O+Q in the Cloud: Removing QLC Write-Amplification through Intel Optane SSD with SPDK WSR

Zhou, Yanbo

Alibaba Group

Alibaba Cloud EBS Local Storage

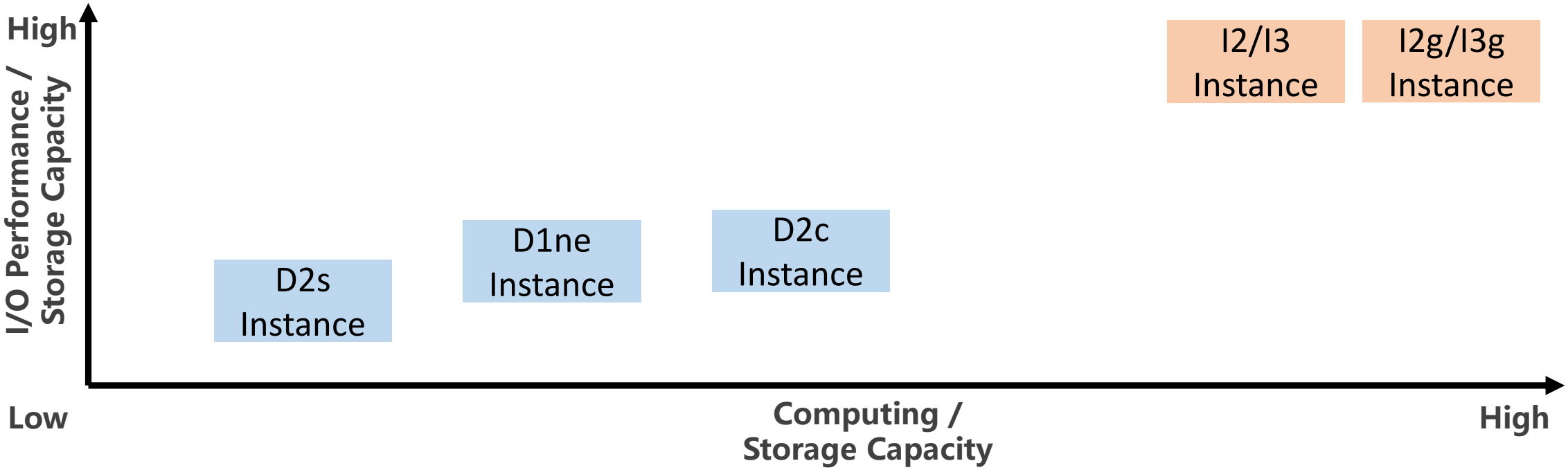
EBS local storage provides local disks that are physical attached to ECS instance.

- I-Series Instances: low latency, very high random and sequential performance

  Designed for services that require low latency, e.g., OLTP/OLAP/NoSQL databases.

