ceph

CEPH

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

Ceph Goes Online at Qihoo 360

Xu Xuehan
xuxuehan@360.cn

Outline

- Motivation
- Ceph RBD
- CephFS

Motivation

Products at Qihoo 360

花椒直播
快视频
360游戏
玩 出 伟 大
360
安全浏览器
安全卫士
小水滴
360
儿童手表
360
开放平台
360搜索
360
手机助手
商业广告

Motivation

- **Virtualization**
- Benefits
  - Avoidance of hardware resource waste
  - Ease of products deployment

- Not all problems solved
  - Long VM failover interval
  - Long VM creation time

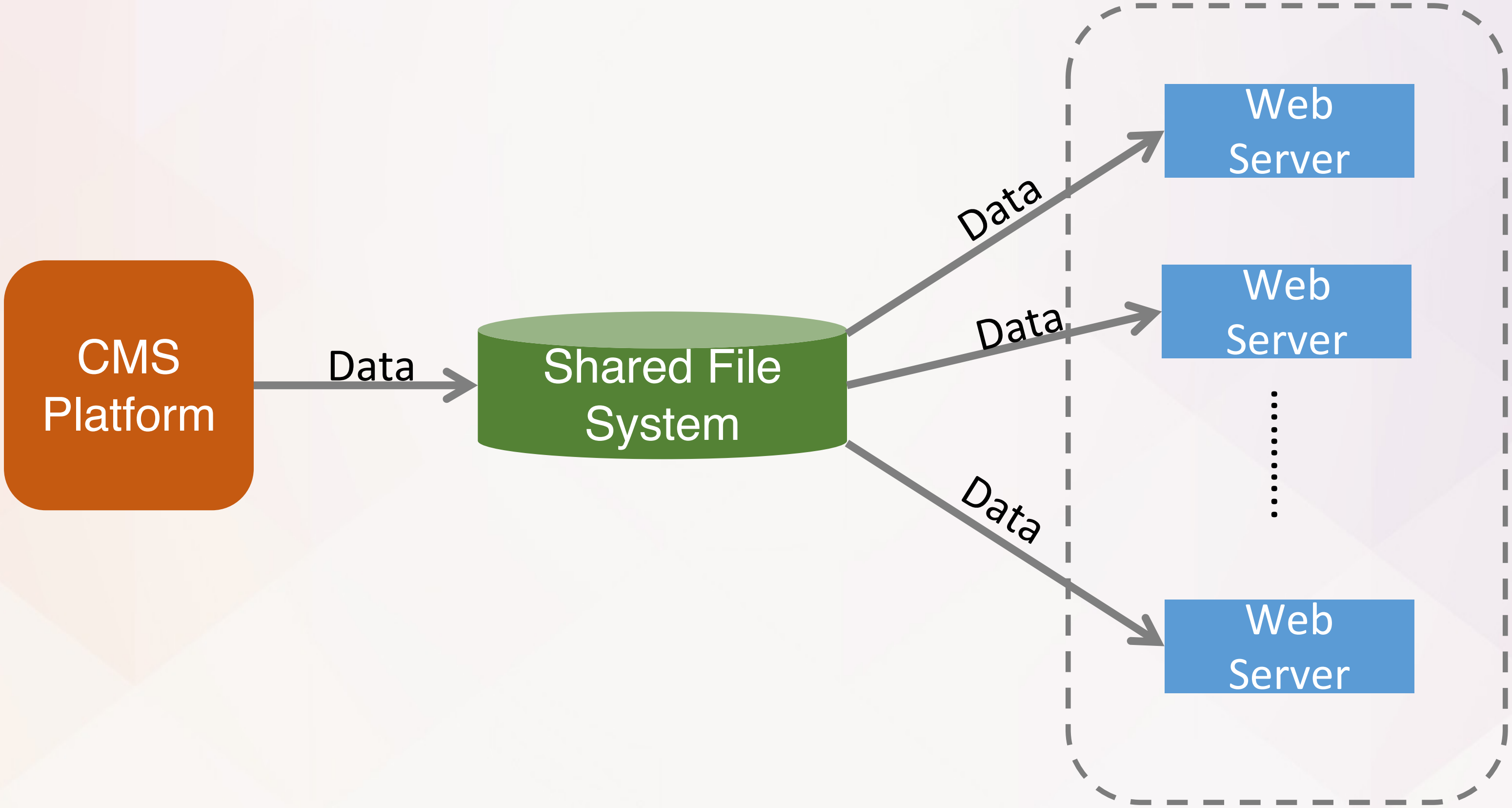
→ **Need for a separated VM image storage backend**

Motivation

- **Need for a separated VM storage backend**
  - Ceph RBD
    - Separation of Computation and Storage.
    - Scalable storage
    - Open source & active community support

