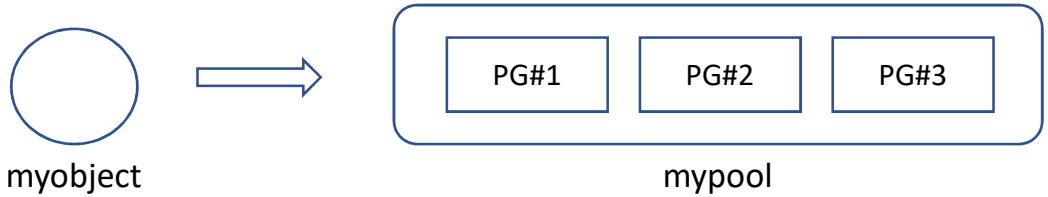
Ceph & RocksDB

변일수(Cloud Storage팀)

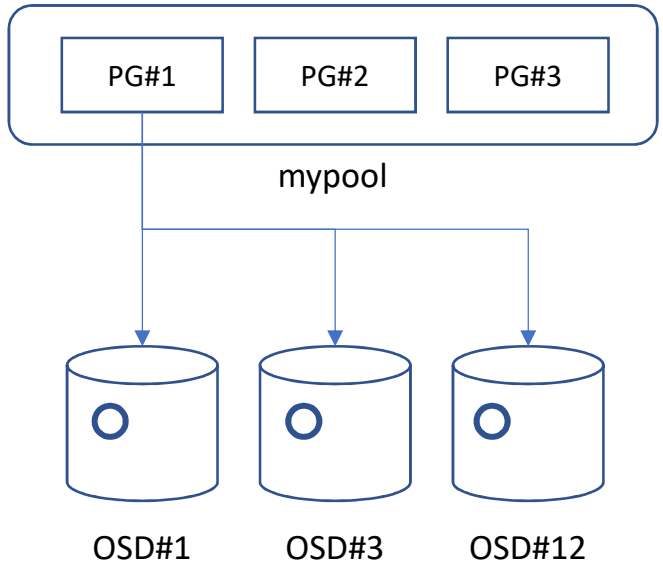
Ceph Basics

Placement Group

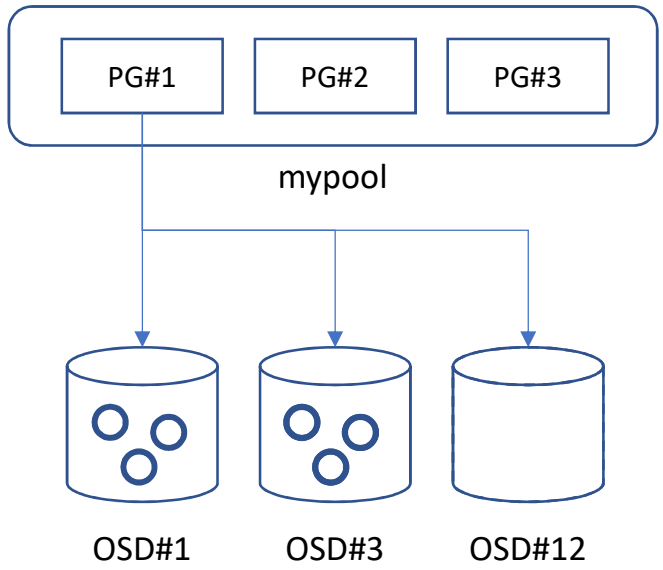


hash(myobject) = 4 % 3(# of PGs) = 1 ← Target PG