- D-Series Instances: high sequential throughput, lower costs

  Designed for mass storage and offline computing, e.g., HDFS/HBase

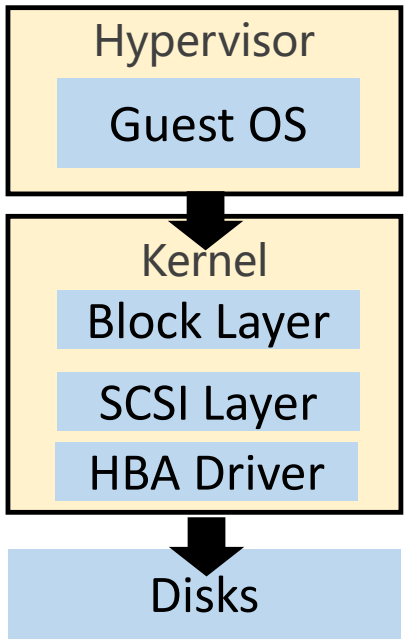


Primary Storage in the Cloud

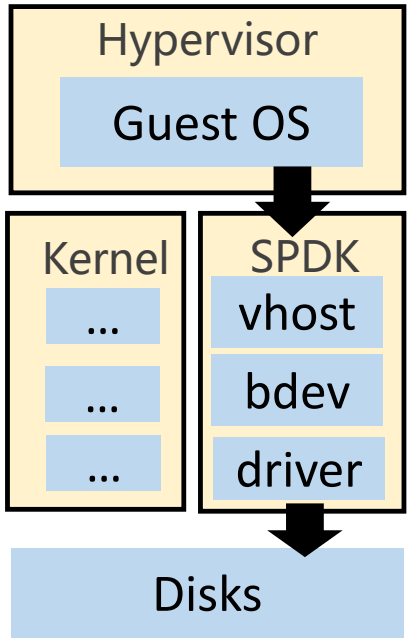
SATA/SAS HDD
high throughput
large capacity and low cost

NVMe 3D NAND TLC SSD
Low latency
very high read performance
Challenge: Traditional Storage Stack Tax

NVMe 3D NAND QLC/PLC SSD
Low latency
high read performance
large capacity and low cost
Challenge: Performance Issues caused by WAF



virtio-blk
Kernel-based Storage
Interrupt-driven I/Os



vhost-user
SPDK-based Storage
Polling-mode I/Os

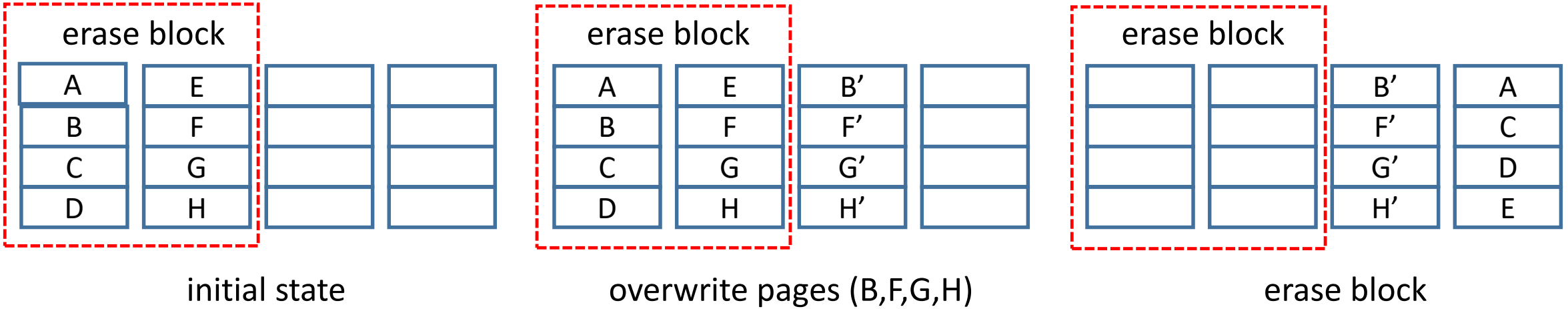
What's the next?

Write Amplification within NAND Flash SSDs

The root cause of write amplification (WAF) is the **mismatch of operation granularity**

- SSD IU(Indirection Unit) becomes larger than block interface.
- SSD erase block size is much larger than applications block size.

