STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

New Developments in Cloud Storage Acceleration Layer (CSAL)

CSAL, A Host Based FTL in SPDK

Kapil Karkra, Sr. Principal Engineer at Solidigm

Legal Disclaimers

All product plans, roadmaps, specifications, and product descriptions are subject to change without notice.

Nothing herein is intended to create any express or implied warranty, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, or any warranty arising from course of performance, course of dealing, or usage in trade.

The products described in this document may contain design defects or errors known as "errata," which may cause the product to deviate from published specifications. Current characterized errata are available on request.

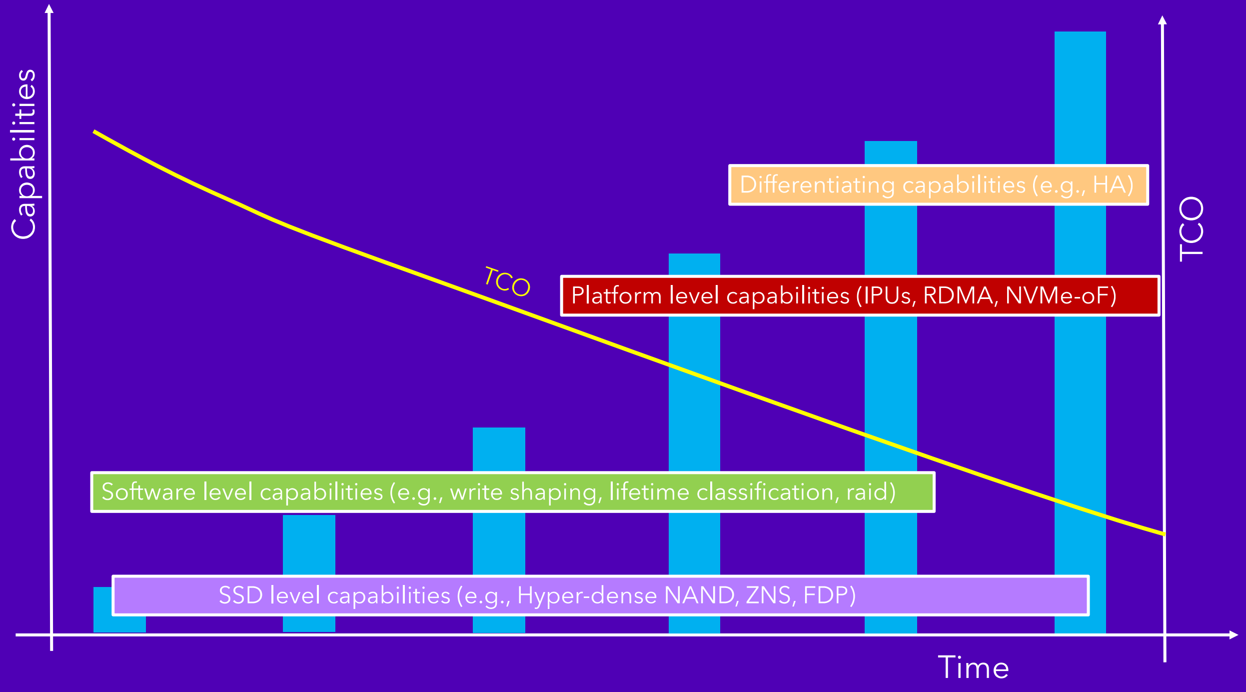
Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Performance results are based on testing as of dates shown in the configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

Your costs and results may vary.

Some results have been estimated or simulated using internal Solidigm analysis or architecture simulation or modeling and provided to you for information purposes only. Any differences in your system hardware, software or configuration may affect your actual performance.

Imagine if we had...



