Red Hat OpenStack Platform 10

Recommendations for Large Deployments

Hardware requirements and configuration for deploying OpenStack Platform at scale

Last Updated: 2019-03-05

Red Hat OpenStack Platform 10 Recommendations for Large Deployments

Hardware requirements and configuration for deploying OpenStack Platform at scale

OpenStack Team
rhos-docs@redhat.com

Legal Notice

Abstract

This guide contains several recommendations for deploying Red Hat OpenStack Platform at scale. These recommendations include hardware recommendations, undercloud tuning, and overcloud configuration.

Table of Contents

CHAPTER 1. INTRODUCTION

This document contains information about the recommended undercloud and overcloud specifications and configuration for deploying a large Red Hat OpenStack Platform environment.

CHAPTER 2. RECOMMENDED SPECIFICATIONS

2.1. UNDERCLOUD

For best performance, install the undercloud node on a physical server. However, if you use a virtualized undercloud node, ensure that the virtual machine has enough resources similar to a physical machine described in the following table.

Table 2.1. Recommended specifications for undercloud node

Counts	1
CPUs	12 cores, 24 threads
Disk	500GB root disk (1x SSD or 2x hard drives with 7200RPM; RAID 1) 500GB disk for swift (1x SSD or 2x hard drives with 7200RPM; RAID 1)
Memory	64 GB
Network	10 Gbps network interfaces

2.2. OVERCLOUD CONTROLLER NODES

When planning for Controller nodes, storage planning is important. OpenStack Telemetry Metrics (gnocchi) and OpenStack Image Storage (glance) services are I/O intensive. For best performance, Red Hat recommends using Ceph Storage for Telemetry and Image Storage because the overcloud moves the I/O load to the Ceph OSD servers. If your deployment does not include Ceph storage, Red Hat recommends a dedicated disk or node for OpenStack Object Storage (swift) that OpenStack Telemetry Metrics (gnocchi) and OpenStack Image Storage (glance) services can use. If using OpenStack Object Storage (swift) on Controller nodes, use an NVMe device separate from the root disk to reduce disk utilization when storing object data.

Table 2.2. Recommended specifications for Controller nodes when you use Ceph Storage nodes

Counts	A least 3 Controller nodes with all controller services contained within a single Controller role. Optionally use composable services and custom roles if required.
CPUs	2 sockets each with 12 cores, 24 threads
Disk	500GB root disk (1x SSD or 2x hard drives with 7200RPM; RAID 1)
Memory	128 GB

Network	25 Gbps network interfaces or 10 Gbps network interfaces. If using 10 Gbps network interfaces, use network bonding to create two bonds: • Provisioning (bond0 - mode4); Internal API (bond0 - mode4); Tenant (bond0 - mode4) • Storage (bond1 - mode4); Storage management (bond1 - mode4)

Table 2.3. Recommended specifications for Controller nodes when you do not use Ceph Storage nodes

Counts	A least 3 Controller nodes with all controller services contained within a single Controller role. Optionally use composable services and custom roles if required.
CPUs	2 sockets each with 12 cores, 24 threads
Disk	500GB root disk (1x SSD or 2x hard drives with 7200RPM; RAID 1) 500GB disk for Swift (1x SSD or 2x hard drives with 7200RPM; RAID 1)
Memory	128 GB
Network	25 Gbps network interfaces or 10 Gbps network interfaces. If using 10 Gbps network interfaces, use network bonding to create two bonds: • Provisioning (bond0 - mode4); Internal API (bond0 - mode4); Tenant (bond0 - mode4) • Storage (bond1 - mode4); Storage management (bond1 - mode4)

2.3. OVERCLOUD COMPUTE NODES

Table 2.4. Recommended Compute node specifications

Counts	Red Hat has tested a scale of 300 nodes.
CPUs	2 sockets each with 12 cores, 24 threads