Example: SSDs handle *"overwrite":*



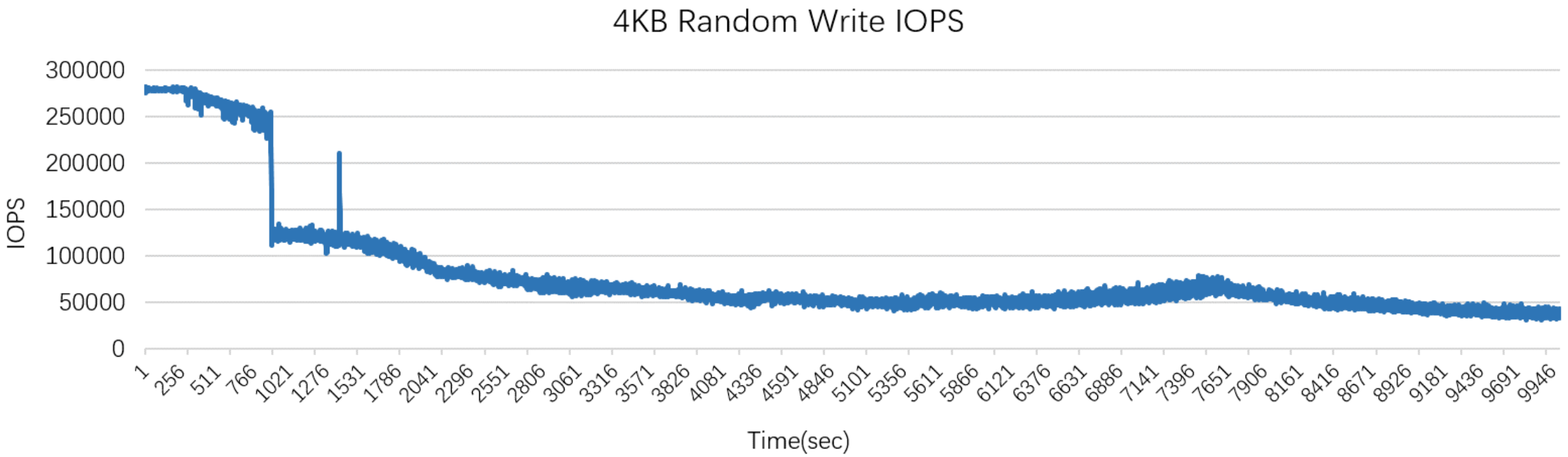
With the increasing of NAND density, the mismatch becomes more serious.

(2020 QLC uses 16K IU, 2021 QLC uses 64K IU since density of memory grows lower than NAND)

Write Amplification within NAND Flash SSDs

Write amplification influences not only performance but also durability

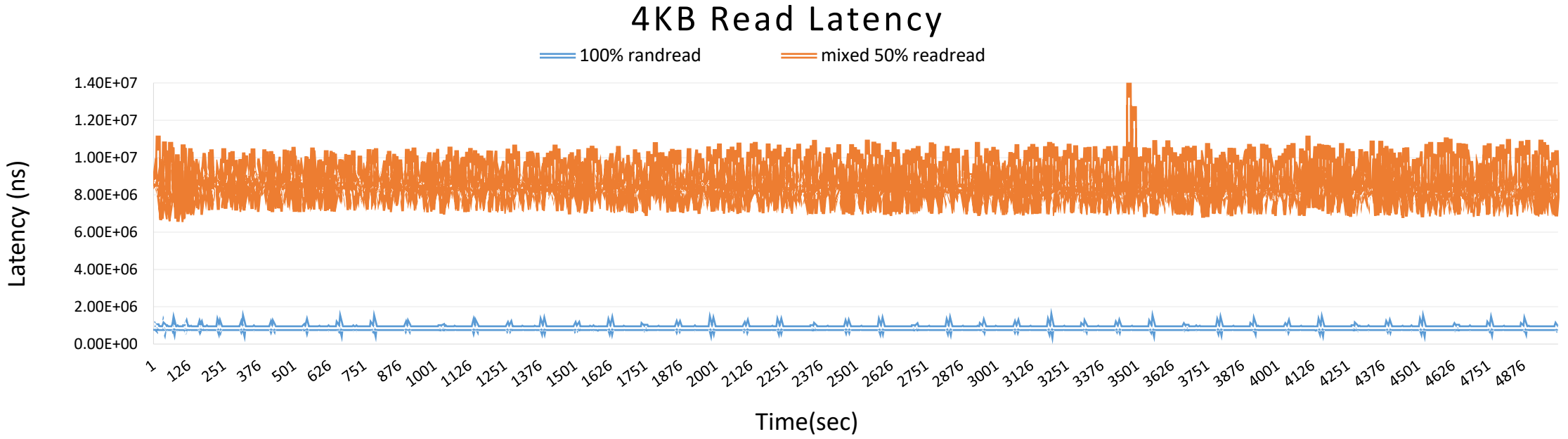
- GC with microsecond-delays can block incoming user requests leading to
  - low stable write performance
  - high tail latency including reads and writes
- Write amplification introduce more extra writes leading to SSDs wear-out.



