STORAGE DEVELOPER CONFERENCE
SDC22
Fremont, CA
September 12-15, 2022
BY Developers FOR Developers
A SNIA Event

Cloud Storage Acceleration Layer (CSAL)

Enabling Unprecedented Performance and Capacity Values with Optane and QLC Flash

Presented by

Qinghua Ye, Staff Engineer, Alibaba Cloud

Kapil Karkra, Principal Engineer, Storage Software Architecture, Intel Corporation

Authors: Yanbo Zhou, Kapil Karkra, Qinghua Ye, Li Zhang, Mariusz Barczak, Wojciech Malikowski, Wayne Gao, Greg Scott and Ron Thornburg

Agenda

- Background & Motivation
  - Alibaba Cloud Local Storage
  - Big Data Trends & Challenges
  - Addressing NAND Density & Scale Challenges
  - New D-Series Big Data Instance
- Architecture & Evaluation
  - CSAL Architecture Overview
  - CSAL Performance and WAF vs. QLC
  - Preliminary Performance Results with ZNS
- Q & A

Background & Motivation

Alibaba Cloud Local Storage

Alibaba Cloud

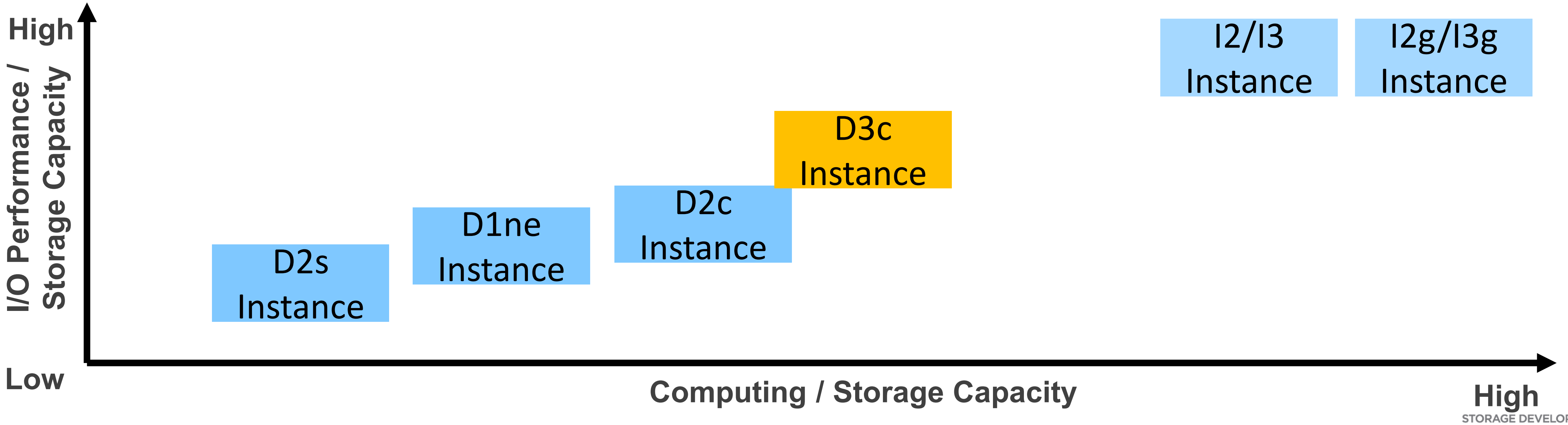
EBS local storage provides local disks that are physical attached to ECS instance.

