Virtual Storage Manager for Ceph

Version: 1.1

Source: 2015-04

Keywords: Ceph, Virtual Storage Management

Supported Combo:

```
OS:         CentOS 6.5 (Basic Server)
Ceph:       Firefly
OpenStack:  Havana/Icehouse

(Other combos might also be working, but we didn't try yet.)
```

Preparation

Before you get ready to install VSM, you should prepare your environment. ***VSM CANNOT manage Ceph Cluster not created by it***. The sections here are helpful for understanding the deployment concepts.

Note: For a Ceph cluster created and managed by VSM you need to prepare at least three storage nodes plus a VSM controller node. VSM requires a minimum of three Ceph storage nodes (physical or virtual) before it will create a Ceph cluster.

Roles

There are two roles for the nodes (servers) on your VSM cresated Ceph cluster.

Controller Node

The controller node is used to run mariadb, rabbitmq, web ui services for the VSM cluster.

Storage Node (a.k.a Agent Node)

The storage or agent node is used to run the vsm-agent service which manages the Ceph and physical storage resources. These nodes are the Ceph storage and monitor nodes.

Network

There are three kinds of networks defined in VSM, and the three networks can all be the same network or separate networks or subnets. VSM does not support split subnets - e.g. two or more different subnets that together make up the management network, or the ceph public network. or the ceph public network.

Management Network

Management Network is used to manage the VSM cluster, and interchanges VSM mangement data between vsm controller and agents.

Ceph Public Network

Ceph Public Network is used to serve IO operations between ceph nodes and clients.

Ceph Cluster Network

Ceph Cluster Network is used to interchange data between ceph nodes like Monitors and OSDs for replication and rebalancing.

Recommendations

Controller node should have connectivity to:

```
>     Management Network
```

Storage Node should have connectivity to:

```
>     Management Network
>     Ceph Public Network
>     Ceph Cluster Network
```

Sample 1

Controller node contains the networks listed below:

```
>     192.168.123.0/24
```

Storage node contains networks below:

```
>     192.168.123.0/24
>     192.168.124.0/24
>     192.168.125.0/24
```

Then we may assign these networks as below:

```
>     Management network: 192.168.123.0/24
>     Ceph public netwok: 192.168.124.0/24
>     Ceph cluster network: 192.168.125.0/24
```

The configuration for VSM in the `cluster.manifest` file should be:

```
>     [management_addr]
>     192.168.123.0/24
>
>     [ceph_public_addr]
>     192.168.124.0/24
>
>     [ceph_cluster_addr]
>     192.168.125.0/24
```

cluster.manifest: Do not worry about this file right now, we will elaborate it later in storage node setup step.

Sample 2

But how about when all the nodes just have two NICs. Such as a controller node and storage node having:

```
>     192.168.123.0/24
>     192.168.124.0/24
```

We can assign these two networks as below:

```
>     Management network: 192.168.123.0/24
>     Ceph public network: 192.168.124.0/24
>     Ceph cluster network: 192.168.123.0/24
```

The configuration for VSM in `cluster.manifest` file would then be:

```
>     [management_addr]
>     192.168.123.0/24
>
>     [ceph_public_addr]
>     192.168.124.0/24
```

```
>
>     [ceph_cluster_addr]
>     192.168.123.0/24
```

Sample 3

It's quite common to have just one NIC in demo environment, then all nodes just have:

```
>   192.168.123.0/24
```

We may assign this network as below:

```
>   Management network: 192.168.123.0/24
>   Ceph public network: 192.168.123.0/24
>   Ceph cluster network: 192.168.123.0/24
```

So all of the three VSM networks use the same subnet, The configurations in `cluster.manifest` file would then be:

```
>     [management_addr]
>     192.168.123.0/24
>
>     [ceph_public_addr]
>     192.168.123.0/24
>
>     [ceph_cluster_addr]
>     192.168.123.0/24
```

Operating System

We have done our development and testing based on a CentOS 6.5 Linux system. For successful installation of VSM and the Ceph cluster it will create, it's best to install system with **CentOS-6.5 Basic Server**.