Write Amplification within NAND Flash SSDs

Write amplification influences not only performance but also durability

- GC with microsecond-delays can block incoming user requests leading to
  - low stable write performance
  - high tail latency including reads and writes
- Write amplification introduce more extra writes leading to SSDs wear-out.



Key Insight – Caching/Tiering with Optane SSD

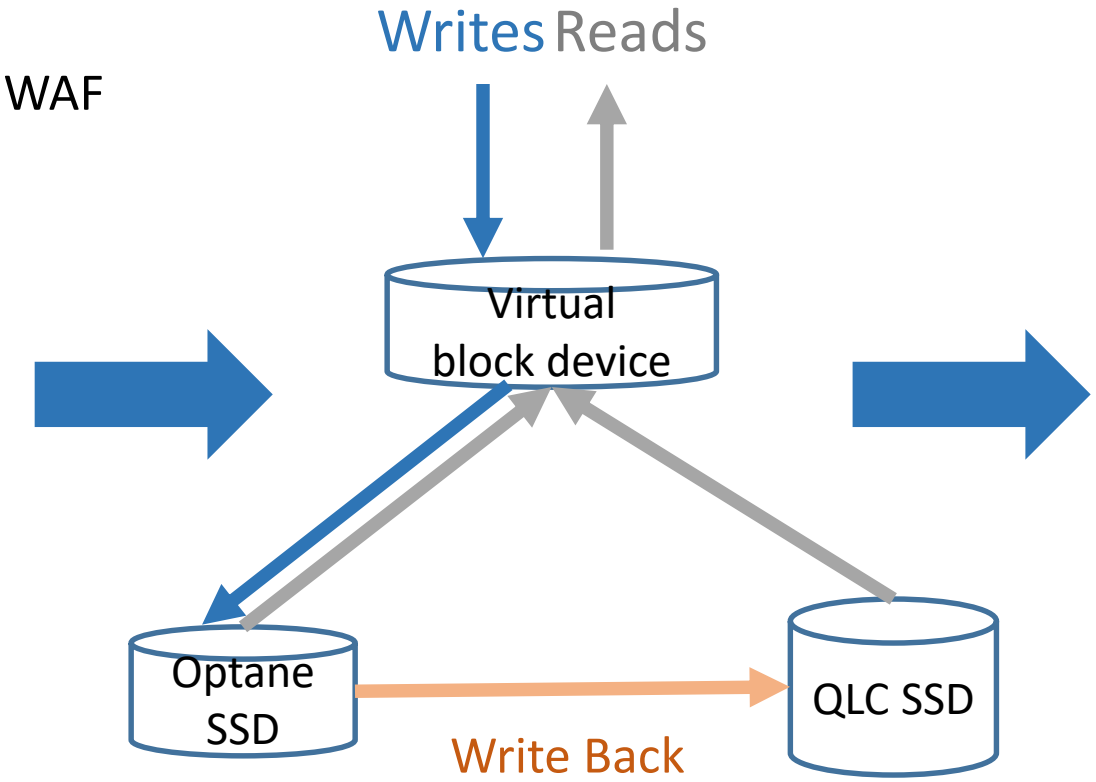
NVMe 3D XPoint Optane SSD

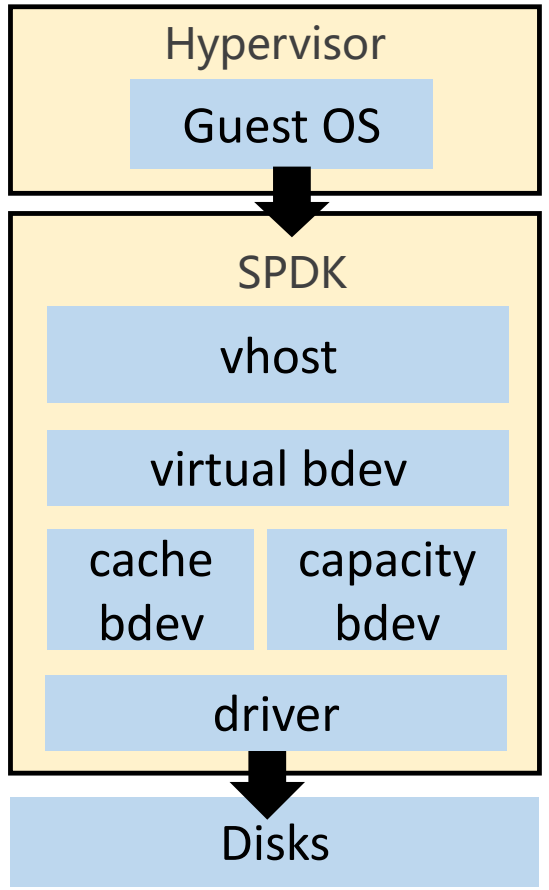
High read/write performance

High write endurance

Support overwrite without WAF

How about 3D XPoint?