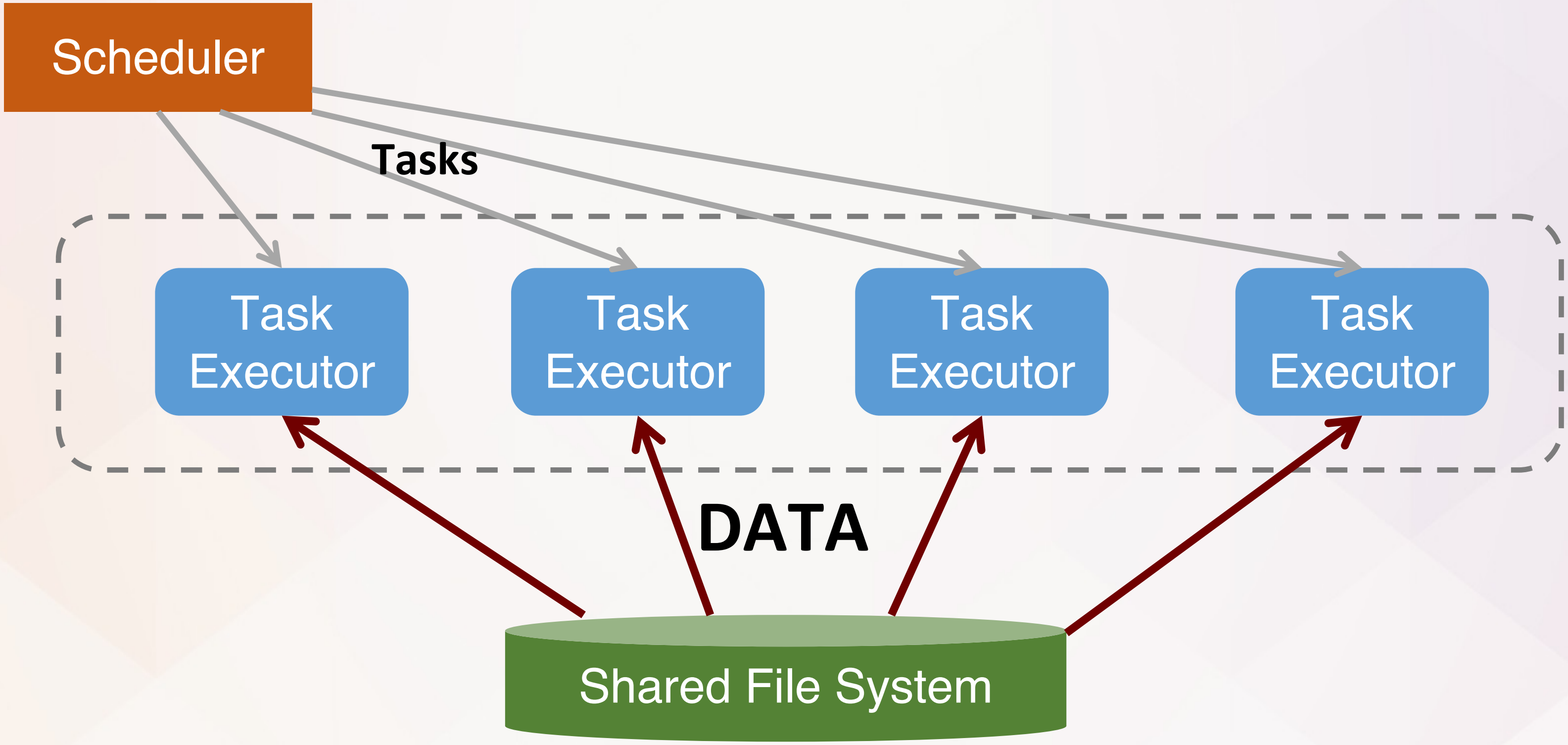
Motivation

- **Need for a shared file system**



Motivation

- **Need for a shared file system**



ceph

- **Need for a shared file system**
  - CephFS
    - POSIX compliance
    - Read-after-write consistency
    - Scalability

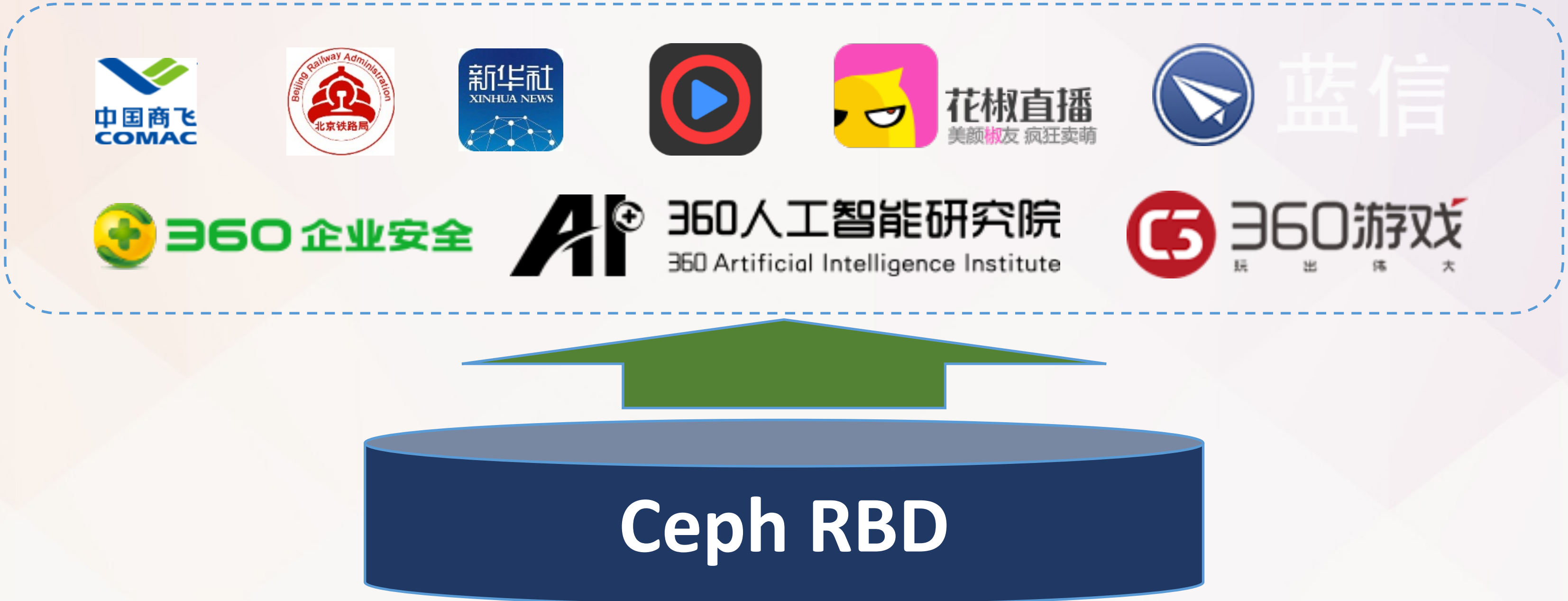
Outline

- Motivation
- Ceph RBD
- CephFS

Ceph RBD

Production Deployment

- 500+ Nodes
- 30+ Clusters
- Largest Cluster: 135 nodes, 1000+ OSDs
- Hammer 0.94.5, Jewel 10.2.5;



Ceph RBD

Ceph RBD

- **Online Clusters**
  - Cost VS Performance
    - Full SSD cluster, for users sensitive to I/O latency(Game Server, etc)

	OSD Nodes
CPU	Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz
RAM	128GB
NIC	10GbE
Hard Drives	8*SSD(SDLF1DAM-800G-1HA1)