- I-Series Instances: low latency, high performance

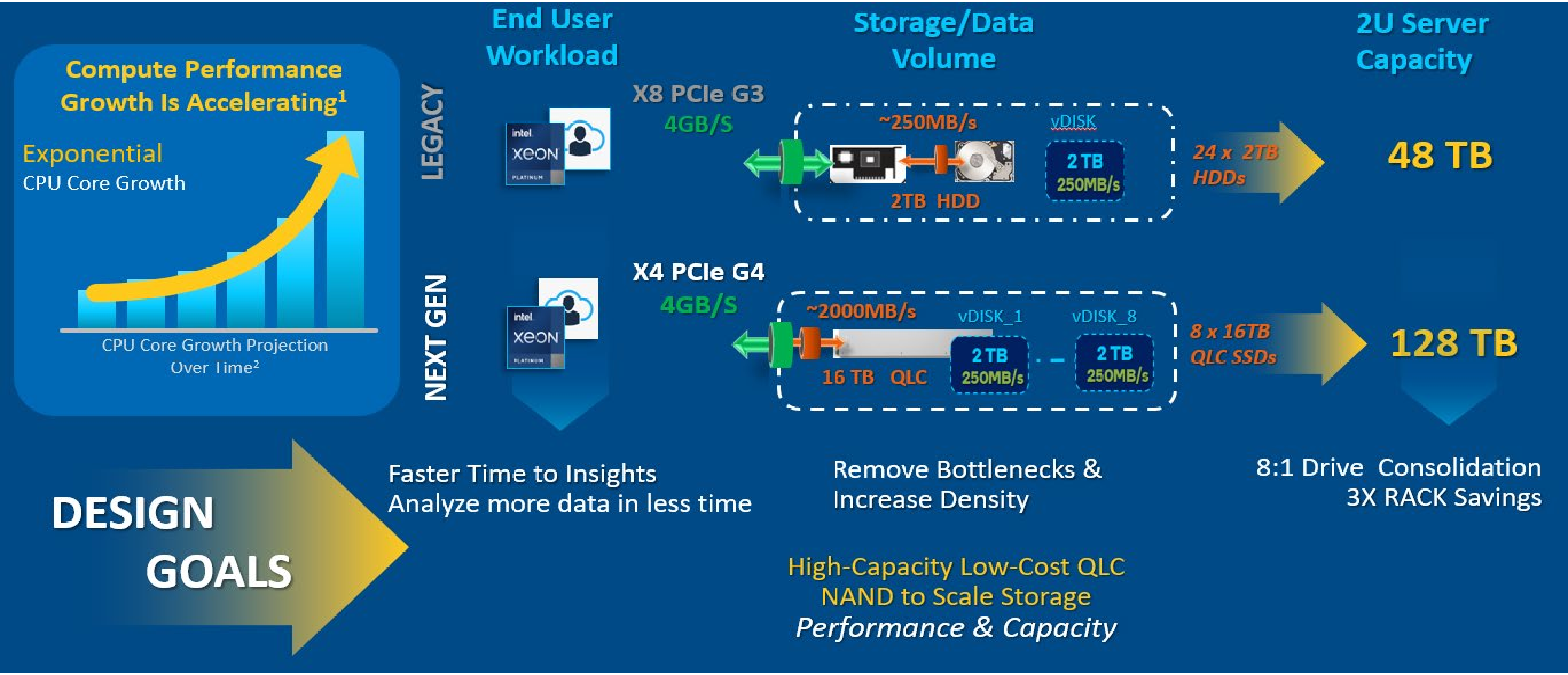
  Designed for OLTP/OLAP databases, e.g., MySQL, Aerospike, OceanBase.

- D-Series Instances: cost-effective, high capacity

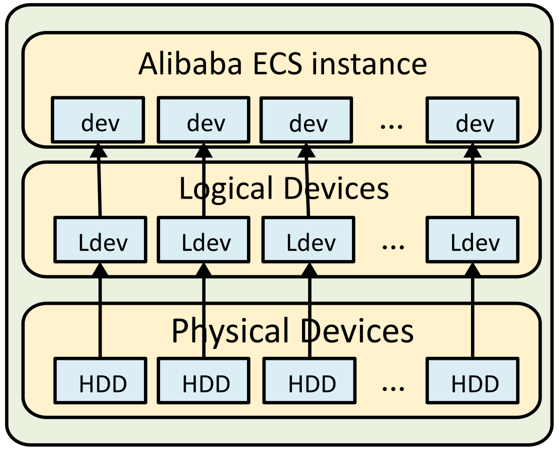
  Designed for big data and analysis, e.g., HDFS, Hbase, Clickhouse, EMR, Spark, Hadoop.



