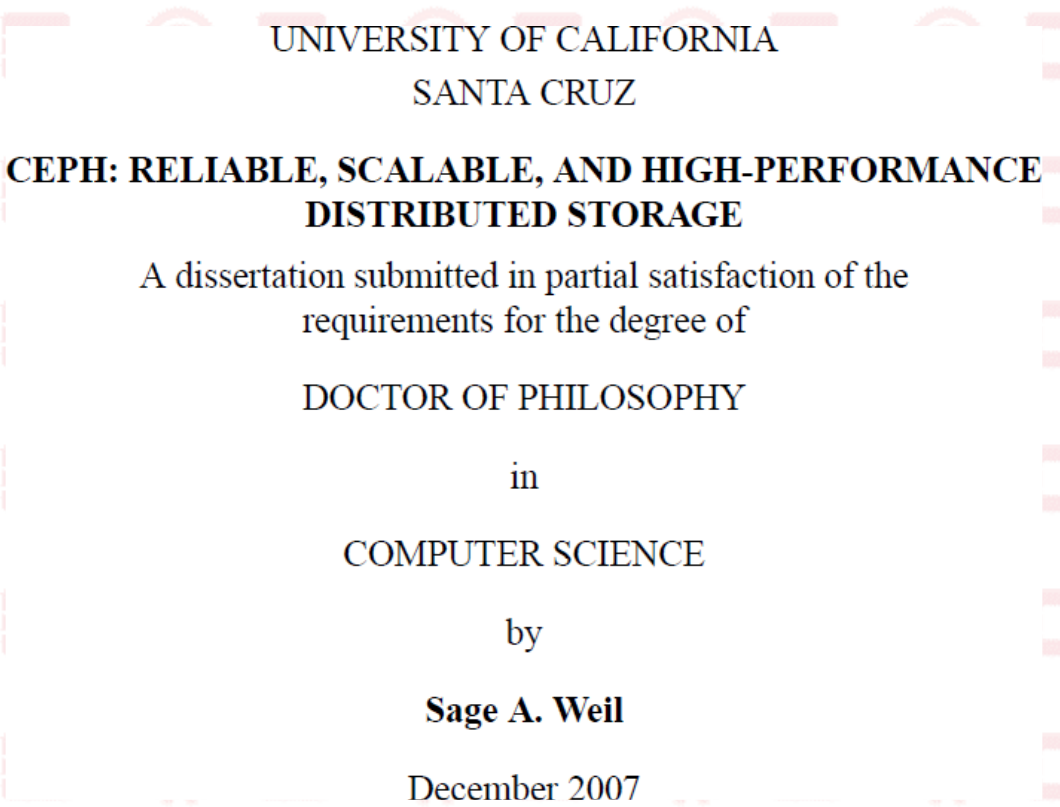
Ceph (& OpenStack) untuk Enterprise

LAZUARDI NASUTION

HTTPS://WWW.LINKEDIN.COM/IN/LAZUARDI-NASUTION-15031717/

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

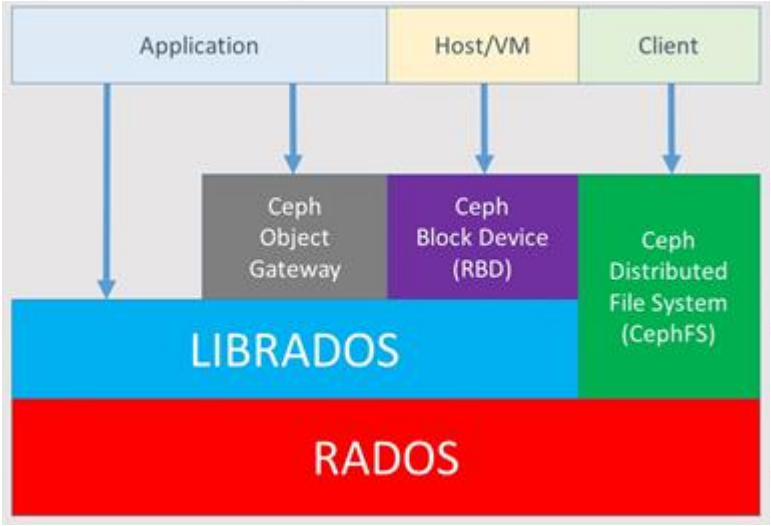
by

Sage A. Weil