A sandbox to explore, add capabilities, and drive down data center TCO…

Agenda

1. TCO Benefits of large capacity (D5-P5536 61.44TB) QLC
2. How to Further Expand the Reach of QLC?

Motivation and Problem

3. Creating an Easy Button with CSAL and a Reference Storage Platform (RSP)
4. Performance Results and TCO Benefits

Solution

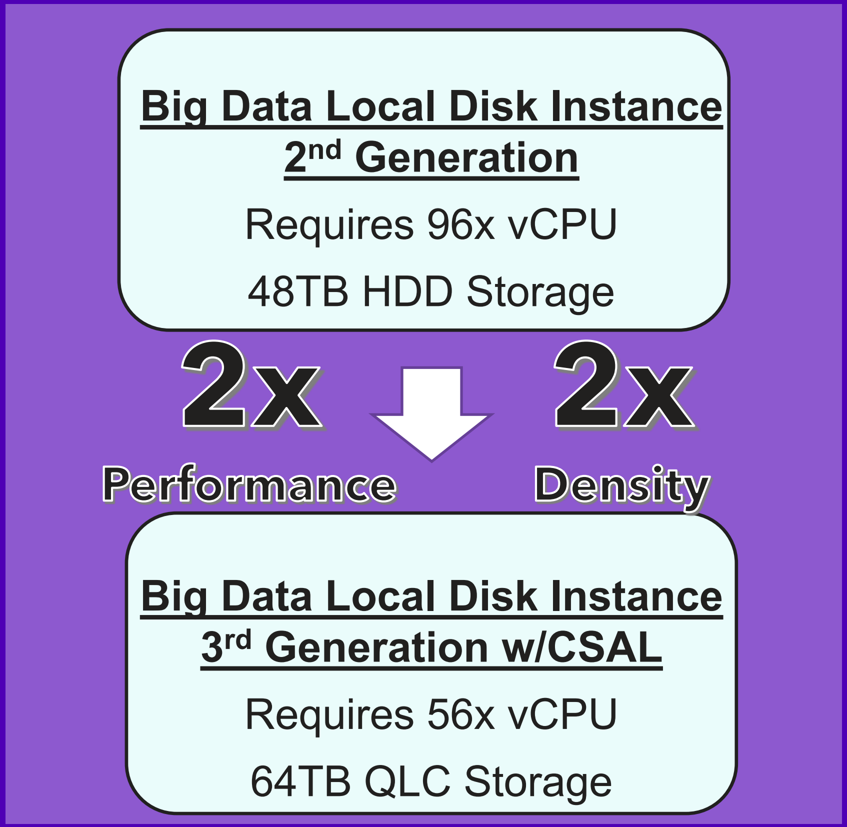
5. Summary and Call to Action

Motivation and Problem

Wide Range of Use Cases/Customers Adopting QLC

Example: Alibaba Local Disk Use Case

- Alibaba replaced HDDs with Solidigm's QLC D5-P5316 QLC SSDs in their 3rd generation big data local disk ECS instances to double the performance vs. 2nd generation while holding the price to their customers constant.
- TCO was the same between the two generations
  - While the CAPEX was higher, the 2x density led to rack tax (building, personnel, land, etc.) and OPEX savings to offset the higher CAPEX.



Please see reference #1 under Sources, References and Test configs section on slide 19

Alibaba local disk use case is a great proof point of QLC successfully replacing HDDs