Big Data Trends & Challenges

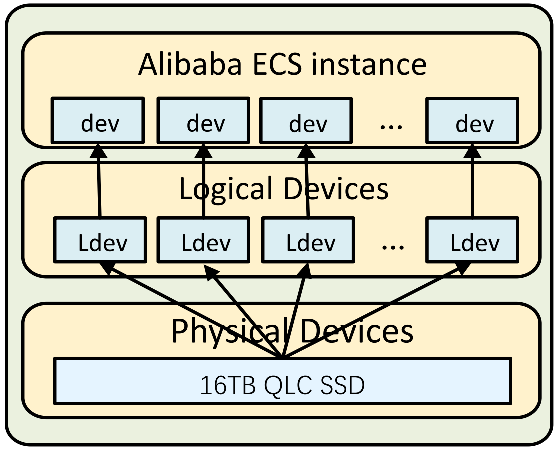


Big Data Trends & Challenges

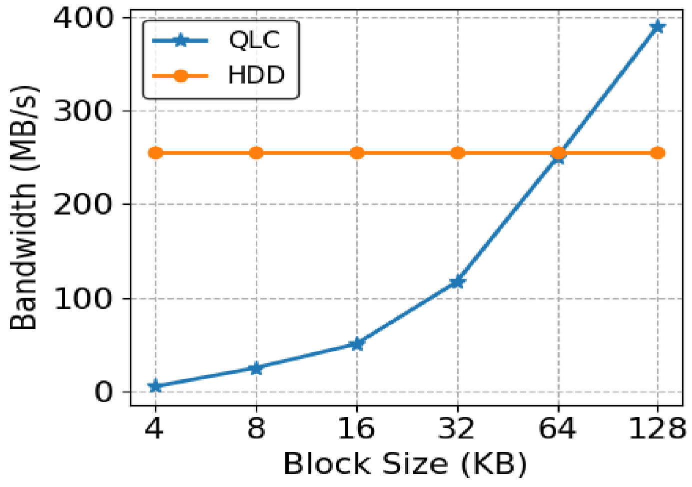


Compute Server

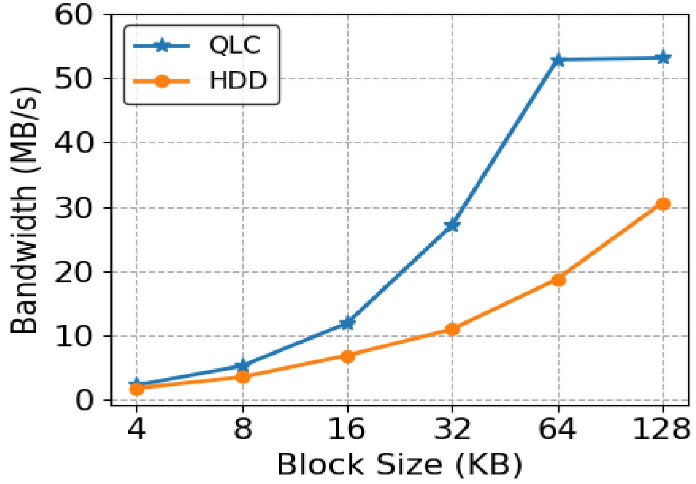
1x dev vs. 1x dev



Compute Server



Sequential writes



Random writes

Write performance per GB is the key challenge of QLC SSD

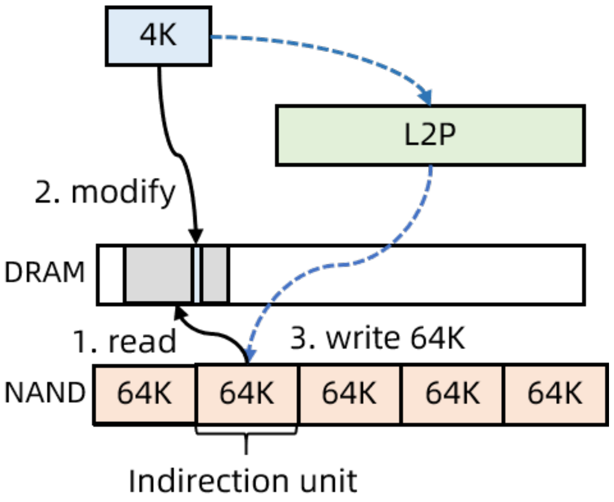
- Sequential writes are even lower than HDD for small block sizes
- Random writes are not optimal especially for small block sizes

Directly applying QLC SSDs into local storage seems hard!

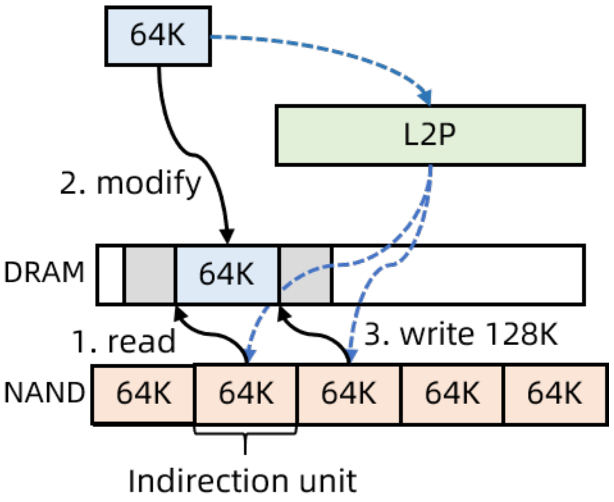
Big Data Trends & Challenges

The root cause is the following two problems that cause extra write amplification (WA):

