Design Best Practices

By
Udayendu Kar
Sr. Technical Architect
Avaya India Pvt. Ltd.

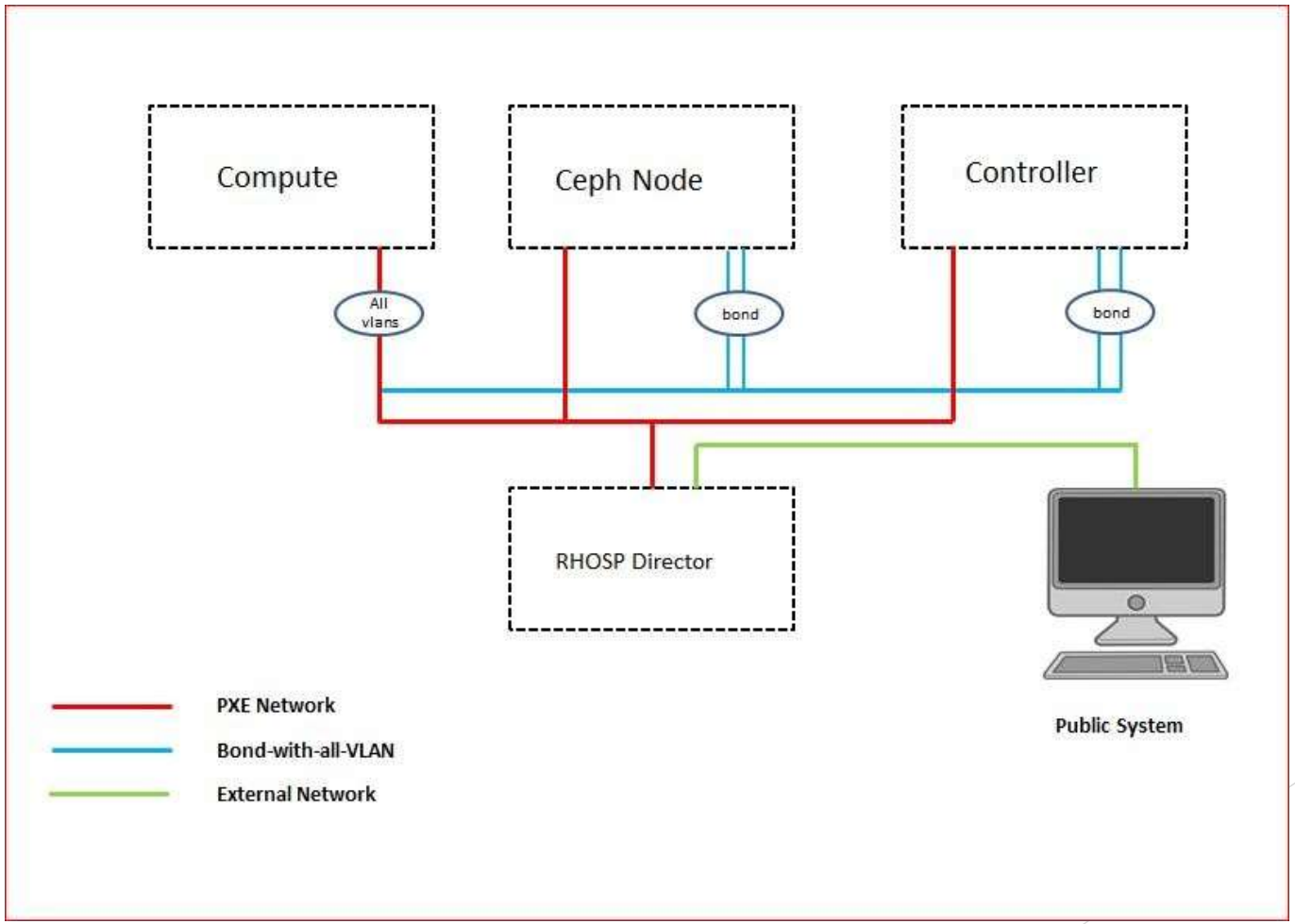
Agenda:

- Best place for CEPH
- Integration diagram of Openstack & CEPH
- How to choose CEPH Storage node
- Disk Mapping of CEPH Storage node (OSD + Journal)
- Networking consideration for CEPH node
- Questions & Answer