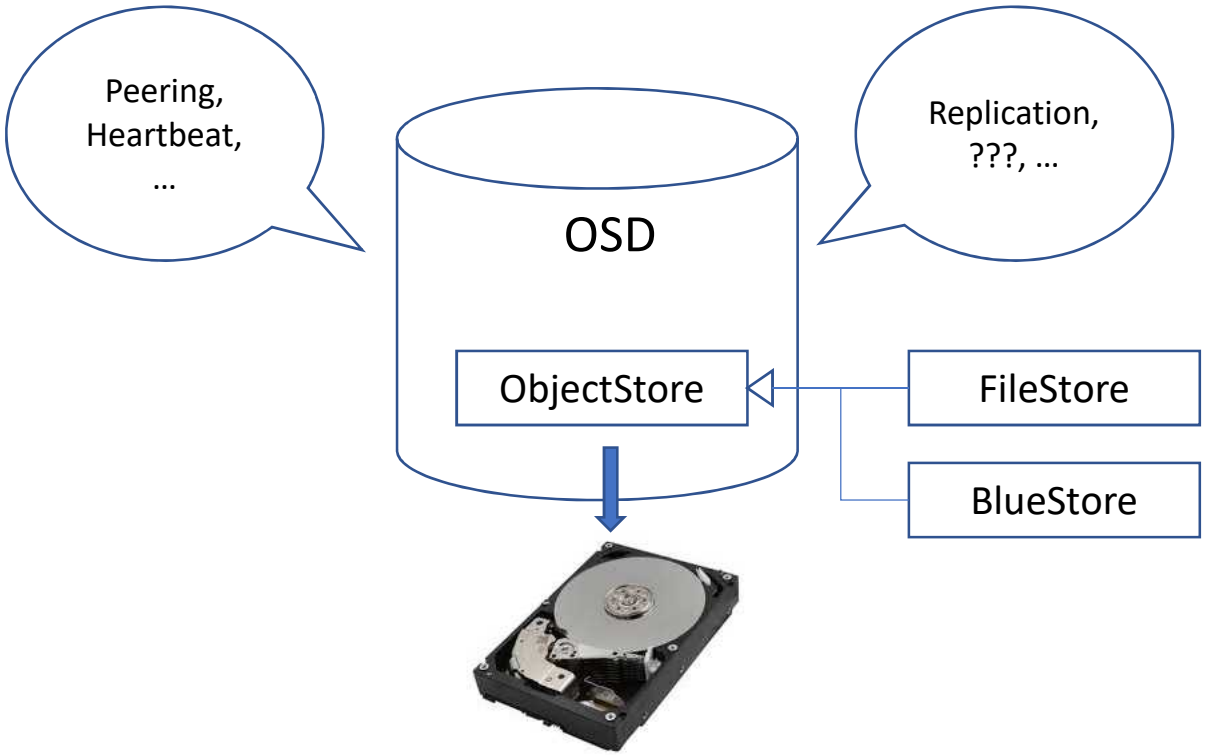
CRUSH



Recovery

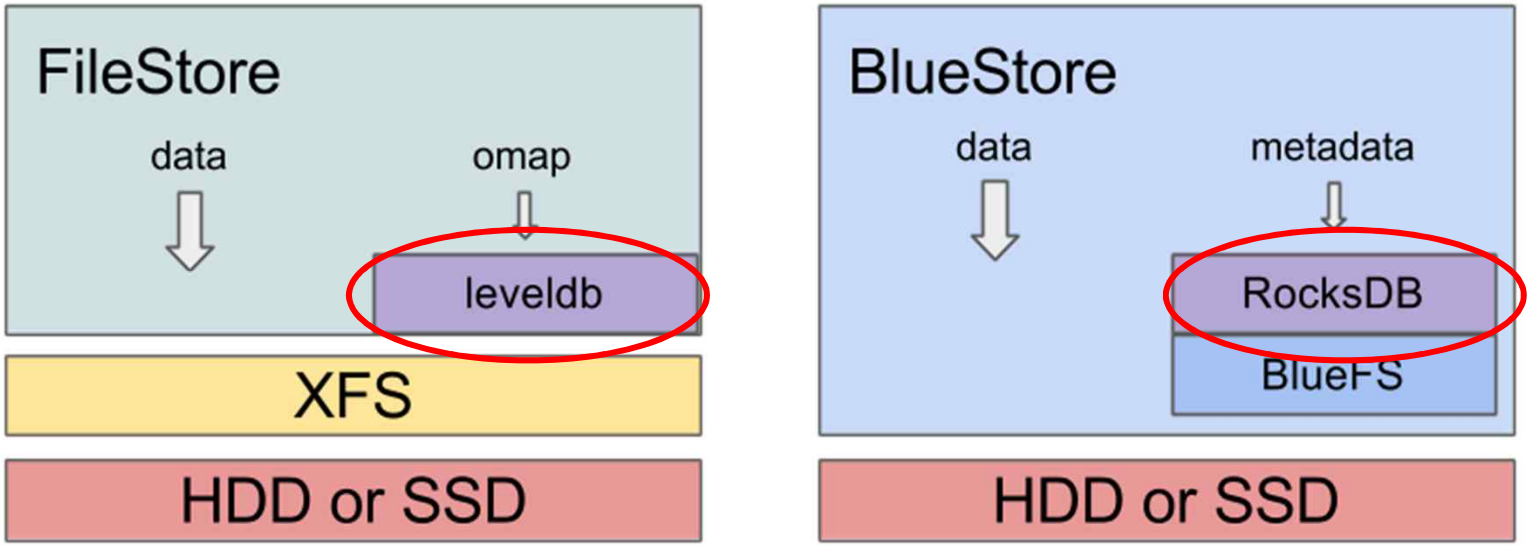


OSD



https://www.scan.co.uk/products/4tb-toshiba-mg04aca400e-enterprise-hard-drive-35-hdd-sata-iii-6gb-s-7200rpm-128mb-cache-oem