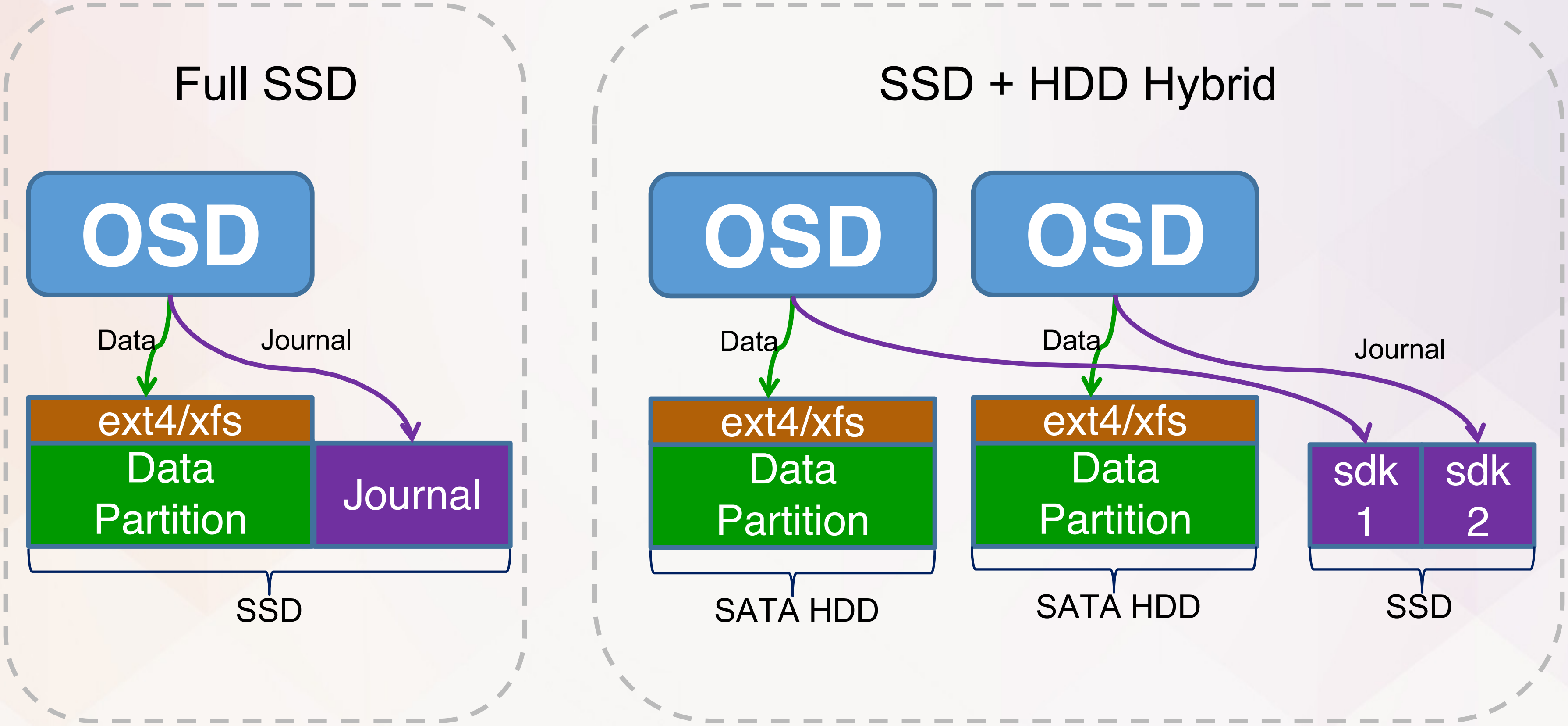
    - SSD + HDD hybrid cluster, for other users

	OSD Nodes
CPU	Intel(R) Xeon(R) CPU E5-2430 v2 @ 2.50GHz
RAM	64GB
NIC	10GbE
Hard Drives	2*SSD(INTEL SSDSC2BB300G4) + 9*HDD(WDC WD4000FYYZ-03UL1B2)

Ceph RBD

Online Clusters

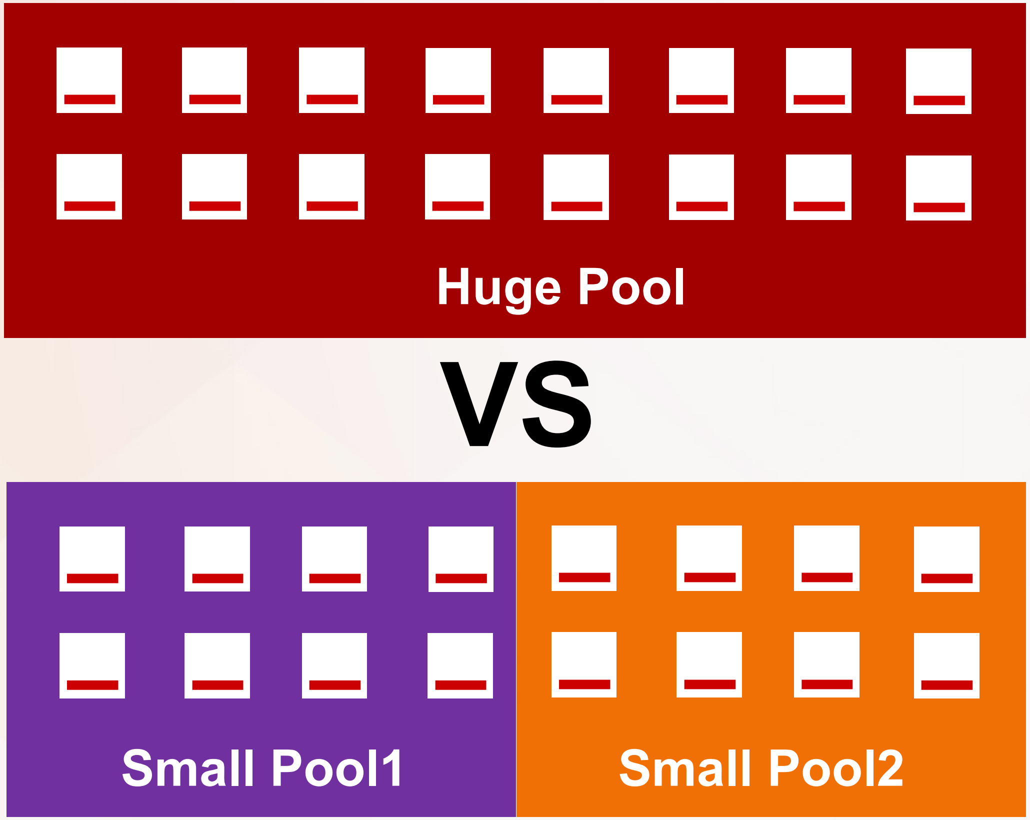
– Cost VS Performance



Ceph RBD

Online Clusters

- One Big Pool *or* Multiple Small Pools?
  - PipeMessenger (Default in Hammer/Jewel): two threads per connection



Max_threads_per_osd =
2*(clients + 1 + 6*pg_num_per_osd)

Too many OSDs in one pool could lead to too many threads in one Machine!