December 2007

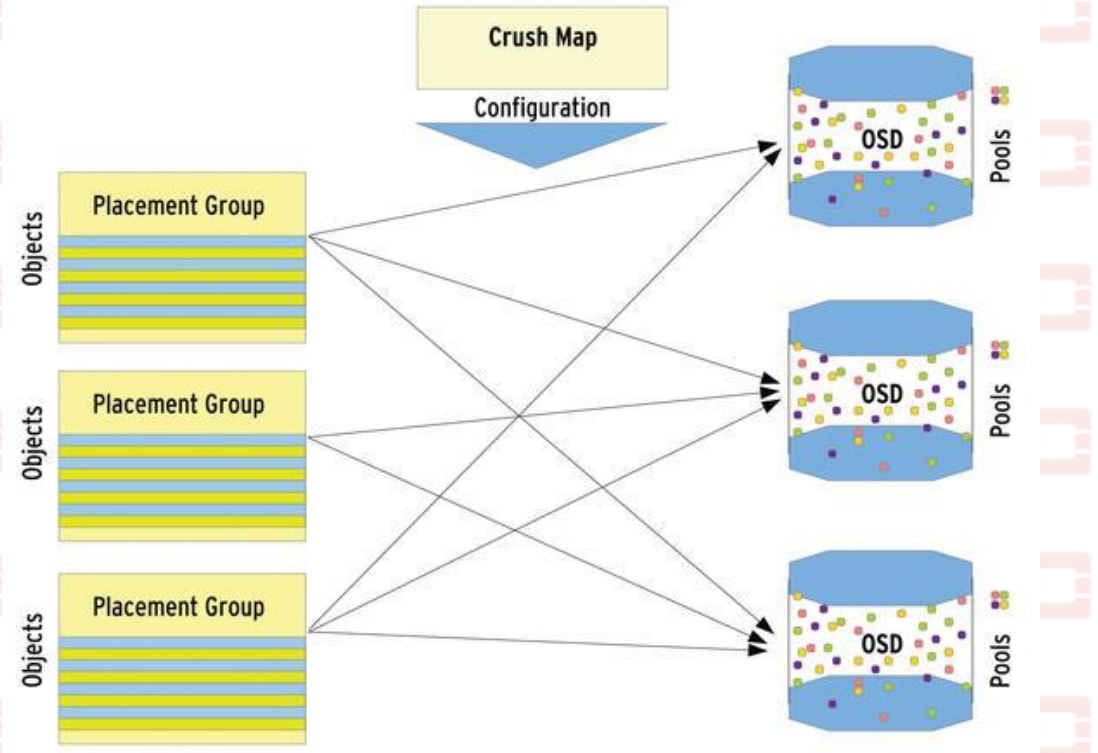
Kok Ceph? – RADOS

- Akses pararel 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%

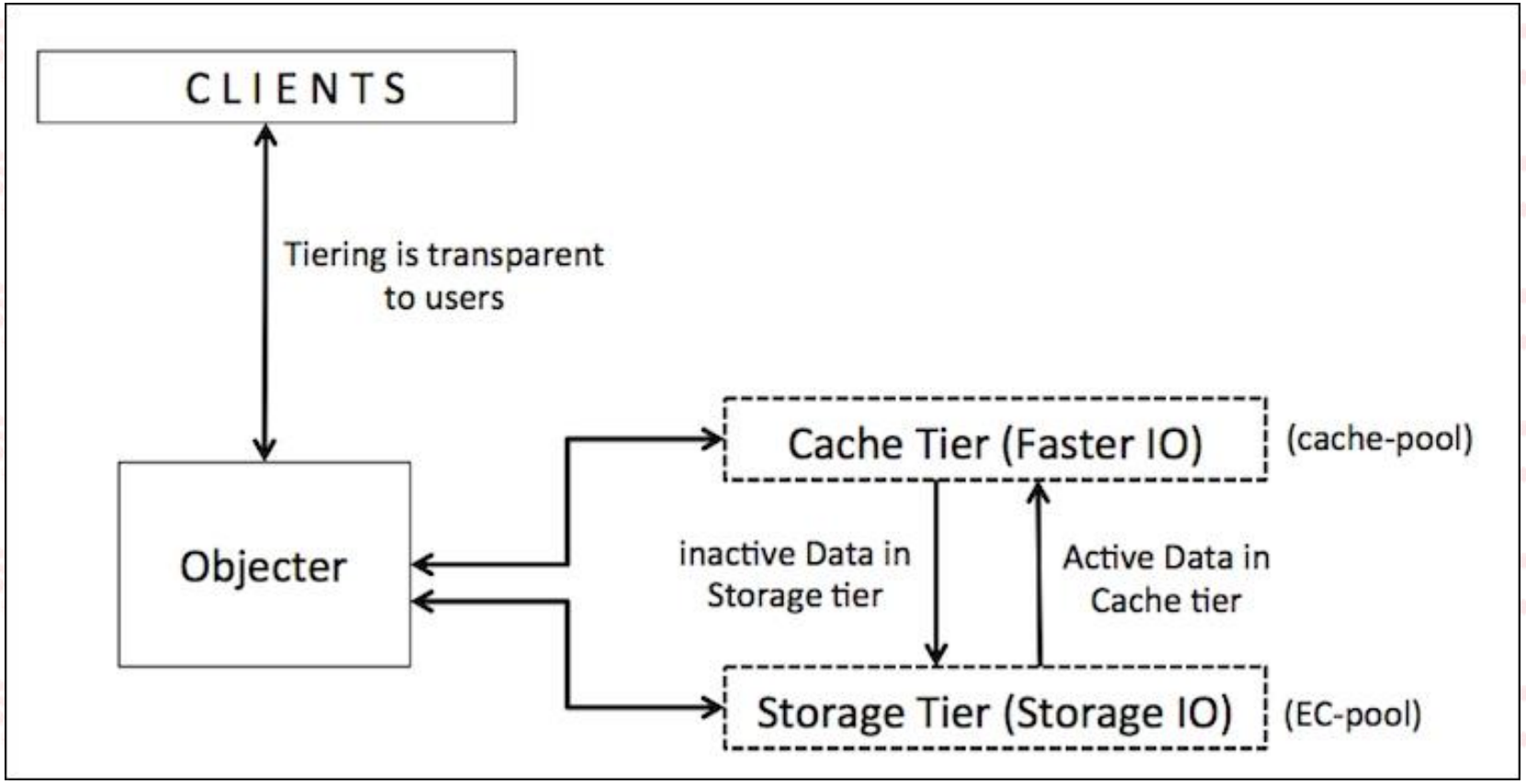


Kok Ceph? – Crush Map

- 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.

