
Ceph (& OpenStack) untuk Enterprise

LAZUARDI NASUTION

[HTTPS://WWW.LINKEDIN.COM/IN/LAZUARDI-NASUTION-15031717/](https://www.linkedin.com/in/lazuardi-nasution-15031717/)

A solid orange horizontal bar at the bottom of the slide.

Kok Ceph? – Awal Perkenalan

- Thesis Sage A. Weil
 - <https://ceph.com/wp-content/uploads/2016/08/weil-thesis.pdf>

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

**CEPH: RELIABLE, SCALABLE, AND HIGH-PERFORMANCE
DISTRIBUTED STORAGE**

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

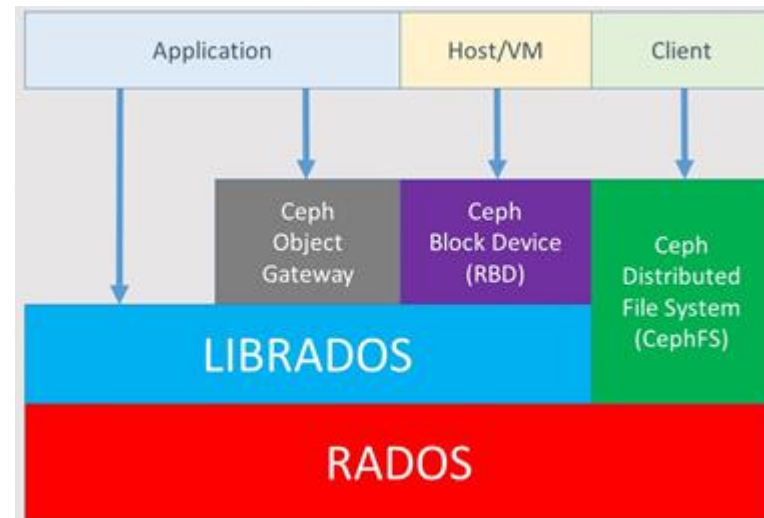
by

Sage A. Weil

December 2007

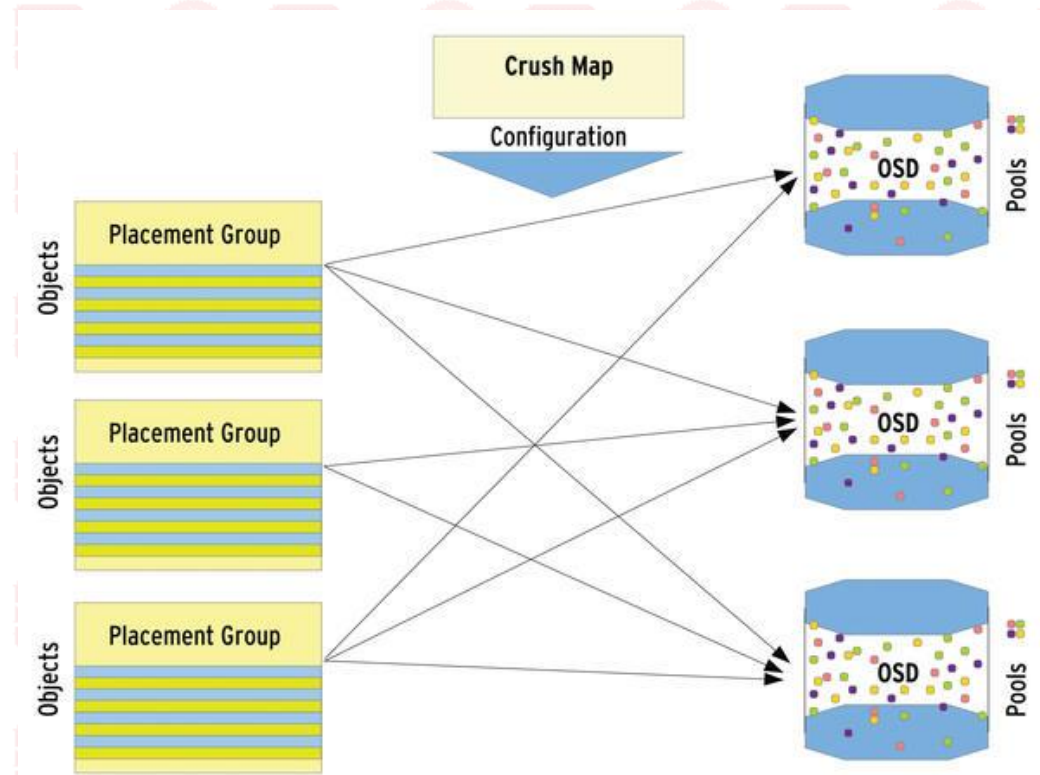