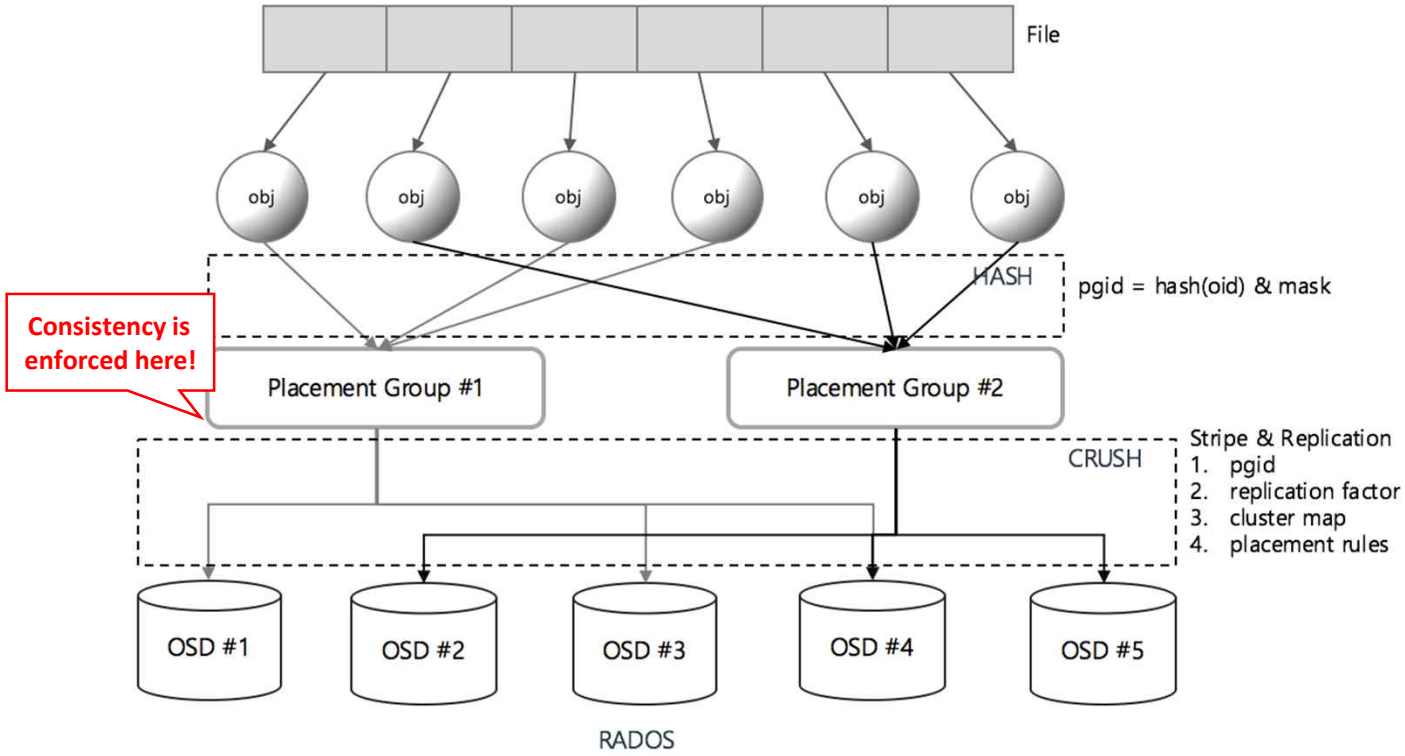
ObjectStore



https://ceph.com/community/new-luminous-bluestore/

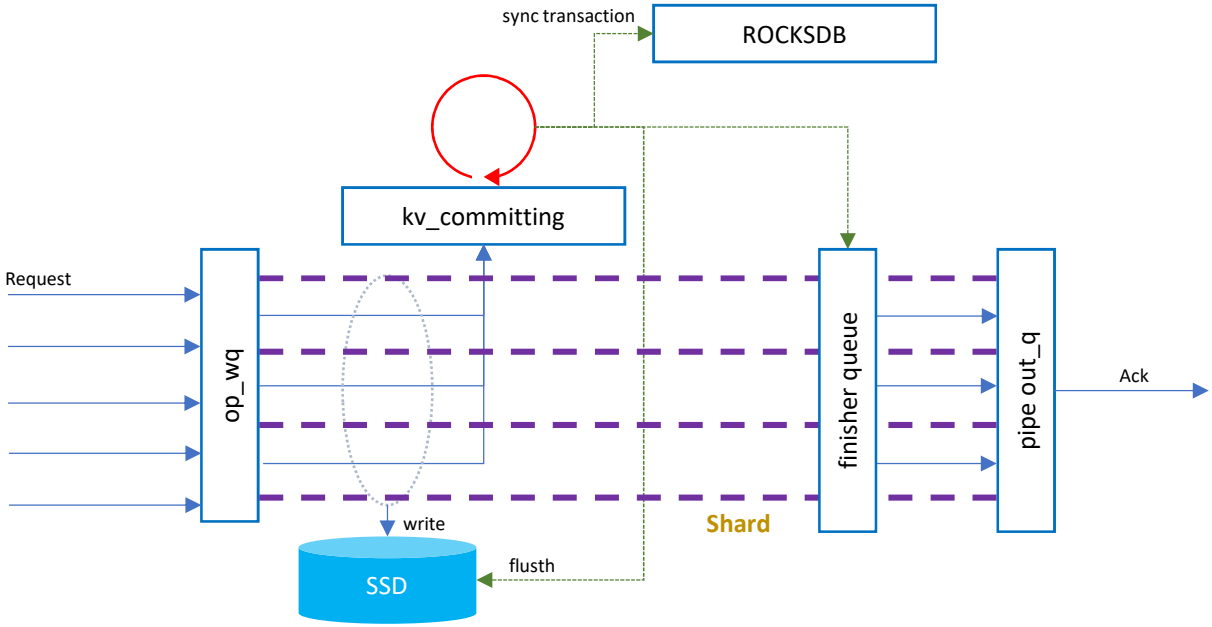
OSD Transaction

CRUSH


File
obj
obj
obj
obj
obj
obj
HASH
pgid = hash(oid) & mask
Consistency is enforced here!
Placement Group #1
Placement Group #2