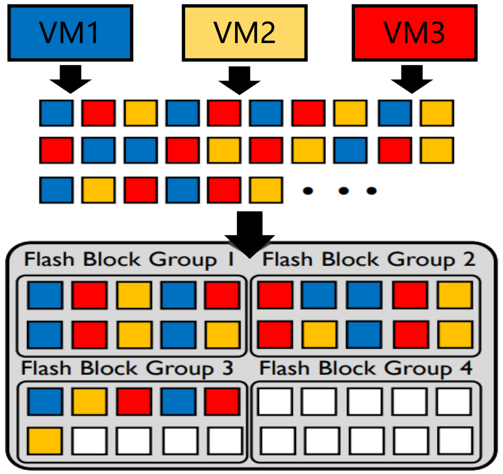
- Missized/Misaligned writes caused by internal Indirection Unit (IU).
  High density SSDs use large IU for cost saving. (e.g., Intel P5316 uses 64K IU)
- Multi-tenancy problem caused by internal Flash-Translation-Layer (FTL).
  FTL mixes I/O requests from different tenants into one stream.



(1) Missized write



(2) Misaligned write



(3) Multi-tenancy

Addressing NAND Density & Scale Challenges

w/ Optane vs QLC

8X 64K ZIPF1.2 RND WR

35X 4K RND WR QLC Performance

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

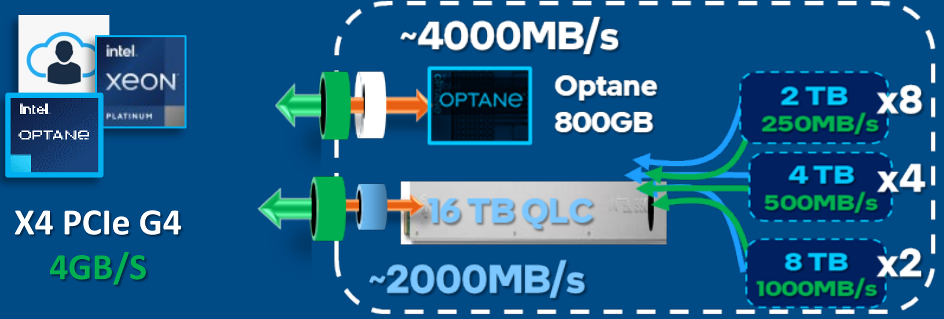
Values in MB/s

#	IO Pattern	8x HDD	1x QLC (10% OP)	1x O+Q w/ CSAL
1	8 job, 64KB SEQ writes	8 * 250	8 * 320	8 * 400
2	8 job, 64KB RND writes	8 * 20	8 * 60	8 * 107
3	8 job, 64KB RND writes, Zipf 0.8	8 * 20	8 * 60	8 * 129
4	8 job, 64KB RND writes, Zipf 1.2	8 * 20	8 * 60	8 * 487
5	8 job, 4KB SEQ writes	8 * 250	8 * 5	8 * 388
6	8 job, 4KB RND writes	8 * 1.2	8 * 3	8 *105
7	4 job, 64KB RND writes 4 job, 64KB RND reads	W: 4 * 20 R: 4 * 20	W: 4 * 170 R: 4 * 100	W: 4 * 190 R: 4 * 250
8	4 job, 64KB RND writes, Zipf 0.8 4 job, 64KB RND reads	W: 4 * 20 R: 4 * 20	W: 4 * 170 R: 4 * 60	W: 4 * 264 R: 4 * 250
9	4 job, 4KB SEQ writes 4 job, 4KB RND reads	W: 4 * 250 R: 4 *1.2	W: 4 * 4 R: 4 *5	W: 4 * 118 R: 4 * 118
10	7 job, 4KB RND writes 1 job, 64KB RND reads	W: 7 * 1.2 R: 1 * 20	W: 7 * 4 R: 1 * 45	W: 7 * 108 R: 1 * 250

Table tags: bad good excellent

Intel & Alibaba Innovation: CSAL

- Flexible scaling of ***NAND Performance & Capacity*** to the user/workload needs
- Optane ultra fast cache device and write shaping ***improves system performance while reducing costs*** scaling QLC value
- Xeon-native storage delivers ***"no-compromises" I/O performance***
- Multi-tenancy QoS software enables ***8X drive density*** resulting in a ***3X rack savings***



128 TB

8 x 16TB QLC SSD
8 x 800GB Optane SSD

New D-Series Big Data Instance

Storage capacity and performance scales with compute

12x 4TB HDD

48x Cascade Lake CPU

48TB Storage

1x

ecs.d2c.24xlarge

96x vCPU

48TB Storage

Old physical configuration

vs

8x 16TB QLC
8x 800GB Optane

64x Ice Lake CPU

128TB Storage

2x

ecs.d3c.14xlarge

56x vCPU

64TB Storage

New physical configuration

- TPCx-HS: storage-intensive: 103% performance improvement in Hsort

TPCx-HS 3TB	d2c.24xlarge	d3c.14xlarge	Improvement
Hsgen (min)	7.11	4.16	70.91%
HSort (min)	20.31	9.96	103.92%
HSValidate (min)	3.46	1.18	193.22%
Total Time (min)	31	15.25	103.28%
HSph@SF	1.9357	3.9354	