Kok Ceph? – RADOS

- Akses paralel per objek
 - Setiap block dipotong menjadi objek-objek kecil
 - Setiap file (dapat) dipotong menjadi objek-objek kecil
 - Setiap objek disebar sesuai PG yang berbeda
 - Default: setara RAID10 tapi direplikasi jadi 3
- Recovery hanya pada objek
 - Cepat: sekitar 1 jam pada utilisasi 50%

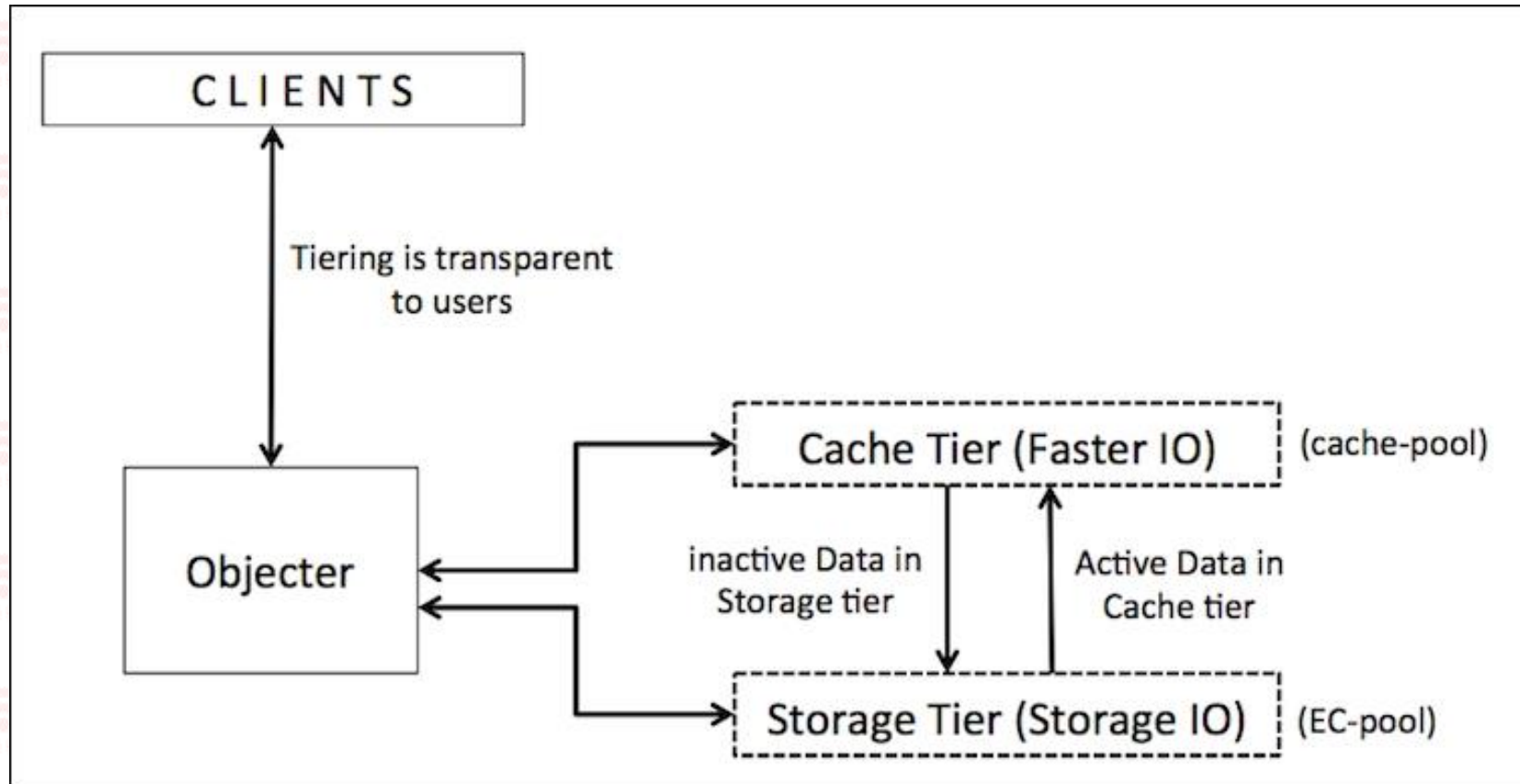


Kok Ceph? – Crush Map

- Setiap objek terasosiasi dengan placement group
- Setiap placement group terdiri dari kombinasi OSD set sesuai profile
- Setiap profil dapat diasosiasikan pada pool yang berbeda
- Setiap OSD dapat berbeda host (default), rack, room, DC, dst.

