス。10 Gbps ネットワークインターフェイスを使用する場合には、ネットワークボンディングを使用して 2 つのボンディングを作成します。 ● プロビジョニング (bond0、mode4)、内部 API (bond0、mode4)、プロジェクト (bond0、mode4) ● ストレージ (bond1、mode4)、ストレージ管理 (bond1、mode4)

表2.3 Ceph Storage ノードを使用しない場合に推奨されるコントローラーノードの仕様

ノード数	Controller ロールに含まれるコントローラーサービスを持つ 3 台のコントローラーノード オプションとして、専用ノードでコントローラーサービスをスケーリングするには、コンポーザブルサービスを使用します。詳しい情報は、**Advanced Overcloud Customization** の Composable services and custom roles を参照してください。

CPU の数	2 ソケット (それぞれ 12 コア、24 スレッド)
ディスク	500 GB のルートディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1)) Swift 用 500 GB のディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1))
メモリー	128 GB
ネットワーク	25 Gbps のネットワークインターフェイスまたは 10 Gbps のネットワークインターフェイス。10 Gbps ネットワークインターフェイスを使用する場合には、ネットワークボンディングを使用して 2 つのボンディングを作成します。 ● プロビジョニング (bond0、mode4)、内部 API (bond0、mode4)、プロジェクト (bond0、mode4) ● ストレージ (bond1、mode4)、ストレージ管理 (bond1、mode4)

2.3. オーバークラウドのコンピュートノード

表2.4 推奨されるコンピュートノードの仕様

ノード数	Red Hat は 300 ノードのスケールをテストしました。
CPU の数	2 ソケット (それぞれ 12 コア、24 スレッド)
ディスク	500 GB のルートディスク (1 x SSD または 2 x 7200 RPM のハードドライブ (RAID 1)) glance イメージキャッシュ用の 500 GB ディスク(1 x SSD または 2 x 7200 RPM のハードドライブ(RAID 1))
メモリー	128 GB (NUMA ノードあたり 64 GB)。2 GB はホスト用に追加設定なしで予約されます。 分散仮想ルーターでは、確保されるメモリーを 5 GB に増やします。

ネットワーク	25 Gbps のネットワークインターフェイスまたは 10 Gbps のネットワークインターフェイス。10 Gbps ネットワークインターフェイスを使用する場合には、ネットワークボンディングを使用して 2 つのボンディングを作成します。 ● プロビジョニング (bond0、mode4)、内部 API (bond0、mode4)、プロジェクト (bond0、mode4) ● ストレージ (bond1、mode4)

2.4. RED HAT CEPH STORAGE ノード

表2.5 推奨される Ceph Storage ノードの仕様

ノード数	3 方向のレプリケーションを持つ最低でも 5 つのノードが必要です。オールフラッシュ設定では、双方向レプリケーションを持つ 3 つ以上のノードが必要です。
CPU の数	1 OSD あたり 1 つの Intel Broadwell CPU コア (ストレージ I/O の要件に対応するため)。I/O 負荷が軽い場合は、Ceph をブロックデバイスの速度で実行する必要がない場合があります。たとえば、一部の NFV アプリケーションの場合、Ceph はデータの持続性、高可用性、および低レイテンシーを提供しますが、スループットはターゲットではないため、CPU 処理能力を下げることが許容されます。
メモリー	OSD ごとに 5 GB の RAM を許可します。これは、OSD プロセスメモリー用だけでなく、パフォーマンスを最適化するために OSD データおよびメタデータをキャッシュするために必要です。ハイパーコンバージドインフラストラクチャー (HCI) 環境では、コンピュートノードの仕様と共に必要なメモリーを計算します。
ネットワーク	ネットワーク容量 (MB/s 単位) を Ceph デバイスの合計 MB/s 容量よりも大きくし、大規模な I/O 転送サイズを使用するワークロードをサポートするようにしてください。OSD 間のトラフィックを別の物理ネットワークポートセットに移動して、クラスターネットワークを使用して書き込みレイテンシーを軽減します。Red Hat OpenStack Platform でこれを行うには、ネットワーク用に別の VLAN を設定し、別の物理ネットワークインターフェイスに VLAN を割り当てます。

