<table>
<thead>
<tr>
<th>Version</th>
<th>Author</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.31</td>
<td>Barnes</td>
<td>Based on VSM Product Specification 2.5 and VSM In-depth training June 2014</td>
<td>June 30th 2014</td>
</tr>
<tr>
<td>0.4</td>
<td>Barnes</td>
<td>Revisions based on customer feedback</td>
<td>July 11th 2014</td>
</tr>
<tr>
<td>0.4.1</td>
<td>Barnes</td>
<td>Various minor updates to clarify Monitor Status, PG Status, RBD Status, and OSD Status</td>
<td>July 14th 2014</td>
</tr>
<tr>
<td>0.7.0</td>
<td>Barnes</td>
<td>Update to conform with new features implemented since VSM0.5: - Support for creation of erasure coded pools - Support for cache tiering - Support for split storage groups - Support for setting pool quotas - Flexible storage class naming - Changes to cluster and server manifest files to support erasure coded pools and cache tiering - Switch to HTTPS for VSM web server</td>
<td>February 8th 2015</td>
</tr>
<tr>
<td>0.7.1</td>
<td></td>
<td>Update to conform with new features implemented since VSM0.5: - Support for creation of new storage groups - Updated description of Manage Pools page to include replica storage group column in pools table</td>
<td>February 20th 2015</td>
</tr>
</tbody>
</table>
Contents

1. Overview ............................................................................................................................................... 7

1.1. VSM .................................................................................................................................................. 7
     VSM Controller ..................................................................................................................................... 8
     VSM Agent .......................................................................................................................................... 8

1.2. Ceph ................................................................................................................................................ 9
     Network Topology ............................................................................................................................... 9
     Ceph Server Nodes .............................................................................................................................. 9
     Ceph Client Nodes .............................................................................................................................. 9

1.3. OpenStack ....................................................................................................................................... 10

1.4. Applications .................................................................................................................................... 10
     CloudStack ........................................................................................................................................... 10
     Object Storage via RADOS Gateway ................................................................................................. 10
     CephFS .............................................................................................................................................. 10

1.5. Cluster Management ....................................................................................................................... 10
     Managing Failure Domains .................................................................................................................. 12
     Cluster and Server Manifest File Configuration for Disk Management ............................................. 13
     Ceph Server Node Roles ...................................................................................................................... 14
     Ceph Server Node Discovery and Authentication ............................................................................... 15
     Pool Data Placement Rules ................................................................................................................ 15

2. System Configuration .......................................................................................................................... 16

2.1. VSM Software Environment ........................................................................................................... 16

2.2. VSM Installation ............................................................................................................................... 16
     Preinstalled Software ........................................................................................................................... 17

2.3. Cluster Manifest ............................................................................................................................... 17
     Storage Class ...................................................................................................................................... 18
     Zone .................................................................................................................................................... 19
     Storage Group .................................................................................................................................... 19
     Cluster Name ...................................................................................................................................... 19
     File System ......................................................................................................................................... 19
1. Overview

Intel Virtual Storage Manager (VSM) is a browser-based management application for Ceph storage systems. VSM supports Ceph clusters comprising a mix of hard disk drives (HDDs), Solid State Drives (SSDs), and SSD-cached HDDs. This chapter provides an overview of how VSM manages a Ceph cluster.

Chapter 2 describes how a VSM-managed Ceph cluster is installed and configured.

Chapter 3 describes the operational services supported by the VSM browser-based interface and API for managing and monitoring the cluster and storage capacity.

Chapter 4 provides low-level details of how VSM controls the Ceph cluster.

1.1. VSM

VSM provides a unified, consistent view of a Ceph storage system. It provides a framework for classifying and organizing storage system assets, and simplifies the initialization, configuration, monitoring, and management of the Ceph storage system.

The following diagram provides a high-level logical view of a VSM-managed Ceph cluster in a cloud datacenter application.
**VSM Controller**

The VSM Storage Controller typically runs on a dedicated server or server instance. The controller maintains system state and manages storage system configuration, including configuration of Ceph data placement rules. The controller also provides and browser-based graphical interface that supports operations managements. The controller provides and interface to OpenStack for managing Ceph storage pools under OpenStack Cinder, and also provides system monitoring of cluster health and capacity utilization.

**VSM Agent**

VSM agent processes run on each storage node. The VSM agent enables automatic discovery of server nodes by the VSM controller, provides server configuration information to the controller, maintains connection between the VSM controller and the Ceph cluster, and monitors server performance metrics.

The key elements of a typical VSM-managed Ceph cluster in a cloud datacenter application are shown in the following diagram:
1.2. Ceph

Ceph is an open-source software-defined storage system. It provides object and block storage in a single platform. Its multi-tenant Object Storage is compatible with Amazon™ S3 and OpenStack™ Swift. Its block storage includes copy-on-write, snapshots, cloning. Ceph implements a self-managing and self-healing distributed storage architecture. A Ceph cluster is comprised of server nodes that that host some combination of Ceph OSD (storage), Monitor, and MDS daemons.

Network Topology

Client nodes attach to the Ceph cluster through via Ethernet using the RADOS protocol. In a VSM-managed cluster, the network is comprised of a public network (client-facing) that carries data traffic between clients and ceph server nodes, a cluster network that carries replication and rebalancing data between server nodes, and an administrative network for VSM and Ceph cluster management.

Ceph Server Nodes

Servers that are members of the Ceph cluster may host a combination of Ceph OSD (storage), monitor, and MDS daemons.

Servers that host storage run Ceph software referred to as Ceph Object Store Daemons (OSDs). There is one OSD process for each HDD or SSD storage drive (or partition) in the system. OSDs use journaling to improve write performance; journal partitions may be placed on a separate partition of the data drive, or may be placed on a separate SSD drive for better performance. OSDs work together in the Ceph cluster to manage data distribution and striping. Data is protected against loss through the use of replication or erasure coding; in either case, data is distributed across drives and servers so as to ensure that a failure of any one drive or server does not result in data loss or interruption of service. When disks or server fail or are removed, OSDs work together to maintain data protection by redistributing data to remaining disk in the cluster. When disks are added to the system, the OSDs work together to redistribute a portion the data to the newly added disks, thereby increasing the overall performance of the cluster.

Ceph monitors provide the cluster with a shared, up-to-date map of storage resources. Ceph clients and servers communicate with the monitor processes in order to maintain an up-to-date cluster map at each node, which allows all clients and cluster servers to independently determine data location. On small clusters, monitor processes may run on storage nodes; on larger clusters, monitor processes are run on dedicated servers. The monitors work together to establish consensus on the state of the cluster map, and the cluster runs multiple Ceph monitors to ensure that cluster state is preserved in the event of a monitor failure.

Ceph Client Nodes

Ceph clients connect to the Ceph cluster via an IP-based network (typically Ethernet or IPoIB) using the RADOS protocol. Ceph clients access storage as block storage (through RADOS Block Device drivers) or as object storage (thought the RADOS RESTfull gateway). Ceph clients determine the location of data in a Ceph cluster by connecting to a Ceph monitor and retrieving a current cluster map, from which the location of any data object can be calculated.
1.3. OpenStack

OpenStack is a well-known cloud orchestration solution. Ceph is frequently used with OpenStack to provide scalable, distributed block and object storage for an OpenStack-managed cloud.

VSM can be optionally connected to an OpenStack Nova server; when so connected, VSM is able to “present” Ceph pools to the OpenStack Cinder multi-backend service. In order to connect VSM to OpenStack, VSM requires a pre-configured SSH connection be established between the server hosting the VSM controller and a server hosting an OpenStack Nova controller.

VSM 1.0 does not manage the interoperation of Ceph object storage with OpenStack Glance or OpenStack Swift. It is expected that pools created in VSM can be used with Glance and Swift, however, this functionality has not been tested, and configuration of Ceph interoperation with Glance and Swift is outside the scope of operations for VSM 1.0.

1.4. Applications

VSM 1.0 has been designed and tested primarily to support the use of Ceph as block storage in cloud datacenter applications using OpenStack. However, it is expected that VSM-managed Ceph clusters can also be used with other cloud orchestration software (e.g. CloudStack™), and in other applications not requiring the use of cloud orchestration software (e.g. object storage via RADOS Gateway or Ceph FS).

CloudStack

Ceph is interoperable with CloudStack cloud orchestration software, and a Ceph cluster managed by VSM 1.0 is expected to be interoperable with CloudStack, however, this configuration has not been tested. Note that VSM 1.0 will not directly interoperate with CloudStack; presentation of pools as block storage for CloudStack clients is outside the scope of operations for VSM 1.0.

Object Storage via RADOS Gateway

A Ceph cluster managed by VSM 1.0 is expected to be interoperable with the Ceph RADOS gateway client, however, this configuration has not been tested. Note that VSM 1.0 does not manage configuration of the Ceph RADOS gateway service; configuration of and interoperability with RADOS gateway is outside the scope of operations for VSM 1.0.

CephFS

To accommodate basic functional test of CephFS in a VSM-managed cluster, VSM 1.0 runs one MDS daemon on one server in the cluster. It is expected that that a Ceph cluster managed by VSM 1.0 is interoperable with CephFS, however, this configuration has not been tested. VSM 1.0 does not manage the operation of CephFS; configuration of and interoperability with CephFS is outside the scope of operations for VSM 1.0.

1.5. Cluster Management

This section describes the basic concepts used by VSM to manage a Ceph cluster.

VSM manages initial Ceph cluster configuration and ongoing operations over the lifetime of the cluster.
Initial cluster configuration parameters are specified in a *cluster manifest* file. Configuration of individual storage servers is specified in a *server manifest* file that is created for each storage server added to the cluster.

VSM manages the creation of the Ceph CRUSH map, placing servers and disks into the CRUSH map in a prescribed manner that facilitates consistent behavior and reporting of capacity utilization.

VSM allows servers (and their associated data disks) to be added and removed from the server as a unit – under VSM the server is the basic unit of addition or removal of capacity to or from the Ceph cluster. VSM 1.0 allows individual disks to be replaced, but *does not* allow individual disks to be added to or removed from the cluster.

VSM enables the creation of storage pools, using either replication or erasure codes to ensure data durability. Pools can be created in a single storage group, or optionally, by placing the primary and secondary replicas in separate storage groups.