CRUSH
Stripe & Replication
1. pgid
2. replication factor
3. cluster map
4. placement rules
OSD #1
OSD #2
OSD #3
OSD #4
OSD #5
RADOS

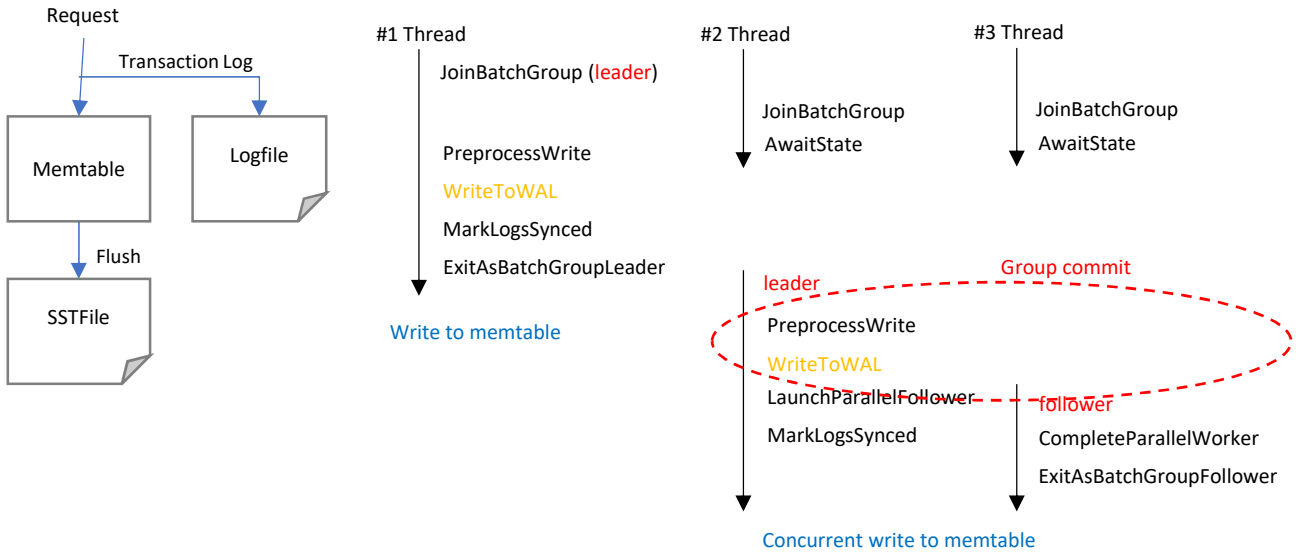
BlueStore Transaction

- To maintain ordering within each PG, ordering within each shard should be guaranteed.