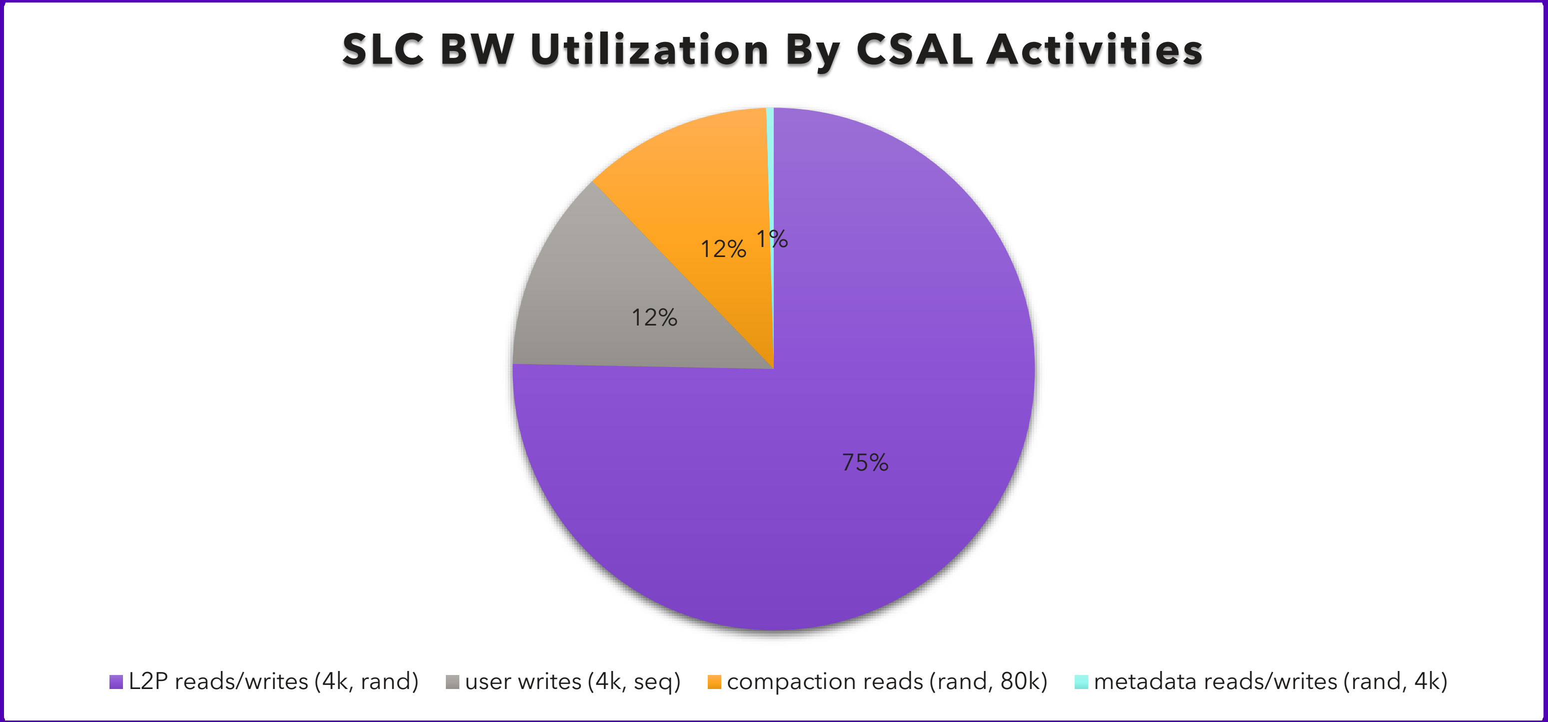
Collaboration with Alibaba, as one of the foundational QLC customers, resulted in co-development of CSAL

Alibaba uses Optane SSDs in their local disk use case. What can we replace Optane SSDs (P5800X) with?

D7-5810 as Optane SSD (P5800X) Replacement for O+Q Deployments

Attributes	Worst-case Use Case Demands	P5800X (Optane) Supplies	P5810 (SLC) Supplies
Endurance	37 DWPD	100 DWPD	50 DWPD
rand 4k write	8 KIOPS	20 KIOPS	10 KIOPS

Based on Solidigm internal analysis



D7-P5810 is engineered to provide best cost-performance-endurance balance to replace P5800X in O+Q deployments

Even Greater TCO Benefits with 61.44TB P5336 and P5810 SSDs

- Applied CSAL to Solidigm's D5-P5336 61.44TB QLC and D7-P5810 SSDs.

Config	8xOptane+ 8xQLC Optane = 400GB QLC = 15.36TB	8xSLC + 8xQLC SLC= 800GB QLC = 61.44TB
1x Drive capacity (GB)	15360	61440
Total storage cap per node (TB)	128	442
Incremental CAPEX for compute (vCPU + host DRAM) and storage (SLC + QLC)	base	+$18K
OPEX per node (5 years)	base	- ($0.5K)
Data center tax per node (5 years)	base	- ($0.3K)
Virtual drive capacity (GB)	16000	16000
Virtual drives per node	8	27
% TCO savings per virtual drive (5 years)	base	**2.5x**

Based on Solidigm internal analysis

How can we bring these TCO benefits of hyper-dense QLC to everyone?

Open-source CSAL is part of the answer!