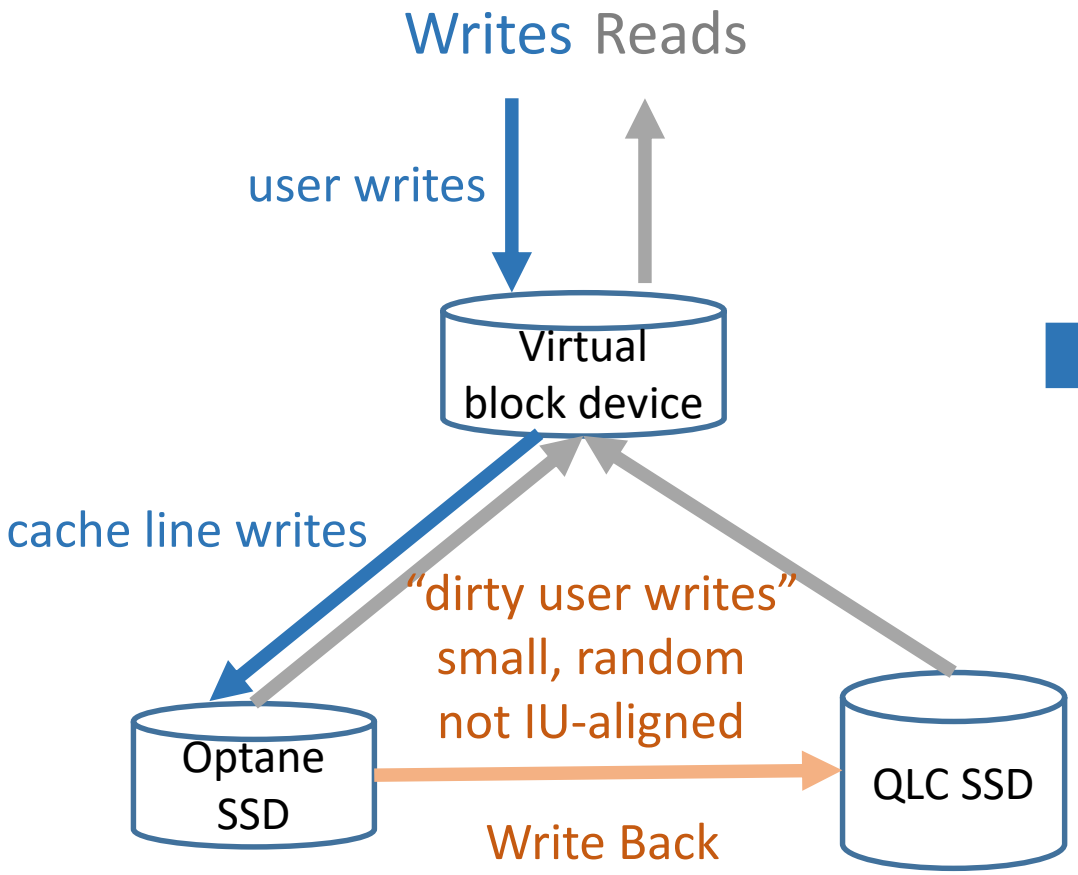
vhost-user

Hybrid storage with Optane SSD

Key Insight – Caching/Tiering with Optane SSD

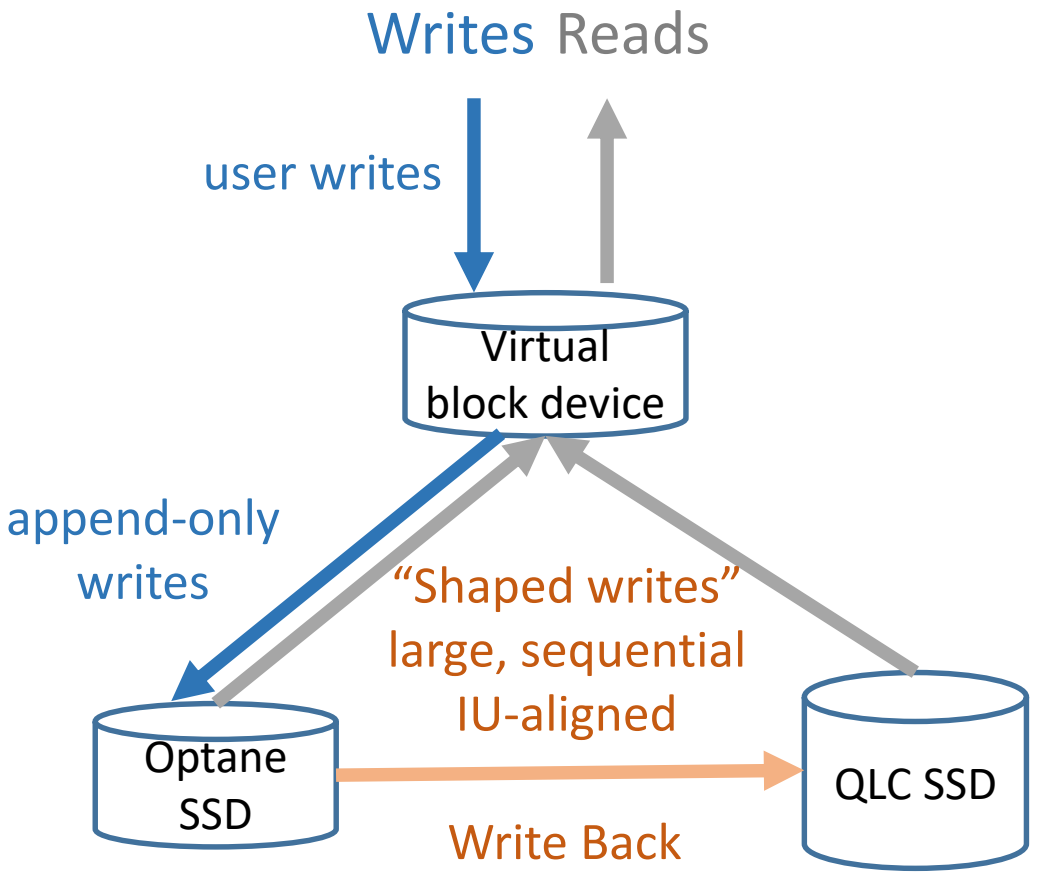
Traditional Cache (bcache, OCF)