- TPC-DS: compute-intensive: Almost same performance in SQL process with less vCPU cores

TPC-DS 3TB	d2c.24xlarge	d3c.14xlarge	Improvement
datagen (min)	40.8	41.93	-2.69%
sql (min)	50.02	50.58	-1.11%
Total Time (min)	90.82	92.51	-1.83%

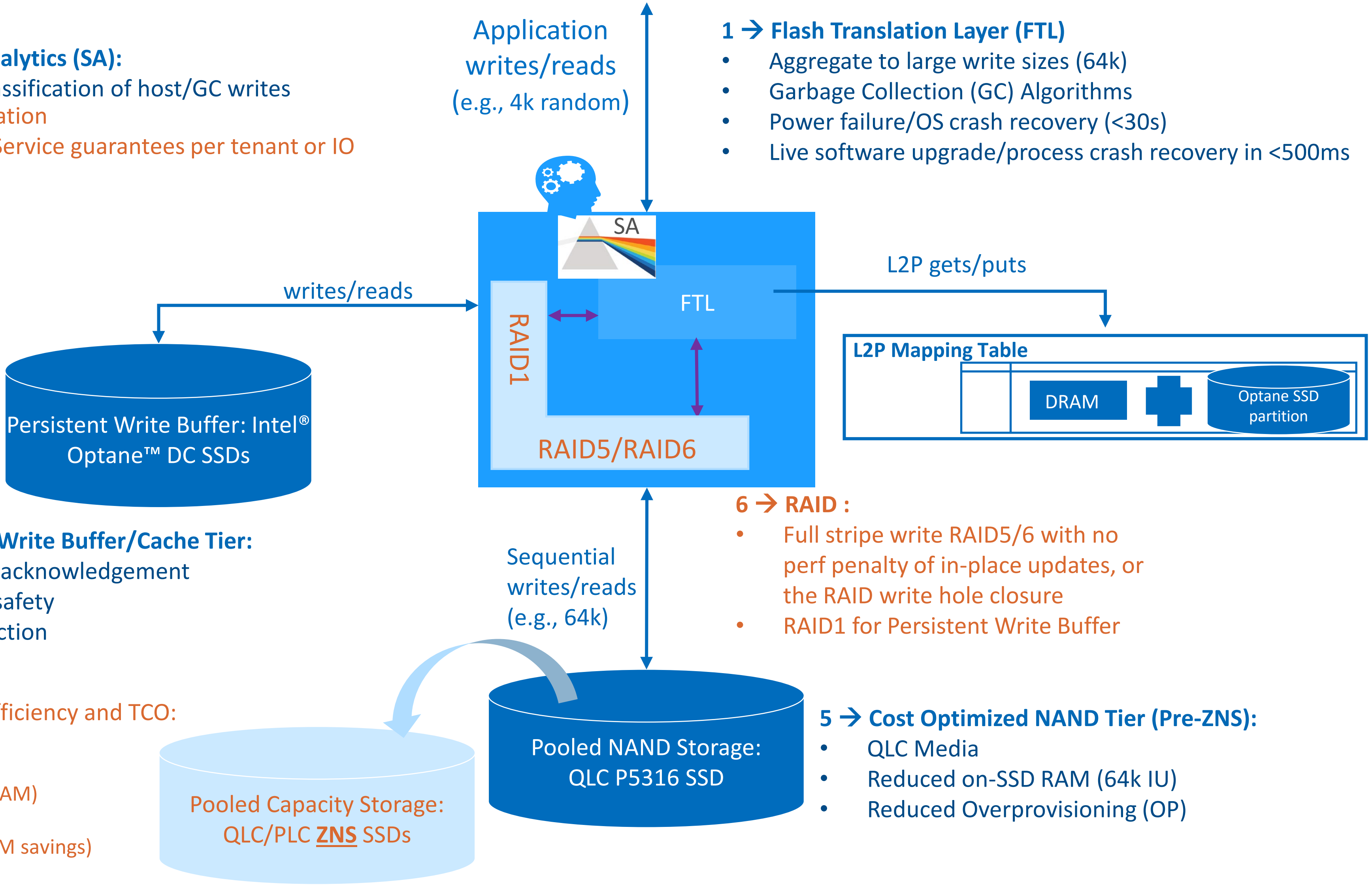
Architecture & Evaluation

CSAL Architecture Overview

Legend:
Blue – Current Implementation
Orange – Target Implementation

4 → Storage Analytics (SA):
- Lifetime classification of host/GC writes
- Tenant isolation
- Quality of Service guarantees per tenant or IO