Disk	500GB root disk (1x SSD or 2x hard drives with 7200RPM; RAID 1) 500GB disk for glance image cache (1x SSD or 2x hard drives with 7200RPM; RAID 1)
Memory	128 GB (64 GB per NUMA node); 2GB is reserved for the host out of the box. With Distributed Virtual Routing, increase the reserved RAM to 5 GB.
Network	25 Gbps network interfaces or 10 Gbps network interfaces. If using 10 Gbps network interfaces, use network bonding to create two bonds: • Provisioning (bond0 - mode4); Internal API (bond0 - mode4); Tenant (bond0 - mode4) • Storage (bond1 - mode4)

2.4. RED HAT CEPH STORAGE NODES

Table 2.5. Recommended Ceph Storage node specifications

Counts	A minimum of 5 nodes with three-way replication is required. With all-flash configuration, a minimum of 3 nodes with two-way replication is required.
CPUs	1 Intel Broadwell CPU core per OSD to support storage I/O requirements. If you are using a light I/O workload, you might not need Ceph to run at the speed of your block devices. For example, for some NFV applications, Ceph supplies data durability, high availability, and low latency but throughput is not really a target, so it is acceptable to supply a little less CPU power.
Memory	Allow 5 GB RAM per OSD. This is required for caching OSD data and metadata to optimize performance, not just for the OSD process memory. For hyper-converged infrastructure (HCI) environments, calculate the required memory in conjunction with the Compute node specifications.

Network	Ensure the network capacity in MB/s is higher than the total MB/s capacity of the Ceph devices to support workloads that use a large I/O transfer size. Use a cluster network to lower write latency by shifting inter-OSD traffic onto a separate set of physical network ports. To do this in Red Hat OpenStack Platform, configure separate VLANs for networks and assigning the VLANs to separate physical network interfaces.
Disk	Solid-State Drive (SSD) Journaling reduces I/O contention on hard disk drives (HDD), which increases the speed of write IOPS, but SSDs have zero effect on read input/output operations per second. If using SATA/SAS SSD journals, you typically need a ratio of SSD:HDD of 1:5. If using NVM SSD journals, you can typically use a SSD:HDD ratio of 1:10 or even 1:15 in cases where the workload is read-mostly. However, if this ratio is too high, the SSD journal device failure can affect the OSDs.

For more information, see Red Hat Ceph Storage for the Overcloud.

For more information on changing the storage replication number, see Pool, PG, and CRUSH Configuration Reference in the *Red Hat Ceph Storage Configuration Guide*.

CHAPTER 3. RECOMMENDED DEPLOYMENT PRACTICES

3.1. DEPLOYMENT PREPARATION CONSIDERATIONS

Set root password for overcloud image

- Set the root password on your overcloud image to allow console access to the overcloud image. Use the console to troubleshoot failed deployments when networking is set incorrectly. See Installing virt-customize to the director and Setting the Root Password in the *Partner Integration Guide*.

Assign specific node IDs

- Use scheduler hints to assign hardware to a role, such as **`Controller`**, **`Compute`**, **`CephStorage`**, and others. Scheduler hints allow for easier identification of deployment issues that affect only a specific piece of hardware.
- The **`nova-scheduler`**, which is a single process, can overexert when scheduling a large number of nodes. Scheduler hints reduce the load on **`nova-scheduler`** when implementing tag matching. As a result, **`nova-scheduler`** encounters fewer scheduling errors during the deployment. The deployment in general takes less time with scheduler hints.
- Do not use profile tagging when using scheduler hints.
- In performance testing, use identical hardware for specific roles in order to reduce variability in testing and performance results.
- See Assigning Specific Node IDs in the *Advanced Overcloud Customization Guide*.

Set root disk hints

- When nodes contain multiple disks, use the introspection data to set the WWN as the root disk hint for each node. This prevents the node from using the wrong disk during deployment and booting. See Defining the Root Disk for Nodes in the *Director Installation and Usage Guide*.

Use OpenStack Bare Metal service (ironic) cleaning

- It is highly recommended to use ironic automated cleaning to erase metadata on nodes that have more than one disk and are likely to have multiple boot loaders. There are some cases where nodes are inconsistent with the boot disk due to the presence of multiple bootloaders on disks, which leads to nodes failing to deploy when attempting to pull the metadata using the wrong URL.