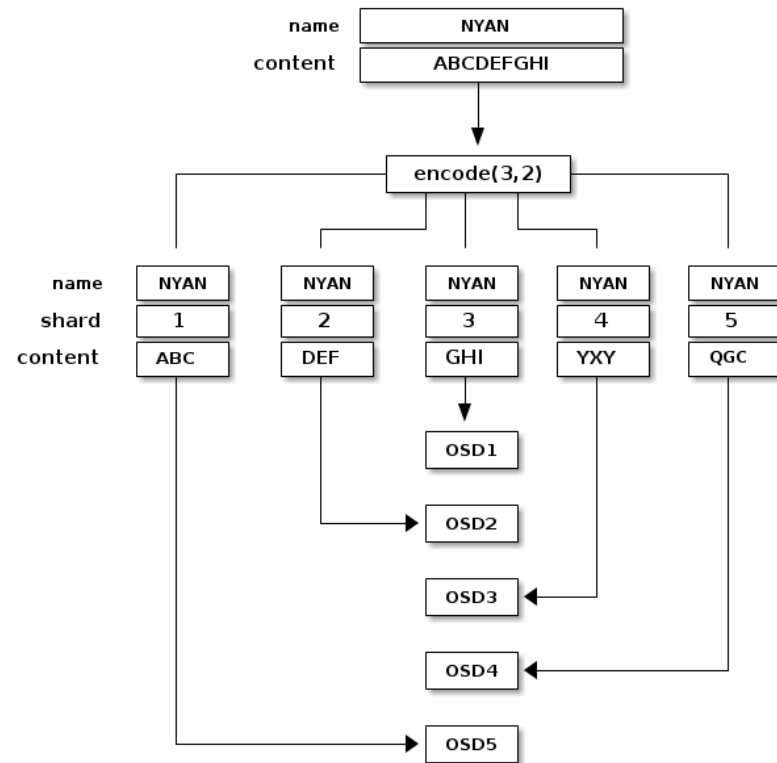


Kok Ceph? – Tiering

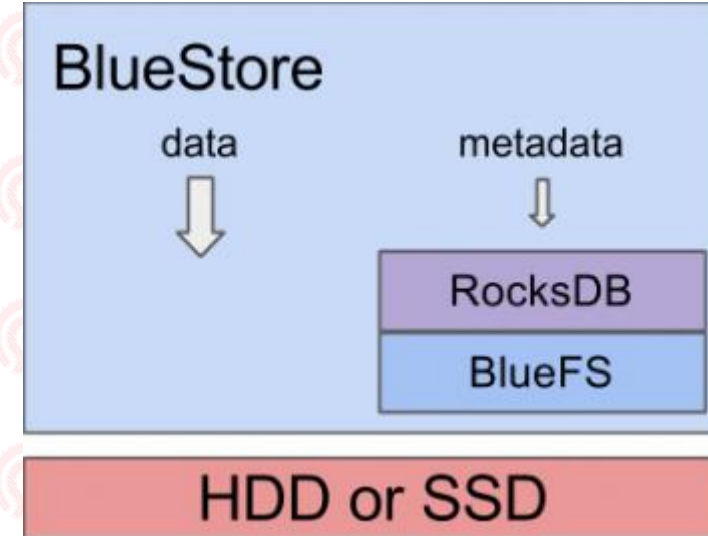
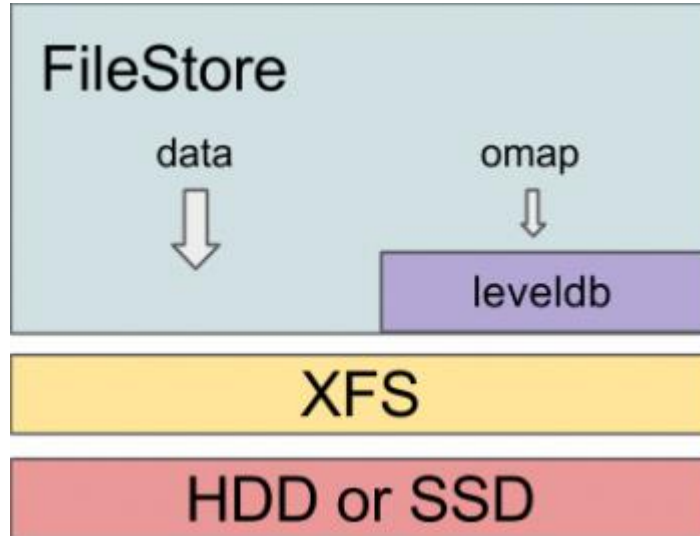