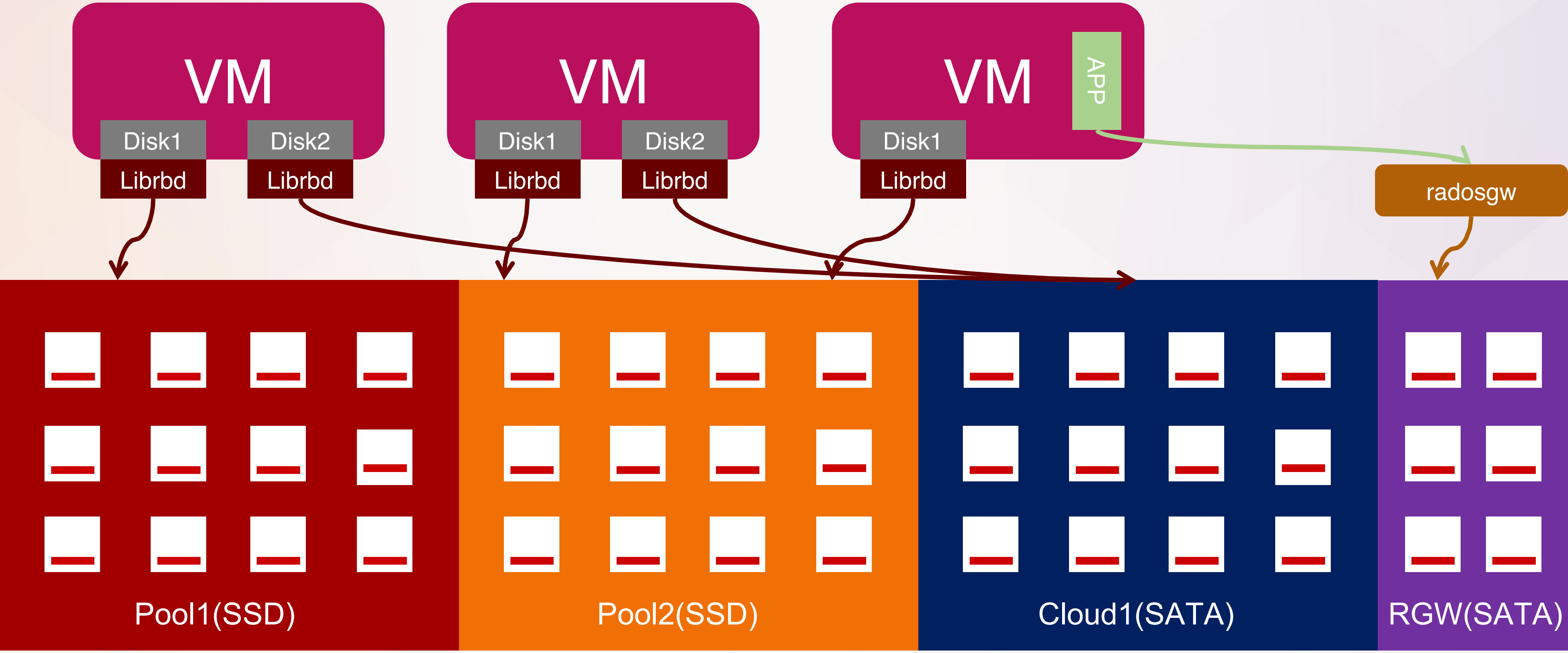
Any other problem with huge pool?

WE DON'T KNOW YET

Ceph RBD

Online Clusters

- One Big Pool *or* Multiple Small Pools?
  - Typical Deployment Pattern



RADOS CLUSTER

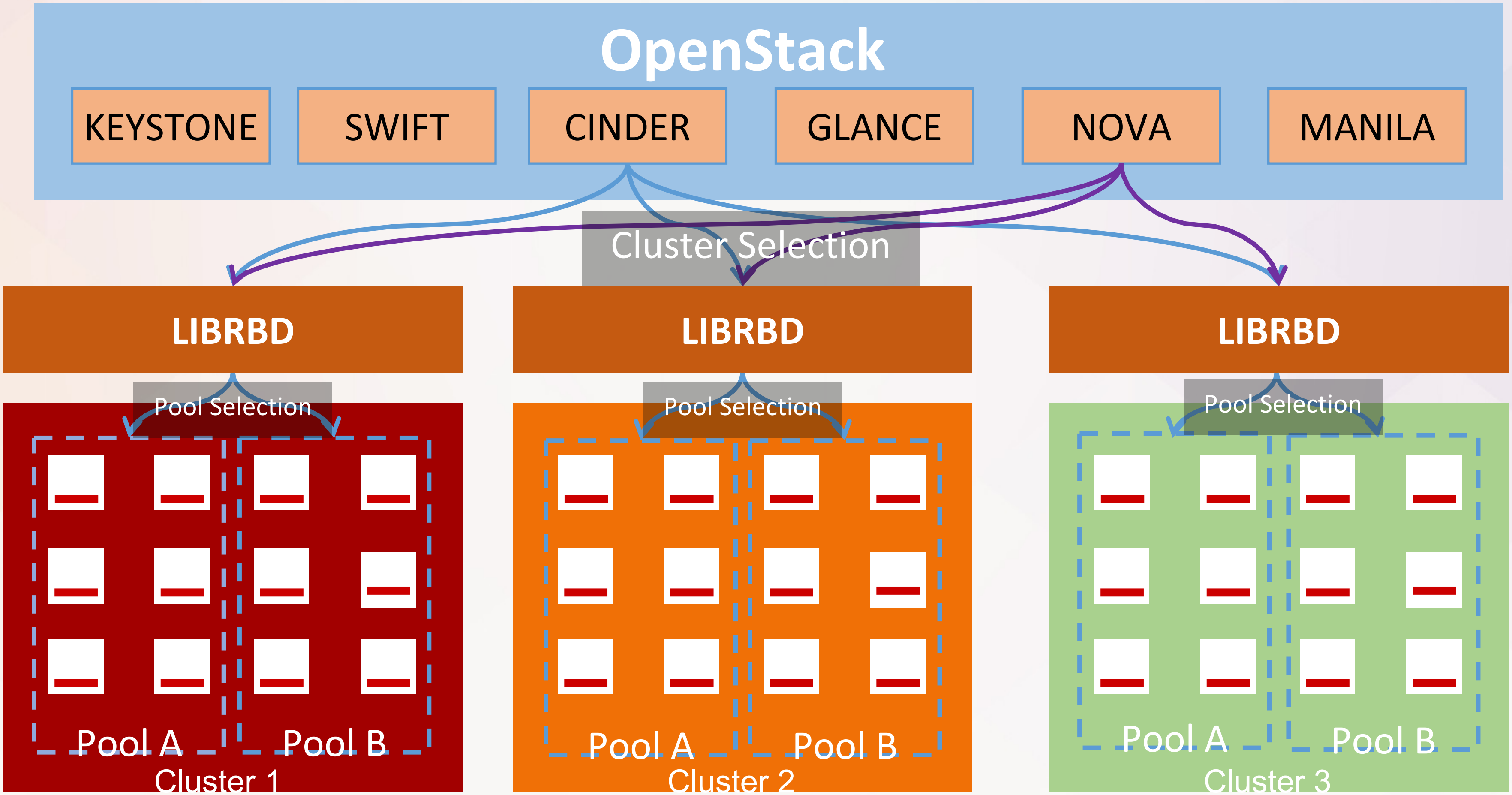
Ceph RBD

Online Clusters

- One Big Pool *or* Multiple Small Pools?(Openstack Modification)
  - OpenStack support for multi-pool(supported upstream now) and multi-cluster
  - VM creation based on rbd image clone(supported upstream now)
  - Flatten after VM image creation(supported upstream now)