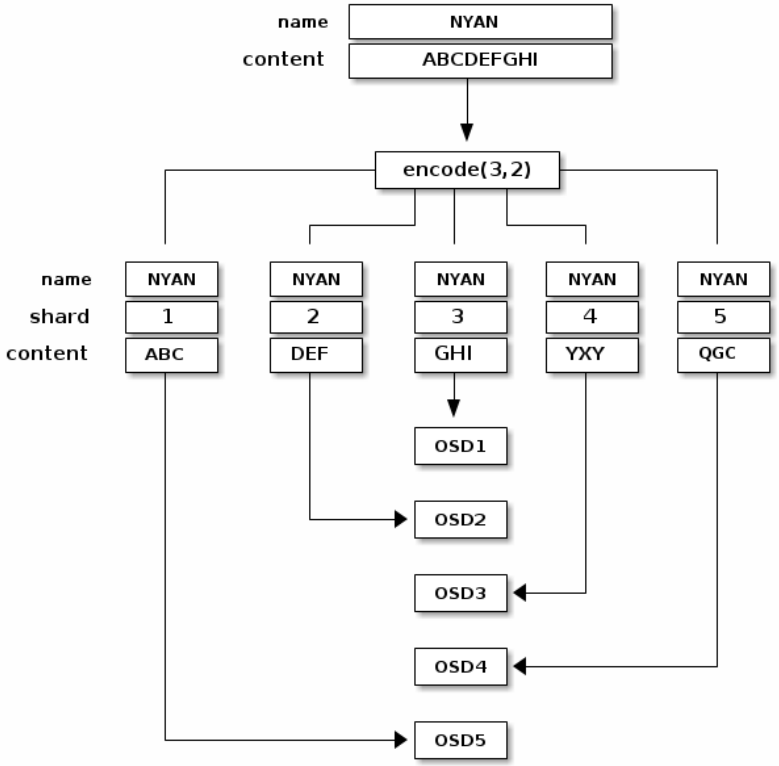


Kok Ceph? – Tiering

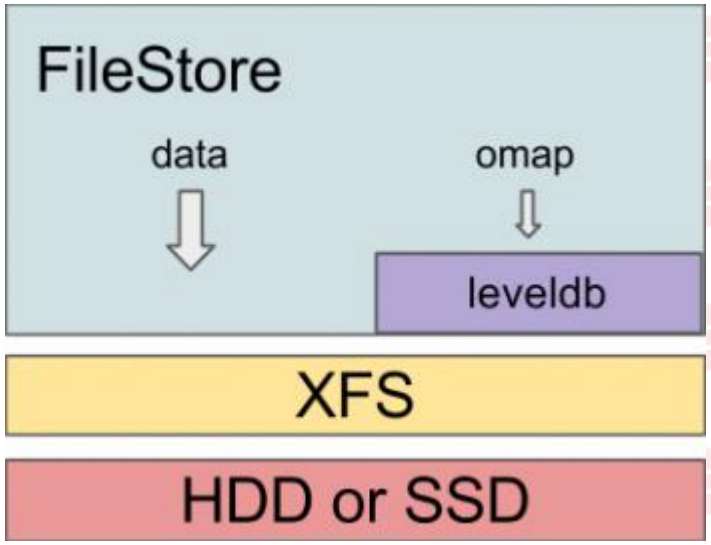


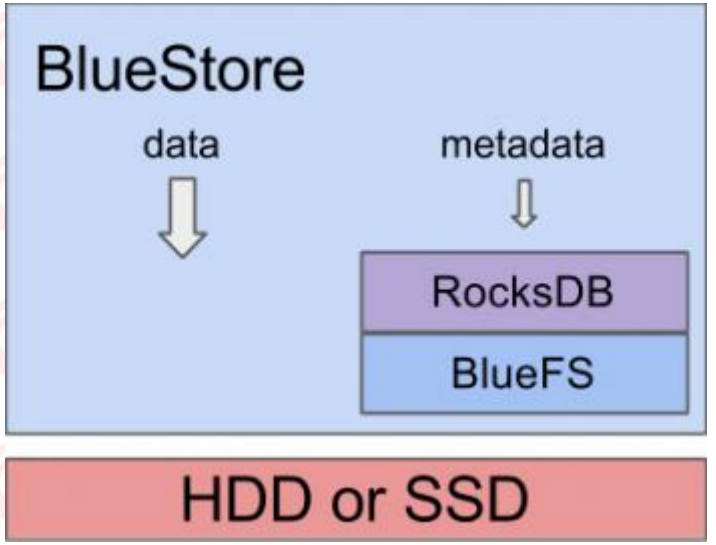
Kok Ceph? – EC

- 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

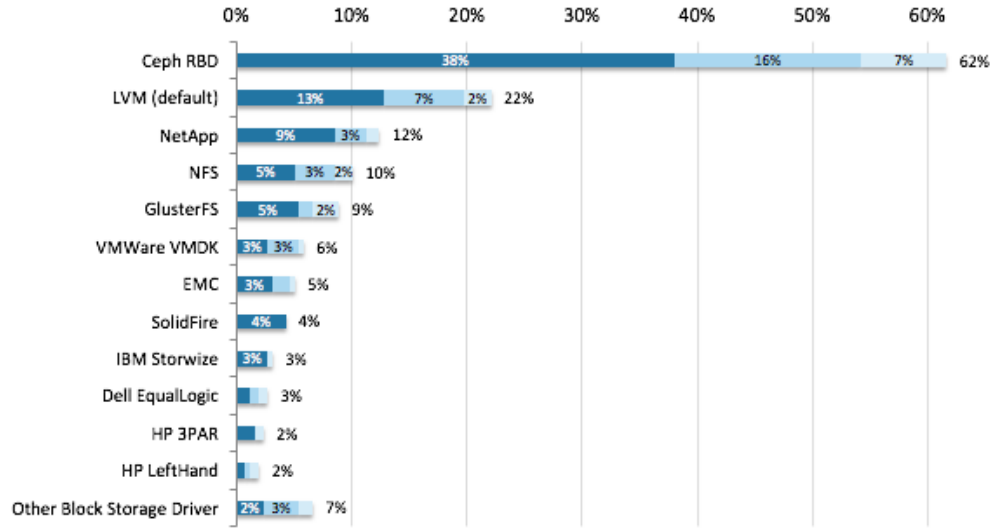




Ceph & OpenStack – Hasil Survey