**Managing Disks - Storage Group and Storage Class**

VSM manages the grouping of drives in the Ceph CRUSH map. Drives can be identified to be placed together under a single CRUSH map bucket. Drives to be placed together under a single CRUSH map bucket are identified by *storage class*; drives belonging to the same *Storage Class* are grouped together in a *Storage Group* bucket.

Storage Groups and associated Storage Class definitions are specified at initial cluster creation time by placing the storage class and storage group definitions in the *cluster manifest* file. New Storage Groups and their associated Storage Classes can be added to the cluster after initial cluster creation by using the Add Storage Group feature on the Manage Storage Group page.

Disk storage class identification is specified in the storage manifest file for each storage server that is to be added to the cluster.

The following diagram illustrates how VSM uses Storage Class and Storage Group to organize drives distributed across multiple servers into common CRUSH map buckets. Servers 01 and 02 contain a mix of SSD and 10K RPM HDDs, while servers 03 and 04 contain 7200 RPM HDDs. In this example, drives with common performance characteristics are assigned to the same Storage Class; drives with the same Storage Class are grouped together under the same Storage Group.
Managing Failure Domains

Ceph allows data to be distributed across physical failure domains. For example, storage servers placed in two different racks with separate top of rack (TOR) Ethernet switches have two network failure domains; in the event of the failure of one switch, data can still be accessed from the servers in the other rack. Ceph can distribute data replicas so that in the event of a failure in one failure domain (e.g. loss of TOR switch), replica data can be accessed from another failure domain. Ceph manages data distribution by placing OSDs in the CRUSH map hierarchy in separate buckets under a root bucket.

VSM manages the placement of servers (and their respective OSDs) that share the same failure domain into the same CRUSH map bucket under their respective Storage Group. In VSM, a zone represents the bucket where the servers and OSDs are placed.

Servers are assigned to a specific failure zone by specifying the zone in the storage manifest file for each storage server to be added to the cluster.

In the following diagram, servers 01 and 02 are placed in zone 1, while servers 03, and 04 are placed in zone 2. VSM creates a CRUSH map with separate 10K RPM and 7200 RPM Storage Group buckets, and places OSDs into their respective zone buckets under the Storage Group bucket.
Note: In VSM 1.0, the zone names specified in the cluster manifest determine the number of zones for *every* Storage Group in the cluster—each storage group has the same number of zones, and each Storage Group must have OSDs from at least one server in each zone in order to maintain cluster stability.

**Cluster and Server Manifest File Configuration for Disk Management**

Storage Class names are defined in the cluster manifest file, as shown in the following example:

```
[storage_class]
7200_rpm_sata_object
7200_rpm_sata_block
10k_rpm_sas_block
ssd_block
```

Storage Group names and Storage Class associations are defined in the cluster manifest file as shown in the following example:

```
[storage_group]
#format: [storage group name] ["user friendly storage group name"] [storage class]
high_performance_SSD_block "High Performance Block Storage" ssd_block
capacity_object "Economy Object Storage" 7200_rpm_sata_object
capacity_block "Economy Block Storage" 7200_rpm_sata_block
performance_block "Performance Block Storage" 10k_rpm_sas_block
```
The Storage Class designation is used in the server manifest file to identify Storage Group membership for the disks of each storage server to be added to the cluster. In the following example the server has eight data drives and corresponding journals – four drives are designated as 7200_rpm_sata_object drives and four drives are designated as 7200_rpm_sata_block drives. Referring to the previous example, they will be placed respectively in the capacity_object and capacity_block storage groups.

```
[7200_rpm_sata_object]
#format [data_device] [journal_device]
%osd-by-path-1% %journal-by-path-1%
%osd-by-path-2% %journal-by-path-2%
%osd-by-path-3% %journal-by-path-3%
%osd-by-path-4% %journal-by-path-4%

[7200_rpm_sata_block]
#format [data_device] [journal_device]
%osd-by-path-5% %journal-by-path-5%
%osd-by-path-6% %journal-by-path-6%
%osd-by-path-7% %journal-by-path-7%
%osd-by-path-7% %journal-by-path-7%
```

Zone names are defined in the cluster manifest file, as shown in the following example:

```
[zone]
zone_a
zone_b
zone_c
```

Each server in the cluster is placed in a specific zone by specifying the zone in the server manifest file, as shown in the following example:

```
[zone]
zone_a
```

Note: In VSM 1.0, the zone names specified in the cluster manifest determine the number of zones for every Storage Group in the cluster – each storage group has the same number of zones, and each Storage Group must have OSDs from at least one server in each zone in order to maintain cluster stability.

Note: This section describes only the cluster and server manifest file settings for disk management. A full description of the cluster manifest and storage manifest files is provided in sections 2.3 and 2.4

**Ceph Server Node Roles**

Each server in the Ceph cluster may run a Ceph monitor daemon and one or more Ceph OSD daemons. The configuration for each server is specified in the server manifest file.
VSM does not perform portioning of drives; drives must be portioned and accurately represented in the cluster manifest prior to adding a server to the cluster.

**Ceph Server Node Discovery and Authentication**

VSM Agent software runs on every Ceph server node managed by VSM.

When VSM Agent runs for the first time, it attempts to contact the VSM controller at the address specified in the server manifest file. If contact is successful, it provides the controller with the authorization key specified in the server manifest. If the key is valid, the agent will be recognized and the server will be added to the list of available servers.

The authorization key is generated by the agent-token utility which can be found on the server hosting the VSM controller at `/usr/bin/agent-token`. The authorization key expires after 24 hours.

Once validated, the VSM agent is always recognized by the VSM controller

**Pool Data Placement Rules**

Ceph uses CRUSH rulesets to determine data placement rules for pools. In VSM 1.0, a ruleset is created for each pool. The ruleset points to the storage group (or storage groups) that have been selected for the pool at the time the pool is created. See section

See section 0 for examples of how the CRUSH ruleset is specified for pools.
2. System Configuration

2.1. VSM Software Environment

VSM has been tested in the following environment:

- CentOS 6.5 and the 3.10 kernel
- Ceph Firefly 0.80.8
- OpenStack Havana

In addition, VSM relies on a number of 3rd party software components for operation. Consult the VSM 1.0 installation instructions for details.

File System

VSM has been tested using the XFS file system for OSD drives.

Ceph

VSM 1.0 has been tested with Ceph Firefly 0.80.8.

Ceph support creation of pools using either replication or erasure coding. Erasure code profiles are specified in the VSM cluster manifest file and may be specified at the time of pool creation.

FSM supports creation of cache tiers. Cache tier default values are specified in the VSM cluster manifest file and may be modified at the time of cache tier creation.

OpenStack

VSM 1.0 has been tested with OpenStack Havana.

VSM 1.0 directly manages presentation of pools to OpenStack Cinder.

Browser

VSM is a web-based application. VSM 1.0 has been tested with Chrome.

2.2. VSM Installation

VSM is provided as an Open Source application under the Apache 2 license.

VSM is hosted by 01.org and may be found at https://01.org/virtual-storage-manager.

VSM source code and installation instructions is available from: https://github.com/01org/virtual-storage-manager.

VSM 1.0 RPMs and installation instructions will be posted as they become available.
**Preinstalled Software**

If storage system hardware has been provided by a storage system vendor (e.g. OEM or System Integrator) with VSM preinstalled (or installed in-premise using vendor-provided configuration automation tools), then it is likely that the cluster and server manifest files have been preconfigured to match the system hardware configuration, and may require only minor modifications (e.g. in-premise network configuration, OpenStack connection) to operate in the target environment. Follow your storage system vendor’s installation and configuration instructions.

**2.3. Cluster Manifest**

The cluster manifest file provides the VSM controller with information about cluster configuration. A cluster manifest file must be placed on the server hosting the VSM controller.
The following is an example of a cluster manifest file:

```plaintext
#1.0 cluster.manifest
[storage_class]
10krpm_sas
7200_rpm_sata
ssd
ssd_cached_7200rpm_sata
ssd_cached_10krpm_sas

[zone]
zone_a
zone_b
zone_c

[storage_group]
#format: [storage group name] [user friendly storage group name] [storage class]
performance Performance_Disk 10krpm_sas
capacity Economy_Disk 7200_rpm_sata
high_performance High_Performance ssd
value_performance High_Performance_Disk_with_ssd_cached_Acceleration ssd_cached_10krpm_sas
value_capacity Capacity_Disk_with_ssd_cached_Acceleration ssd_cached_7200rpm_sata

[cluster]
abc_100

[file_system]
xfs

[management_addr]
192.168.20.0/24

[ceph_public_addr]
field_public_addr

[ceph_cluster_addr]
192.168.40.0/24

[settings]
storage_group_near_full_threshold 65
storage_group_full_threshold 85
ceph_near_full_threshold 75
ceph_full_threshold 90
pg_count_factor 100
heartbeat_interval 5
osd_heartbeat_interval 6
osd_heartbeat_grace 20

[ec_profiles]
#format: [profile-name] [path-to-plugin] [plugin-name] [pg_num value] [json format key/value]
#the key/value strings should not have spaces
default profile /usr/lib64/ceph/erasure-code jerasure 3 {"k":2,"m":1,"technique":"reed_sol_van"}
profile_name2 /usr/lib64/ceph/erasure-code jerasure 6 {"k":4,"m":2,"technique":"reed_sol_van"}

[cache_tier_defaults]
ct_hit_set_count 1
cct_hit_set_period_s 3600
cct_target_max_mem_mb 1000000
cct_target_dirty_ratio 0.4
```

**Storage Class**

Used at cluster create time to identify drives to be placed together under a single CRUSH map bucket are identified by *storage class*. See section 1.5 for a detailed discussion of disk management.
Zone

The [zone] section of the cluster manifest specifies the failure zones to which servers may be members. If there is only one failure domain in the cluster, the [zone] section may be omitted from the cluster manifest.

See section 1.5 for a detailed discussion of zones and failure domains.

Storage Group

Used at cluster create time to identify storage classes. Drives belonging to the same Storage Class are grouped together in a Storage Group bucket. See section 1.5 for a detailed discussion of disk management.

