Flash Memory Summit

Open-Channel SSDs for Host-Based Optimization

Yu Du, Feng Zhu, Sheng Qiu, Shu Li

Alibaba Group

Alibaba Group 阿里巴巴集团

A brief of AliFlash

AliFlash V1

- Host-Based PCIe SSD
- Deployed since 2016
- > 50k pcs

AliFlash V2

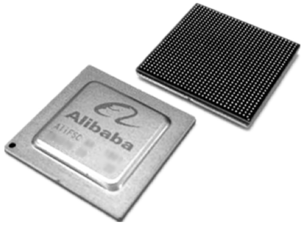
- Device based NVMe SSD
- Deployed since 2017

AliFlash V3

- Open Channel SSD
- Volume ramping up
- Targeting DB/RDS/ Search/EBS etc.

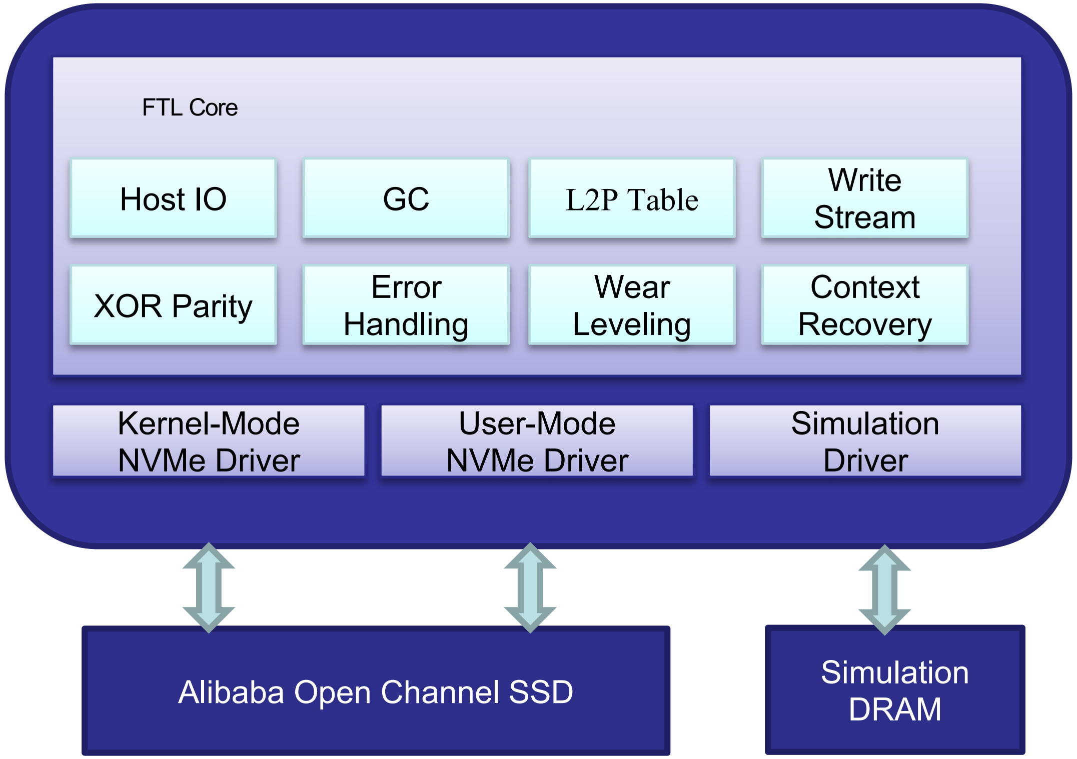
First Productionized OC-SSD

- Alibaba's home-developed Open Channel SSD - AliFlash V3
- Deployment ongoing in data centers
- Major milestone since the announcement of Alibaba's Open Channel SSD Architecture in FAST'2018
- Collaborating with multiple SSD vendors to build an ecosystem

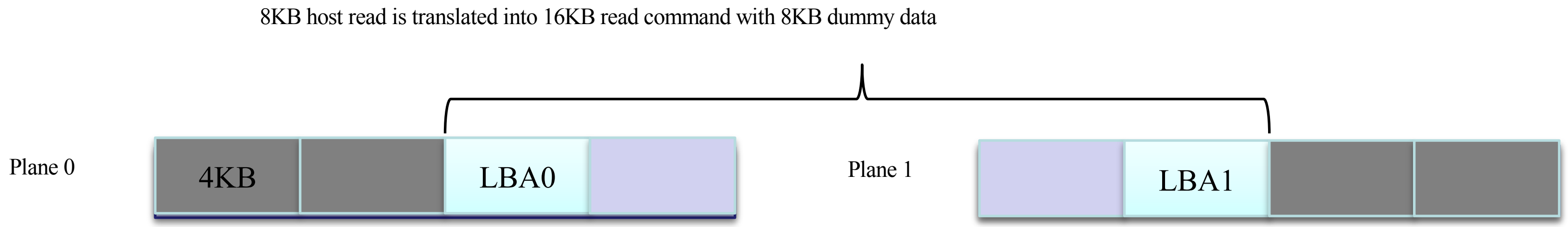


Block FTL Driver Overview

- Simplified FTL design using Alibaba Open Channel (AOC) Command Set
- FTL Core: LOC < 50K
- A single code base to support kernel/user modes
- Accelerate regression test with simulation mode