Online Clusters

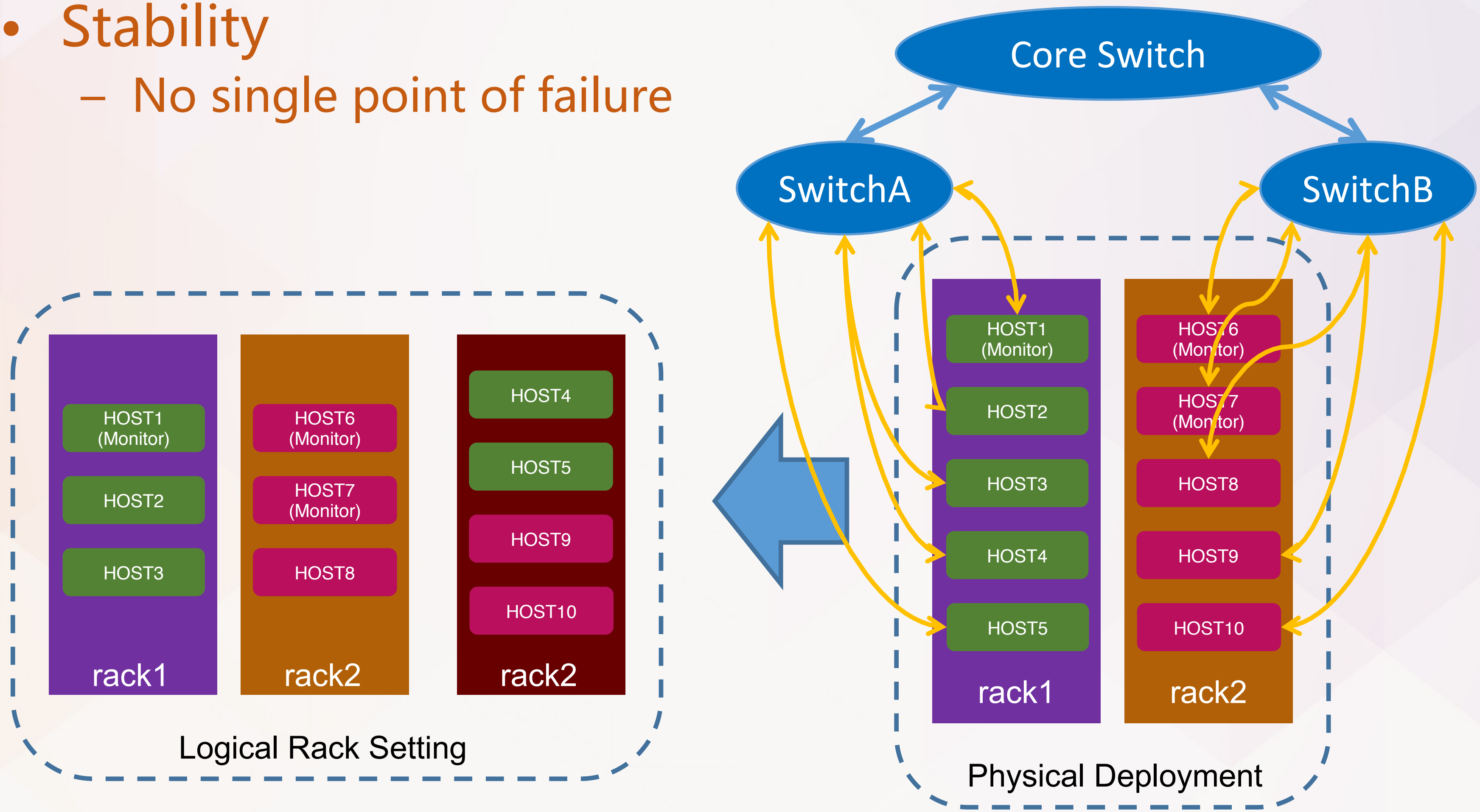
– Final Deployment Pattern



Ceph RBD

Online Clusters

- Stability
  - No single point of failure



Ceph RBD

Online Clusters

- QoS
  - Done in QEMU
  - Full SSD clusters:
    - IOPS: 10 iops/GB
    - Throughput: 260 MB/s
  - SSD + HDD Hybrid clusters:
    - READ iops: 1400
    - WRITE iops: 300~600 iops
    - Throughput: 70 MB/s

Ceph RBD

Online Clusters

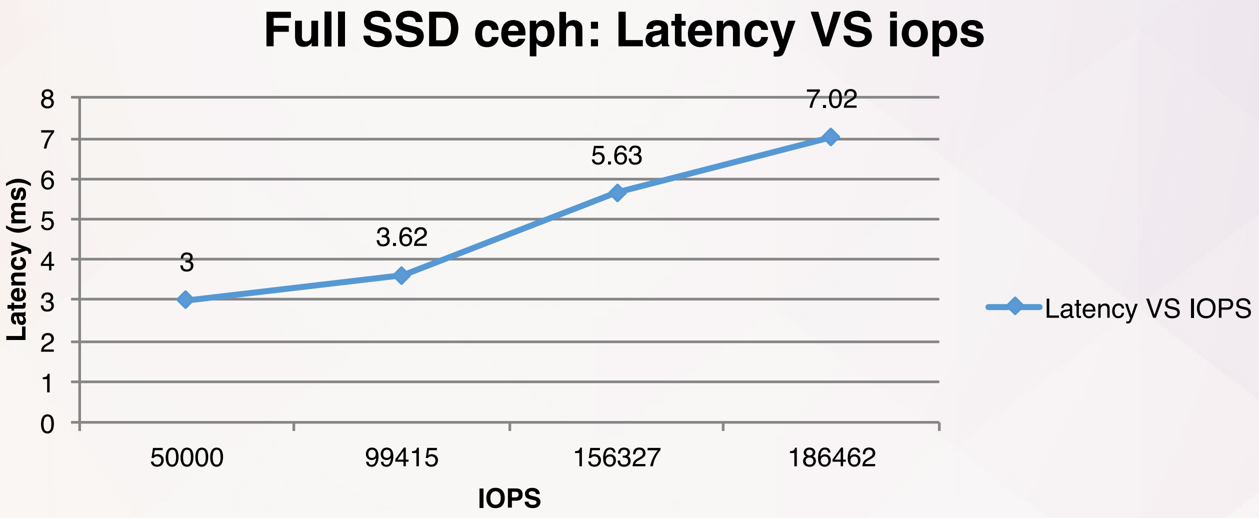
- Capacity
  - Thin provisioning
  - Capacity requirement prediction:

    $C_{total} = (N_{VM\_num} * C_{capacity\_per\_vm}) / (T_{thin\_provision\_ratio} * 0.7)$
  - Create new pool instead of pool expansion