The user friendly name is the test displayed in the VSM user interface.

Cluster Name

The cluster name is specified in the [cluster] section of the cluster manifest.

File System

The file system to be used for OSD data drives is specified in the [file_system] section of the cluster manifest. VSM 1.0 has been tested with the XFS file system.

Network Configuration

VSM 1.0 expects three network subnets. The specific subnets are specified in the [management_addr], [ceph_public_addr] and [ceph_cluster_addr] sections of the cluster manifest.

The network configuration sections of the manifest file should be modified to match the configuration of the network in the environment in which the cluster will be operated.
**VSM Settings**

VSM settings are specified in the `[settings]` section of the cluster manifest.

In VSM 1.0, the following settings are supported.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>storage_group_near_full_threshold</code></td>
<td>65%</td>
<td>Near full threshold for Storage Groups. VSM Cluster Status and Storage Group Status pages indicate when utilization of any Storage Group exceeds this threshold.</td>
</tr>
<tr>
<td><code>storage_group_full_threshold</code></td>
<td>85%</td>
<td>Full threshold for Storage Groups. VSM Cluster Status and Storage Group Status pages indicate when utilization of any Storage Group exceeds this threshold.</td>
</tr>
<tr>
<td><code>ceph_near_full_threshold</code></td>
<td>75%</td>
<td>Near full threshold for Ceph. VSM Cluster Status page (Cluster Health Summary) will provide a warning when used capacity of any OSD in the cluster exceeds this threshold.</td>
</tr>
<tr>
<td><code>ceph_full_threshold</code></td>
<td>90%</td>
<td>Full threshold for Ceph. VSM Cluster Status page (Cluster Health Summary) will provide a warning when used capacity of any OSD in the cluster exceeds this threshold.</td>
</tr>
<tr>
<td><code>pg_count_factor</code></td>
<td>100</td>
<td>VSM uses <code>pg_count_factor</code> when calculating the PG count for pools. See section 3.2 – Managing Pools for more information regarding VSMs use of <code>pg_count_factor</code>.</td>
</tr>
</tbody>
</table>

**Erasure Code Profiles**

Ceph erasure code profiles are specified in the `[cache_tier_defaults]` section of the cluster manifest. The profiles are displayed for selection (but not modification) when the operator creates an erasure coded pool. Erasure coded pool creation is described in section 4.5.

In VSM 1.0, the following settings are supported for each specified erasure code profile:

<table>
<thead>
<tr>
<th>[profile-name]</th>
<th>User-visible name of erasure code profile that VSM displays to the operator when creating erasure coded pools.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[path-to-plugin]</td>
<td>Directory name from which the erasure code plugin is loaded.</td>
</tr>
<tr>
<td>[plugin-name]</td>
<td>Name of erasure code plugin to use to compute coding chunks and recover missing chunks.</td>
</tr>
<tr>
<td>[pg_num value]</td>
<td>Specifies the value of pg_num used when an erasure coded pool is created using this profile.</td>
</tr>
<tr>
<td>[json format key/value]</td>
<td>Sequence of json-formatted key/value pairs used to configure the erasure code plug-in. Key/value pairs must match the semantics expected by the specified erasure code plugin.</td>
</tr>
</tbody>
</table>
See the Ceph documentation for details regarding the behavior of parameters that govern cache tier operations.

**Cache Tier Defaults**

Cache tier default values are specified in the `[cache_tier_defaults]` section of the cluster manifest. These values are displayed as the default values when the operator creates a cache tier, and can be modified by the operator to satisfy the intended use of the cache tier. Cache tier creation is described in section 4.6.

In VSM 1.0, the following cache tier default settings are supported.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ct_hit_set_count</code></td>
<td>1</td>
<td>Specifies the time that each HitSet covers</td>
</tr>
<tr>
<td><code>ct_hit_set_period_s</code></td>
<td>3600</td>
<td>Specifies the number of HitSets to store</td>
</tr>
<tr>
<td><code>ct_target_max_mem_mb</code></td>
<td>1000000</td>
<td>Specifies the ratio of dirty objects to pool size at which the cache tier</td>
</tr>
<tr>
<td></td>
<td></td>
<td>agent will flush dirty objects to the storage pool.</td>
</tr>
<tr>
<td><code>ct_target_dirty_ratio</code></td>
<td>0.4</td>
<td>Specifies the percentage of cache pool capacity at which the cache tier</td>
</tr>
<tr>
<td></td>
<td></td>
<td>agent will evict objects to maintain free capacity.</td>
</tr>
<tr>
<td><code>ct_target_max_objects</code></td>
<td>1000000</td>
<td>Specifies the maximum number of objects held in cache tier</td>
</tr>
<tr>
<td><code>ct_target_min_flush_age_m</code></td>
<td>10</td>
<td>Specifies the minimum age of an object before the cache tiering agent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>flushes a recently modified (i.e. dirty) object to the backing storage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pool.</td>
</tr>
<tr>
<td><code>ct_target_min_evict_age_m</code></td>
<td>20</td>
<td>Specifies the minimum age of an object before it will be evicted from the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cache tier.</td>
</tr>
</tbody>
</table>

See the Ceph documentation for details regarding the behavior of parameters that govern cache tier operations.

**2.4. Server Manifest**

A server manifest file must be placed on each server to be included in the VSM-managed Ceph cluster. The server manifest file provides the VSM controller with information about the configuration of the server on which it resides.
The following is an example of a server manifest file:

```
[vsm_controller_ip]
# 10.239.82.168
[role]
storage
monitor
[zones]
zone_a
[auth_key]
token-tenant
[ssd]
#format [ssd_device] [journal_device]
[7200-rpm-sata]
#format [sata_device] [journal_device]
[10k-rpm-sas]
#format [data_device] [journal_device]
%osd-by-path-1 %journal-by-path-1
%osd-by-path-2 %journal-by-path-2
%osd-by-path-3 %journal-by-path-3
%osd-by-path-4 %journal-by-path-4
%osd-by-path-5 %journal-by-path-5
%osd-by-path-6 %journal-by-path-6
%osd-by-path-7 %journal-by-path-7
[ssd-cached_7200-rpm-sata]
#format [intel_cache_device] [journal_device]
[ssd-cached_10k-rpm-sas]
```

**VSM Controller Address**

The IP address of the VSM controller on the management network is specified in the `[vsm_controller_ip]` section of the server manifest.

VSM Agents uses this address to establish contact with the VSM controller.

The controller address should section of the server manifest file should be modified to match the actual IP address of the server hosting the VSM controller in the environment in which he cluster will be operated.

**Role**

The `[roles]` section of the server manifest specifies which Ceph daemons to run on the server. Each server may run a Ceph monitor daemon and one or more Ceph OSD daemons.
• If a Ceph monitor is to be run on the server, include `monitor` in the roles section.

• If Ceph OSDs are to be run on the server, include `storage` in the section

Roles that are not to be run on the serve may be removed from the manifest file or commented out.

**Zone**

The `[zone]` section of the server manifest specifies the server’s membership in a failure domain. A server may be a member of only one zone. If there is only one failure domain in the cluster, the `[zone]` section may be omitted from the server manifest.

See section 1.5 for a detailed discussion of zones and failure domains.

**Authorization Key**

The VSM agent authorization key is specified in the `[auth_key]` section of the server manifest.

When server activation is executed, VSM Agent attempts to contact the CSM controller at the address specified in the `[vsm_controller_ip]` section of the server manifest file. If contact is successful, it provides the controller with the authorization key provided in the server manifest. If the key is valid, the agent will be recognized and the server will be added to the list of available servers.

The authorization key is manually generated using the `agent-token` utility, which can be found on the server hosting the VSM controller at `/usr/bin/agent-token`. Authorization keys generated by the `agent-token` utility expire after 24 hours.

Copy the key generated by the `agent-token` utility into the server manifest prior to initializing the cluster.

**Storage Configuration**

The configuration of storage on the server is specified in the storage class sections.

Each entry (line) in a storage class section identifies the path to a data drive (or drive partition) and the path to its corresponding journal drive (or drive partition).

In the example above, seven data drive paths and corresponding journal drive paths are specified for the `[10krpm_sas]` storage class, while no drives are specified for the other storage classes.

See section 1.5 for a detailed discussion of disk management.
3. Operations

VSM provides a browser-based user interface that provides centralized management of common administrative tasks, and provides real-time monitoring of storage system metrics and health.

The following administrative tasks are supported:

- Initial cluster creation
- Adding and removing servers
- Adding and removing monitors
- Replacing failed disks
- Replacing failed storage servers
- Maintaining servers
- Creating storage pools
- Creating cache tiers
- Monitoring cluster health
- Monitoring capacity utilization
- Presenting storage pools to OpenStack Cinder
- Managing VSM users

The remainder of this section is organized as follows:

1. **Getting Started**: The operator logs in for the first time, accepts the end-user license agreement, and creates the initial cluster from the servers supplied with the system.

2. **Managing Capacity**: The operator creates pools in which data can be stored, and monitors capacity utilization and performance

3. **Monitoring Cluster Health**: The operator monitors overall cluster status, and can inspect the status of individual ceph subsystems to verify correct operation or diagnose problems

4. **Managing Servers**: Cluster capacity and performance can be increased by adding servers to the cluster. Servers can be stopped for maintenance, and then restarted. Failed or problematic servers can be removed from the cluster.

5. **Managing Storage Devices**: Failed or problematic OSD drives can be restarted or replaced.

6. **Working with OpenStack**: VSM can connect with a server hosting an OpenStack Nova controller, and tell OpenStack about pools that are available to OpenStack Cinder for block storage

7. **Managing VSM**: VSM users can be created and deleted and passwords set and changed.

3.1. Getting Started

**VSM Secure Web Server**

The VSM 1.0 web server is accessed at the [HTTPS://default address]. The VSM web server uses a self-signed certificate to support a secure HTTPS confection between the operator and VSM. Because the
certificate is self-signed, it will not be recognized by the browser when it is first accessed, and the browser will provide a warning. For example, Chrome displays the following warning:

You may continue to the VSM log-in by clicking on Advanced and the click on Proceed as shown below:

Log In

When logging in to VSM, VSM will display the VSM Log-In screen:

When logging in for the first time, the operator uses the user name “admin” and the first-time password. The first-time password is auto-generated on the server hosting the VSM Controller. It can be found at the following location:

#cat /etc/vsmdeploy/deployrc | grep ADMIN >vsm-admin-dashboard.passwd.txt
The operator is encouraged to change the password after logging in – this is done on the Manage Users page – see section 3.7 for details.