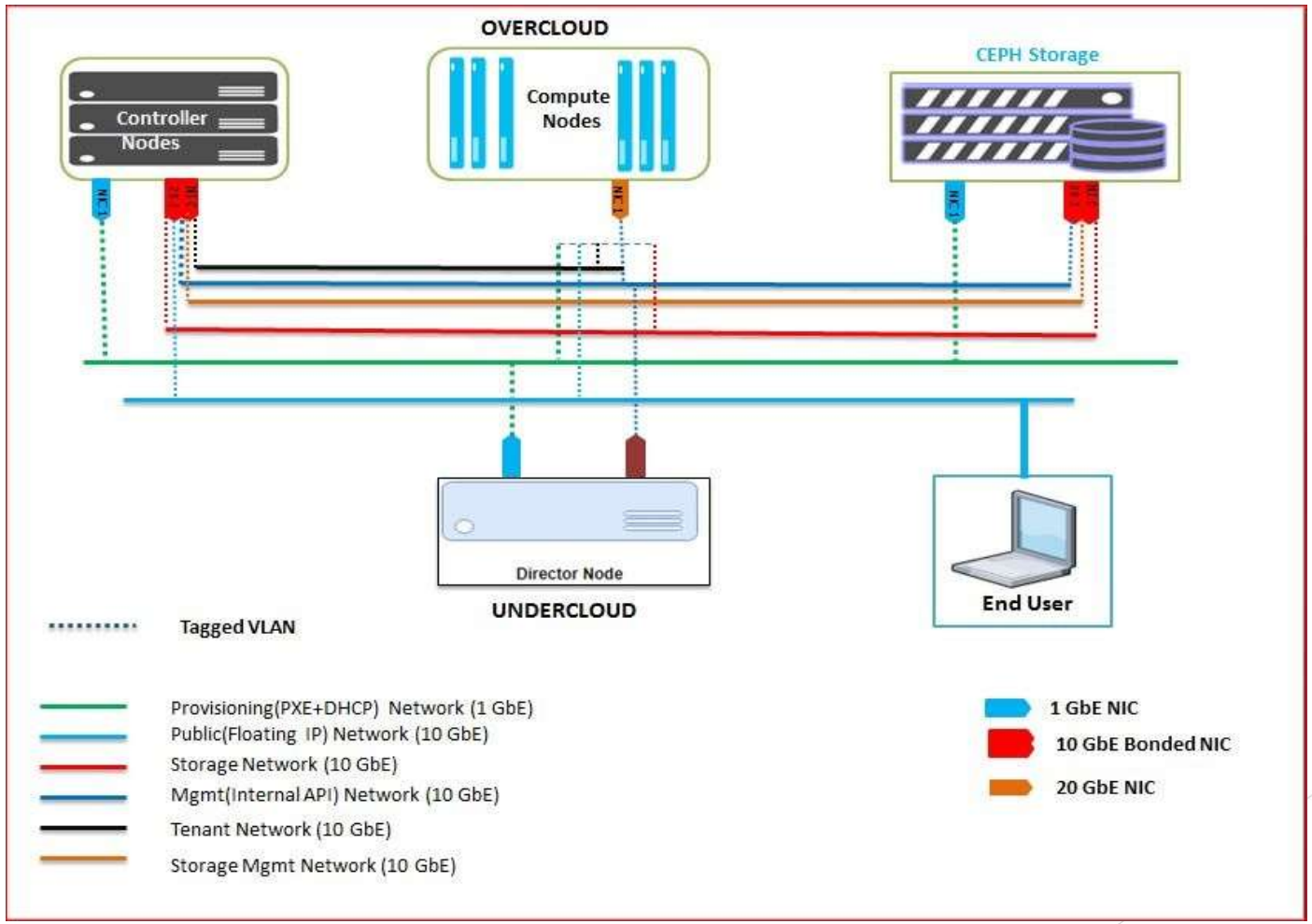
Best Place for CEPH:

+

Integration diagram of Openstack with CEPH:



Integration diagram of Openstack with CEPH:



How to choose CEPH Storage node:

Storage:

- 7.2K/12k/15K RPM based SAS drive for OSD disk
  - 6TB or 10TB HDD to be used
- SSD for Journal Disk (256 GB)
  - 5GB to 15GB per OSD