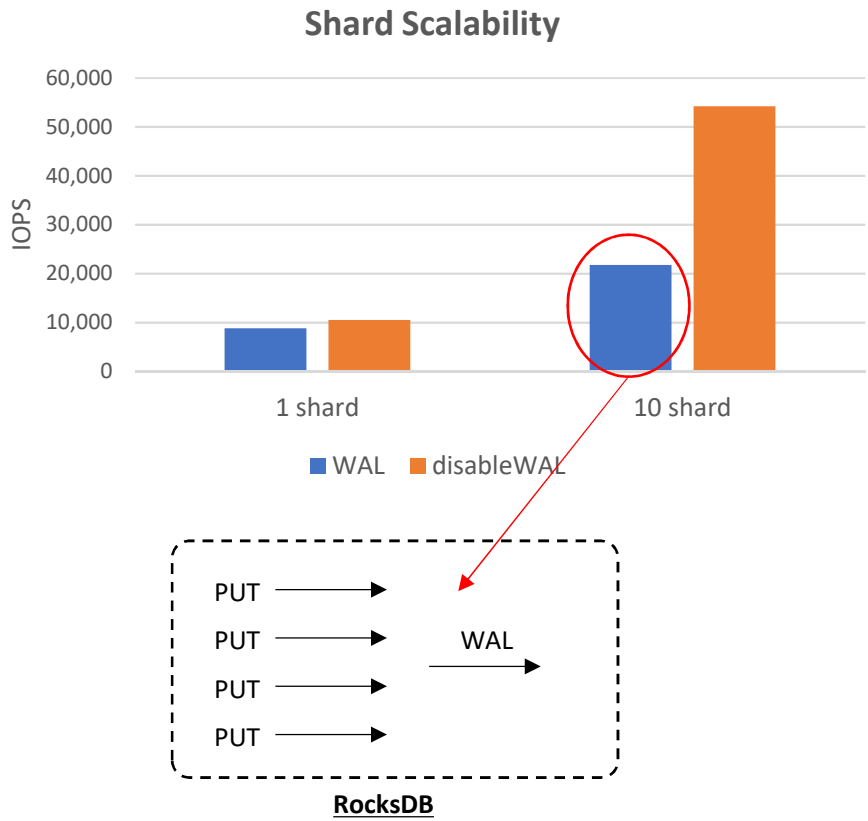


RocksDB Group Commit

- Metadata is stored in RocksDB.
- After storing metadata atomically, data is available to users.