1 → Flash Translation Layer (FTL)
- Aggregate to large write sizes (64k)
- Garbage Collection (GC) Algorithms
- Power failure/OS crash recovery (<30s)
- Live software upgrade/process crash recovery in <500ms



2 → Tiered L2P Mapping Table
- Async L2P in DRAM + Optane SSD
- Paging Logic

3 → Persistent Write Buffer/Cache Tier:
- Early write acknowledgement
- Power fail safety
- Write reduction

6 → RAID :
- Full stripe write RAID5/6 with no perf penalty of in-place updates, or the RAID write hole closure
- RAID1 for Persistent Write Buffer

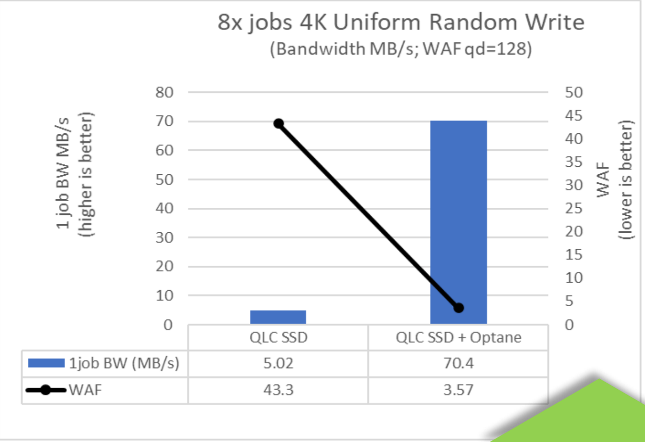
5 → Cost Optimized NAND Tier (Pre-ZNS):
- QLC Media
- Reduced on-SSD RAM (64k IU)
- Reduced Overprovisioning (OP)

Additional Storage Efficiency and TCO:
- QLC/PLC Media
- No on-SSD DRAM
- No PLI (capacitor, SRAM)
- Zero OP
- No XOR (NAND, SRAM savings)

CSAL Performance and WAF vs QLC

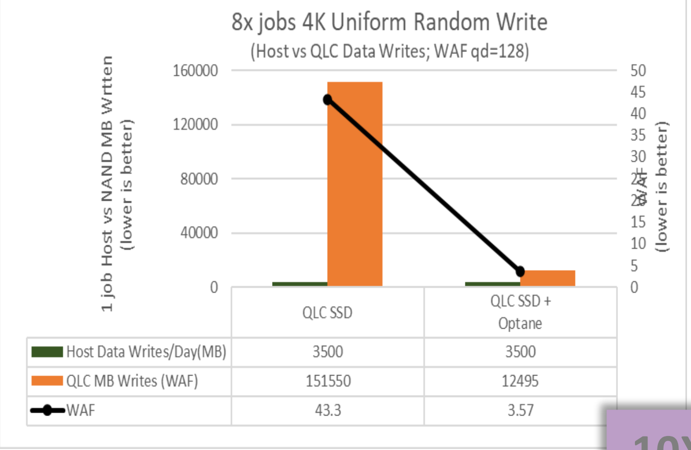
Increase CPU Performance ***18X~35X*** *AND* Extend NAND Usable Life:***10X***

For the performance of 4K uniform random write, single job O+Q BW is 70.4 MiB/s, 14x of QLC BW of 5.02 MiB/s, while WAF is only 3.57, 8.2% of QLC only WAF of 43.3.



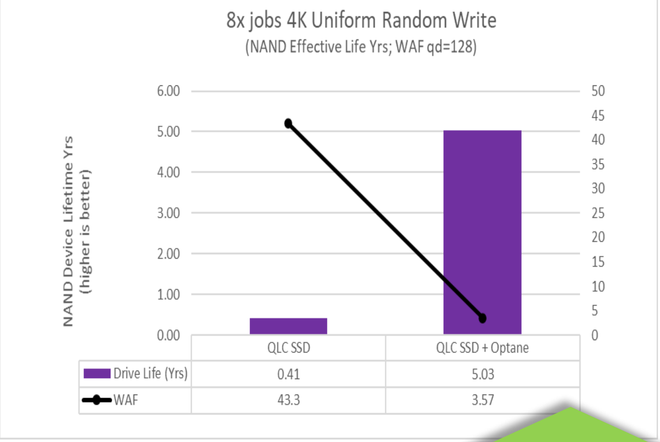
	QLC SSD	QLC SSD + Optane
1job BW (MB/s)	5.02	70.4
WAF	43.3	3.57