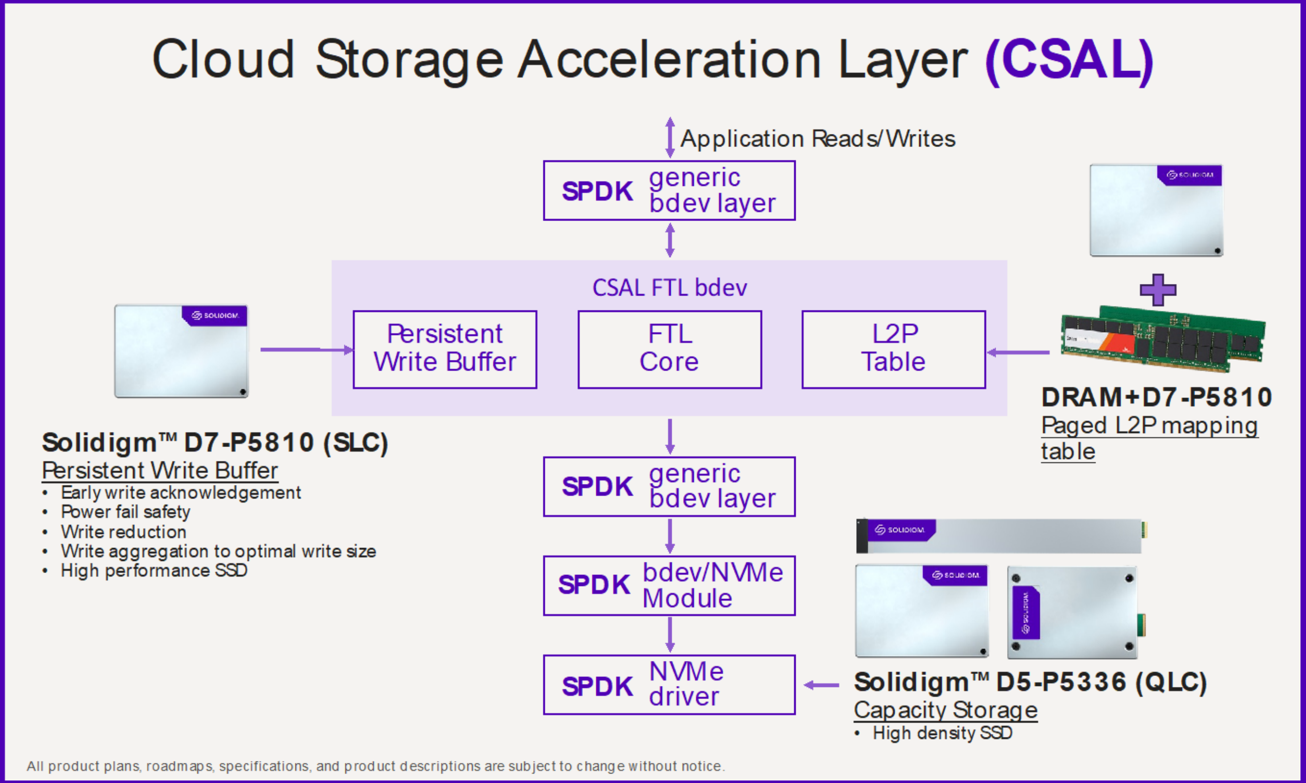
Solution

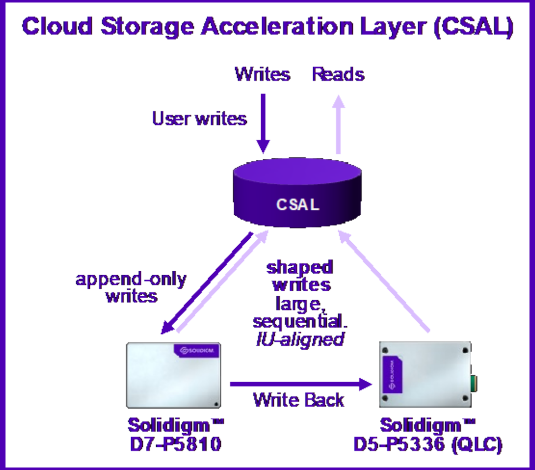
Cloud Storage Acceleration Layer (CSAL)

Refresher from last SDC: What is Cloud Storage Acceleration Layer (CSAL)

What is CSAL?

- Open-source cloud-scale shared-nothing Flash Translation Layer (FTL bdev) in Storage Performance Development Kit (SPDK)
- Ultra fast cache and write shaping tier to improve performance and endurance to scale QLC value
- Consistent performance in multi-tenant environment
- Flexible scaling of NAND performance and capacity to the user/workload needs





Functionality SLC Provides in CSAL?	Benefits
Cache Region (C)	• Write shaping to adapt large IU/ZNS/FDP to 4k block interface. • Boosts perf of workloads with temporal locality, also extends QLC endurance. • Write buffering to absorb bursts, early write completion. • Full-stripe RAID cache for RAID backends
L2P Region (L)	• DRAM cost savings with paged L2P
FTL Metadata Region (M)	• Crash consistency • FTL consistency and TTR after power failure • Superblock