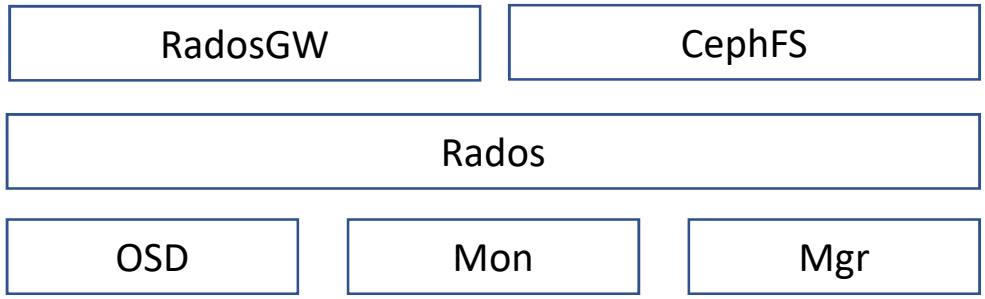
Thread Scalibility



RadosGW

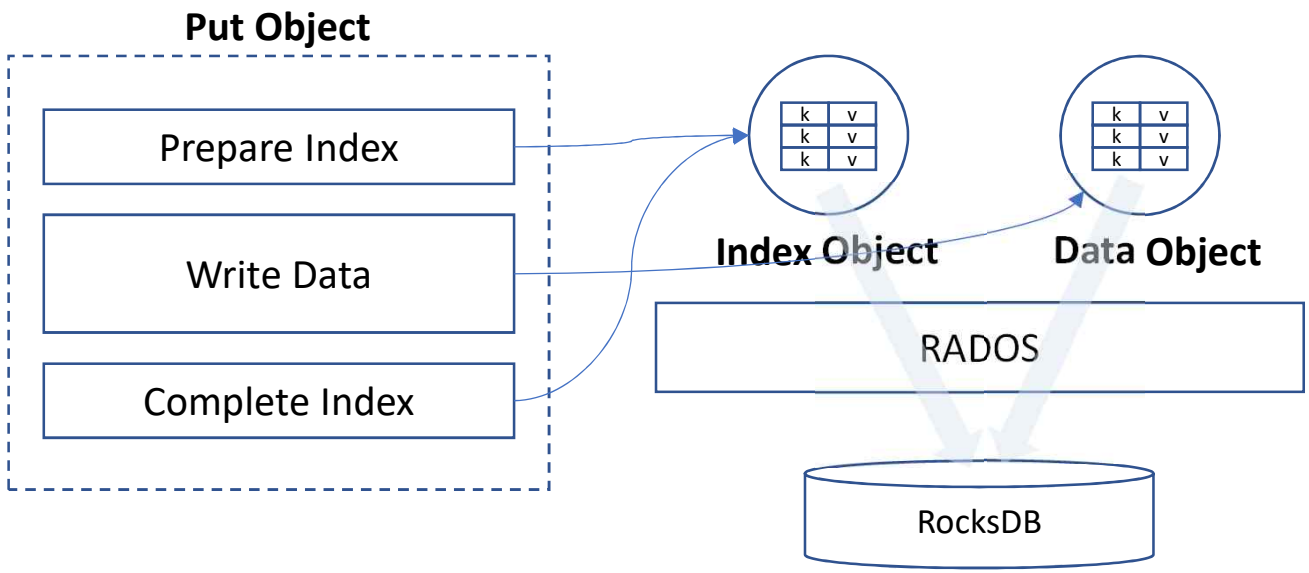
RadosGW

- RadosGW is an application of RADOS



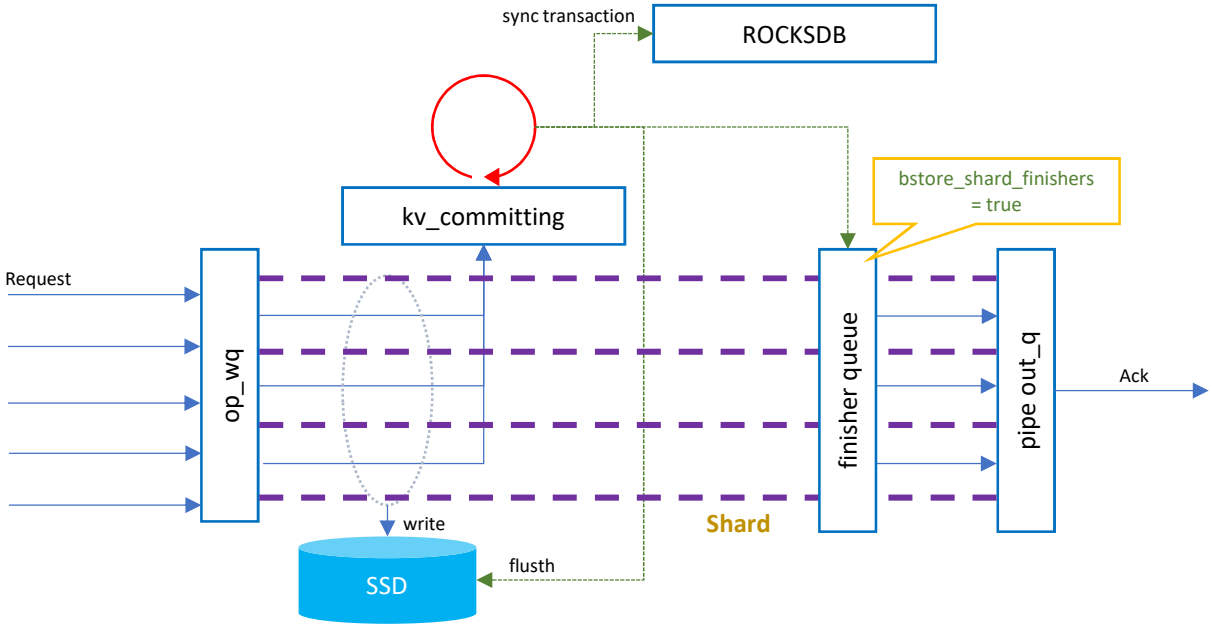