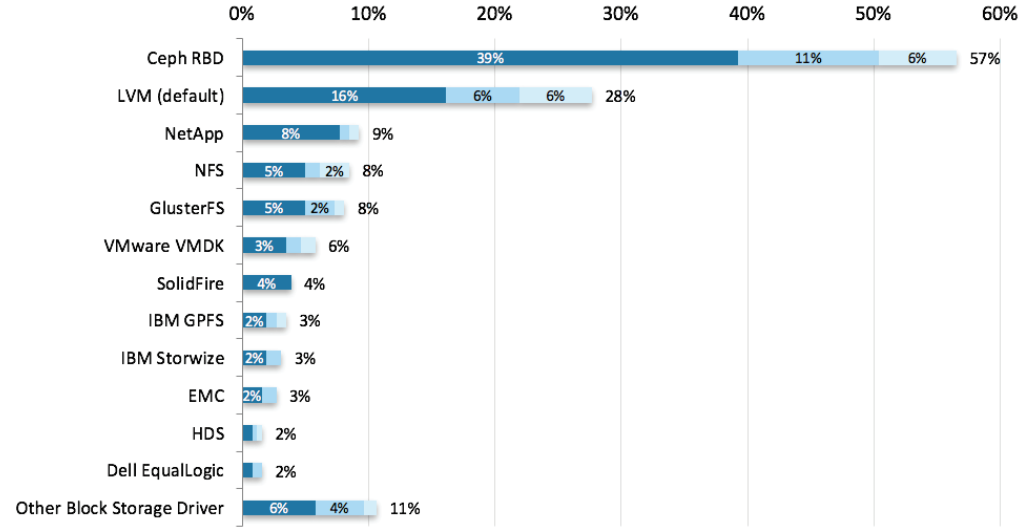
OPENSTACK USER SURVEY - OKTOBER 2015

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



OPENSTACK USER SURVEY - APRIL 2016

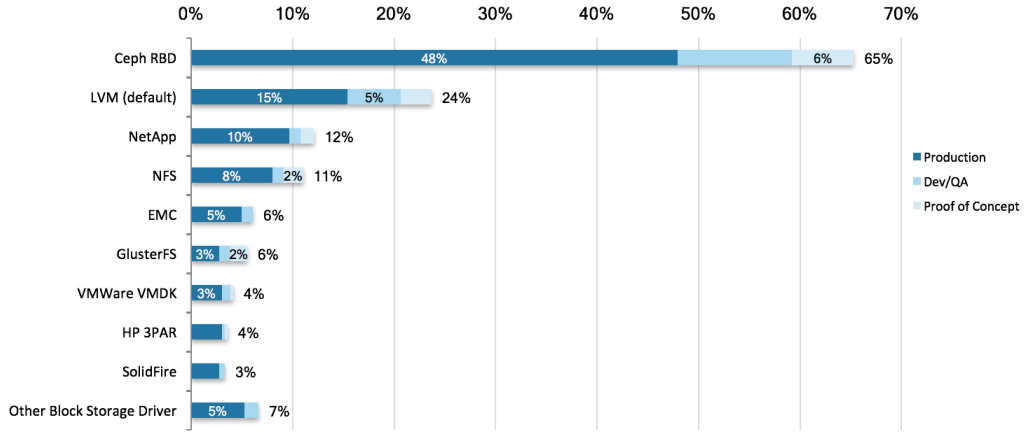
https://www.openstack.org/assets/survey/April-2016-User-Survey-Report.pdf