Initial Cluster Creation

After logging in for the first time, VSM will display the Create Cluster page:

All servers available for inclusion in the cluster will be listed. In order for servers to appear in the list of available servers, VSM Agent software must be properly installed, configured, and activated – see section 2.2 for an overview of this process.

The VSM controller will have collected information from each of the VSM agents in the cluster. This information is based on the configuration information provided in each server’s server manifest file. The operator may verify the following before initiating cluster creation:

- All expected servers are present
- Servers are located at the expected IP addresses and subnets
- Each server contains the expected number of drives
- Each server is placed in the expected zone
- Each server’s status is shown as “available”, indicating that the server’s VSM agent has established contact with the VSM controller, and that the server is currently responsive.

Cluster creation is initiated by clicking on the “Create Cluster” button. VSM will open a Create Cluster dialog box listing the servers that will be added to the cluster. Click on the Create Cluster button to confirm.
The dialog box will close and a message box will confirm that initiation of cluster creation is successful. The Create cluster page will update the status of the servers as the cluster is created:

When all servers report Active, the cluster is fully initialized, and is ready for pools to be created.
VSM Navigation

The VSM navigation bar enables the operator to navigate through the pages of the VSM application interface. The navigation bar and function of the VSM pages are described in the following table:

<table>
<thead>
<tr>
<th>Page</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Status</td>
<td>Displays summary information for all cluster subsystems</td>
</tr>
<tr>
<td>Manage Servers</td>
<td>Displays configuration information for all servers in the cluster. Allows Servers to be started and stopped. Allows Monitors to be added and removed. Allows servers to be added and removed.</td>
</tr>
<tr>
<td>Manage Devices</td>
<td>Displays detailed configuration and status information for each data disk in the cluster, including OSD number and up/in/auto-out status, capacity utilization, and data disk and associated journal partition path. Allows data disks to be restarted, removed, and restored.</td>
</tr>
<tr>
<td>Create Cluster</td>
<td>Initial cluster creation is initiated from this page.</td>
</tr>
<tr>
<td>Manage Pools</td>
<td>Lists pools, manages pool and cache tier creation</td>
</tr>
<tr>
<td>Manage Storage Groups</td>
<td>Lists storage groups and associates storage classes Allows new Storage Groups and associated Storage Classes to be created</td>
</tr>
<tr>
<td>Storage Group Status</td>
<td>Displays capacity utilization and node with highest capacity utilization for each Storage Group in the cluster. Provides warning when capacity utilization exceeds near full or full thresholds</td>
</tr>
<tr>
<td>Pools Status</td>
<td>Displays detailed information for each pool in the cluster, including Storage Group, replication, PG count, object statistics, client read and write bytes per second, and client read and write operations per second.</td>
</tr>
<tr>
<td>OSD Status</td>
<td>Displays summary information for all OSDs in the cluster, and detailed information for each OSD in the cluster, including OSD number, OSD up/in/auto-out status, and capacity utilization.</td>
</tr>
<tr>
<td>Monitor Status</td>
<td>Displays summary information for all monitors in the cluster, and detailed information for each monitor in the cluster, including address, health, health detail, skew, latency, and disk capacity utilization.</td>
</tr>
<tr>
<td>MDS Status</td>
<td>Displays summary information for all MDSs in the cluster, and detailed information for each MDS in the cluster, including name, Gid, State, and address.</td>
</tr>
<tr>
<td>PG Status</td>
<td>Displays detailed summary information for all PGs in the cluster, including PG state, object, capacity and performance summaries</td>
</tr>
<tr>
<td>RBD Status</td>
<td>Displays detailed information for the first 100 RBD images reported by Ceph, including pool, image name, allocated size, and format.</td>
</tr>
<tr>
<td>Manage RBD Pools</td>
<td>Connect RBD pools to OpenStack Cinder.</td>
</tr>
</tbody>
</table>
OpenStack Access

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect VSM to an instance of OpenStack Nova so that RBD pools can be connected to OpenStack Cinder.</td>
<td></td>
</tr>
</tbody>
</table>

Add/Remove Users

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add and remove VSM users, change user passwords.</td>
<td></td>
</tr>
</tbody>
</table>

Settings

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inspected and modify selected VSM settings.</td>
<td></td>
</tr>
</tbody>
</table>

VSM Dashboard – Cluster Status

The VSM dashboard provides a comprehensive summary of Cluster status, as shown in the following diagram:

1) The Storage Group Summary section displays summary information for all Storage Groups in the cluster. The operator can quickly verify normal operation, e.g. that no Storage Groups have exceeded the near full or full threshold. The operator can inspect detailed status information for each Storage Group in the cluster by clicking on the Storage Group Status button, which will open the Storage Group Status page.

2) The Monitor Summary section displays summary information for all Ceph monitors in the cluster. The operator can quickly verify normal operation, e.g. that the correct number of monitors are running and that there is a quorum. The operator can inspect detailed status
information for each monitor in the cluster by clicking on the Monitor Status button, which will open the Monitor Status page.

3) The VSM status section shows the uptime for the VSM controller.

4) The OSD Summary section displays the summary status for all OSDs in the cluster. The operator can quickly verify normal operation, e.g. that all OSDs are present, up and in, and that no OSDs have exceeded the near full or full threshold. The operator can inspect detailed status information for each OSD in the cluster by clicking on the OSD Status button, which will open the OSD Status page.

5) The MDS Summary section displays the summary information for all MDS daemons in the cluster. The operator can inspect detailed status information for each MDS daemon in the cluster by clicking on the MDS Status button, which will open the MDS Status page.

6) The PG Summary section displays the summary information for all Placement Groups in the cluster. The operator can quickly verify normal operation, e.g. that the vast majority of PGs are Active + Clean. The operator can inspect detailed summary information for all Placement Groups in the cluster by clicking on the PG Status button, which will open the PG Status page.

7) The Cluster Health Summary section displays all messages reported by the Ceph Health command.

8) The Warnings and Errors section displays all warnings and errors reported by the ceph status -f json_pretty command.

Status information displayed in the Monitory Summary, OSD Summary, MDS Summary, PG Summary and Warnings and Errors sections is collected from the ceph status -f json_pretty command.

Information on the Cluster Status page is updated approximately once per minute, but may vary based on what operations are underway on the cluster. The time since the last update is shown for each summary section. The operator may view the most recent information by refreshing the page.

3.2. Managing Capacity

VSM provides detailed information showing how capacity is utilized in Storage Groups, provides the ability to create new storage groups, provides the ability to inspect and create pools, and provides the ability to inspect RBD images created by Ceph RBD clients.
**Storage Group Status**

The Storage Group Status page provides a summary of capacity utilization for each Storage Group in the cluster.

1) **Storage Groups:** VSM calculates the storage utilization for each Storage Group separately. Used capacity is shown graphically for each Storage Group in the cluster, and includes primary and replica data copies in each Storage Group. To determine the used capacity for each pool in a Storage Group, navigate to the Pool Status page.

2) **Capacity Warning Thresholds:** VSM provides separate near full and full threshold warnings for each Storage Group. The current settings are displayed below the summary graphs. VSM also configures Ceph OSD near full and full settings; current settings are displayed below the summary graphs.

   If any OSD in the Ceph cluster reaches full threshold, the cluster will no long accepts writes until the situation is remedied. Storage Group near full and full thresholds are typically set below Ceph near full and full thresholds so that the operator is warned in advance of any OSD reaching the full threshold.

   When the used capacity of a Storage Group exceeds the near full or full threshold, the graph color will change and a waning message will be displayed in the table below.

   Ceph near full and full status is displayed in the Cluster Status page (OSD Summary section) and OSD Status pages.

3) **Storage Group List:** VSM displays detailed information for each Storage Group:

   - Number of attached pools
• Total capacity, used capacity, and remaining capacity in Storage Group
• The largest amount of used capacity on any one server with used capacity in the Storage Group.
• VSM provides the following warnings:
  o Storage Group used capacity exceeds the Storage Group near full or full thresholds
  o Storage Group available capacity is insufficient to allow the Storage Group to rebalance in the event that the server with the largest amount of used capacity in the Storage Group fails.

Largest Node Capacity Used: When a server fails or becomes unresponsive for an extended period of time, Ceph will create new copies of the data that resided on the failed (or unresponsive) server on the remaining available capacity in the Storage Group. If the Storage Group does not have enough available capacity to absorb the new copies, the Storage Group will become full and will not be able to accept new data. VSM will provide a warning if the Storage Group available capacity is insufficient to allow the Storage Group to rebalance in the event that the server with the largest amount of used capacity in the Storage Group fails.
Managing Storage Groups

The Manage Storage Group page provides a summary of currently defined storage groups and their associated storage classes, and provides the ability to create new storage groups and storage class pairings.

Creating New Storage Groups

New storage groups can be created by clicking the Add Storage Group button on the Manage Storage Groups page. Clicking the Add Storage Group button will open the Create Storage Group dialog:

To create a new storage group, enter the storage group name, storage group friendly name, and associated storage class, and then click on the Create Storage Group button.
Note: The recommended sequence for adding new storage group capacity to the cluster is as follows: First, add the new storage group and storage class as described above. Once the new storage group is created, then add new servers using the Manage Servers page. Each new server to be added must have disks identified as belonging to the new storage groups new storage class in each servers storage manifest file.

Note: VSM will not allow pools to be created in storage groups containing less than three servers. This ensures that data can be distributed across a minimum of three servers, consistent with VSMs minimum replication level of 3 for replicated pools.

Managing Pools

The Manage Pools page provides configuration information for each pool in the cluster, and manages creation of new pools.