Non-Contiguous Read Optimization

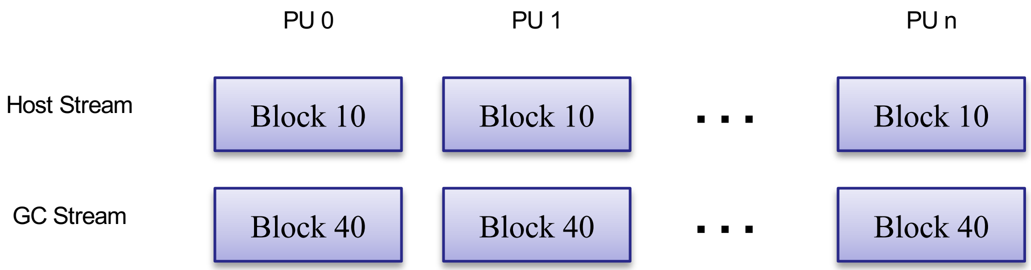

8KB host read is translated into 16KB read command with 8KB dummy data
Plane 0
4KB
LBA0
Plane 1
LBA1

- 8KB/16KB read request
- LBAs are mapped to the same multi-plane page, but not contiguous
- Non-contiguous vector read is not support yet (HW limit)
- Read extra dummy data to avoid multiple 4KB reads

Write Coalescing and Padding

- Reduce # of write commands for better IOPS and latency
- **Write Coalescing**: multiple 4KB/8KB host write requests are combined into a single write commands
- **Write Padding**: periodical padding to ensure host writes are translated into PU-aligned write commands

Host / GC Write Stream



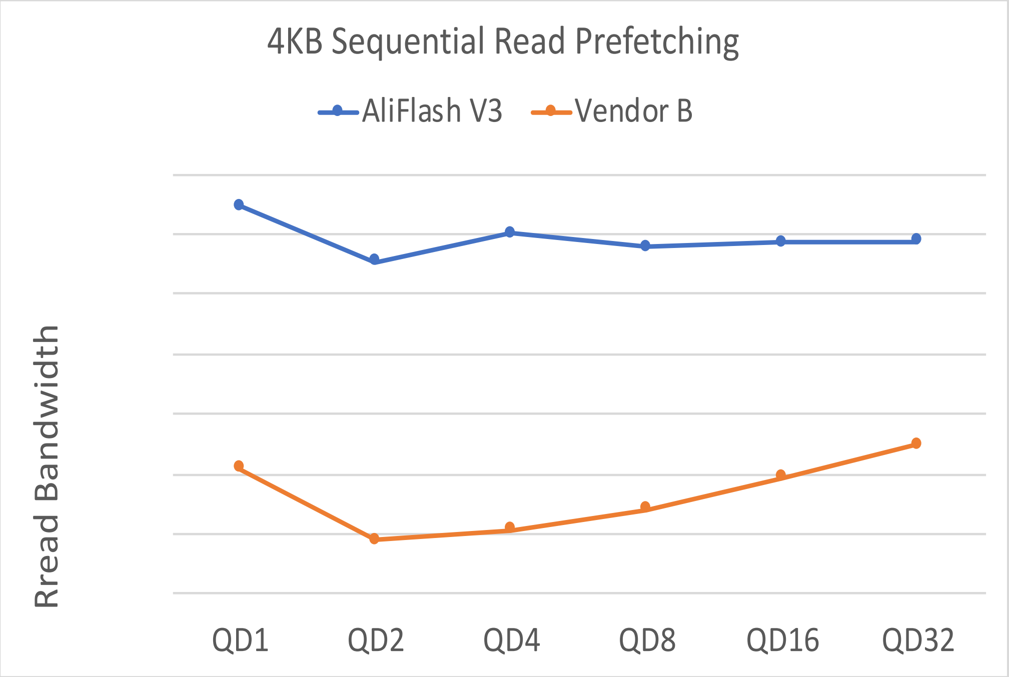
	PU 0	PU 1		PU n
Host Stream	Block 10	Block 10	. . .	Block 10
GC Stream	Block 40	Block 40	. . .	Block 40