CPU Perf (MB/s) 35X



	QLC SSD	QLC SSD + Optane
Host Data Writes/Day(MB)	3500	3500
QLC MB Writes (WAF)	151550	12495
WAF	43.3	3.57

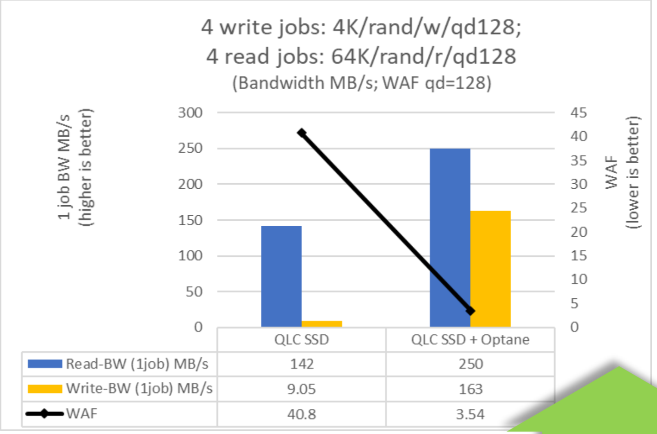
WAF Reduction 10X



	QLC SSD	QLC SSD + Optane
Drive Life (Yrs)	0.41	5.03
WAF	43.3	3.57

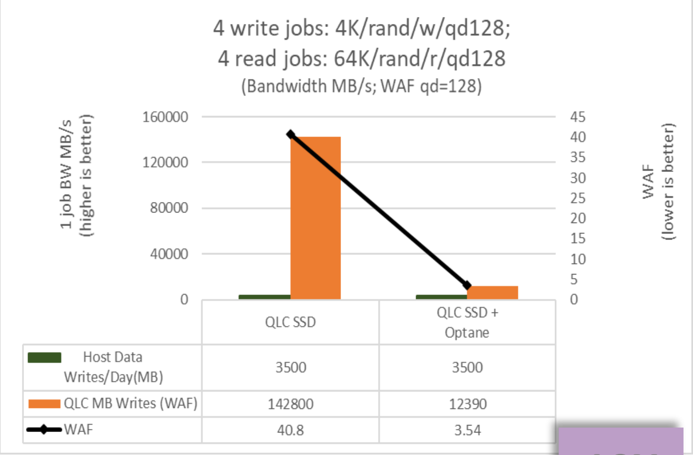
QLC NAND Life 10X

Single job QLC read BW is only 142MiB/s, cannot meet 250MiB/s target; Single job O+Q write BW is 163MiB/s, 18 times of QLC BW of 9.05MiB/s, while WAF is 3.54, only 8.7% of QLC WAF of 40.8.



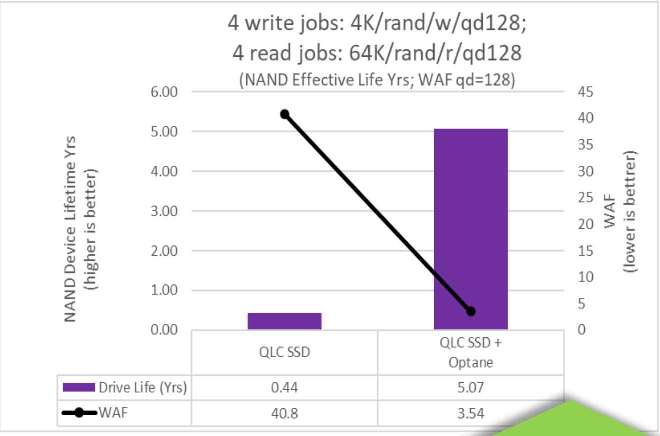
	QLC SSD	QLC SSD + Optane
Read-BW (1job) MB/s	142	250
Write-BW (1job) MB/s	9.05	163
WAF	40.8	3.54

CPU Perf (MB/s) 18X



	QLC SSD	QLC SSD + Optane
Host Data Writes/Day(MB)	3500	3500
QLC MB Writes (WAF)	142800	12390
WAF	40.8	3.54