Kok Ceph? – EC

- Profil: A (artis) + B (bodyguard), A & B bebas
- $A+B$ = jumlah minimal sebaran OSD
- Setara RAID
 - RAID5: $B = 1$
 - RAID6: $B = 2$
- Sebaiknya menggunakan replicated tiering (terutama versi lama)



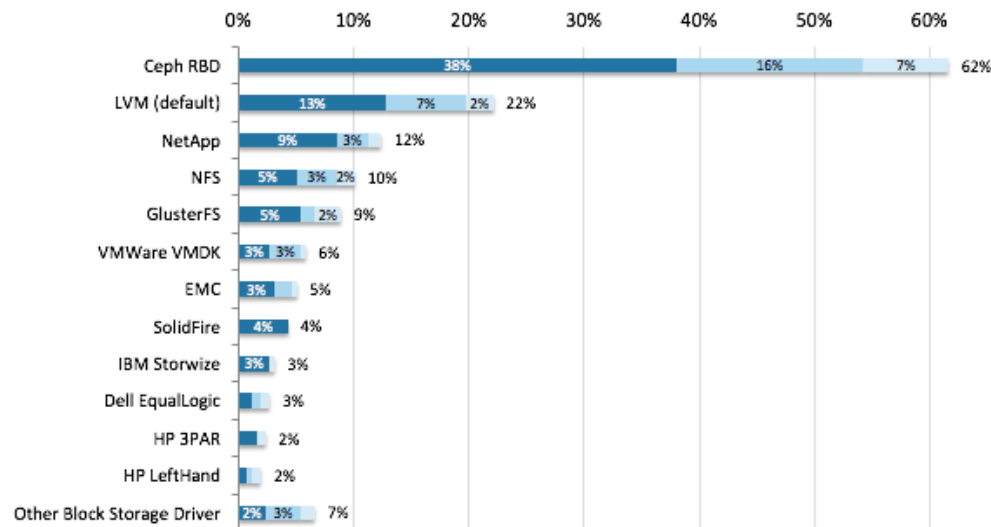
Kok Ceph? – FileStore vs BlueStore



Ceph & OpenStack – Hasil Survey