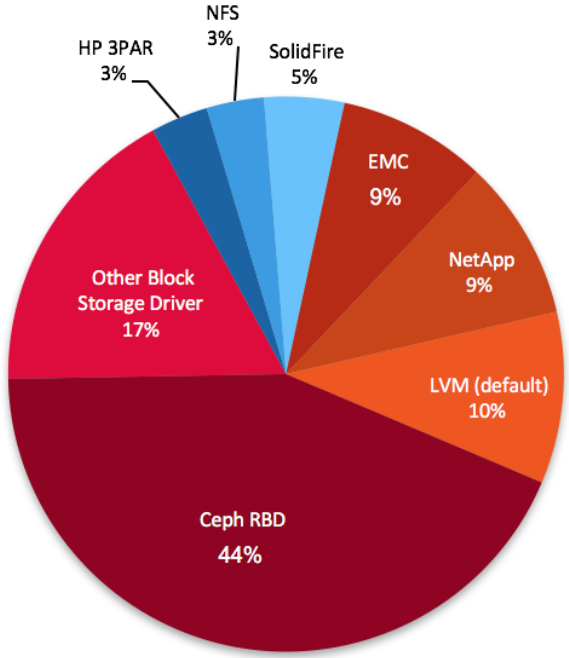


Ceph & OpenStack – Hasil Survey

OPENSTACK USER SURVEY - APRIL 2017

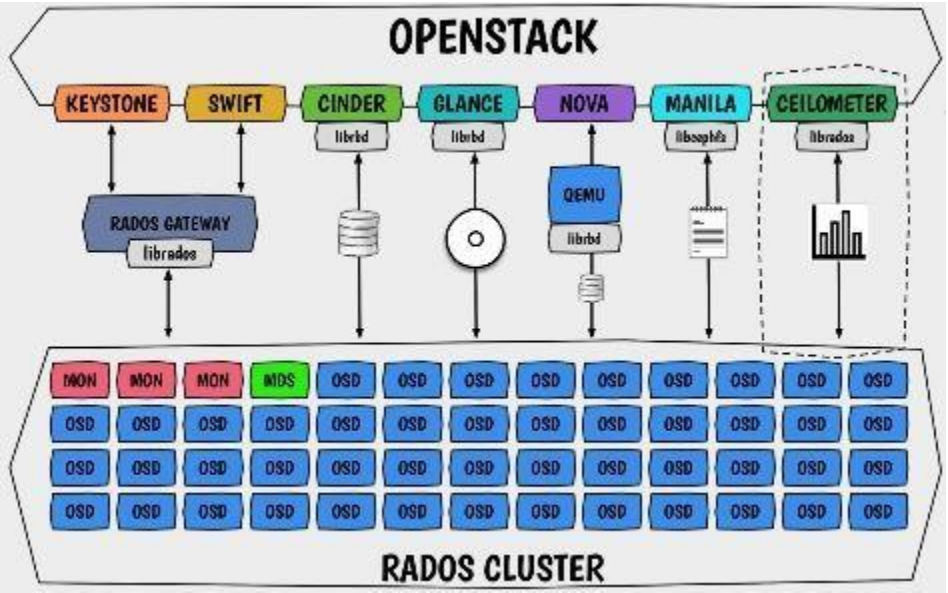
https://www.openstack.org/assets/survey/April2017SurveyReport.pdf