Ceph RBD

Online Experience

- High I/O util on Full SSD cluster
  - I/O utils: 10%+(Full SSD Ceph) **VS** 1%-(Local disk)
  - Users may complain, but NOT a problem



Ceph RBD

Online Experience

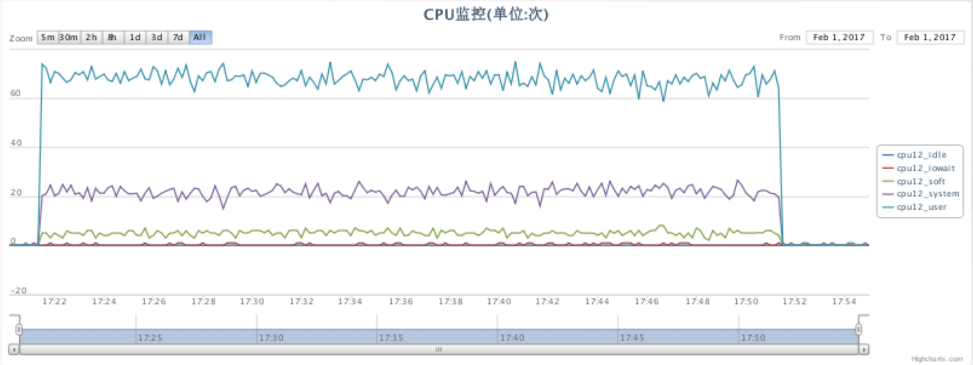
- Burst image deletion
  - Users remove massive images all at once
  - Cluster almost not available
  - Solution:
    - Modify openstack to remove images asynchronously, do concurrency control
- Scrub
  - Could severely impact I/O performance
  - Only between 2:00 and 6:00 AM

Ceph RBD

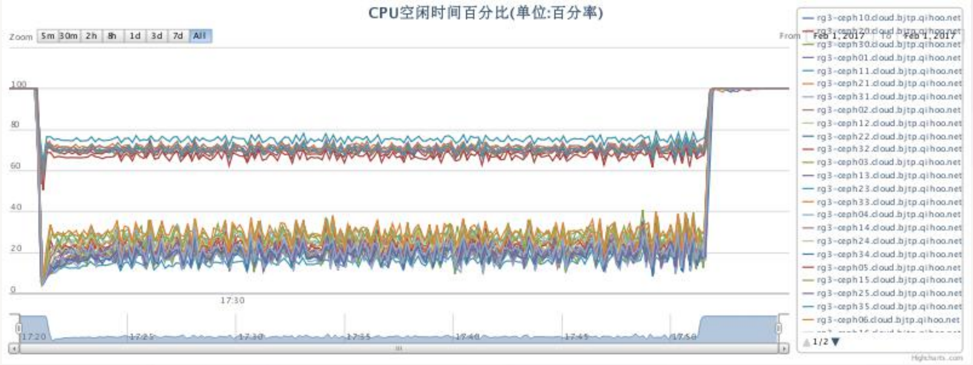
Online Experience

- Full SSD ceph (Hammer): really cpu consuming
  - 10^4 IOPS per CPU (CPU: Xeon E5-2630 v3)
  - The more SSDs per cpu, the less IOPS per cpu

Single CPU states



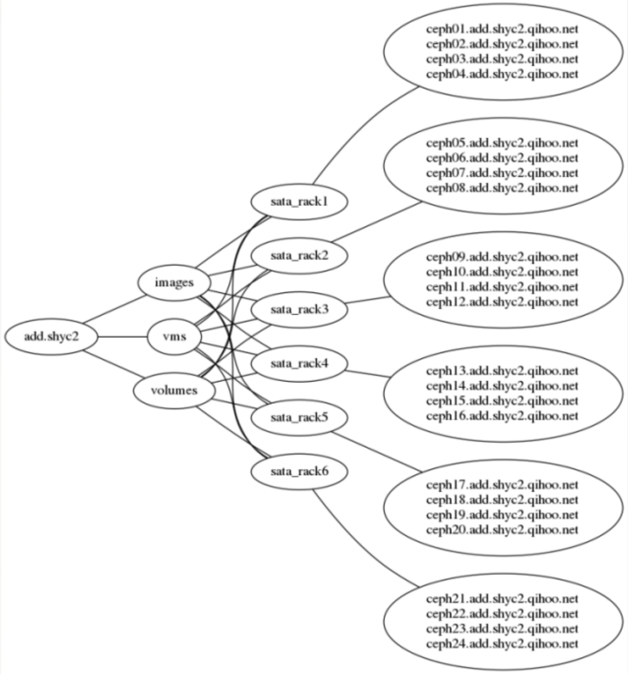
CPU %idle



Ceph RBD

Online Experience

- One OSD full == Cluster full (Hammer, Jewel)
- Daily Inspection: An intuitional way to observe cluster topology maybe needed
  - For now, we use a script to draw a topology graph


add.shyc2
images
vms
volumes
sata_rack1
sata_rack2
sata_rack3
sata_rack4
sata_rack5
sata_rack6
ceph01.add.shyc2.qihoo.net
ceph02.add.shyc2.qihoo.net
ceph03.add.shyc2.qihoo.net
ceph04.add.shyc2.qihoo.net
ceph05.add.shyc2.qihoo.net
ceph06.add.shyc2.qihoo.net
ceph07.add.shyc2.qihoo.net
ceph08.add.shyc2.qihoo.net
ceph09.add.shyc2.qihoo.net
ceph10.add.shyc2.qihoo.net
ceph11.add.shyc2.qihoo.net
ceph12.add.shyc2.qihoo.net
ceph13.add.shyc2.qihoo.net
ceph14.add.shyc2.qihoo.net
ceph15.add.shyc2.qihoo.net
ceph16.add.shyc2.qihoo.net
ceph17.add.shyc2.qihoo.net
ceph18.add.shyc2.qihoo.net
ceph19.add.shyc2.qihoo.net
ceph20.add.shyc2.qihoo.net
ceph21.add.shyc2.qihoo.net
ceph22.add.shyc2.qihoo.net
ceph23.add.shyc2.qihoo.net
ceph24.add.shyc2.qihoo.net