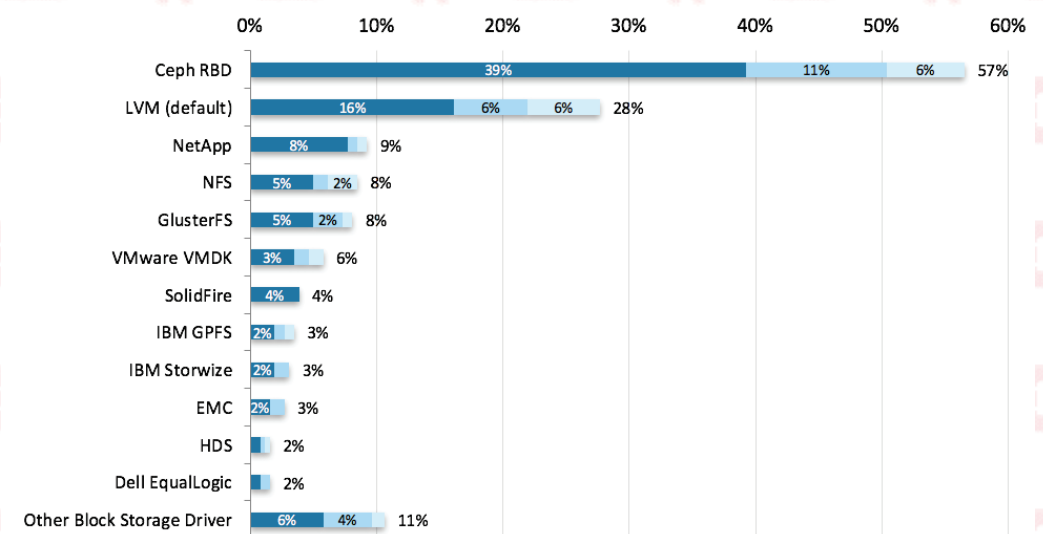
OPENSTACK USER SURVEY - OKTOBER 2015

<https://www.openstack.org/assets/survey/Public-User-Survey-Report.pdf>



OPENSTACK USER SURVEY - APRIL 2016

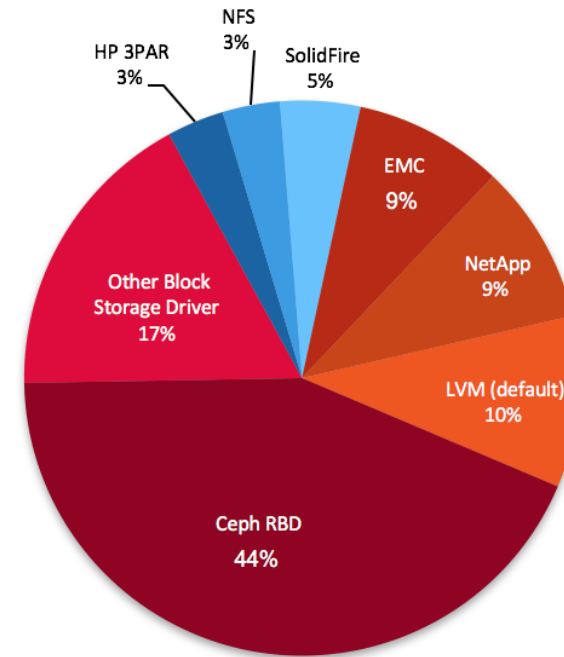
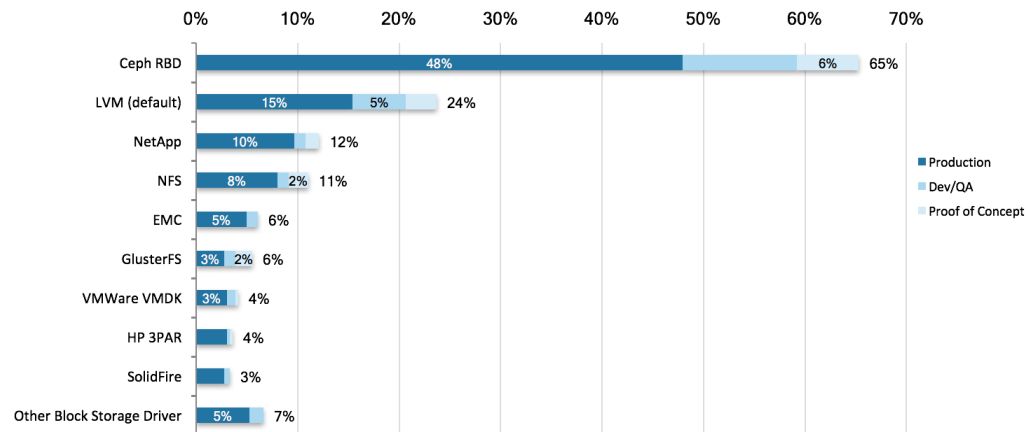
<https://www.openstack.org/assets/survey/April-2016-User-Survey-Report.pdf>