VSM supports the creation of pools using either replication or erasure coding to protect against data loss and data availability in the event of a disk or server failure.
For each pool in the cluster, the Manage Pools page displays the following information:

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Pool ID – Ascending in order of creation.</td>
</tr>
<tr>
<td>Primary Storage Group</td>
<td>Identifies the storage group in which the pool is placed. For pools with separate primary and replica storage groups, indicates the storage group in which the primary copy is placed.</td>
</tr>
<tr>
<td>Replica Storage Group</td>
<td>For pools with separate primary and replica storage groups, indicates the storage group in which the replica copy is placed. If the pool is placed in a single storage group, “-” is displayed.</td>
</tr>
<tr>
<td>Placement Group Count</td>
<td>Displays the current number of placement groups assigned to the pool.</td>
</tr>
<tr>
<td>Size</td>
<td>For replicated pools, displays the number of replicas. For erasure coded pools, displays the number of placement groups.</td>
</tr>
<tr>
<td>Quota</td>
<td>Displays the quota (if set) when the pool was created. If no quota was set when the pool was created, “-” is displayed.</td>
</tr>
<tr>
<td>Cache Tier Status</td>
<td>For pools participating in a cache tier, displays the pools tier role (cache or storage) and the storage group to which it is paired. If the pool is not participating in a cache tier, “-” is displayed.</td>
</tr>
<tr>
<td>Erasure Code Status</td>
<td>If the pool is an erasure coded pool, displays the profile name used to create the pool. If the pool is replicated, “-” is displayed.</td>
</tr>
<tr>
<td>Status</td>
<td>“Running” status indicates that the pool is currently available and operational</td>
</tr>
<tr>
<td>Created By</td>
<td>Displays “VSM” if the pool was created by VSM. Displays “Ceph” if the pool was created outside of VSM (e.g., via Ceph command line).</td>
</tr>
<tr>
<td>Tag</td>
<td>Informative tag that may be specified when pool is created.</td>
</tr>
</tbody>
</table>
Creating Replicated Storage Pools

New replicated pools can be created by clicking the Create Replicated Pool button on the Manage Pools page. Clicking the Create Replicated Pool button will open the Create Replicated Pool dialog:

To create a new replicated pool, enter the pool name, and select a storage group where the pool’s primary copy of data will be placed from the Primary Storage Group pull-down menu. If the pool’s primary and secondary replicas are to be placed in the same storage group, leave the Replicated Storage Group pull-down menu set to “Same as Primary”.

The non-primary replica data may be optionally placed in a separate storage group. This feature can be used, for example, to place the primary copy in a Storage Group comprised of SSDs, while placing the replica copies in a Storage Group comprised of 7200 RPM rotating disks; in this configuration, data will be read with very high performance from the primary OSDs placed on SSDs, while will be written with the usual performance of HDDs (because all copies must be written before the write operation is signaled as complete). Select a storage group where the pool’s replica data will be placed from the Replicated Storage Group pull-down menu.

Note: When using separate primary and replicas storage groups, it is required that disks belonging to the Primary Storage Group and disks belonging to the Replica Storage Group be located on separate servers in order to ensure that primary and replica data copies do not reside on the same server.

Specify the replication factor for the pool. The minimum replication factor permitted by VSM is 3. To ensure the data replicas reside on separate servers, VSM will fail to create the storage pool if the number of servers in each of the selected storage group(s) equals or exceeds the replication factor, and will display the following error message:
Optionally specify an informative tag for the pool.

Optionally set a quota for the pool by selecting the “enable Pool Quota” checkbox and specifying the desired pool quota.

Complete the operation by clicking on the “Create Replicated Pool” button.

When a replicated pool is created, VSM sets the placement group count using the following formula:

\[
\text{Target PG Count} = \frac{\text{pg\_count\_factor} \times \text{number OSDs in pool}}{\text{replication factor}}
\]

The parameter `pg_count_factor` is specified in the cluster manifest file; the default value is 100, consistent with Ceph community guidelines.

When additional OSDs are added to a Storage Group, VSM recalculates the target PG count for each affected replicated pool and updates the pool’s PG count when it falls below one half of the target PG count.

**Creating Erasure Coded Storage Pools**

New erasure coded pools can be created by clicking the Create Replicated Pool button on the Manage Pools page. Clicking the Create Replicated Pool button will open the Create Replicated Pool dialog:

To create a new replicated pool, enter the pool name, and select the storage group where the pool’s erasure coded data will be placed from the Storage Group pull-down menu.
Select the erasure code profile from the Storage Group pull-down menu. The pull down menu lists the storage profiles that are specified in the cluster manifest file.

Select the erasure code failure domain from the Storage Group pull-down menu. Select OSD failure domain to place erasure coded across OSDs. Select Zone to place erasure coded data across failure zones. Select Host to place erasure coded data across servers.

Note: The erasure code failure domain must be equal to or larger than the K+M value of the profile. For example, the failure domain for a K=3 M=2 erasure code must be 5 or greater; for a cluster with three servers with each server containing 10 disks, a K=3 M=2 erasure code cannot be placed across three servers (Host failure domain or Zone failure domain); the only valid erasure code failure domain is OSD. To support Host failure domain, 5 servers are required, and to support Zone failure domain, five zones are required.

Optionally specify an informative tag for the pool.

Optionally set a quota for the pool by selecting the “enable Pool Quota” checkbox and specifying the desired pool quota.

Complete the operation by clicking on the “Create Erasure Coded Pool” button.
Creating Cache Tiers

Cache Tiers can be created by clicking the Add Cache Tier button on the Manage Pools page. Clicking the Add Cache Tier button will open the Add Cache Tier dialog:

To create a cache tier, select a pool to act as cache from the Cache Tier Pool pull-down menu, and select and pool to act as the storage tier from the Storage Tier Pool pull-down menu. VSM will only display pools that are not currently participating in existing cache tiers. VSM will fail to create the cache tier if the same pools are selected for the cache and storage tiers, and will display the following message:

Select the cache mode from the Cache Mode pull-down menu:

Writeback Mode: Data is written to the cache tier and the write is signaled as complete. Written data older than the minimum flush age is eventually flushed from the cache tier and written to the storage tier. When a read cannot be
completed from the cache tier, the data is migrated to the cache tier, and the read is then completed. This mode is preferred for mutable data (e.g., RBD data, photo/video editing, transactional data, etc.).

Read-only Mode: Data is written to the backing tier. When data is read, Ceph copies the corresponding object(s) from the storage tier to the cache tier. Stale objects get removed from the cache tier based on the minimum evict age. This approach is suitable for immutable data (e.g., data that will not be modified, such as pictures/videos on a social network, DNA data, X-Ray imaging, etc.). Do not use read-only mode for mutable data.

Check the “FORCE NONEMPTY” checkbox.

Configure the remaining cache tier parameters as required for your application.

Complete the operation by clicking on the “Create Cache Tier” button.

See the Ceph documentation on cache tiering for additional information regarding cache mode, FORCE NONEMPTY, and the cache tier configuration parameters.

3.3. Monitoring Cluster Health

Ceph is generally self-repairing. However, when problems persist, inspecting the status of the Ceph subsystems will help identify and resolve problems.

VSM provides detailed status information for Ceph pools, OSD daemons, Placement Groups, Monitors, MDS daemons, and RBD images created by the Ceph RBD client.

Pool Status

The Pool Status page provides configuration, capacity utilization, object state summary, and performance information for each pool in the cluster.
Scrolling left, the Pool Status page provides configuration information similar to that found on the Manage Pools page:

For each pool in the cluster, the left hand side of the page displays the pool name, the informative tag, the Storage Group in which the pool was created, the replication factor, the current number of Placement Groups (PGs) and Placement Group Pages, status, and how the pool was created (inside or outside of VSM).
Scrolling right, the Pool Status page displays capacity utilization, object state summary, and performance information for each pool in the cluster.

For each pool in the cluster, the right hand side of the page displays the following information:

- Capacity used by the pool – this includes primary and replica data
- The number of objects and cloned objects in the pool
- The number of degraded objects. Degraded objects have less than the number of replicas specified for the pool. Objects may be temporarily degraded in the event of a disk or server failure; Ceph will attempt to recreate replicas elsewhere in the cluster and this number should return to zero.
- Number of unfound objects. Unfound objects are objects for which all copies of data have been lost. This may occur if failures occur such that all copies of an object are lost or corrupted and Ceph cannot restore data replicas from the remaining data in the cluster.
- The total number of read and write operations, and the total amount of data read from and written to the pool since pool creation.
- The current client read and write bandwidth for the pool, and the current client operations per second for the pool
**OSD Status**

The OSD Status page provides a summary of status for all OSDs in the clusters, as well as detailed information for each OSD in the cluster:

The OSD Summary displays the summary status for all OSDs in the cluster; this is the same information as is displayed in the OSD Summary section of the Cluster Status page.

**OSD List**

The OSD List displays all OSDs in the cluster. For each OSD, the following information is displayed:

- **OSD name**: The OSD name assigned by Ceph when the OSD is created.
- **VSM Status**: The OSD is either attached to a disk (present) or not attached to a disk (removed). The normal operating state is “present”. An OSD will have VSM Status ‘removed” when the OSD has been removed from the cluster using the “Remove OSD” operation on the Manage Devices page; this is typically done when replacing a failed or problematic disk.
- **OSD State**: Ceph-reported OSD status - see below for a detailed discussion of OSD Status.
- **CRUSH weight**: In VSM 1.0, all disks in a Storage Group are assumed to have the same capacity, and so all OSDs in the Storage Group are assigned a weight of 1.
- **Total, used, and available capacity**: This information enables the operator to inspect the capacity utilization of each OSD.
- **Server**: The server on which the OSD is located.
- **Storage Group**: The Storage Group in which the OSD resides.
- **Zone**: The zone (failure domain) in which the OSD resides.
**OSD Status**

An OSD’s status is either up and running (up) or down and not running (down), and is either in the cluster (in) or out of the cluster (out).

Under normal operation, all OSDs should be up and in.

A problem is indicated when the OSD Summary shows that the number of OSDs that are in is greater than the number of OSDs that is up; inspect the OSD List to identify OSD daemons that are not running.

Under certain error conditions, Ceph will place an OSD out of the cluster and stop the OSD daemon (out-down-autoout). The operator can attempt to restart OSDs that are down using the “Restart OSD” operation on the Manage Devices page.