Ceph RBD

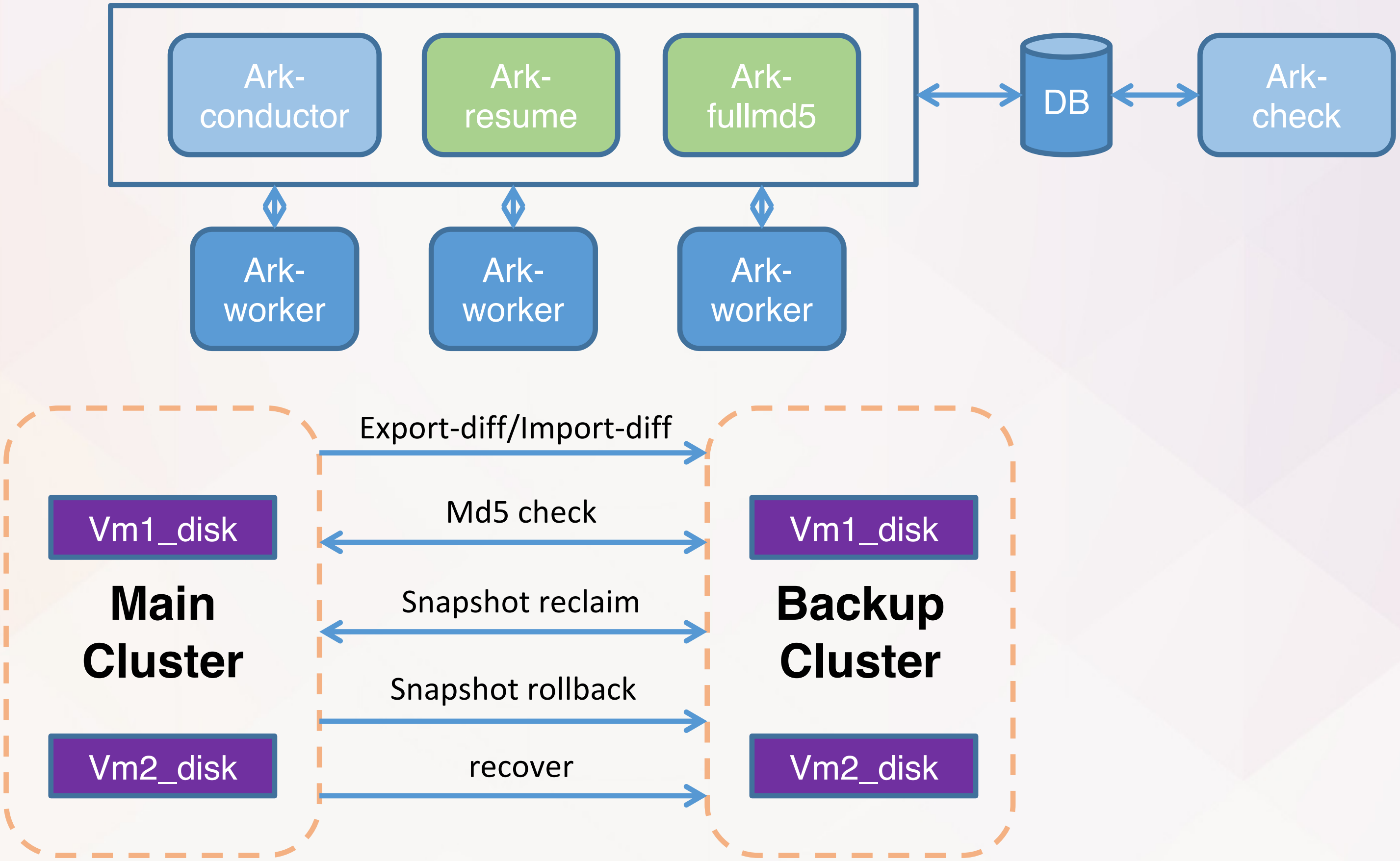
Online Experience

- Tracing
  - Hard to reproduce some online problems
  - Can’t turn on high priority log online
- Alerting
  - Integrate with other alerting services like Nagios?
  - A new alerting module? Or at least some alerting interface for users to capture exceptional events?

Ceph RBD

Online Experience

- RBD image Backup



Outline

- Motivation
- Ceph RBD
- CephFS

CephFS

- **MDS Performance Evaluation(mdtest)**
  - MDS machine

	MDS nodes
CPU	Intel(R) Xeon(R) CPU E5-2630 0 @ 2.30GHz
RAM	192GB
NIC	10GbE
OS	CentOS 7.1.1503

  - 1 active MDS, 1 standby-replay MDS
  - 27 OSDs in data pool, 6 SSD OSDs in metadata pool
  - 7*10^6 files/directories, 70 clients

ceph

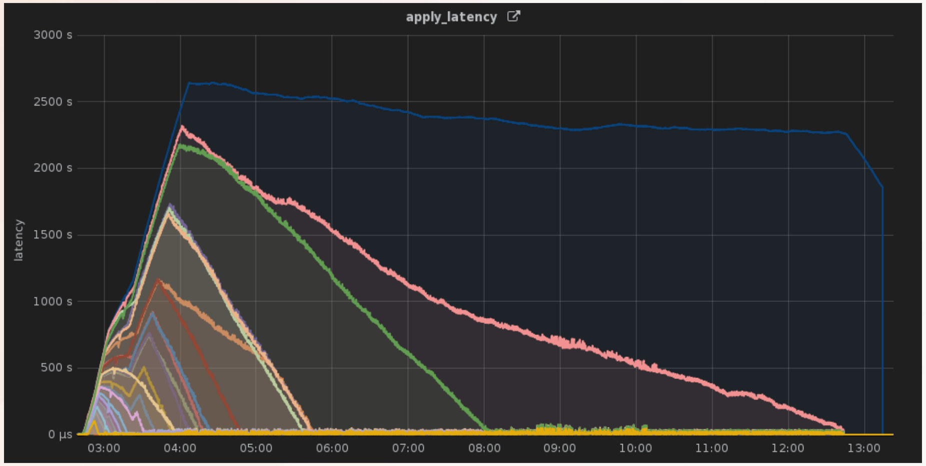
CephFS

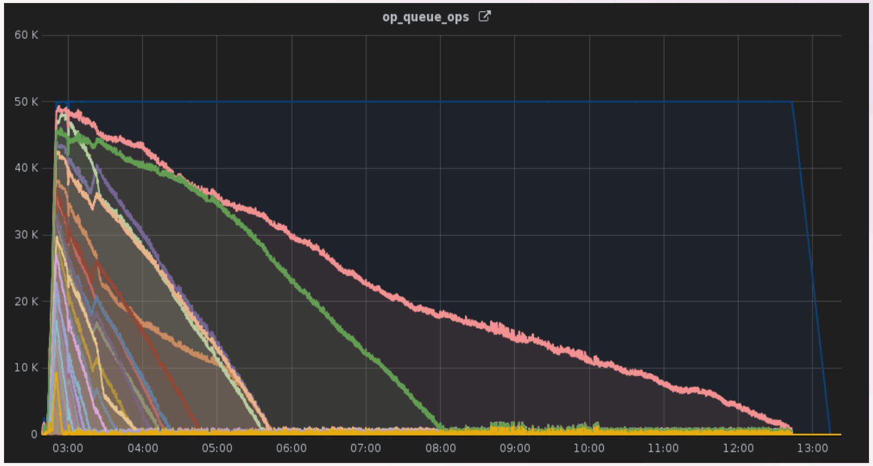
- ## MDS Performance Evaluation

Metadata Operation	Result (ops/sec)
File Creation(shared directory)	2624.09
File Creation(job separated directories)	4311.339
Stat	11000
File Removal(shared directory)	788.960
File Removal(job separated directories)	2531.538
Directory Creation(shared directory)	794.030
Directory Creation(job separated directories)	3497.949
Directory Removal(shared directory)	697.333
Directory Removal(job separated directories)	2848.889
File Open	6757.269269
File Rename(shared directory)	485.083123
File Rename(job separated directories)	3073.370671
Utime	2947.364765
Readdir	243844.3312