Network:

- 2 x 1Gb NIC for Mgmt Network
- 2 x (10Gb/20Gb/40Gb) NIC for Storage Network

Compute:

- 128GB RAM
- 16 Core (1socket x 2cpu x 8core)

Disk Mapping of CEPH node (OSD + Journal)

- In this design we have used 10 x 6TB (7.2K RPM SAS) HDD as OSD drive
- 2 x 480GB SSD per CEPH node. 5 HDD are mapped to one SSD.

```
# fdisk -l /dev/sdb
WARNING: fdisk GPT support is currently new, and therefore in an experimental phase. Use at your own discretion
Disk /dev/sdb: 400.1 GB, 400054902784 bytes, 781357232 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 262144 bytes / 262144 bytes
Disk label type: gpt
Disk identifier: 60687129-0205-4DA0-9627-ABEC4733D9F0

 #         Start          End    Size  Type            Name
 1          2048     10487807      5G  unknown         ceph journal
 2      10487808     20973567      5G  unknown         ceph journal
 3      20973568     31459327      5G  unknown         ceph journal
 4      31459328     41945087      5G  unknown         ceph journal
 5      41945088     52430847      5G  unknown         ceph journal
```

```
# fdisk -l /dev/sdc
WARNING: fdisk GPT support is currently new, and therefore in an experimental phase. Use at your own discretion
Disk /dev/sdc: 400.1 GB, 400054902784 bytes, 781357232 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 262144 bytes / 262144 bytes
Disk label type: gpt
Disk identifier: F687B7FF-8395-4909-826E-4B5686C7449A

 #         Start          End    Size  Type            Name
 1          2048     10487807      5G  unknown         ceph journal
 2      10487808     20973567      5G  unknown         ceph journal
 3      20973568     31459327      5G  unknown         ceph journal
 4      31459328     41945087      5G  unknown         ceph journal
 5      41945088     52430847      5G  unknown         ceph journal
```

All the OSD mapping per SSD per Node:

```
# ls -l /dev/disk/by-partuuid/* | grep -i sdb
lrwxrwxrwx. 1 root root 10 Feb 26 12:49 /dev/disk/by-partuuid/1d59b255-260b-4622-9780-50d88e5d643a -> ../../sdb5
lrwxrwxrwx. 1 root root 10 Feb 26 12:49 /dev/disk/by-partuuid/3492526a-2af7-4973-97be-c68f870b2b84 -> ../../sdb2
lrwxrwxrwx. 1 root root 10 Feb 26 12:49 /dev/disk/by-partuuid/4fac716b-9f71-4f16-ae90-687ad248b46c -> ../../sdb1
lrwxrwxrwx. 1 root root 10 Feb 26 12:49 /dev/disk/by-partuuid/693e9b04-1896-470a-a47c-0df32e59ea0f -> ../../sdb3
lrwxrwxrwx. 1 root root 10 Feb 26 12:49 /dev/disk/by-partuuid/e74c1d42-fca0-4506-b86c-b541622ba085 -> ../../sdb4

# ls -l /dev/disk/by-partuuid/* | grep -i sdc
lrwxrwxrwx. 1 root root 10 Feb 26 12:50 /dev/disk/by-partuuid/041defb6-35a6-435f-bd0c-31c9a5025971 -> ../../sdc1
lrwxrwxrwx. 1 root root 10 Feb 26 12:50 /dev/disk/by-partuuid/0721e0c9-631b-4a19-85ac-d7468e9741b7 -> ../../sdc4
lrwxrwxrwx. 1 root root 10 Feb 26 12:50 /dev/disk/by-partuuid/2c9bc342-d7a9-4530-9759-508519c0ce50 -> ../../sdc3
lrwxrwxrwx. 1 root root 10 Feb 26 12:50 /dev/disk/by-partuuid/71586cee-b144-4623-b347-cd7769a526f5 -> ../../sdc2
lrwxrwxrwx. 1 root root 10 Feb 26 12:50 /dev/disk/by-partuuid/d1a3f6ae-18b5-437e-93be-ee003312c1e6 -> ../../sdc5
```