After installation of a clean CentOS 6.5 Basic Server operating system, do not run:

```
>     yum update
```

Otherwise you may get conflicts between yum packages when you install VSM.

Automatic Deployment

Starting with VSM 1.1, an automatic deployment tool is provided which can simplify the deployment. This tool is still in development, so your feedback and Jira reports of any problens are very welcome.

This section will describe how to use the tool to conduct automation.

1. Firstly, a VSM binary release package should be acquired. It may be downloaded from binary repository, or built from source (see Build VSM). Then unpack the release package, the folder structure looks as following:

> ```
> .
> ├── CHANGELOG
> ├── hostrc
> ├── INSTALL.md
> ├── install.sh
> ├── LICENSE
> ├── manifest
> │   ├── cluster.manifest.sample
> │   └── server.manifest.sample
> ├── NOTICE
> ├── README
> └── vsmrepo
>     ├── python-vsmclient-2015.03.10-1.1.el6.noarch.rpm
>     ├── repodata
> ```

```
    ├── vsm-2015.03.10-1.1.el6.noarch.rpm
    ├── vsm-dashboard-2015.03.10-1.1.el6.noarch.rpm
    └── vsm-deploy-2015.03.10-1.1.el6.x86_64.rpm
```

2. Changing the *hostrc* file, set the *storageiplist* and the *controller_ip*, the ip addresses in storage*ip*list is delimitered by space, and all ip addresses are used in management subnet. e.g.:

   ```
   storage_ip_list="192.168.123.21 192.168.123.22 192.168.123.23"
   controller_ip="192.168.123.10"
   ```

3. Under the *manifest* folder, you should create the folders named by the management ip of the controller and storage nodes, and then the structure looks as follows:

   ```
   .
   ├── 192.168.123.10
   ├── 192.168.123.21
   ├── 192.168.123.22
   ├── 192.168.123.23
   ├── cluster.manifest.sample
   └── server.manifest.sample
   ```

4. Copy the *cluster.manifest.sample* to the folder named by the management ip of controller node, then change the filename to *cluster.manifest* and edit it as required, refer Setup Controller Node for details.

5. Copy the *server.manifest.sample* to the folders named by the management ip of storage nodes, then change the filename to *server.manifest* and edit it as required, refer Setup Storage Node for details.

6. Finally, the manifest folder structure looks as follows:

   ```
   .
   ├── 192.168.123.10
   │   └── cluster.manifest
   ├── 192.168.123.21
   │   └── server.manifest
   ├── 192.168.123.22
   │   └── server.manifest
   ├── 192.168.123.23
   │   └── server.manifest
   ├── cluster.manifest.sample
   └── server.manifest.sample
   ```

7. If you want to upgrade vsm rpm packages, one approach is to build rpm packages separately (see Build RPMs), then put the generated rpm packages into *vsmrepo* folder.

8. Now we are ready to start the automatic procedure by executing this command line:

   ```
   bash +x install.sh -v <version>
   ```

   The version looks like 1.1, 2.0.

9. If execution is blocked at any point, please try to enter "y" and move ahead.

10. If all goes well, you can then login to the VSM Web UI.

Manual Deployment

Install Dependencies

VSM depends on a few third party packages, and resolving those dependencies is often a headache. To mitigate this we are maintaining another repository called vsm-dependencies, which includes the rpm package list and corresponding binary packages. You can get those packages through command as following:

```
>   wget https://github.com/01org/vsm-dependencies/archive/<version>.zip
```

where is the vsm version like 1.1.

After obtaining this zip file, just unpack it and install the included rpm packages as follows:

```
>   yum install -y unzip
>   unzip <version>.zip
>   cd vsm-dependencies-<version>/repo
>   yum localinstall -y *.rpm
```

Build Packages

There are a few ways to get a VSM release package, a direct way is to download release package from github, or you can build rpm packages only or a full release package from source code. Below are the two approaches to build packages:

Build RPM Packages

After you download the source code from the VSM github, the first step is to build the VSM RPMs.

```
>   ./buildrpm
```

After building, all the rpms are located in $source*code*path/vsmrepo directory. If you already have the VSM RPMs, you can jump to VSM RPM Install.