CephFS

- **MDS Performance Evaluation**
  - Slow metadata modification writeback
    - Caused by O(n) list::size() in gcc earlier than 5.0
    - https://github.com/ceph/ceph/commit/7e0a27a5c8b7d12d378de4d700ed7a95af7860c3





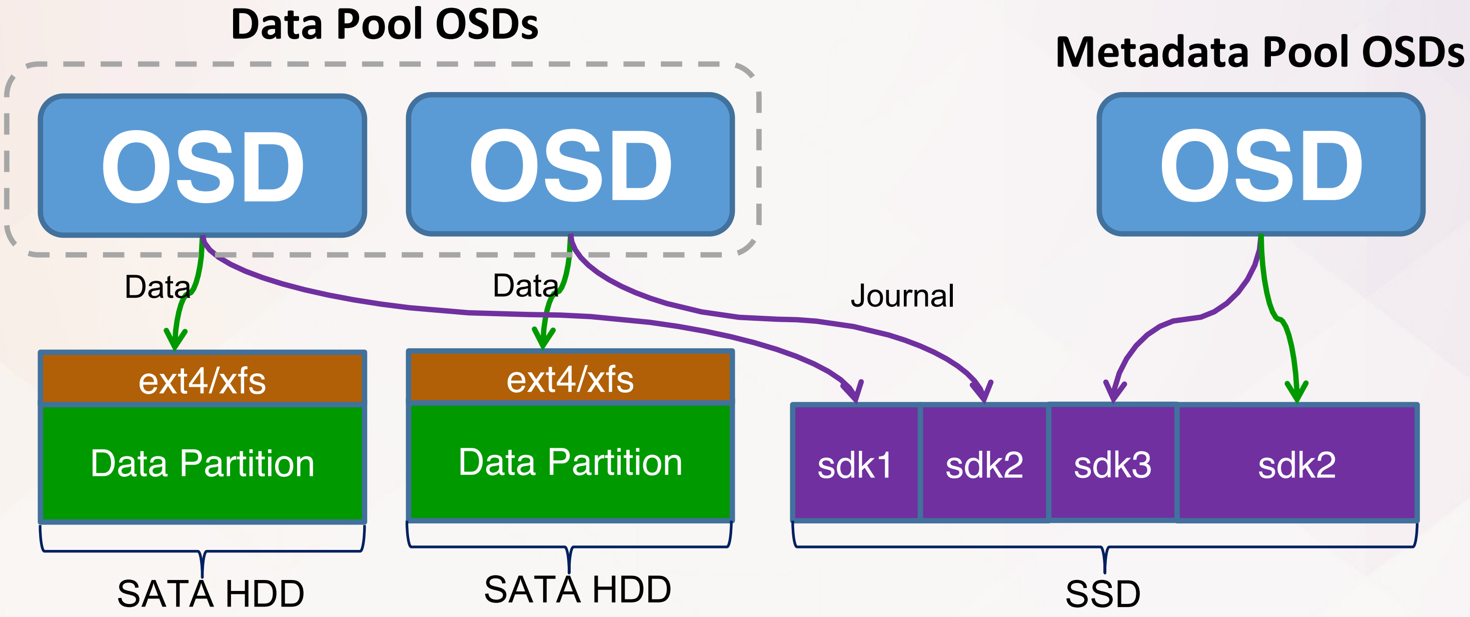
  - Single Thread MDS, low cpu utilization

CephFS

- **Considerations about putting CephFS online**
  - Access Control
    - Namespace?
    - Kernel limitation → One pool for each user☹
  - Active-Active MDS or Active-Standby MDS?
  - QoS
    - No available solution yet

CephFS

- **Online Clusters**
  - 3 small Clusters
    - 1 active MDS, 1 standby-replay MDS
    - 3 OSD/MON machines, 27 OSDs in data pool, 6 SSD OSDs in metadata pool



CephFS

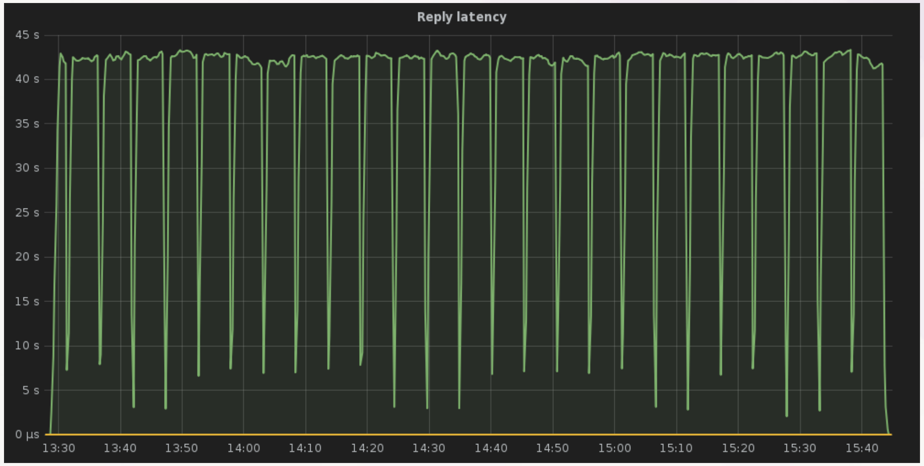
Online Experience

- mds “r” cap must be given to every user
  - Users may see directory subtree structures of each others.
- Kernel limitations
  - Most users use CentOS 7.4, kernel 3.10.0-693, many patches are **NOT** backported
  - Some users run kernel 2.6.32....☹
  - Could NFS or Samba be a solution

CephFS

Online Experience

- Slow “getattr” when lots of clients are issuing reads/writes
  - http://tracker.ceph.com/issues/22925
  - getattrs are blocked in filelock’s LOCK_SYNC_MIX state
  - When filelock gets out LOCK_SYNC_MIX state, mds has to reprocess all blocked getattrs one by one
  - Almost every getattr request make filelock go into LOCK_SYNC_MIX state again



ceph

Thanks
Q&A