RadosGW Transaction

- All atomic operations depen on RocksDB



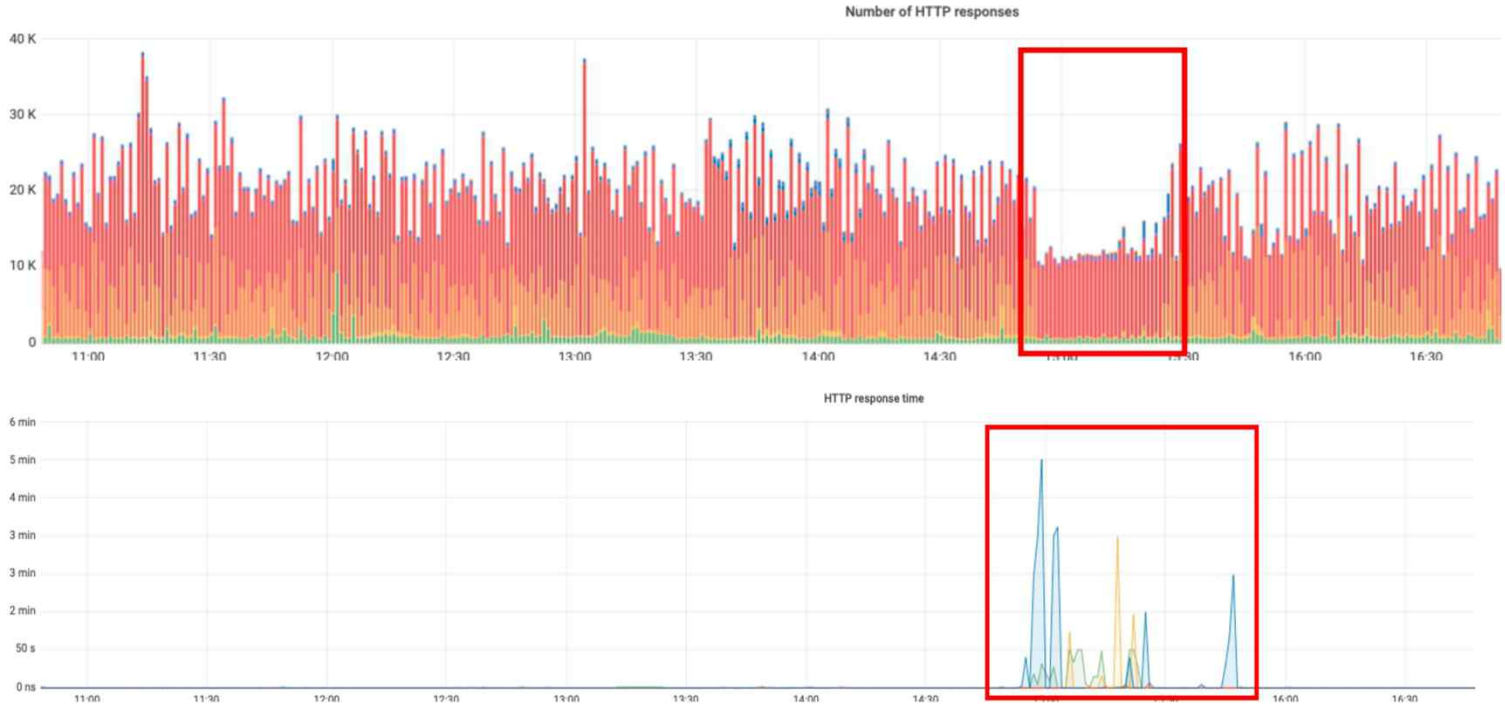
BlueStore Transaction

- To maintain ordering within each PG, ordering within each shard should be guaranteed.