Optimized for low-QD random write latency

- Up to 3 host streams and 1 GC stream
- Host and GC writes are scheduled independently
- Quota-based GC policy to balance host and gc writes
- Fewer free space → faster GC reads

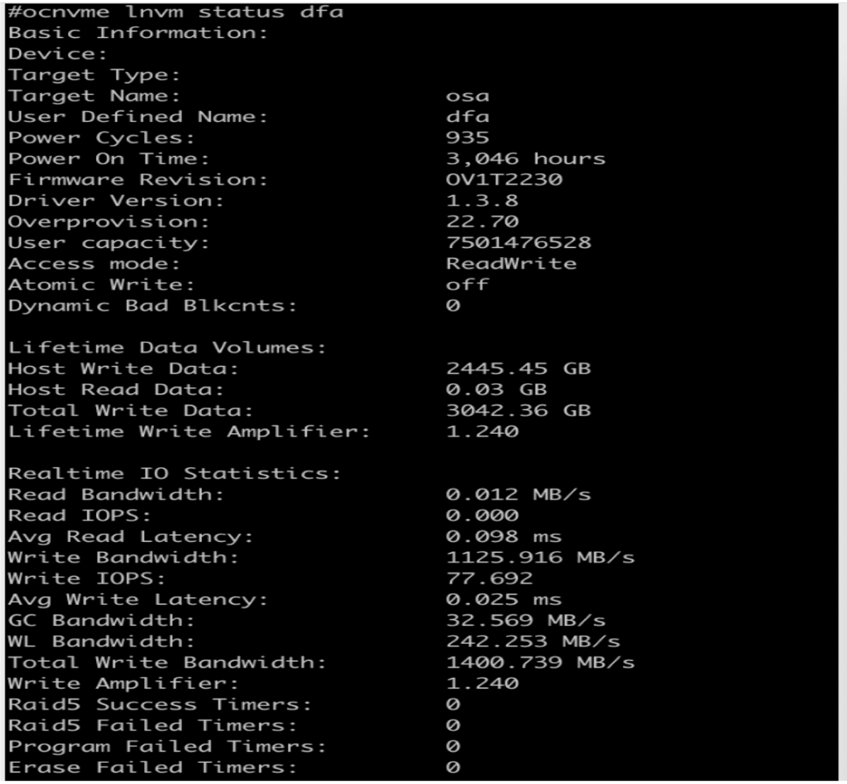
Sequential Prefetching

- Sequential read IO patterns in some production workloads
- State machine based detector
- Host DRAM prefetching buffer
- Up to 5X read bandwidth improvement.



Diagnostic Support

- 300+ runtime diagnostic parameters:
  - IOPS
  - Latency/QoS
  - GC/WL
  - Media Error
  - FTL driver parameters
  - FTL key data structures



```
#ocnvme lnvm status dfa
Basic Information:
Device:
Target Type:
Target Name:                    osa
User Defined Name:              dfa
Power Cycles:                   935
Power On Time:                  3,046 hours
Firmware Revision:              OV1T2230
Driver Version:                 1.3.8
Overprovision:                  22.70
User capacity:                  7501476528
Access mode:                    ReadWrite
Atomic Write:                   off
Dynamic Bad Blkcnts:            0

Lifetime Data Volumes:
Host Write Data:                2445.45 GB
Host Read Data:                 0.03 GB
Total Write Data:               3042.36 GB
Lifetime Write Amplifier:       1.240

Realtime IO Statistics:
Read Bandwidth:                 0.012 MB/s
Read IOPS:                      0.000
Avg Read Latency:               0.098 ms
Write Bandwidth:                1125.916 MB/s
Write IOPS:                     77.692
Avg Write Latency:              0.025 ms
GC Bandwidth:                   32.569 MB/s
WL Bandwidth:                   242.253 MB/s
Total Write Bandwidth:          1400.739 MB/s
Write Amplifier:                1.240
Raid5 Success Timers:           0
Raid5 Failed Timers:            0
Program Failed Timers:          0
Erase Failed Timers:            0
```

Application Mode Support

- Fine-tune FTL driver parameters and policies for different usage scenarios
  - Database
  - Distributed Block Storage Service
  - …
- Dynamic Configuration
  - Runtime adjustment support to a subset of driver parameters.

Conclusion Remarks

- AliFTL: FTL driver implementation based on Alibaba open channel command set
- Read/write optimizations to reduce # of commands
- Multiple write stream optimized for low-QD random write latency
- Diagnostic support and application-based tuning
- Alibaba is open to industry collaboration

Flash Memory Summit

THANK YOU