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

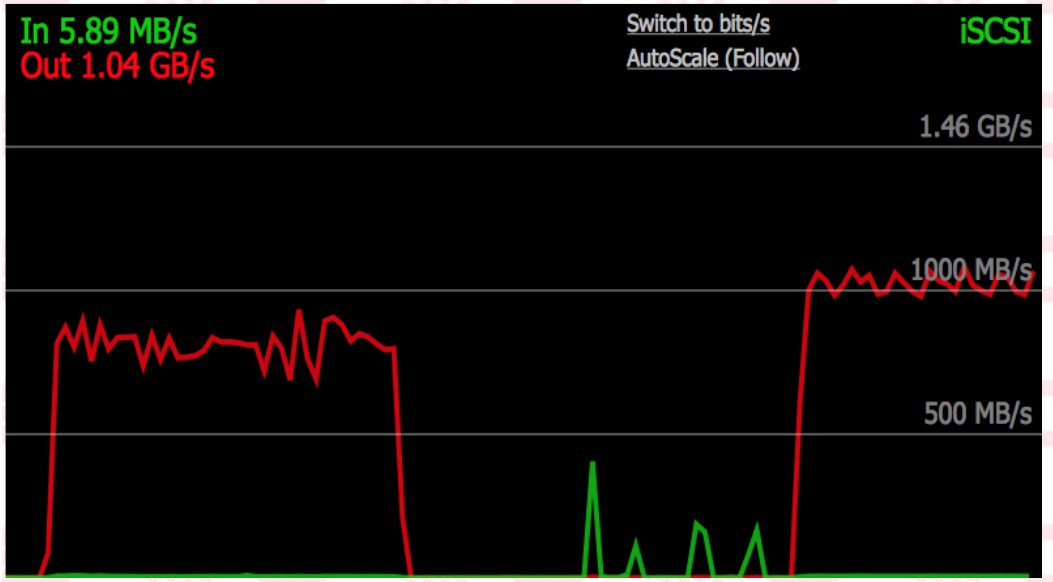


Ceph & OpenStack – Pengujian

SPESIFIKASI

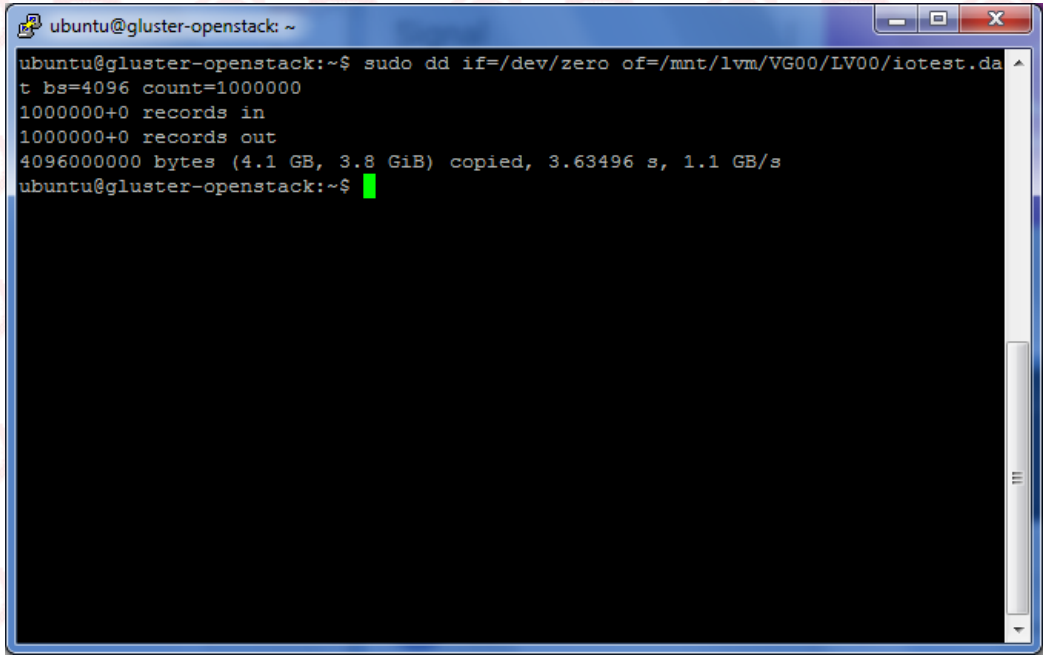
- Fisik
  - CPU: 2x E5-2667v4
  - RAM: 16x 16GB
  - Link: 2x 10GBase-T
- VM
  - vCPU: 8
  - vRAM: 16GB
  - Volume: 40GB + 5x 10TB