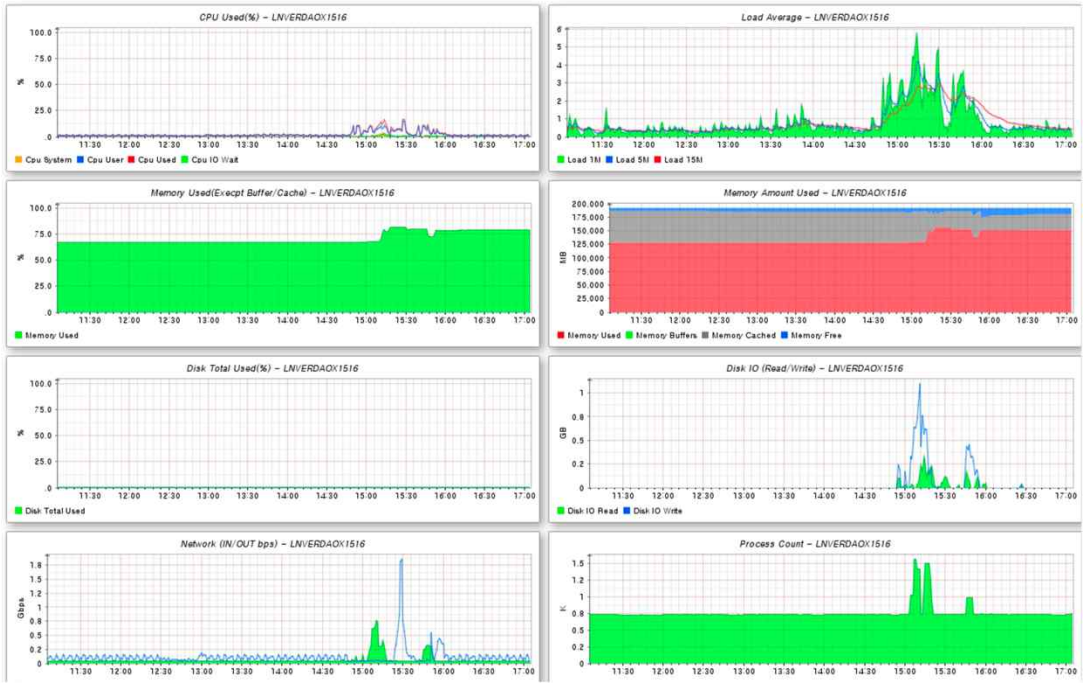


Performance Issue

Tail Latency

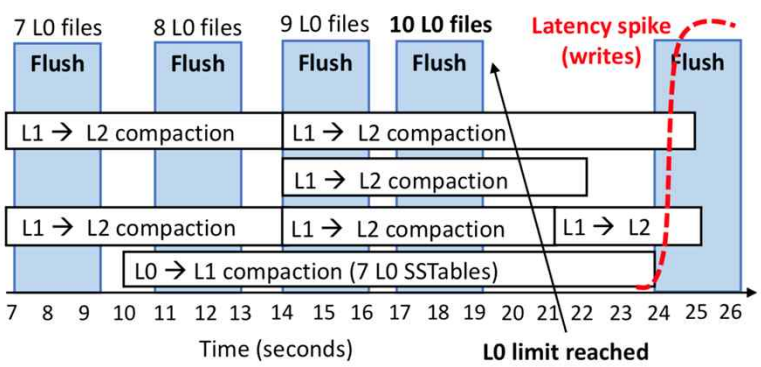


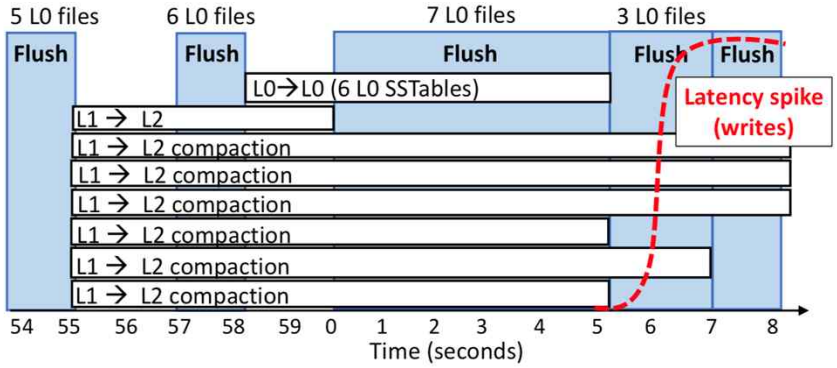
Performance Metrics



RocksDB Compaction Overhead

- "SILK: Preventing Latency Spikes in Log-Structured Merge Key-Value Stores" (ATC'19)





Conclusions

- Ceph highly depends on RocksDB
  - Strong consistency of Ceph is implemented using RocksDB transactions
- The performance of ceph also depends on RocksDB
  - Especially for Small IO
- But RocksDB has some performance Issues
  - Flushing WAL
  - Compaction

- ilsoobyun@linecorp.com

THANK YOU