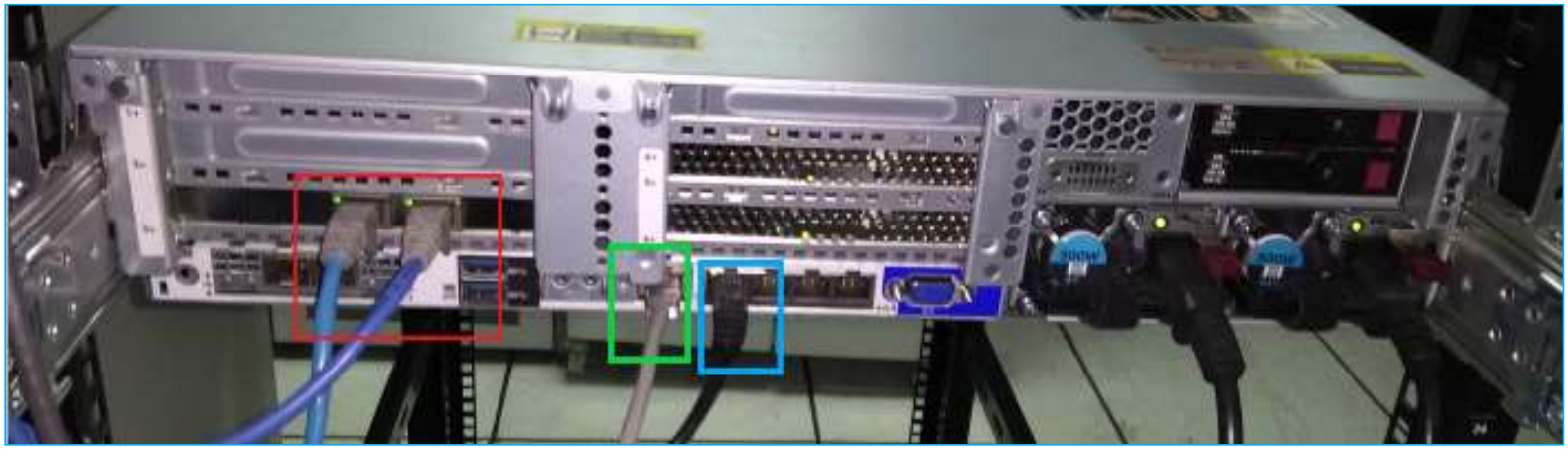
```
sdd      8:48   0   5.5T  0 disk
└─sdd1   8:49   0   5.5T  0 part /var/lib/ceph/osd/ceph-22
sde      8:64   0   5.5T  0 disk
└─sde1   8:65   0   5.5T  0 part /var/lib/ceph/osd/ceph-25
sdf      8:80   0   5.5T  0 disk
└─sdf1   8:81   0   5.5T  0 part /var/lib/ceph/osd/ceph-16
sdg      8:96   0   5.5T  0 disk
└─sdg1   8:97   0   5.5T  0 part /var/lib/ceph/osd/ceph-20
sdh      8:112  0   5.5T  0 disk
└─sdh1   8:113  0   5.5T  0 part /var/lib/ceph/osd/ceph-12


sdi      8:128  0   5.5T  0 disk
└─sdi1   8:129  0   5.5T  0 part /var/lib/ceph/osd/ceph-14
sdj      8:144  0   5.5T  0 disk
└─sdj1   8:145  0   5.5T  0 part /var/lib/ceph/osd/ceph-6
sdk      8:160  0   5.5T  0 disk
└─sdk1   8:161  0   5.5T  0 part /var/lib/ceph/osd/ceph-9
sdl      8:176  0   5.5T  0 disk
└─sdl1   8:177  0   5.5T  0 part /var/lib/ceph/osd/ceph-0
sdm      8:192  0   5.5T  0 disk
└─sdm1   8:193  0   5.5T  0 part /var/lib/ceph/osd/ceph-3
```

```
# ceph-disk list
/dev/sda :
 /dev/sda1 other, iso9660
 /dev/sda2 other, xfs, mounted on /
/dev/sdb :
 /dev/sdb1 ceph journal, for /dev/sdd1
 /dev/sdb2 ceph journal, for /dev/sde1
 /dev/sdb3 ceph journal, for /dev/sdf1
 /dev/sdb4 ceph journal, for /dev/sdg1
 /dev/sdb5 ceph journal, for /dev/sdh1
/dev/sdc :
 /dev/sdc1 ceph journal, for /dev/sdi1
 /dev/sdc2 ceph journal, for /dev/sdj1
 /dev/sdc3 ceph journal, for /dev/sdk1
 /dev/sdc4 ceph journal, for /dev/sdl1
 /dev/sdc5 ceph journal, for /dev/sdm1
/dev/sdd :
 /dev/sdd1 ceph data, active, cluster ceph, osd.22, journal /dev/sdb1
/dev/sde :
 /dev/sde1 ceph data, active, cluster ceph, osd.25, journal /dev/sdb2
/dev/sdf :
 /dev/sdf1 ceph data, active, cluster ceph, osd.16, journal /dev/sdb3
/dev/sdg :
 /dev/sdg1 ceph data, active, cluster ceph, osd.20, journal /dev/sdb4
/dev/sdh :
 /dev/sdh1 ceph data, active, cluster ceph, osd.12, journal /dev/sdb5
/dev/sdi :
 /dev/sdi1 ceph data, active, cluster ceph, osd.14, journal /dev/sdc1