ディスク	SSD （ソリッドステートドライブ）のジャーナリングにより、ハードディスクドライブ(HDD)での I/O 競合が減少し、書き込み IOPS の速度が向上しますが、SSD は 1 秒あたりの読み取りの入出力操作には影響がありません。SATA/SAS SSD ジャーナルを使用する場合には、通常 SSD:HDD の比率を 1:5 にする必要があります。NVM SSD ジャーナルを使用する場合には、通常は SSD:HDD の比率を 1:10 または 1:15 に指定し、ワークロードが最も読み取られる場合にも使用できます。ただし、この比率が高すぎると、SSD ジャーナルデバイスの障害により OSD に影響が及ぶ可能性があります。

詳しくは、Deploying an Overcloud with Containerized Red Hat Ceph を参照してください。

ストレージレプリケーション番号の変更に関する詳細は、**Red Hat Ceph Storage 設定ガイド**の プール、PG、および CRUSH 設定リファレンス を参照してください。

第3章 推奨されるデプロイメントプラクティス

3.1. デプロイメントの準備に関する考慮事項

オーバークラウドイメージの root パスワードを設定します。

- オーバークラウドイメージの root パスワードを設定して、オーバークラウドイメージへのコンソールアクセスを許可する。ネットワークが正しく設定されていない場合に、コンソールを使用して失敗したデプロイメントのトラブルシューティングを行います。**パートナー統合ガイド** の director への virt-customize のインストール および root パスワード の設定 を参照してください。

特定のノード ID の割り当て

- スケジューラーヒントを使用して、**Controller**、**Compute**、**CephStorage** などのロールにハードウェアを割り当てます。スケジューラーヒントにより、特定のハードウェアのみに影響するデプロイメントの問題をより簡単に特定できます。

- 単一プロセスである **nova-scheduler** は、多数のノードをスケジュールする際に酷使される可能性があります。タグの照合を実装する際に、スケジューラーヒントは **nova-scheduler** への負荷を軽減します。その結果、**nova-scheduler** はデプロイメント時のスケジューリングエラーが少なくなりました。スケジューラーヒントにより、デプロイメントが一般的に短縮されます。

- スケジューラーヒントを使用する場合は、プロファイルのタグ付けを使用しないでください。

- パフォーマンステストでは、特定のロールに同じハードウェアを使用して、テストおよびパフォーマンスの結果の差異を軽減します。

- **オーバークラウドの高度なカスタマイズの 特定の ノード ID の割り当て を参照して** ください。

ルートディスクヒントの設定

- ノードに複数のディスクが含まれる場合は、イントロスペクションデータを使用して、各ノードのルートディスクヒントとして WWN を設定します。これにより、デプロイメントおよび起動時にノードが誤ったディスクを使用しないようになります。**director のインストールと使用方法の ルートディスク の定義 を参照してください。**

OpenStack Bare Metal サービス(ironic)のクリーニングを使用する

- ironic の自動クリーニングを使用して、複数のディスクを持ち複数のブートローダーを持つ可能性があるノードでメタデータを削除することを強く推奨します。ディスクに複数のブートローダーが存在することが原因でノードがブートディスクと一貫性がなくなる場合があり、これにより、誤った URL を使用してメタデータをプルする際にノードがデプロイに失敗する場合があります。

ironic イントロスペクションのノード数を制限する

- 一度に全ノードをイントロスペクションすると失敗します。イントロスペクションの場合は、一度に 20 ノードが推奨されます。**undercloud.conf** ファイルの **dhcp_start** および **dhcp_end** の範囲が、環境にあるノードの数に対して十分な大きさになるようにしてください。十分な IP が利用できない場合は、同時にイントロスペクション操作の数を制限するために、範囲のサイ

ズを超えて発行します。イントロスペクションの DHCP リースが期限切れになるのを許可するには、イントロスペクションが完了してから数分間は IP アドレスをさらに発行しないでください。

Ceph の準備

- 以下の一覧は、異なるタイプの設定の推奨事項のセットです。

  オールフラッシュ OSD 設定

  それぞれの OSD には、デバイス種別の IOPS 能力に応じて追加の CPU が必要になるため、Ceph IOPS は少数の OSD で CPU に制限されます。これは、従来の HDD よりも 2 桁大きい IOPS 能力を持つことのできる NVM SSD の場合に言えます。SATA/SAS SSD の場合、HDD よりも 1 桁大きいランダム IOPS/OSD が予想されますが、シーケンシャル IOPS は 2 - 4 倍しか増えません。OSD デバイスが必要とするよりも少ない CPU リソースを Ceph に提供できますが、すべてのフラッシュ設定はコストがかかります。

  ハイパーコンバージドインフラストラクチャー (HCI)