An OSD that cannot be restarted, or that is repeatedly set to out-down-autoout by Ceph may indicate that the associated disk has failed or is failing, and should be replaced. An OSD disk can be replaced by using the “Remove OSD” and “Restore OSD” operations on the Manage Devices page.

**Monitor Status**

The Monitor Status page provides a summary of status for all Monitors in the cluster, as well as detailed information for each Monitor in the cluster:

The Monitor Summary section displays the summary status for all Monitors in the cluster; this is the same information as is displayed in the Monitor Summary section of the Cluster Status page.

The Monitor List displays all OSDs in the cluster. For each OSD, the following information is displayed:
• Name: Monitor Rank – Monitor rank is assigned starting at zero in order from lowest [IP address:port number] to highest [IP address:port number]. Monitor rank is recalculated whenever monitors are added or removed.

• Address: The IP address associated with the monitor.

• Health: The overall health status of the monitor

• Detail: Detailed health messages as provided by `ceph status -f json_pretty`

• Skew: Offset of local server time relative to primary monitor (leader). In the example above, note that monitor 3 is the leader (see Monitor Summary – the first monitor listed in the quorum is leader); skew for monitor 3 is reported as zero.

• Latency: Round trip delay between this monitor and the primary monitor (leader). In the example above, note that monitor 3 is the leader (see Monitor Summary – the first monitor listed in the quorum is leader); latency for monitor 3 is reported as zero.

• System (OS) disk capacity use for the server on which the monitor daemon is running. In VSM 1.0, you can determine which server each monitor is on by matching the monitor address to the management address on the Manage Servers page.

If monitor health warnings persist, it may indicate a more serious problem in the cluster.

**PG Status**

The PG Status page provides summary information for all Placement Groups in the cluster:

![PG Status](image)

**Object Summary**

In Ceph, the unit of data storage is called an object. Ceph places copies of objects on different disks and servers in the cluster; the number of copies is determined by the replication factor of the pool that the object (data) is placed in.

In order to track object placement and object metadata more efficiently, Ceph aggregates objects into Placement Groups, and maps Placement groups across a series of OSDs.
When a disk or server fails or cannot be accessed by the cluster, copies of objects are not accessible; the affected objects are referred to as “degraded”. Under normal operation, it is possible to have some small percentage of objects degraded. This may be due either due to disk or server failure, or may occur if a server is stopped for maintenance and is temporarily inaccessible by the cluster.

Under certain conditions, Ceph may be unable to find any valid copies of an object; when this occurs, the object is referred to as “unfound”. This condition may be temporary or may be permanent. I/O to lost objects will block and wait for the cluster to find a valid copy of the object.

The Object Summary section displays the following information:

- Degraded Objects: The total number of degraded objects in the cluster
- Degraded Total: The number of degraded xxx in the cluster
- Degraded Ratio: The ratio of degraded Placeman Groups to Total Placement Groups
- Unfound Objects: The total number of degraded objects in the cluster
- Unfound Total: The number of degraded Placement Groups in the cluster
- Unfound Ratio: The ratio of degraded Placeman Groups to Total Placement Groups

**PG Summary**

The PG summary displays the number of placement groups (PGs) in the cluster and the count of PGs in various state combinations – note that a placement group may be in more than one state at the same time.

Under normal operation, the majority of PGs should be “active+clean”. This indicates that the majority of PGs are processing requests (“active”), and that all object replicas exist and are up to date (“clean”).

A small number of objects may report “active+clean+scrubbing”. This indicates that Ceph is in the process of verifying the consistence of object replicas, which it performs on a continuous basis in order to detect and correct disk data errors.

It is normal for placement groups to enter states like “degraded” (object replicas are missing) or “peering” following a disk, server, or network failure – Ceph will automatically attempt to self-heal following a failure. However if this condition persists, it may indicate a more serious problem in the cluster.

See section 0 for a brief description of all possible individual PG states.

**Performance Summary**

The Performance Summary provides the following performance statistics for the entire cluster: total reads and writes per second, and total operations per second.

**Capacity Summary**
The Capacity Summary provides the following capacity statistics for the entire cluster: total data capacity used (primary copy), total capacity used (primary + replica data), capacity available, and total cluster capacity.

**MDS Status**

The MDS Status page provides summary information for all MDS daemons in the cluster:

VSM 1.0 does not support management of CephFS, however, to allow evaluation of CephFS functionality, VSM enables one MDS daemon in the cluster.

For each MDS in the cluster, the MDS List displays the MSD name, Gid, state, and address.
**RBD Status**

The RBD Status page provides summary information for RBD images created by the Ceph RBD block storage client:

The Object Summary section displays the following information:

- Pool that the RBD image is created in
- RBD Image name (assigned when image is created by RBD client)
- Allocated size (not used capacity) of RBD image
- Number of objects that comprise the image. The RBD client stripes the image across multiple objects so that the image is distributed (striped) across the cluster, improving overall performance.
- Order (size) of the object, where size = 2^order. Default is 22 (4MB).
- Format: Format 1 is original format and is understood by all versions of Ceph, but does not support cloning. Format 2 is the more recent format; it adds support for cloning, and is the version typically used for VM images.

In VSM 1.0, display is limited to 100 RBD images. If the cluster is hosting more than 100 RBD images, the RBD Status page will display the first 100 images reported by:

```
  rbd ls -l {pool name} --format json --pretty-format
```
3.4. Managing Servers

Managing Servers

The Manage Servers page provides configuration information for each server detected by VSM. It also supports stopping and starting servers, adding and removing monitors, and adding and removing servers.

1) The Manage Servers page supports stopping and starting servers, adding and removing monitors, and adding and removing servers.

2) The Cluster Server List displays all servers that have been reported to the VSM Controller by VSM Agents. It displays the following information for each server:
   - Server name
   - Management IP address, Ceph public address, and Ceph cluster address
   - Number of OSDs on the server
   - Whether a monitor is running on the servers
   - The zone in which the server is placed
   - The server status

Server Status

Each server will display one of the following:

- **Available**: The server has been reported to VSM controller by a VSM Agent, and is available to be added to the cluster
- **Active**: The server has been added to the cluster; an OSD has been added to the cluster and started for each data drive on the server
- **Stopped**: All OSDs attached to data drives on the server have been stopped (but not removed from the cluster) and the cluster has been placed in “noout”. The server can be turned off for maintenance without causing the cluster to rebalance
- **Unavailable**: The server is currently not reachable by the VSM controller – the server may have experienced a failure, been stopped, disconnected from the network, or physically removed from the storage system
• **Stopping**: The server is transitioning from *active* to *stopped* status as a result of a “stop server” operation.

• **Starting**: The server is transitioning from *stopped* to *active* status as a result of a “start server” operation

• **Removing**: The server is being removed from the cluster as the result of a “remove server” operation

• **Adding**: The server is transitioning from *available* to *active* status as the result of an “add server” operation.

### Add and Remove Servers

Servers that have not been added to the cluster may be added to the cluster by clicking on the “Add Server” button on the Manage Servers page, which will open the “Add Server” dialog:

![Add Server Dialog](image)

All servers with *available* status are eligible to be added to the cluster and will be listed. In order for servers to appear in the list of available servers, VSM Agent software must be properly installed, configured, and activated – see section 2.2 for an overview of this process.

Select the servers to be added to the cluster and click on the “Add Server” button on the dialog box to confirm the operation.

OSDs will be created, started, and placed in the appropriate Storage Group for each of the data drives on each added server, as specified by the server’s server manifest file.

A monitor daemon will be started if so specified in the server’s server manifest file.

The server status will transition to *active* status once all operation are complete. The time required to complete the operation is a function of the number of OSDs to be added and the load on the cluster – estimate approximately 30 seconds per OSD.
A server that has been added to the cluster may be removed from the cluster by clicking the “Remove Server” button on the Manage Servers page, which will open the “Remove Servers” dialog:

![Remove Servers dialog](image)

Servers with *active* or *stopped* status, or that have previously been had *active* or *stopped* status and currently have *unavailable* status are eligible to be removed from the cluster and will be listed. Select the servers to be removed from the cluster and click on the “Remove Server” button on the dialog box to confirm the operation.

OSD and monitor daemons associated with the server will be stopped (if the server status is active) and removed from the cluster.

The server will transition to “available” status, or will revert to “unavailable” if the server is turned off or disconnected from the storage system. The time required to complete the operation is a function of the number of OSDs to be removed and the load on the cluster – estimate approximately 15 seconds per OSD.

It may take up to an additional minute for the results of the “Add Server” and “Remove Server” operations to be reflected on the Cluster Status page.

**NOTE**

To ensure stable operation, allow “Add Server” or “Remove Server” operations to complete before initiating other VSM operations.

Verify the count of OSDs, OSDs *up*, OSDs *in*, and monitors on the Cluster Status page to confirm that “Add Server” and “Remove Server” operations have completed as expected.
Add and Remove Monitors

A monitor daemon may be started on a server that is not currently running a monitor daemon by clicking the “Add Monitor” button on the manage Servers page, which will open the “Add Monitor” Dialog”

![Add Monitors Dialog](image)

Servers that are not currently running a monitor daemon are eligible to have a monitor daemon added and will be listed. Select the servers on which to add a monitor and click on the “Add Monitor” button on the dialog box to confirm the operation.

To ensure continued stable cluster operation, it is recommended to maintain an odd number of monitors in the cluster. If adding the selected monitors will result in an even number of monitors, a warning dialog will be opened; click on OK to continue the operation, and return the monitor count to an odd number in the near future to ensure continues stable cluster operation.

A monitor daemon may be removed on a server that is currently running a monitor daemon by clicking the “Remove Monitor” button on the manage Servers page, which will open the “Remove Monitor” Dialog”
Servers that are currently running a monitor daemon are eligible to have a monitor daemon removed and will be listed. Select the servers on which to remove a monitor and click on the “Remove Monitor” button on the dialog box to confirm the operation.

To ensure stable cluster operation in the event of the failure of a server hosting a monitor, it is recommended to maintain a minimum of three monitors in the cluster. If removing the selected monitors will result in less than three monitors in the cluster, a warning dialog will be opened, and the operation will not be permitted to continue; click the Cancel button to close the dialog. Add monitors before removing monitors so that the monitor count is never less than three.