WAF Reduction 10X



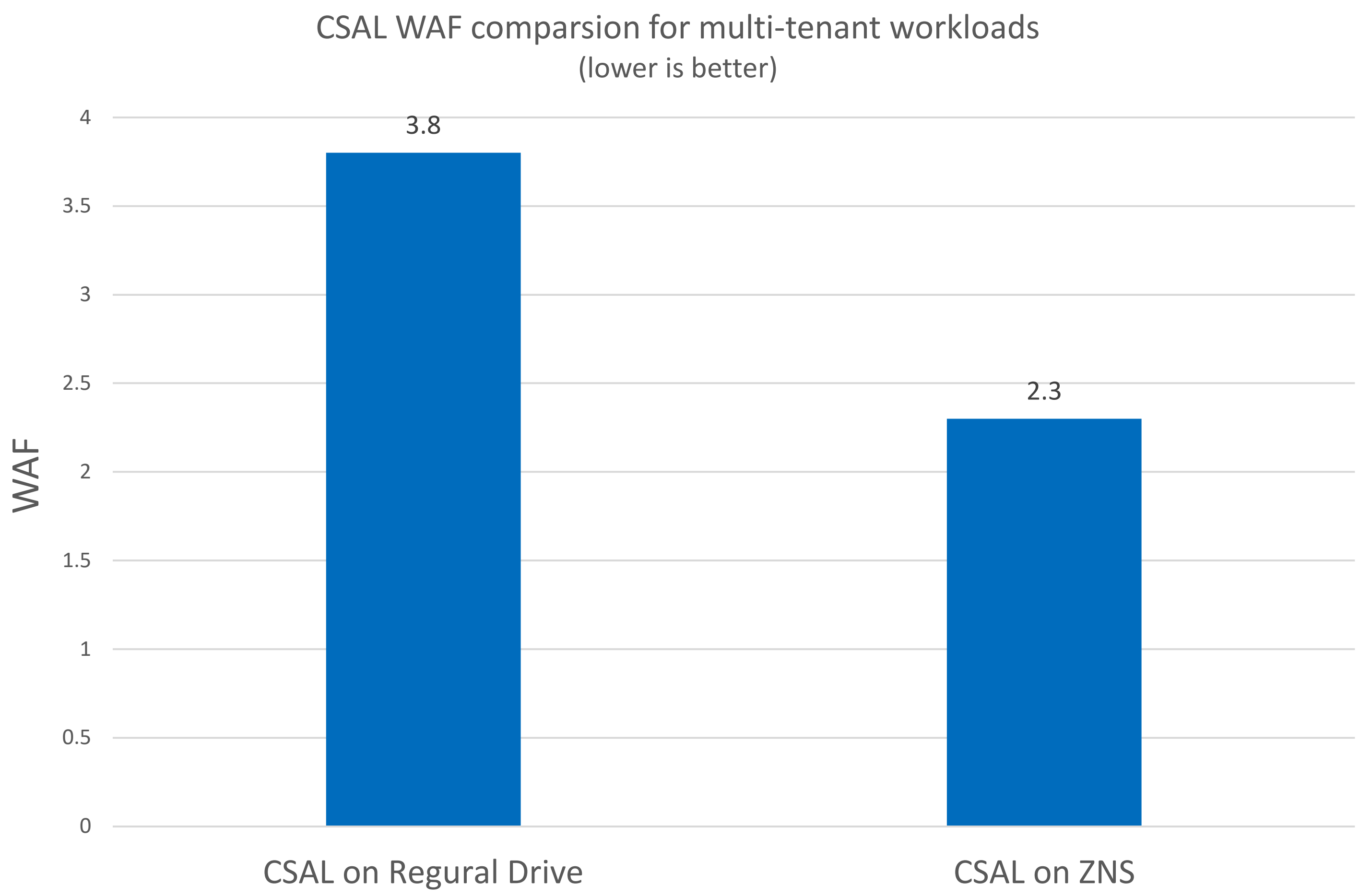
	QLC SSD	QLC SSD + Optane
Drive Life (Yrs)	0.44	5.07
WAF	40.8	3.54

QLC NAND Life 10X

Preliminary Performance Results with ZNS

Multiple Tenants

- 1 write job: 4K/seq/qd128
- 1 write job1: 4K/rand/qd128
- 1 write job:4K/zipf0.8/qd128
- 1 write job:4K/zipf1.2/qd128



ZNS SSD: Ultrastar DC ZN540 4TB from Western Digital
Regular Drive: (used for ZNS WAF comparison) – Ultrastar DC SN640 7.68TB from Wester Digital

Looking Forward

Future plan

1. CSAL Upstream to SPDK
    - bdev modules for SPDK
    - Community review in process
    - Future support for:
        - RAID, ZNS, PLC
2. NVMeOF Ref Solution

References

- Alibaba D3c Instance

https://help.aliyun.com/document_detail/25378.html#d3c

- SPDK PRC Summit

https://spdk.io/news/2021/12/22/prc_virtual_forum_presentations/

- System level benchmarking & white paper coming soon

See Sources slide for all source details.

Q & A

Thank you!