  OpenStack Compute (nova) ゲスト用に、CPU パワー、メモリー、およびネットワークの半分以上を確保することが推奨されます。OpenStack Compute (nova)ゲストと Ceph Storage の両方をサポートするのに十分な CPU およびメモリーを確保することを計画します。Ceph Storage のメモリー消費は弾力的ではないため、メモリー消費を確認します。マルチ CPU ソケットシステムでは、NUMA ピニングされた Ceph で Ceph の CPU 消費を 1 つのソケットに制限します。たとえば、**numactl -N 0 -p 0** コマンドを使用します。Ceph のメモリー消費を 1 つのソケットにハードピニングしないでください。

  NFV 等のレイテンシーに敏感なアプリケーション

  異なる NUMA ソケットおよびネットワークカード上で動作するネットワークアプリケーションにより、可能であれば Ceph を Ceph が使用するネットワークカードと同じ CPU ソケットに配置し、ネットワークカードの割り込みをその CPU ソケットに制限します。

- デュアルブートローダーを使用する場合は、OSD マップに disk-by-path を使用することを推奨します。これにより、デバイス名を使用するのとは異なり、ユーザーは一貫性のあるデプロイメントを行うことができます。以下のスニペットは、disk-by-path マッピングの **CephAnsibleDisksConfig** の例です。

```
CephAnsibleDisksConfig:
  osd_scenario: non-collocated
  devices:
    - /dev/disk/by-path/pci-0000:03:00.0-scsi-0:2:0:0
    - /dev/disk/by-path/pci-0000:03:00.0-scsi-0:2:1:0
  dedicated_devices:
    - /dev/nvme0n1
    - /dev/nvme0n1
  journal_size: 512
```

3.2. デプロイメントに関する考慮事項

小規模なスケールでデプロイメントコマンドを検証する

- 3 つ以上のコントローラーノード、1 つのコンピュートノード、および 3 つの Ceph Storage ノードで設定される小規模な環境をデプロイします。この設定を使用して、すべての heat テンプレートが正しいことを確認することができます。ノードをさらに追加すると、デプロイにかかる時間が長くなるため、この推奨ノードレイアウトとその他のノード種別を使用して、Heat テンプレートに問題が存在するかどうかを確認することができます。

同時にプロビジョニングされるノードの数を制限します。

- Red Hat は、32 ノードを同時にデプロイすることを推奨します。32 は、平均的なエンタープライズレベルのラックユニット内に収まる一般的なサーバーの量です。これにより、平均の 1 つのラックを同時にデプロイできます。デプロイメントでの問題を診断するために必要なデバッグを最小限にするには、一度に 50 を超えるノードをデプロイしないでください。より多くのノードをデプロイすることに気づいた場合は、Red Hat は最大 100 ノードを同時に成功した状態でテストしました。

未使用の NIC を無効にする

- デプロイ中にオーバークラウドに未使用の NIC がある場合には、NIC 設定テンプレートで未使用のインターフェイスを定義して、インターフェイスを **use_dhcp: false** および **defroute: false** に設定する必要があります。これを行わないと、イントロスペクションおよびスケーリング操作中にルーティングの問題および IP の割り当ての問題が発生します。デフォルトでは、NIC は BOOTPROTO=dhcp を設定します。つまり、未使用のオーバークラウド NIC は、PXE プロビジョニングに必要な IP アドレスを消費します。これにより、ノードで利用可能な IP アドレスのプールが減少する場合があります。

未使用の ironic ノードの電源をオフにする

- メンテナーンスモードで未使用の ironic ノードの電源をオフにしてください。Red Hat は、以前のデプロイメントからのノードが電源オンの状態でメンテナーンスモードのままになるケースを特定しています。これは、クリーニングに失敗したノードがメンテナーンスモードになる OpenStack Bare Metal (ironic)の自動クリーニングで発生する可能性があります。ironic はメンテナーンスモードのノードの電源状態を追跡しないため、ironic は電源状態を誤ってオフとして報告します。これにより、進行中のデプロイメントで問題が発生する可能性があります。デプロイメントの失敗後に再デプロイする場合には、ノードの電源管理デバイスを使用して未使用のノードの電源をオフにしてください。

3.3. アンダークラウドのチューニングに関する考慮事項

Keystone ワーカー数の増加

- Red Hat では、8 つ以上の keystone 管理プロセスと、アンダークラウド上に 4 つの keystone のメインプロセスを使用することを推奨しています。設定ファイルは、**/etc/httpd/conf.d/10-keystone_wsgi_admin.conf** および **/etc/httpd/conf.d/10-keystone_wsgi_main.conf** です。