Build VSM Release Package

Starting from 1.1, a tool is provided to generate a VSM release package, which covers what [Build_RPM]("Build RPMs") does. The script executes as following:

```
>    ./buildvsm.sh -v <version>
```

where is the vsm version like 1.1. If all execute smoothly, a binary package named version-.tar.gz will be generated in *release* folder if all execute well.

VSM RPM Install

Go to the directory that you placed your VSM RPMs in, or the /vsmrepo directory if you just built them from source. Then you can install vsm packages by:

```
>    cd vsmrepo
>    yum localinstall -y *.rpm
```

Note: vsm-dashboard will use the httpd service to setup the Web UI. Sometimes it conflicts with the OpenStack dashboard, so try to install the OpenStack dashboard and the VSM dashboard onto different nodes.

Configuration

Here is the information about the sample installation environment and its roles:

- 1. test1-control: 192.168.123.10 (this is the vsm controller node)
- 2. test1-storage1: 192.168.123.21, 192.168.124.21, 192.168.125.21 (ceph storage server 1)
- 3. test1-storage2: 192.168.123.22, 192.168.124.22, 192.168.125.22 (ceph storage server 2)
- 4. test1-storage3: 192.168.123.23, 192.168.124.23, 192.168.125.23 (ceph storage server 3)

So we configure the networks as below in VSM. While separate networks are recommended, they can all be the same network for demo or functionality testing.

```
- Management network: 192.168.123.0/24
- Ceph public network: 192.168.124.0/24
- Ceph cluster network: 192.168.125.0/24
```

NoteYou should set network appending on your network environment or check the network settings mentioned before.

Firewall and SELinux

Solution 1

1> Disable SELinux in the file /etc/selinux/config. A reboot is required to apply it.

```
>    SELINUX=disabled
```

2> Close the firewall

```
>    /etc/init.d/iptables save
>    /etc/init.d/iptables stop
>    chkconfig iptables off
```

Solution 2

1> If you want to open selinux, you should run commands below to add policies httpd. A reboot is required to apply it.

```
>    setsebool -P httpd_can_network_connect 1 &
>    chcon -R -h -t httpd_sys_content_t /var/www/html/
>    chmod -R a+r /var/www/html/
```

2> Settings for iptables. You should open these ports on every nodes in VSM.

```
22 ssh
80 http
443 https for future use
6789 Ceph Monitor
6800:8100 Ceph
123 ntp
8778  vsm
5673  rabbitmq
35357 keystone
5000  keysone
3306 mariadb
```

Here is one sample configuration `iptables`, take it as references.

```
*filter
:INPUT ACCEPT [0:0]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
-A INPUT -m state --state ESTABLISHED,RELATED -j ACCEPT
-A INPUT -p icmp -j ACCEPT
-A INPUT -i lo -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 22 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 80 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 443 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp --dport 6789 -j ACCEPT
-A INPUT -p tcp -m multiport --dports 6800:8100 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 123 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 8778 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 5673 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 35357 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 5000 -j ACCEPT
-A INPUT -p tcp -m tcp --dport 3306 -j ACCEPT
-A INPUT -j REJECT --reject-with icmp-host-prohibited
-A FORWARD -j REJECT --reject-with icmp-host-prohibited
COMMIT
```

Hosts file

VSM will sync /etc/hosts file from the controller node. Make sure your controller node's /etc/hosts file follows these rules:

```
- Lines with `localhost`, `127.0.0.1` and `::1` should not contains the actual hostname.
```

Take the correct version as an example to set your /etc/hosts file to on the controller node:

```
127.0.0.1       localhost localhost.localdomain localhost4 localhost4.localdomain4
::1             localhost localhost.localdomain localhost6 localhost6.localdomain6

192.168.124.10 test1-control

192.168.123.21 test1-storage1
192.168.124.21 test1-storage1
192.168.125.21 test1-storage1

192.168.123.22 test1-storage2
192.168.124.22 test1-storage2
192.168.125.22 test1-storage2

192.168.123.23 test1-storage3
192.168.124.23 test1-storage3
192.168.125.23 test1-storage3
```

Setup controller node

cluster.manifest

Next edit the cluster.manifest file. It's in /etc/manifest folder for manual deployment , or current manifest/ folder for automatic deployment.

modify three IP addresses

1> For the VSM controller, edit the cluster.manifest and modify it as described below:

Modify the three IP addresses according to your environment. `management_addr` is used by VSM to communicate with different services, such as using rabbitmq to transfer messages, rpc.call/rpc.cast etc.`ceph_public_addr` is a public (front-side) network address. `ceph_cluster_addr` is a cluster (back-side) network address.