Ceph & OpenStack – Hasil Survey

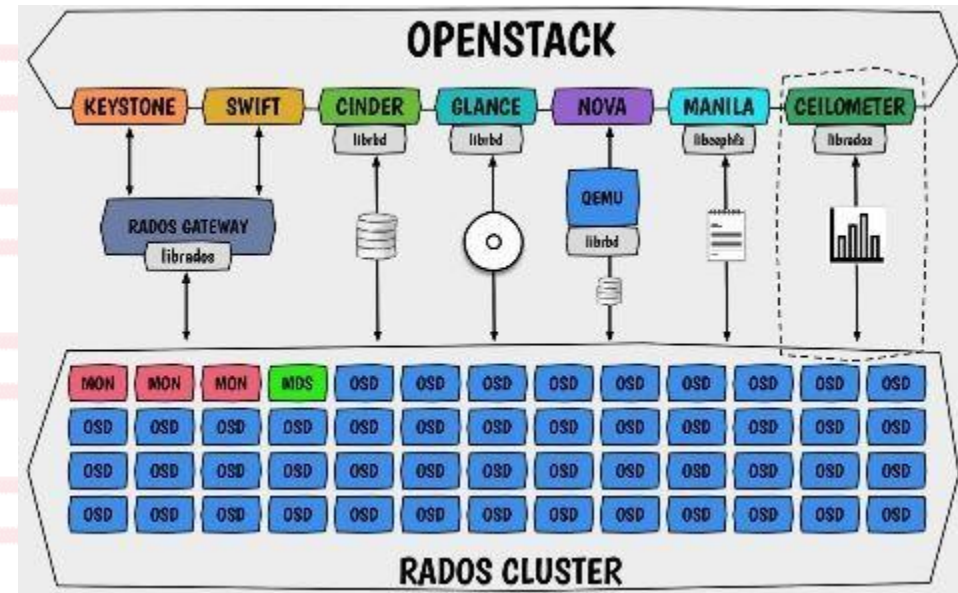
OPENSTACK USER SURVEY - APRIL 2017

<https://www.openstack.org/assets/survey/April2017SurveyReport.pdf>



Ceph & OpenStack – Integrasi

- Keystone: local, backup ke CephFS
- Swift: tidak digunakan
- Cinder: RBD
- Glance: RBD
- Nova: RBD
- Manila (jarang): CephFS
- Ceilometer (jarang): local, backup ke CephFS



Ceph & OpenStack – Pengujian

SPEKIFIKASI

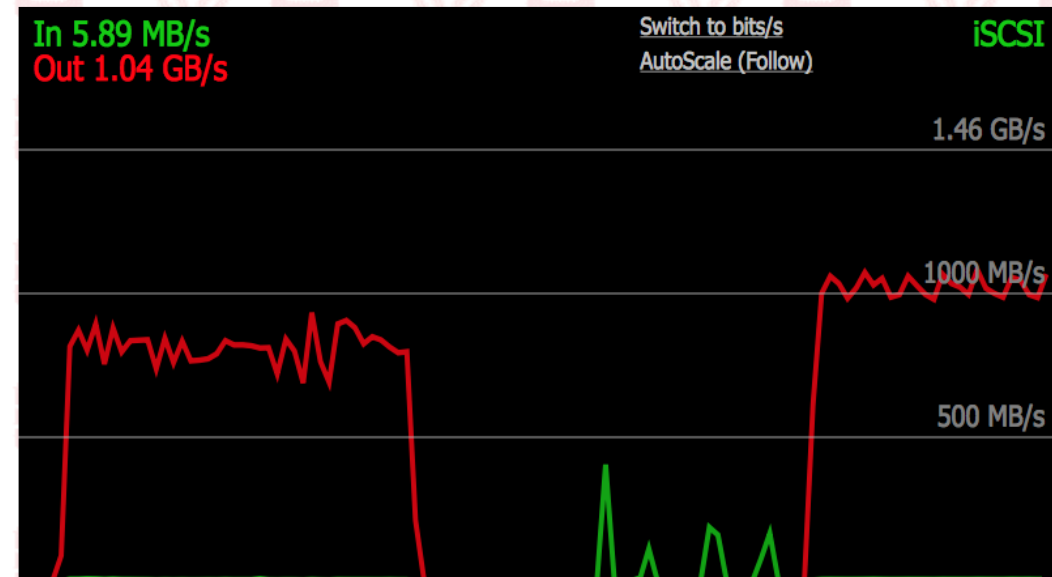
Fisik

- CPU: 2x E5-2667v4
- RAM: 16x 16GB
- Link: 2x 10GBase-T

VM

- vCPU: 8
- vRAM: 16GB
- Volume: 40GB + 5x 10TB

XIGMANAS DASHBOARD



Ceph & OpenStack – Pengujian

DD DI OPENSTACK + CEPH

```
ubuntu@gluster-openstack: ~  
ubuntu@gluster-openstack:~$ sudo dd if=/dev/zero of=/mnt/lvm/VG00/LV00/iotest.dat  
t bs=4096 count=1000000  
1000000+0 records in  
1000000+0 records out  
4096000000 bytes (4.1 GB, 3.8 GiB) copied, 3.63496 s, 1.1 GB/s  
ubuntu@gluster-openstack:~$
```

