Co-Design software and hardware for SSD storage in Alibaba Data Center

Fei Liu, Sheng Qiu, Pan Liu, Shu Li, Zhongjie Wu
Alibaba
Multitenancy, fast-changing workloads
Service-level agreement, QoS
Continuous pressure for TCO reduction
Demand for “white box” of I/O path – more control and determinism
Demand for SW/HW co-optimization
Smart data placement with AliFlash

- Detect data write pattern
- Separate “Hot” and “Cold” data
  - AliFlash provide interface to control data placement
- Benefit
  - Reduce WA and GC
  - Improve QoS
AliFlash

- Direct access to physical media
- Fully control of data placement and I/O scheduling
- FTL/GC customization based on application requirement
Multi-stream performance

- Seqwriter1: log write
- Randwriter2: metadata update
- Randwriter3: data update
- Reader: data read
Mismatch: Network BW & SSD BW

- Public cloud storage is sold by IOPS/GB
- SSD IOPS/GB decreasing
- To match network BW, storage density in a server will rise
- Take 4T SSD, 100Gb Network as example:
  - 4T SSD: 24 Disks, 96TB
  - 8T SSD: 20 Disks, 160TB
- Actual cost will rise.

© 2019 Alibaba Group
Heterogeneous Storage Pool

- Write three replicas to fast pool
- Transfer the data from fast pool to slow pool
  - EC + compress
  - Only 0.5 replica
- DWPD and IOPS requirement is 6 times less than before
- Avoid the bad latency of direct EC write.

EC + Compress:
write 0.5 replica

Read one replica

Transfer the data from fast pool to slow pool

© 2019 Alibaba Group
Heterogeneous Storage Pool

Storage Class Memory

Fast Speed Medium

TLC

QLC

HDD

Slow Speed Medium
What’s Next

- Customized FTL for storage engine
- QLC deployment in open channel
- Computational capability in open channel