To ensure continued stable cluster operation, it is recommended to maintain an odd number of monitors in the cluster. If removing the selected monitors will result in an even number of monitors, a warning dialog will be opened; click on OK to continue the operation, and return the monitor count to an odd number in the near future to ensure continued stable cluster operation.

To remove a monitor on a server that has failed or been removed from the system, use the “remove server” operation.

It may take up to an additional minute for the results of the “Add Monitor” and “Remove Monitor” operations to be reflected on the Cluster Status page.

NOTE
To ensure stable operation, allow “Add Monitor” or “Remove Monitor” operations to complete before initiating other VSM operations.

Verify the count of monitors on the Cluster Status page to confirm that “Add Monitor” and “Remove Monitor” operations have completed as expected.
Stop and Start Servers

Servers that are currently active (i.e. host storage with OSDs that are up and in) may be stopped by clicking on the “Stop Server” button, which will open the “Stop Server” dialog:

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Management Address</th>
<th>Ceph Public Address</th>
<th>Ceph Cluster Address</th>
<th>OSDs (Data Drives)</th>
<th>Monitor</th>
<th>Zone</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>htf1gceph2</td>
<td>192.168.22.42</td>
<td>192.168.102.42</td>
<td>10.0.13.42</td>
<td>16</td>
<td>yes</td>
<td>zone_one</td>
<td>Active</td>
</tr>
<tr>
<td>2</td>
<td>htf1gceph1</td>
<td>192.168.21.41</td>
<td>192.168.101.41</td>
<td>10.0.12.41</td>
<td>16</td>
<td>yes</td>
<td>zone_one</td>
<td>Active</td>
</tr>
<tr>
<td>3</td>
<td>htf1gceph3</td>
<td>192.168.23.43</td>
<td>192.168.103.43</td>
<td>10.0.14.43</td>
<td>16</td>
<td>yes</td>
<td>zone_one</td>
<td>Active</td>
</tr>
<tr>
<td>4</td>
<td>htf1gceph4</td>
<td>192.168.20.44</td>
<td>192.168.100.44</td>
<td>10.0.12.44</td>
<td>16</td>
<td>yes</td>
<td>zone_one</td>
<td>Active</td>
</tr>
<tr>
<td>5</td>
<td>htf1gceph5</td>
<td>192.168.21.45</td>
<td>192.168.101.45</td>
<td>10.0.13.45</td>
<td>16</td>
<td>yes</td>
<td>zone_one</td>
<td>Active</td>
</tr>
<tr>
<td>6</td>
<td>htf1gceph6b</td>
<td>192.168.21.46</td>
<td>192.168.103.46</td>
<td>10.0.13.46</td>
<td>no</td>
<td>no</td>
<td>zone_one</td>
<td>Active</td>
</tr>
</tbody>
</table>

Servers that are currently active (running OSD or monitor daemons) are eligible to be stopped. Select the servers to be stopped and click on the “Remove Monitor” button on the dialog box to confirm the operation.

The “Stop Server” operation stops each OSD hosted by the selected server and will also stop a monitor daemon if one is running. VSM directs Ceph to put “noout” in effect, which directs Ceph to suspend rebalancing. In this way, a server can be stopped for maintenance without incurring degraded performance due to unwanted cluster rebalance.

NOTE

While “noout” is in effect, the cluster will not rebalance failed drives or servers, and objects may become degraded because object replicas on the stopped server become out of date. It is therefore recommended to complete maintenance and return all stopped servers to active status (using the “Start Server” operation) in the shortest possible interval in order to restore full data replication.

Servers that are stopped may be restarted by clicking on the “Start Server” button, which will open the “Start Server” dialog:
Servers that are currently *stopped* are eligible to be started. Select the servers to be started and click on the “Start Servers” button on the dialog box to confirm the operation.

The “Start Server” operation starts each OSD hosted by the selected server. It also starts a monitor daemon if the server is so configured. If no other servers remain that are stopped, VSM will direct Ceph to cancel “noout”.

It may take up to an additional minute for the results of the “Stop Server” and “Start Server” operations to be reflected on the Cluster Status page.

**NOTE**

To ensure stable operation, allow “Stop Server” and “Start Server” operations to complete before initiating other VSM operations.

Verify the count of down OSDs on the Cluster Status page to confirm that “Stop Server” and “Start Server” operations have completed as expected.

---

3.5. Managing Storage Devices

**Manage Devices**

The Manage Devices page provides configuration information for each data disk identified by VSM. It also supports restarting stopped OSDs, and replacing failed or problematic disks.
1) The Manage Servers page supports restarting stopped OSDs. Removing OSDs, and Restoring OSDs

2) The Device List displays the OSDs that have created for all of the disks that VSM has identified. It displays the following information for each OSD:

- **OSD name**
- **OSD Status** – Present or Removed
- **OSD State** – See section 3.3 - OSD Status for a description of OSD state
- **OSD weight** – Ceph uses OSD weight to control relative data distribution across disks and servers in the cluster. In VSM 1.0, all drives (or drive partitions) in a Storage Group are expected to have the same capacity, and so OSD weight is set to 1.
- **Server** – The server on which the OSD and associated data disk resides.
- **Storage Group** – The Storage Group to which the OSD and associated data disk belong.
- **Zone** – The Zone (failure domain) to which the OSD and associated data disk belong.
- **Data Device Path** – The location specified in the server manifest where the data drive (or drive partition) is located on the host server. It is recommended that the location of the data disk (or disk partition) be specified by path to ensure that drive paths are not changed in the event that a drive fails or is removed from the system.
- **Data Device Status** – VSM periodically verifies that the data drive path exists, indicating OK if found present and FAILED if found missing. This is useful for identifying removed or completely failed drives.
- **Drive total capacity, used capacity, and available (remaining) capacity.**
- **Journal Device Path** – The location specified in the server manifest where the data drives, corresponding journal drive (or drive partition) is located on the host server. It is recommended that the location of the journal disk (or disk partition) be specified by path to ensure that drive paths are not changed in the event that a drive fails or is removed from the system.
• Journal Device Status – VSM periodically verifies that the journal drive path exists, signaling OK if found present and FAILED if not found. This is useful for identifying removed or completely failed drives.

**Restart OSDs**

Under certain error conditions, Ceph will place an OSD out of the cluster and stop the OSD daemon (out-down-autoout). The operator can attempt to restart OSDs that are down using the “Restart OSD” operation on the Manage Devices page.

The operator checks the checkbox for each OSD to be removed in the Devices List, and initiates the restart operation by Clicking on the “Restart OSDs” button, which opens the “Restart OSDs” dialog:

![Warning! Restart Osds](image)

The operator confirms the operation by clicking on “Restart OSDs” in the dialog box. VSM then displays the waiting icon until it has attempted to start all selected OSDs:

![Restarting ...](image)

When the operation is complete, inspection of the Device List will show which OSDs were successfully restarted.

An OSD that cannot be restarted, or that is repeatedly set to out-down-autoout by Ceph may indicate that the associated disk has failed or is failing, and should be replaced. An OSD disk can be replaced by using the “Remove OSD” and “Restore OSD” operations on the Manage Devices page.

**Remove OSDs**

A failed or failing OSD can be replaced by using the “Remove OSD” and “Restore OSD” operations

The operator starts the disk replacement process by checking the checkbox for each OSD to be removed in the Devices List, and initiates the remove operation by Clicking on the “Remove OSDs” button, which opens the “Remove OSDs” dialog:

![Warning! Remove Osds](image)
The operator confirms the operation by clicking on “Remove OSDs” in the dialog box. VSM then displays the waiting icon until it has removed all selected OSDs:

![Image of waiting icon]

When the operation is complete, inspection of the Device List will show that the VSM status of removed OSDs have been changed to Removed. The OSD name of removed OSDs will be changed to “OSD.xx” until a new OSD is assigned to the device path using the Restore OSD operation.

**Restore OSDs**

Once OSDs have been removed, the corresponding physical disks can be replaced using the process recommended for the server and disk controller hardware.

*Note: The replaced drive must be located on the same path as originally specified in the server’s server manifest file.*

The operator completes the disk replacement process by checking the checkbox for each OSD to be restored in the Devices List, and initiates the restore operation by clicking on the “Restore OSDs” button, which opens the “Restore OSDs” dialog:

![Image of restore OSDs dialog]

The operator confirms the operation by clicking on “Remove OSDs” in the dialog box. VSM then displays the waiting icon until it has restored all selected OSDs:

![Image of waiting icon]

When the operation is complete, inspection of the Device List will show that the VSM status of restored OSDs have been changed to Present. The OSD name of restored OSDs will be changed to show the OSD number that has been assigned by Ceph. The OSD Status should be **up and in**.
3.6. Working with OpenStack

VSM can be optionally connected to an OpenStack Nova server; when so connected, VSM is able to “present” Ceph pools to the OpenStack Cinder multi-backend service. In order to connect VSM to OpenStack, VSM requires a pre-configured SSH connection be established between the server hosting the VSM controller and a server hosting an OpenStack Nova controller.

**Accessing OpenStack**

When VSM is not connected to an OpenStack Nova controller, the Manage OpenStack Access page displays the “Add OpenStack Nova Controller” button. The operator initiated connection by clicking on the button:

![Image of Add OpenStack Nova Controller dialog]

Clicking on the “Add OpenStack Nova Controller” button opens the “Add OpenStack Nova Controller” dialog:

![Image of Add OpenStack Nova Controller dialog]

The operator enters the IP address of a server hosting the OpenStack Nova controller, and clicks on the “Add IP” button.

**NOTE**

To ensure stable operation, allow the “Remove OSD” or “Restore OSD” operations to complete before initiating other VSM operations.

Verify the count of OSDs, OSDs up, and OSDs in on the Cluster Status page to confirm that “Remove OSD” and “Restore OSD” operations have completed as expected.
Once a connection is successfully established, the Manage OpenStack Access page displays the server IP address and “reachable” connection status:

In VSM 1.0, VSM can connect to only one OpenStack controller.

**Manage RBD Pools**

The Manage RBD pools page displays all pools in the cluster and allows pools to be presented to the Cinder block storage multi-backend service of a connected OpenStack cluster.