/dev/sdj :
 /dev/sdj1 ceph data, active, cluster ceph, osd.6, journal /dev/sdc2
/dev/sdk :
 /dev/sdk1 ceph data, active, cluster ceph, osd.9, journal /dev/sdc3
/dev/sdl :
 /dev/sdl1 ceph data, active, cluster ceph, osd.0, journal /dev/sdc4
/dev/sdm :
 /dev/sdm1 ceph data, active, cluster ceph, osd.3, journal /dev/sdc5
```

```
# lsblk
NAME   MAJ:MIN RM   SIZE RO TYPE MOUNTPOINT
sda      8:0    0 279.4G  0 disk
├─sda1   8:1    0     1M  0 part
└─sda2   8:2    0 279.4G  0 part /
sdb      8:16   0 372.6G  0 disk
├─sdb1   8:17   0    14G  0 part
├─sdb2   8:18   0    14G  0 part
├─sdb3   8:19   0    14G  0 part
├─sdb4   8:20   0    14G  0 part
└─sdb5   8:21   0    14G  0 part
sdc      8:32   0 372.6G  0 disk
├─sdc1   8:33   0    14G  0 part
├─sdc2   8:34   0    14G  0 part
├─sdc3   8:35   0    14G  0 part
├─sdc4   8:36   0    14G  0 part
└─sdc5   8:37   0    14G  0 part
sdd      8:48   0   5.5T  0 disk
└─sdd1   8:49   0   5.5T  0 part /var/lib/ceph/osd/ceph-22
sde      8:64   0   5.5T  0 disk
└─sde1   8:65   0   5.5T  0 part /var/lib/ceph/osd/ceph-25
sdf      8:80   0   5.5T  0 disk
└─sdf1   8:81   0   5.5T  0 part /var/lib/ceph/osd/ceph-16
sdg      8:96   0   5.5T  0 disk
└─sdg1   8:97   0   5.5T  0 part /var/lib/ceph/osd/ceph-20
sdh      8:112  0   5.5T  0 disk
└─sdh1   8:113  0   5.5T  0 part /var/lib/ceph/osd/ceph-12
sdi      8:128  0   5.5T  0 disk
└─sdi1   8:129  0   5.5T  0 part /var/lib/ceph/osd/ceph-14
sdj      8:144  0   5.5T  0 disk
└─sdj1   8:145  0   5.5T  0 part /var/lib/ceph/osd/ceph-6
sdk      8:160  0   5.5T  0 disk
└─sdk1   8:161  0   5.5T  0 part /var/lib/ceph/osd/ceph-9
sdl      8:176  0   5.5T  0 disk
└─sdl1   8:177  0   5.5T  0 part /var/lib/ceph/osd/ceph-0
sdm      8:192  0   5.5T  0 disk
└─sdm1   8:193  0   5.5T  0 part /var/lib/ceph/osd/ceph-3
```

Networking consideration for CEPH node:



iLO Network
Storage Network
Mgmt Network


Q&A
How
Why
Where
When
What
Who

Thank You !