DD DI VMWARE + HP MSA 2040

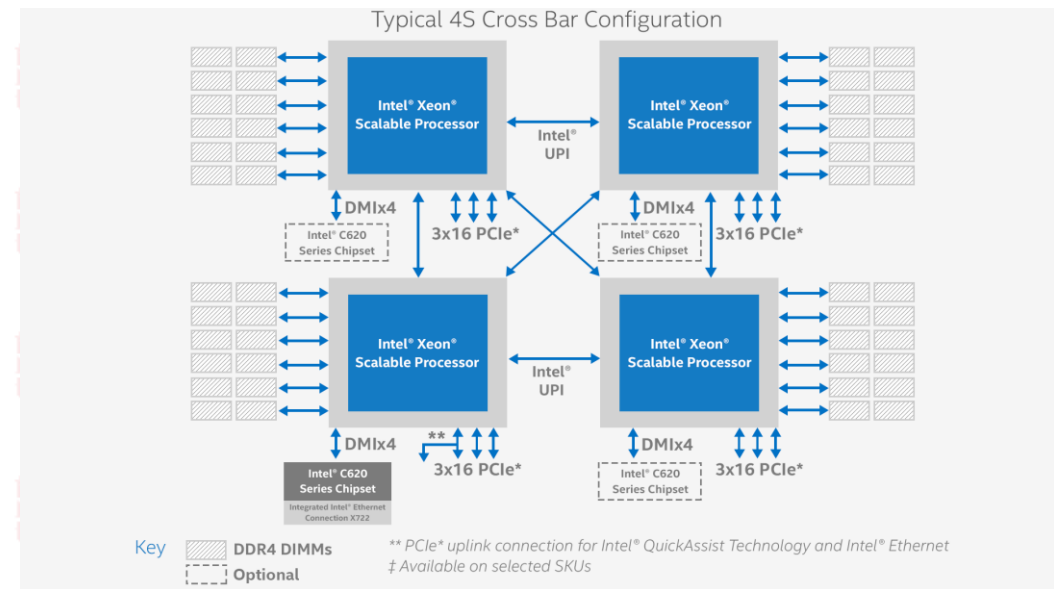
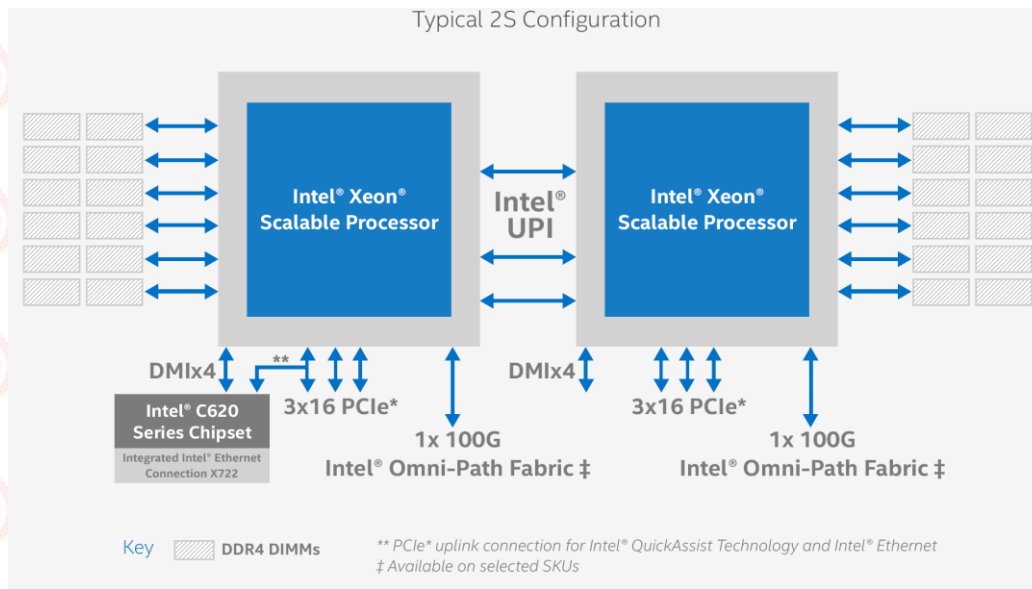
```
ubuntu@gluster-vmware: ~  
ubuntu@gluster-vmware:~$ sudo dd if=/dev/zero of=/iotest.dat bs=4096 count=10000  
00  
1000000+0 records in  
1000000+0 records out  
4096000000 bytes (4.1 GB, 3.8 GiB) copied, 11.2572 s, 364 MB/s  
ubuntu@gluster-vmware:~$ sudo rm -f /iotest.dat  
ubuntu@gluster-vmware:~$
```