Limit the number of nodes for ironic introspection

- Introspecting all nodes at once result in failure. The recommendation is 20 nodes at a time for introspection. Make sure that the **dhcp_start** and **dhcp_end** range in the **undercloud.conf** file is large enough for the number of nodes you expect to have in the environment. If not enough IPs are available, issue no more than the size of the range to limit the number of simultaneous introspection operations. Do not issue more IP addresses for a few minutes after the introspection completes to allow introspection DHCP leases to expire.

Ceph preparation

- The following list is a set of recommendations for different types of configurations:

  All-flash OSD configuration

  Each OSD requires additional CPU according to the IOPS capacity of the device type, so Ceph IOPS are CPU-limited at a lower number of OSDs. This is true for NVM SSDs, which can have two orders of magnitude higher IOPS capacity than traditional HDDs. For SATA/SAS SSDs, expect one order of magnitude greater random IOPS/OSD than HDDs, but only about two to four times the sequential IOPS increase. You can supply less CPU resources to Ceph than Ceph needs for OSD devices, but all-flash configurations are expensive.

  Hyper Converged Infrastructure (HCI)

  It is recommended to reserve at least half of your CPU, memory, and network for the OpenStack Compute (nova) guests. Plan on having enough CPU and memory to support both OpenStack Compute (nova) guests and Ceph Storage. Observe memory consumption because Ceph Storage memory consumption is not elastic. On a multi-CPU socket system, limit Ceph CPU consumption with NUMA-pinning Ceph to a single socket. For example use the **`numactl -N 0 -p 0`** command. Do not hard-pin Ceph memory consumption to 1 socket.

  Latency-sensitive applications such as NFV

  Place Ceph on the same CPU socket as the network card Ceph uses and limit the network card interruptions to that CPU socket if possible, with a network application running on a different NUMA socket and network card.

- If using dual bootloaders, it is recommended to use disk-by-path for the OSD map. This gives the user consistent deployments, unlike using the device name. The following snippet is an example of the Ceph hieradata for a disk-by-path mapping.

```
ceph::profile::params::osds: +
  '/dev/disk/by-path/pci-0000:03:00.0-scsi-0:2:0:0': +
    journal: '/dev/nvme0n1'
  '/dev/disk/by-path/pci-0000:03:00.0-scsi-0:2:1:0':
    journal: '/dev/nvme0n1'
```

3.2. DEPLOYMENT CONSIDERATIONS

Validate the deployment command with small scale

- Deploy a small environment that consists of at least 3 Controllers, 1 Compute, and 3 Ceph Storage nodes. Use this configuration to ensure that all of your Heat templates are correct. Adding more nodes increases the amount of time to deploy, so running a small deployment with this recommended node layout and any other node types you might have confirms if an issue exists in your Heat templates.

Limit the number of nodes provisioned at the same time

- Red Hat recommends deploying 32 nodes at the same time. 32 is the typical amount of servers that can fit within a average enterprise-level rack unit, which allows you to deploy an average of one rack of nodes simultaneously. Deploy no more than 32 nodes at a time to minimize the debugging necessary to diagnose issues with the deployment. If you feel comfortable deploying a higher number of nodes, Red Hat has tested up to 100 nodes simultaneously with high success.

Disable unused NICs

- If the overcloud has any unused NICs during the deployment, you must define the unused interfaces in the NIC configuration templates and set the interfaces to **`use_dhcp: false`** and **`defroute: false`**. Failing to do so causes routing issues and IP allocation problems during introspection and scaling operations. By default, the NICs set **`BOOTPROTO=dhcp`**, which means the unused overcloud NICs consume IP addresses meant for the PXE provisioning. This can reduce the pool of available IP addresses for your nodes.