Also, make sure the netmask is correctly set. In this sample, *netmask*=24 is fine, but with AWS instances, normally, *netmask*=16 are required.

```
[management_addr]
192.168.123.0/24

[ceph_public_addr]
192.168.124.0/24

[ceph_cluster_addr]
192.168.125.0/24
```

2> Now check the correctness of your cluster.manifest file by running the manifest checker:

```
cluster_manifest
```

Install

3> Install the vsm controller.

```
vsm-controller
```

NoteAfter executing this command, it will generate a configuration file located in /etc/vsmdeploy/deployrc owned by root. If you want to use the old version of /etc/vsmdeploy/deployrc, you may run `vsm-controller -f /etc/vsmdeploy/deployrc`.

Warning Do not set proxy env during installation.

Setup storage node

server.manifest

step 1 For VSM storage nodes, edit the file `/etc/manifest/server.manifest` and modify it as described below:

```
[vsm_controller_ip]
controller_ip
```

Update `vsm_controller_ip` to the VSM controller's IP address under subnet `management_addr`

```
[vsm_controller_ip]
192.168.123.10  #refer to test1-control node.
```

*_step 2_

Generate the `auth_key` by running the following command on the VSM controller node:

```
[root@test1-control manifest]# agent-token
9291376733ec4662929eadcf9eda3b44-e38aeba41c884fc88321ac84028792e4
```

Then insert the string generated by agent token under the [auth-key] section in the server.manifest file, as in:

```
[auth-key]
9291376733ec4662929eadcf9eda3b44-e38aeba41c884fc88321ac84028792e4
```

Or run below command to do the replacement:

```
>    replace-str 9291376733ec4662929eadcf9eda3b44-e38aeba41c884fc88321ac84028792e4
```

step 3 The storage you use for your Ceph cluster must have previously been provisioned by you with a label and a partition. For example:

```
>    parted /dev/sdb -- mklabel gpt
>    parted -a optimal /dev/sdb -- mkpart xfs 1MB 100%
```

Enter your primary and associated journal storage information in the `/etc/manifest/server.manifest` file.

For example, change the lines below:

```
[10krpm_sas]
#format [sas_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%
```

to be:

```
[10krpm_sas]
#format [sas_device]  [journal_device]
/dev/sdb1 /dev/sdc1
/dev/sdd1 /dev/sdc2
/dev/sde1 /dev/sdf
```

Then delete the redundant lines with %osd-by-path%, if you have fewer disks.

step 4 We recommend though that you use disk-by-path instead for the disk paths. Use the command below to find the true by-path:

```
>    ls -al /dev/disk/by-path/* | grep `disk-path` | awk '{print $9,$11}'
```

For example:

```
>    ls -al /dev/disk/by-path/* | grep sdb | awk '{print $9,$11}'
/dev/disk/by-path/pci-0000:00:0c.0-virtio-pci-virtio3 ../../sdb
```

Then replace the /dev/sdb with `/dev/disk/by-path/pci-0000:00:0c.0-virtio-pci-virtio3` in `/etc/manifest/server.manifest` file. Do this also for all the other disks listed in this file.

Warning It may cause an error when you add a disk without by-path. So, If you can not find the by-path for a normal disk, you should not use it. Or if you use it to create the cluster, and the create cluster fails, please delete it from the `/etc/manifest/server.manifest` file.