Hati-Hati! — Networking

- Beban recovery traffic sangat besar (default config)
 - Solusi: perbesar &/| perbanyak (bonding) koneksi
 - Solusi: recovery throttling
 - Solusi: jumbo frame (khusus $\geq 10\text{GbE}$)
- Setiap OSD membuka 1 TCP port
 - Solusi: L3+L4 hash bonding
 - Solusi: koneksi (setidaknya VLAN) terpisah & trusted

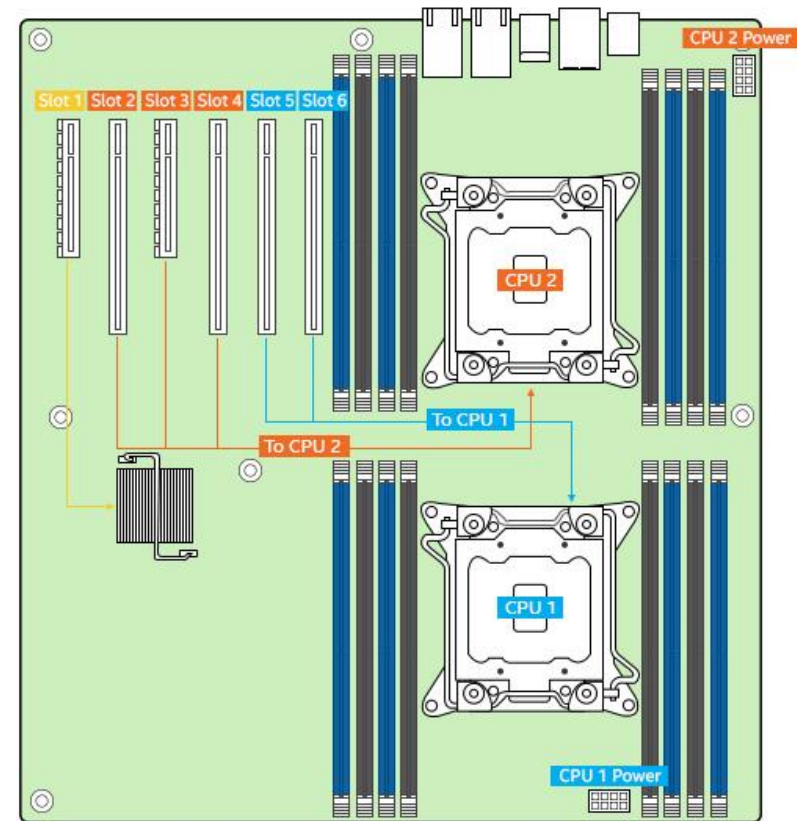
```
root@management-b:~  
Every 5.0s: ceph status  
  
cluster 9c24f3f5-bbc6-4f66-bd82-8fae5e4acaa3  
health HEALTH_WARN  
47 pgs backfilling  
48 pgs stuck unclean  
recovery 28/48135979 objects degraded (0.000%)  
recovery 731543/48135979 objects misplaced (1.520%)  
monmap e2: 3 mons at {management-a=192.168.1.1:6789/0,management-b=192.168.1.2:6789/0,management-c=192.168.1.3:6789/0}  
election epoch 330, quorum 0,1,2 management-a,management-b,management-c  
osdmap e46745: 135 osds: 134 up, 134 in; 47 remapped pgs  
flags sortbitwise,require_jewel_osds  
pgmap v48362287: 10880 pgs, 10 pools, 62429 GB data, 15529  
183 TB used, 259 TB / 443 TB avail  
28/48135979 objects degraded (0.000%)  
731543/48135979 objects misplaced (1.520%)  
10832 active+clean  
47 active+remapped+backfilling  
1 active+remapped  
recovery io 4187 MB/s, 1053 objects/s  
client io 0 B/s rd, 3245 kB/s wr, 32 op/s rd, 83 op/s wr
```

Hati-Hati! – CPU



Hati-Hati! – CPU

- Setiap OSD berbentuk 1 process
 - Solusi: multi core CPU
- Inter CPU link (QPI/UPI) terbatas
 - Solusi: NUMA & slotting
- Latensi
 - Solusi: perbesar RAM (storage & client)
 - Solusi: tiering
 - Solusi: high frequency CPU (terutama pada EC)



Pengalaman (Iklan Mode On)

- Dimulai 2015
- 4 (Ceph & OpenStack) + 1 (Ceph saja) di lingkungan K/L
- Sisanya: “***They*** Who Must Not Be Named”
- Kapasitas:
 - Implementasi: 100TB-500TB (raw)
 - Perencanaan: 1PB-3PB (raw)

Terima Kasih