- アップグレード間で永続的な変更を行うには、または **openstack undercloud install** を再実行する際に、**undercloud.conf** ファイルに **hieradata_override** を設定してカスタムの hieradata ファイルを注入します。以下の行をカスタムの hieradata ファイルに追加します。

```
keystone::wsgi::apache::custom_wsgi_process_options_admin: { processes : "8" }
keystone::wsgi::apache::custom_wsgi_process_options_main: { processes : "4" }
```

Heat API 呼び出しの応答タイムアウトを増やします。

- デフォルトの **rpc_response_timeout** は、**/etc/heat/heat.conf** で 600 秒に設定されています。深刻なリソースの競合がある場合は、タイムアウトを増やします。デプロイメントがメッセージングタイムアウトで終了した場合、これはこの設定を増やすためのインジケーターです。これは一般的な問題にしないでください。

- アップグレード間で永続的な変更を行うか、**openstack undercloud install** を再実行する際には、以下の行をカスタムの hieradata ファイルに追加し、適切なタイムアウト時間を指定します。

```
heat::rpc_response_timeout: 600
```

Keystone トークンのタイムアウト時間を増やします。

- オーバークラウドのデプロイのタイムアウト時間を 14,400 秒以上に増やす場合は、**keystone.conf** の keystone トークンの有効期限のタイムアウトを、秒単位で同等の値に更新する必要があります。デフォルトの Keystone トークンのタイムアウト時間は 14400 秒です。

- アップグレード間で永続的な変更を行うか、**openstack undercloud install** を再実行する際には、以下の行をカスタムの hieradata ファイルに追加し、適切なタイムアウト時間を指定します。

```
* keystone::token_expiration: 14400
```

Telemetry が使用されていない場合は、無効にします。

- 請求に使用するメトリックデータが必要ない場合は、Telemetry を無効にします。アンダークラウド上で Telemetry を無効にするには、**undercloud.conf** ファイルを編集し、**enable_telemetry** の値を false に変更し、 **openstack undercloud install** コマンドを再実行します。

- **openstack overcloud deploy** 中に Telemetry を無効にするには、 **Deployment Recommendations for Specific Red Hat OpenStack Platform Services Guide** の Telemetry を参照してください。

第4章 デバッグのヒント

4.1. イントロスペクションのデバッグ

- **undercloud.conf** ファイルでイントロスペクション用 DHCP 範囲と NIC を確認するこれらの値のいずれかが誤りである場合は修正し、openstack undercloud install コマンドを再度実行します。

- DHCP 範囲のノードが許可できるよりも多くのノードのイントロスペクションを試みないようにしてください。また、イントロスペクション完了後、各ノードの DHCP リースが約 2 分間アクティブになることに注意してください。

- 全ノードでイントロスペクションに異常が発生した場合には、設定済みの NIC を使用してネイティブ VLAN 経由でターゲットノードを ping できること、および帯域外インターフェイスの認証情報およびアドレスが正しいことを確認します。

- 特定のノードをデバッグする際には、ノードのブート時にコンソールを監視し、ノードのイントロスペクションコマンドを確認します。PXE プロセスの完了前にノードが停止した場合は、接続、IP の割り当て、およびネットワーク負荷を確認します。ノードが BIOS を終了し、イントロスペクションイメージでブートする場合は、障害はまれで、ほぼ接続性の問題に関連します。イントロスペクションイメージからのハートビートが、アンダークラウドへの伝送中に中断されないようにします。

4.2. デプロイメントのデバッグ

- プロビジョニングネットワーク上のアドレスを提供する追加の DHCP サーバーがあると、director がマシンを検査およびプロビジョニングできなくなります。

- DHCP または PXE の問題の場合：

  - イントロスペクションの問題については、以下のコマンドを実行します。

    ```
    sudo tcpdump -i any port 67 or port 68 or port 69
    ```

  - デプロイメントの問題については、以下を実行します。

    ```
    sudo ip netns exec qdhcp tcpdump -i <interface> port 67 or port 68 or port 69
    ```

- 障害の発生したディスクまたは外部ディスクの場合、マシンの帯域外管理に応じて **Up** 状態を持たないディスクに注意してください。ディスクは、デプロイメントサイクル中に **Up** の状態を終了し、ディスクがベースオペレーティングシステムに表示される順番を変更する可能性があります。

- **openstack stack failures list overcloud**、**heat resource-list -n5 overcloud | grep -i fail** を実行します。出力を確認し、障害が発生するノードにログインして **/var/log/** および **/var/log/containers/** のログを確認し、**journalctl -u os-collect-config** を実行します。