After that the disk list appears like this:

```
[10krpm_sas]
#format [sas_device]  [journal_device]
/dev/disk/by-path/pci-0000:00:0c.0-virtio-pci-virtio3   /dev/disk/by-path/pci-0000:00:0d.0-virtio-pci-virtio4
/dev/disk/by-path/pci-0000:00:0e.0-virtio-pci-virtio5   /dev/disk/by-path/pci-0000:00:0f.0-virtio-pci-virtio6
/dev/disk/by-path/pci-0000:00:10.0-virtio-pci-virtio7   /dev/disk/by-path/pci-0000:00:11.0-virtio-pci-virtio8
```

step 5 If you have several kinds of storage media, and you want these disks organized into different storage groups in VSM, then you may follow the operations below. Otherwise, you may skip this step and just put all the disks into the [10krpm_sas] section.

You may want to add disks into other sections in the `/etc/manifest/server.manifest` file after the [10krpm_sas] section. Take [ssd] as an example:

1> Add storage class in `/etc/manifest/cluster.manifest` in controller node.

```
[storage_class]
ssd # add this line
10krpm_sas
```

2> Add storage group in `/etc/manifest/cluster.manifest` on the controller node.

```
[storage_group]
high_performance_test  High_Performance_SSD_test ssd
```

Note No extra spaces in word, use _ to instead spaces.

3> Add disks under `/etc/manifest/server.manifest` on the storage node(s) which have SSD, such as:

```
[ssd]
/dev/disk/by-path/pci-0000:00:0c.0-virtio-pci-virtio9   /dev/disk/by-path/pci-0000:00:0d.0-virtio-pci-virtio1
/dev/disk/by-path/pci-0000:00:0e.0-virtio-pci-virtio11   /dev/disk/by-path/pci-0000:00:0f.0-virtio-pci-virtio14
/dev/disk/by-path/pci-0000:00:10.0-virtio-pci-virtio23  /dev/disk/by-path/pci-0000:00:11.0-virtio-pci-virtio10
```

4> Now check the correctness of your server.manifest file by running the manifest checker:

```
>    server_manifest
```

Setup VSM for the storage node

After the configuration of `/etc/manifest/server.manifest`, you may run:

```
>    vsm-node
```

to complete setup of the storage node.

If you want to check that the vsm agent started correctly, you can look at the two files in */var/log/vsm* on each storage node. The vsm.physical.log should have no errors and end with the line:

```
INFO [vsm.openstack.common.rpc.common] Connected to AMQP server on <vsm-controller-IP-address>:5673
```

and you should see a similar message at the start of vsm.agent.log and ending with:

```
INFO [vsm.agent.manager] agent/manager.py update ceph.conf from db. OVER
```

Likewise, you should see no errors in the three log files in /var/log/vsm on the controller node.

VSM Web UI

After the command has finished execution, performing the following steps to check if you have setup the controller correctly:

1. Access https://vsm controller IP/dashboard/vsm.(for example *https://192.168.123.10/dashboard/vsm*)
2. User name: admin, and password can be obtained from: */etc/vsmdeploy/deployrc* in the ADMIN_PASSWORD field:

   > cat /etc/vsmdeploy/deployrc |grep ADMIN_PASSWORD

3. Then you can switch to the `Create Cluster` Panel, and push the create cluster button to create a ceph cluster. At this point please refer to the VSM Manual, which is located at *https://01.org/virtual-storage-manager*

Frequently Asked Questions

** Q: Executing "agent-token" is hung.**

```
A: Please check http proxy setting to make sure no http_proxy variable is set in the enviornment.
```

** Q: "An error occurred authenticating. Please try again later." appears on the controller web ui after fresh installation.**

```
A: Firstly, please make sure the right password is entered, the password can be obtained from /etc/vsmdeploy/deployrc in "ADMIN_PA
```

** Q: keyring error on cluster creation.**

```
A: The root cause is that the vsm controller has already updated a new token, but it is not applied on all agents.
```

** Q: Negative update time is showing on RBD list page.**

```
A: Before creating the ceph cluster, please make sure all ceph nodes are time synchronized via NTP.
```

** Q: vsm-agent process causes one disk to be saturated with i/o load.**

```
A: A known case causes i/o saturation if multiple OSDs are defined on the same physical device, which is normally used in the demo
```

** Q: Can't replace node if ceph cluster contains only 3 nodes.**

```
A: This is an expected safeguard. A 3 node cluster minimum is needed to meet availability requirements.
```