XIGMANAS DASHBOARD

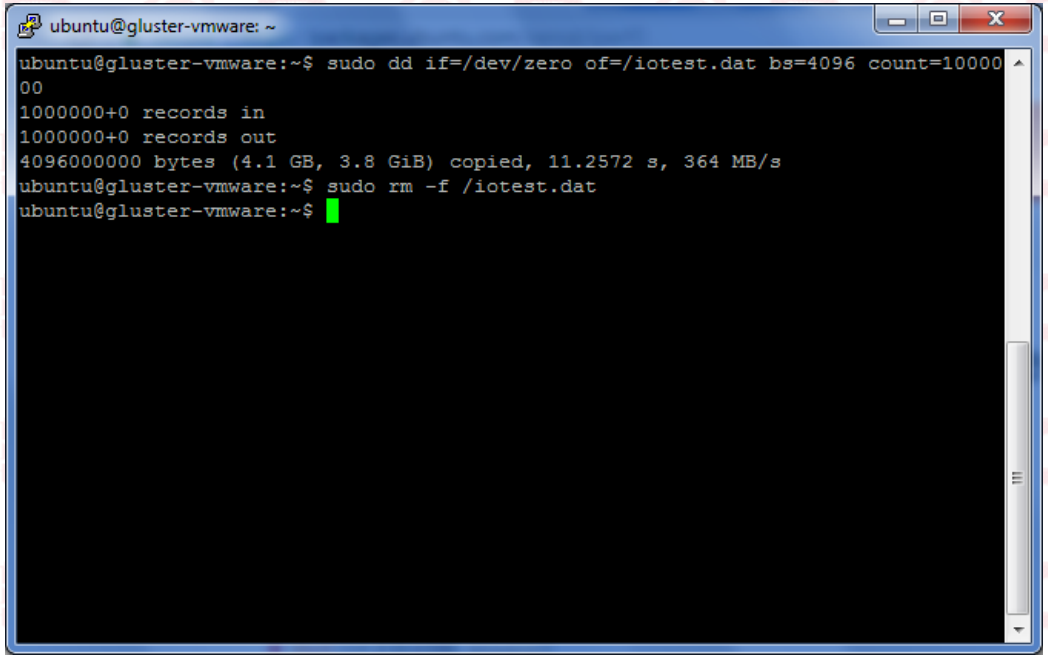


Ceph & OpenStack – Pengujian

DD DI OPENSTACK + CEPH

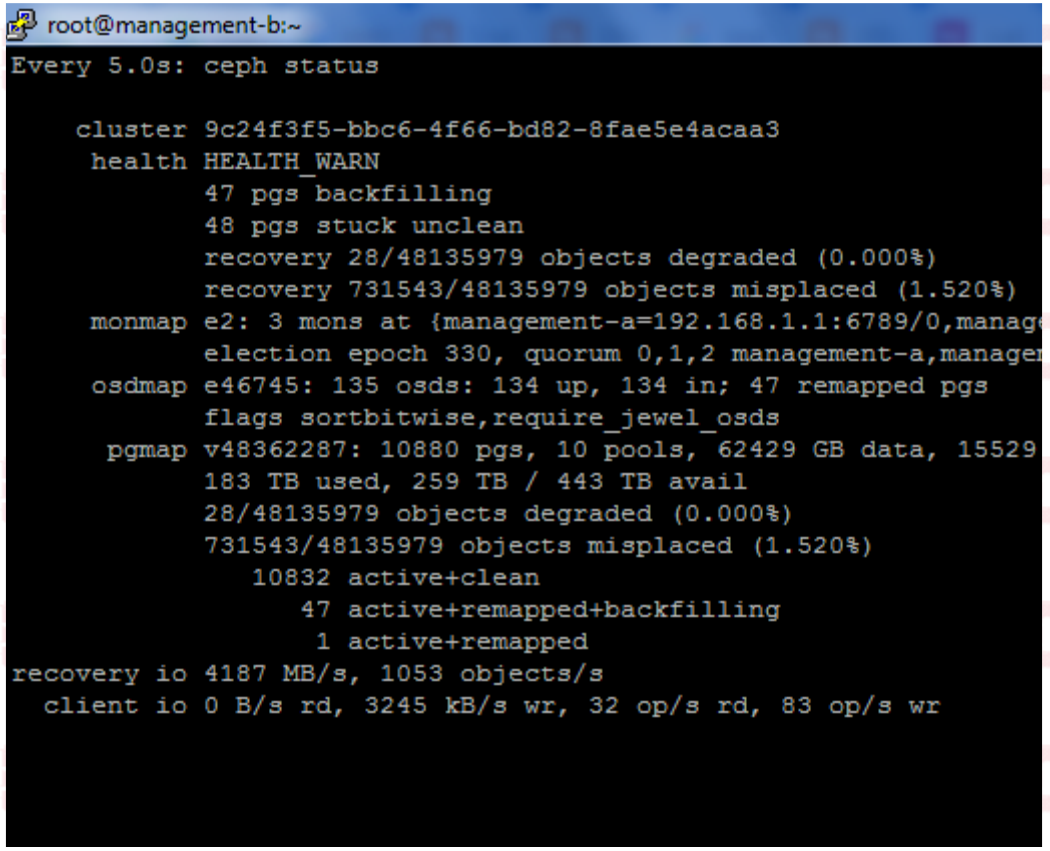


DD DI VMWARE + HP MSA 2040

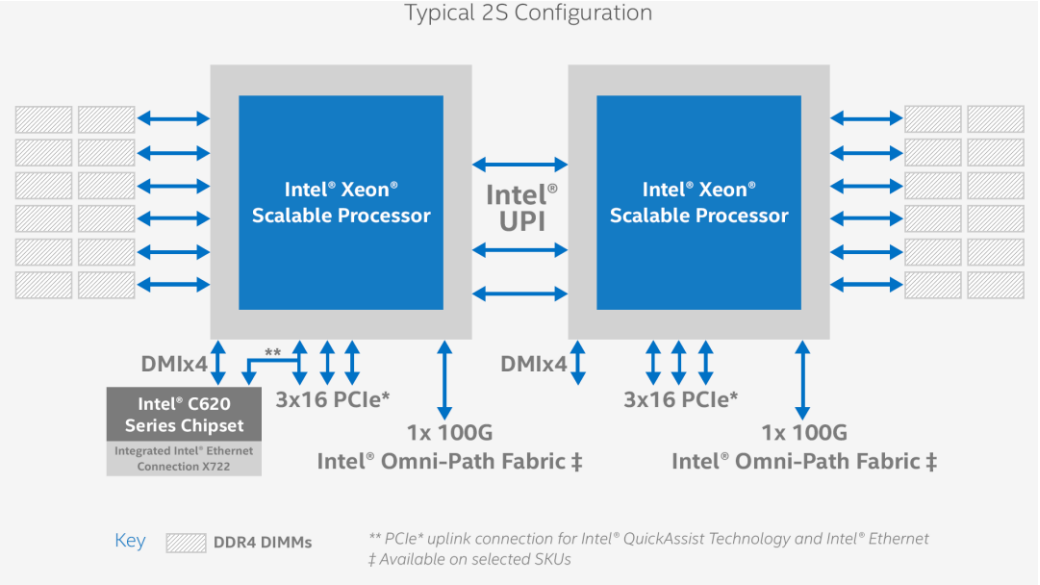