In order to present pools to OpenStack, VSM must first be connected to the server hosting the OpenStack Nova Controller using the “Add OpenStack Nova Controller” operation on the Manage OpenStack Access page.

The Manage RBD Pools page lists all pools in the cluster; for each pool it displays the pools name, Storage Group in which it resides, placement group count, replication factor, status, creation source, tag, and attach status. Attach Status indicates whether the pool has been presented (attached) to OpenStack Cinder backend service; pools that are not attached to OpenStack Cinder are shown as “no” in the Attach Status column; pools that are attached to OpenStack Cinder are shown as “success”.
The operator may present an unattached pool to OpenStack Cinder backend service by clicking on the “Preset Pools” button on the Manage RBD Pools page, which opens the “Present Pools” dialog box:

The Present RBD Pools dialog box lists all pools available to be presented to OpenStack Cinder (e.g. pools not already attached). Pools to be presented are selected by clicking the corresponding checkbox and clicking the “Present Pools” button on the dialog box to initiate the operation.

When the operation is complete, the Manage RBD Pools page displays the updated status.

3.7. Managing VSM

Manage VSM Users

The Manage Users page lists all VSM users, and enables VSM users to be added and removed, and enables VSM user passwords to be changed:

A new user can be added by clicking the New User button, which opens the New User dialog:
To create a new user, enter the new user’s user name and password. Note that VSM requires the password to contain at least eight characters and to include at least one uppercase character, one lowercase character, one numeric character, and one special character. Enter the password a second time to confirm and then click on the “Create User” button in the dialog.

A user’s password can be changed by clicking on the “Change Password” button for the selected user, which opens the “Update User” dialog:

Enter the new password, following the requirements listed above. Enter the new password a second time to confirm and then click on the “Update User” button in the dialog.

A user can be deleted by selecting more and then selecting “Delete User” from the pull-down:

Clicking on “Delete User” will execute the operation.

The default “admin” user cannot be deleted.
Manage VSM Configuration

The VSM Settings page allows selected VSM configuration parameters to be viewed and modified:
4. Troubleshooting

4.1. Stopping Servers without Rebalancing

Components of the system (e.g. servers, networking) may periodically require maintenance, or you may need to resolve a problem that affects a failure domain (i.e. zone).

If you shut down a server in the Ceph cluster or disconnect the server from the cluster, Ceph will determine that OSDs have become unresponsive, and will attempt to repair (rebalance) the cluster by creating new copies of the data that resided on the unresponsive OSDs. This may result in substantial data movement and may reduce cluster performance.

To prevent rebalancing in the case where server disconnection is expected to be temporary, it is recommended to stop run the Stop Server operation (located on the Manage Servers page) on each of the affected servers. When the Stop Server operation is executed, VSM sets Ceph to “noout”, and then stops each OSDs on the selected server. Setting the cluster to “noout” tells Ceph to not rebalance the cluster if OSDs become unresponsive. Setting “noout” affects the entire cluster, and remains in effect while any servers are in the “stopped” state.

When maintenance is complete, stopped servers are restarted using the Start Server operation on the Manage Servers page; VSM restarts the OSDs on the affected servers and clears Ceph from “noout”.

Because “noout” affects the entire cluster, it is recommended to limit the time that servers are stopped to the minimum required to perform maintenance or resolve other problems.

4.2. OSDs Not Up and In

OSDs may become responsive due to transient issues such as an overloaded OSD, or may become unresponsive because an associated data or journal disk has failed or is failing. When Ceph detects that an OSD has become unresponsive, it will place the OSD down and out – this is referred to as “auto-out”.

The operator can identify OSDs that are down an out by inspecting the OSD Summary section of the Cluster Status or OSD Status pages – the number of OSDs up and in should equal the total number of OSDs. In the following example, the total number of OSDs is 96, while the number of OSDs up and in is 94, indicating that two OSDs are down and out:
The operator can identify OSDs that have been placed in “autoout” on the Device Management or OSD Status pages by sorting by OSD State.

The operator can attempt to restart OSDs that have been placed “autoout” using the “restart OSD” operation on the Manage Devices page. OSDs that will not restart or that repeated return to “autoout” may indicate a failure in the associated data or journal disk.

A set of auto out OSDs that share the same journal SSD may indicate that the journal disk is the common source of failure:

VSM periodically checks to verify the presence of the data and journal devices. An “Error” Device Status indicates that the OS no longer thinks the drive is present, and may be indicative of a drive failure, cabling failure, or physical connection issue:
4.3. OSDs Near Full or Full

If the capacity utilization of any one OSD exceeds the Ceph full threshold (default = 90%), ceph will stop accepting writes until capacity is added to the cluster. OSDs at near full (default 75%) and full threshold are reported in the Cluster Health Summary on the Cluster Status page:

![Cluster Health Summary](image)

Cluster will stop accepting writes when OSD exceeds full ratio

4.4. Replacing Failed Drives

Replacing Failed Data Drives

See section 3.5 – Remove OSDs and Restore OSDs – for a description of how to replace OSD data drives.

Replacing Failed Journal Drives

To replace a failed journal drive, first remove all OSDs using the journal drive, then replace the journal drive, ensuring that the journal drive (or drive partitions) is restored to the same path (or paths) as originally specified in the server manifest file, and finally, restore all of the removed OSDs.

4.5. Monitor Clock Skew

High skew indicates that the clocks the servers hosting the monitors are out of synchronization. Clocks can be synchronized using the NTP protocol.
5. Technical Reference

5.1. Data Sources

The following table shows the Ceph commands that each of VSM pages use to collect and display cluster information, and their respective update intervals.

<table>
<thead>
<tr>
<th>Page</th>
<th>Source – Ceph Command</th>
<th>Update Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Status</td>
<td>ceph status -f json pretty&lt;br&gt;ceph status -f json pretty&lt;br&gt;ceph health</td>
<td>1 minute</td>
</tr>
<tr>
<td>Warnings and Errors Health</td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td>Storage Group Status</td>
<td>ceph pg dump osds -f json-pretty</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Pool Status</td>
<td>ceph pg dump pools -f json-pretty&lt;br&gt;ceph osd dump -f json-pretty&lt;br&gt;ceph osd pool stats -f json-pretty</td>
<td>1 minute</td>
</tr>
<tr>
<td>Object and R/W stats</td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td>Size &amp; PG Count</td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td>Client I/O rates</td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td>OSD Status</td>
<td>ceph status -f json pretty&lt;br&gt;ceph osd dump -f json-pretty&lt;br&gt;ceph osd tree -f json-pretty&lt;br&gt;ceph pg dump osds -f json-pretty</td>
<td>1 minute</td>
</tr>
<tr>
<td>Summary data</td>
<td></td>
<td>10 minutes</td>
</tr>
<tr>
<td>OSD State</td>
<td></td>
<td>10 minutes</td>
</tr>
<tr>
<td>CRUSH State</td>
<td></td>
<td>10 minutes</td>
</tr>
<tr>
<td>Capacity stats</td>
<td></td>
<td>10 minutes</td>
</tr>
<tr>
<td>Monitor Status</td>
<td>ceph status -f json pretty</td>
<td>1 minute</td>
</tr>
<tr>
<td>PG Status</td>
<td>ceph status -f json pretty</td>
<td>1 minute</td>
</tr>
<tr>
<td>Summary data</td>
<td></td>
<td>1 minute</td>
</tr>
<tr>
<td>RBD Status</td>
<td>rbd ls -l (pool name) --format json --pretty-format</td>
<td>30 minutes</td>
</tr>
<tr>
<td>MDS Status</td>
<td>ceph mds dump -f json-pretty</td>
<td>1 minute</td>
</tr>
</tbody>
</table>
5.2. PG States

The following table summarizes the possible Ceph PG states:

<table>
<thead>
<tr>
<th>PG State</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating</td>
<td>Ceph is still creating the placement group.</td>
</tr>
<tr>
<td>Active</td>
<td>Ceph will process requests to the placement group.</td>
</tr>
<tr>
<td>Clean</td>
<td>Ceph replicated all objects in the placement group the correct number of times.</td>
</tr>
<tr>
<td>Down</td>
<td>A replica with necessary data is down, so the placement group is offline.</td>
</tr>
<tr>
<td>Replay</td>
<td>The placement group is waiting for clients to replay operations after an OSD crashed.</td>
</tr>
<tr>
<td>Splitting</td>
<td>Ceph is splitting the placement group into multiple placement groups.</td>
</tr>
<tr>
<td>Scrubbing</td>
<td>Ceph is checking the placement group for inconsistencies.</td>
</tr>
<tr>
<td>Degraded</td>
<td>Ceph has not replicated some objects in the placement group the correct number of times yet.</td>
</tr>
<tr>
<td>Inconsistent</td>
<td>Ceph detects inconsistencies in the one or more replicas of an object in the placement group (e.g. objects are the wrong size, objects are missing from one replica after recovery finished, etc.).</td>
</tr>
<tr>
<td>Peering</td>
<td>The placement group is undergoing the peering process.</td>
</tr>
<tr>
<td>Repair</td>
<td>Ceph is checking the placement group and repairing any inconsistencies it finds (if possible).</td>
</tr>
<tr>
<td>Recovering</td>
<td>Ceph is migrating/synchronizing objects and their replicas.</td>
</tr>
<tr>
<td>Backfill</td>
<td>Ceph is scanning and synchronizing the entire contents of a placement group instead of inferring what contents need to be synchronized from the logs of recent operations. Backfill is a special case of recovery.</td>
</tr>
<tr>
<td>Wait-backfill</td>
<td>The placement group is waiting in line to start backfill.</td>
</tr>
<tr>
<td>Backfill-toofull</td>
<td>A backfill operation is waiting because the destination OSD is over its full ratio.</td>
</tr>
<tr>
<td>Incomplete</td>
<td>Ceph detects that a placement group is missing a necessary period of history from its log. If you see this state, report a bug, and try to start any failed OSDs that may contain the needed information.</td>
</tr>
<tr>
<td>Stale</td>
<td>The placement group is in an unknown state - the monitors have not received an update for it since the placement group mapping changed.</td>
</tr>
<tr>
<td>Remapped</td>
<td>The placement group is temporarily mapped to a different set of OSDs from what CRUSH specified.</td>
</tr>
</tbody>
</table>