Writes are not NAND-friendly

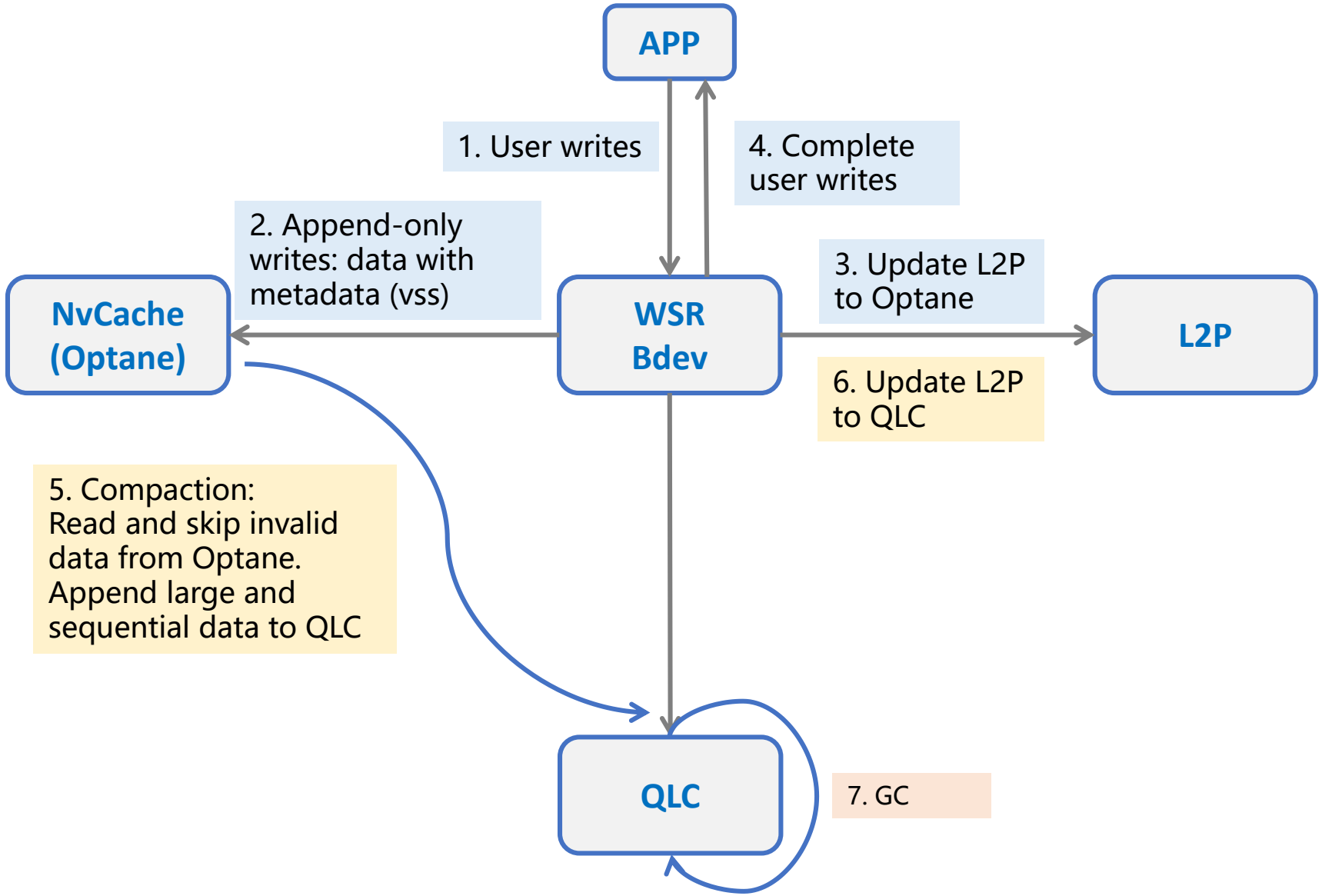


Write-Shaping RAID (WSR)

Writes are NAND-friendly



WSR Data Flow



Performance

SPDK 21.04, Optane: P5800X 800GB, QLC: P5316 16TB

Values in MB/s

#	IO Pattern	QLC (10% OP)	O+Q OCF	O+Q WSR
1	8 job, 64KB SEQ writes	8 * 390	8 * 400	8 * 400
2	8 job, 64KB RND writes	8 * 60	8 * 120	8 * 107
3	8 job, 64KB RND writes, Zipf 0.8	8 * 60	8 * 107	8 * 129
4	8 job, 64KB RND writes, Zipf 1.2	8 * 60	8 * 205	8 * 487
5	8 job, 4KB SEQ writes	8 * 5	8 * 83	8 * 388
6	8 job, 4KB RND writes	8 * 3	8 * 12	8 *105
7	4 job, 64KB RND writes 4 job, 64KB RND reads	W: 4 * 170 R: 4 * 100	W: 4 * 107 R: 4 * 89	W: 4 * 190 R: 4 * 250
8	4 job, 64KB RND writes, Zipf 0.8 4 job, 64KB RND reads	W: 4 * 170 R: 4 * 60	W: 4 * 191 R: 4 * 89	W: 4 * 264 R: 4 * 250
9	4 job, 4KB SEQ writes 4 job, 4KB RND reads	W: 4 * 4 R: 4 *5	W: 4 * 6 R: 4 * 5	W: 4 * 118 R: 4 * 118
10	7 job, 4KB RND writes 1 job, 64KB RND reads	W: 7 * 4 R: 1 * 45	W: 7 * 5 R: 1 * 74	W: 7 * 108 R: 1 * 250