Power off unused ironic nodes

- Ensure that you power off any unused ironic nodes in maintenance mode. Red Hat has identified cases where nodes from previous deployments are left in maintenance mode in a powered on state. This can occur with OpenStack Bare Metal (ironic) automated cleaning where a node that fails cleaning is put into maintenance mode. Since ironic does not track the power state of nodes in maintenance mode, ironic incorrectly reports the power state as off. This can cause problems with ongoing deployments. When redeploying after a failed deployment, ensure that you power off any unused nodes using the node's power management device.

3.3. UNDERCLOUD TUNING CONSIDERATIONS

Increase Keystone Worker count

- Red Hat recommends that you have more than 8 keystone admin processes and 4 keystone main processes on your undercloud. The configuration files are **`/etc/httpd/conf.d/10-keystone_wsgi_admin.conf`** and **`/etc/httpd/conf.d/10-keystone_wsgi_main.conf`**.

- To make a persistent change across upgrades or when you rerun **`openstack undercloud install`**, inject a custom hieradata file by setting **`hieradata_override`** in the **`undercloud.conf`** file. Add the following lines to the custom hieradata file:

```
keystone::wsgi::apache::custom_wsgi_process_options_admin: { workers
=> "8" }
keystone::wsgi::apache::custom_wsgi_process_options_main: { workers
=> "4" }
```

Increase the response timeout for Heat API calls

- The default **`rpc_response_timeout`** is set to 600 seconds in **`/etc/heat/heat.conf`**. In cases with severe resource contention, increase the timeout. If you see the deployment exiting with messaging timeouts, that is an indicator to increase this setting. This should not be a common issue.

- To make a persistent change across upgrades or when you rerun **`openstack undercloud install`**, add the following line to the custom hieradata file and specify a suitable timeout time:

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

Increase Keystone token timeout time

- If you increase the overcloud deploy timeout time to more than 14,400 seconds, you must update the keystone token expiration timeout in **`keystone.conf`** to the equivalent value in seconds. The default Keystone token timeout time is 14400 seconds.
- To make a persistent change across upgrades or when you rerun **`openstack undercloud install`**, add the following line to the custom hieradata file and specify a suitable timeout time:

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

If Telemetry is not used, disable it

- If you do not require metric data, which is used for billing purposes, disable Telemetry. To disable Telemetry on the undercloud, edit the **`undercloud.conf`** file, change the **`enable_telemetry`** value to false, and rerun the **`openstack undercloud install`** command.
- To disable Telemetry during **`openstack overcloud deploy`**, see Telemetry in the *Deployment Recommendations for Specific Red Hat OpenStack Platform Services Guide* for more information.

CHAPTER 4. DEBUGGING TIPS

4.1. INTROSPECTION DEBUGGING

- Check your introspection DHCP range and NICs in your **`undercloud.conf`** file. If either of these values are incorrect, fix them and rerun the **`openstack undercloud install`** command.
- Ensure you are not trying to introspect more than your DHCP range of nodes can allow. Also remember that the DHCP lease for each node will still be active for approximately two minutes after introspection finishes.
- If all nodes fail introspection, ensure that you can ping target nodes over the native VLAN using the configured NIC and that the out-of-band interface credentials and addresses are correct.
- For debugging specific nodes, watch the console when the node boots and observe introspection commands to the node. If the node stops before completing the PXE process, check the connectivity, IP allocation, and the network load. When a node exits the BIOS and boots the introspection image, failures are rare and almost exclusively connectivity issues. Ensure that the heartbeat from the introspection image is not interrupted on its way to the undercloud.

4.2. DEPLOYMENT DEBUGGING

- Any additional DHCP servers that supply addresses on the provisioning network can prevent director from inspecting and provisioning machines.
- For DHCP or PXE issues:
  - For introspection issues, run the following command:

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

  - For deployment issues, run:

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

- For failed or foreign disks, be aware of disks that do not have an **`Up`** state according to the machine's out-of-band management. Disks can exit the **`Up`** state during a deployment cycle and change the order that your disks appear in the base operating system.
- Run **`openstack stack failures list overcloud`**, and **`heat resource-list -n5 overcloud | grep -i fail`**. Review the output, then log into the node where the failure occurs and review **`/var/log/messages`** or **`journalctl -u os-collect-config`**.