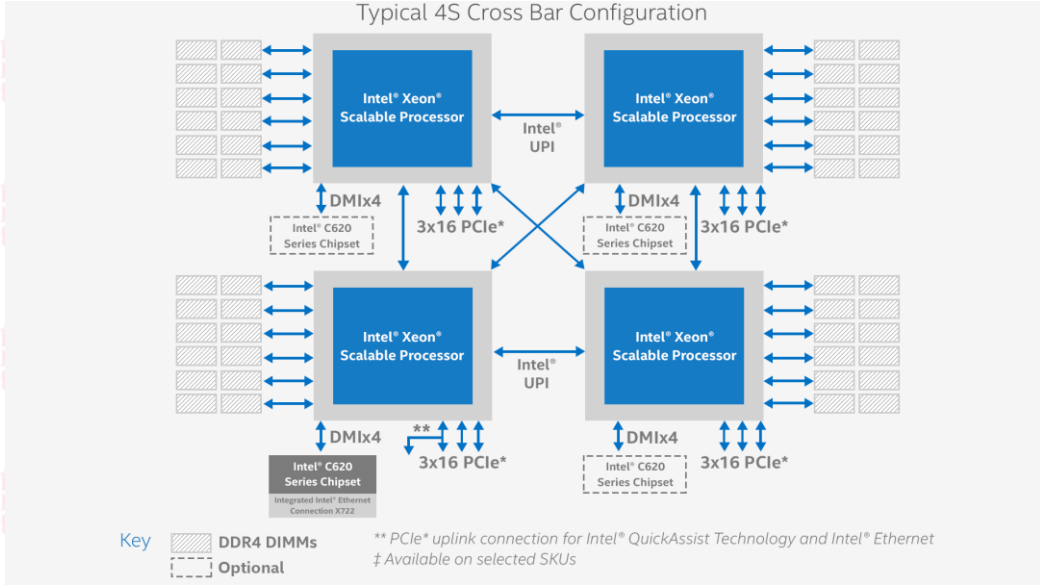
Hati-Hati! – Networking

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



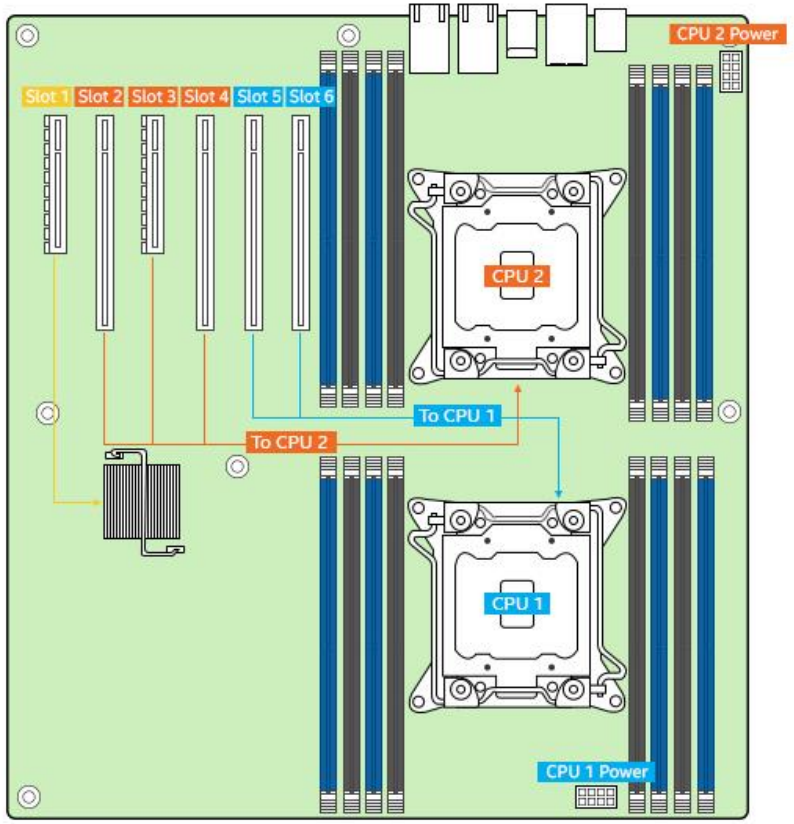
Hati-Hati! – CPU





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