Table tags: good excellent

WSR Challenges

- Memory Consumption
  - Metadata size depends on capacity devices

    -> Tiered L2P Table: DRAM for LRU metadata cache
- Failure Recovery
  - Persistent metadata

    -> SSD VSS (Variable Sector Size): data with metadata
  - Fast recovery

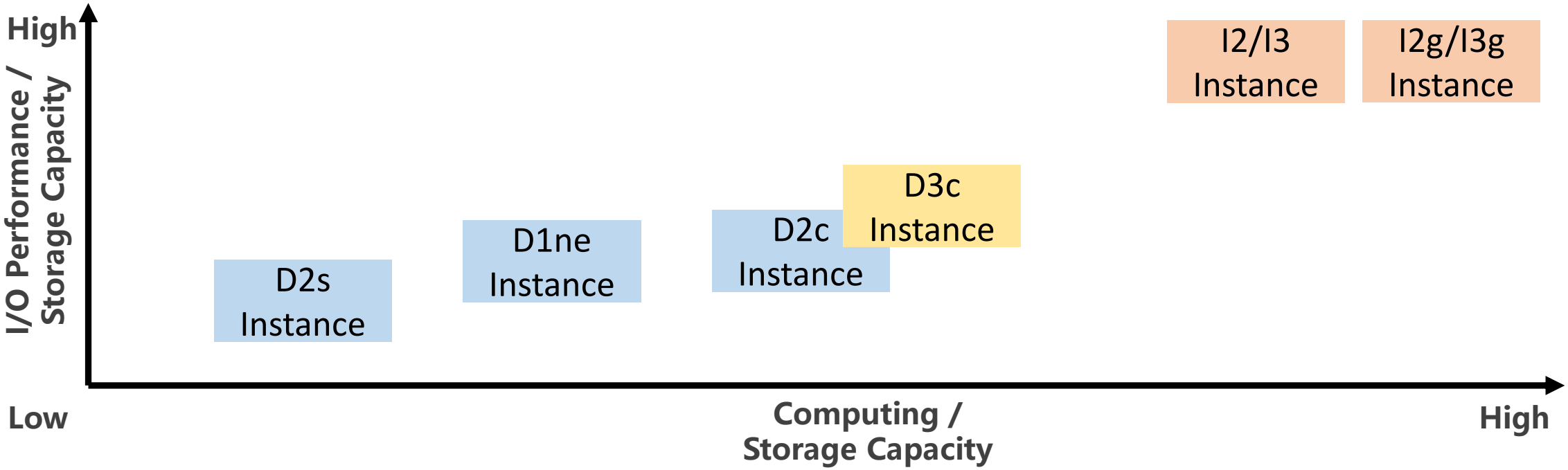
    -> Shared memory and checkpointing
- Resource Duplication
  - Internal FTL and GC

    -> Leverage ZNS to remove internal FTL layer and capacity over-provisioning.

Summary

New D-Series local storage instance based on SPDK WSR

- Exploit the extreme performance of ultra low latency SSDs (Optane SSD)
  - User-space and CPU-efficient polling-mode stack (SPDK vhost and NVMe driver)
- Overcome the limitations of NAND flash SSDs (QLC/PLC SSD)
  - Hybrid storage with "*write-shaping*" (WSR)



Alibaba Group
阿里巴巴集团

THANKS

--------- Q&A Section --------