What Changed in CSAL since last SDC

- CSAL open sourced (SPDK v22.09)
- Solidigm acquired CSAL team (Feb. 2023)
- New CSAL capabilities:
    - SLC as Optane Replacement
    - Mitigated in CSAL the need for VSS for crash consistency and power fail safety
    - RAID5F (in progress)
    - ZNS/FDP (in progress)

CSAL's Core Capabilities Expand the QLC Benefits

Capability #1: Write shaping to enable a reduced DRAM footprint with Large IU drives and SLC caching tier can provide additional ~2x endurance and perf benefit for locality workloads vs. TLC

	1xTLC	1xSLC + 1xQLC	Unit
64k rand write zipf 1.2	1875	3317	MiB/s

Please see Test Configuration #2 under Sources, References and Test configs section on slide 22

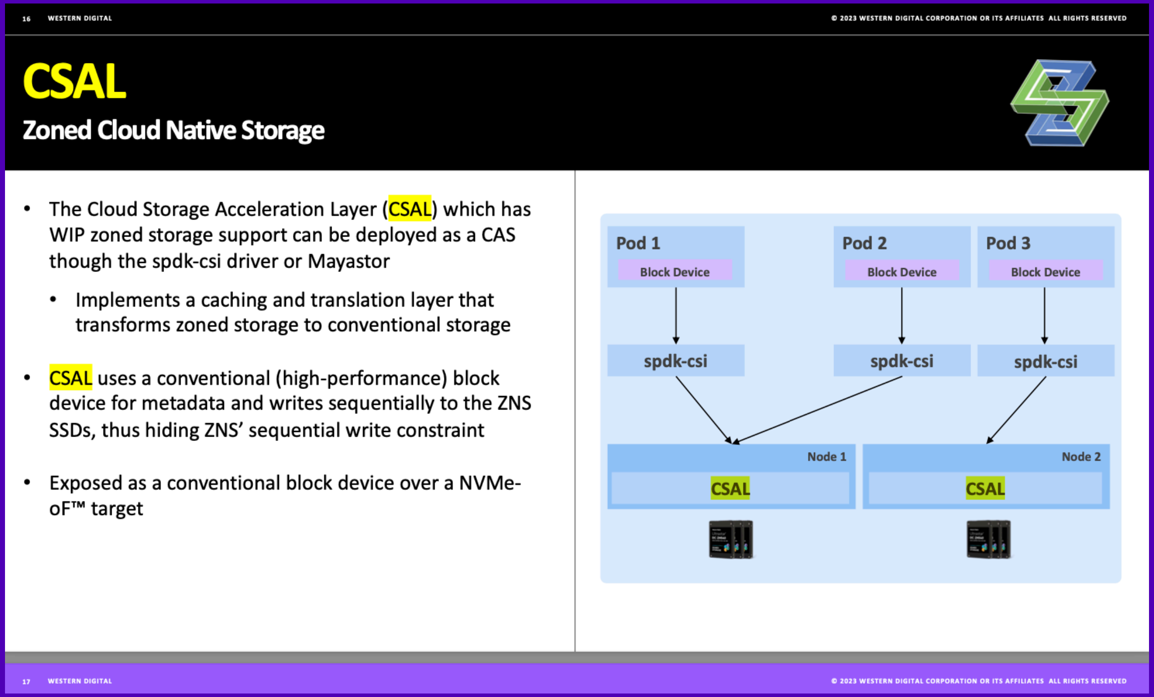
Capability #2: CSAL tiered arch enables full-stripe RAID5F ~2x more efficient than traditional RAID5 to improve system fault tolerance

$$raid5fWritePerf = (N-1) \times diskWritePerf$$

$$raid5WritePerf = N \times \frac{diskWritePerf}{\left(2 + 2 \times \frac{diskWritePerf}{diskReadPerf}\right)} \rightarrow \frac{raid5fWritePerf}{raid5WritePerf} = \sim 2$$

Capability #3: CSAL enables pooling a large QLC capacity that can be shared across multiple cloud tenants to increase capacity and performance utilization.

Capability #4: CSAL writes sequentially to QLC to adapt to emerging interfaces e.g., ZNS to a regular 4k block, and enable multi-tenant isolation.



Please see reference #2 under Sources, References and Test configs section on slide 19

Workload	CSAL on Standard SSD WAF	CSAL on ZNS SSD WAF
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	3.8	2.3

Please see Test Configuration #3 under Sources, References and Test configs section on slide 24

CSAL has key abstractions to extend the use of high-density QLC NAND and adapt the emerging interfaces (ZNS/FDP) to 4k block interface

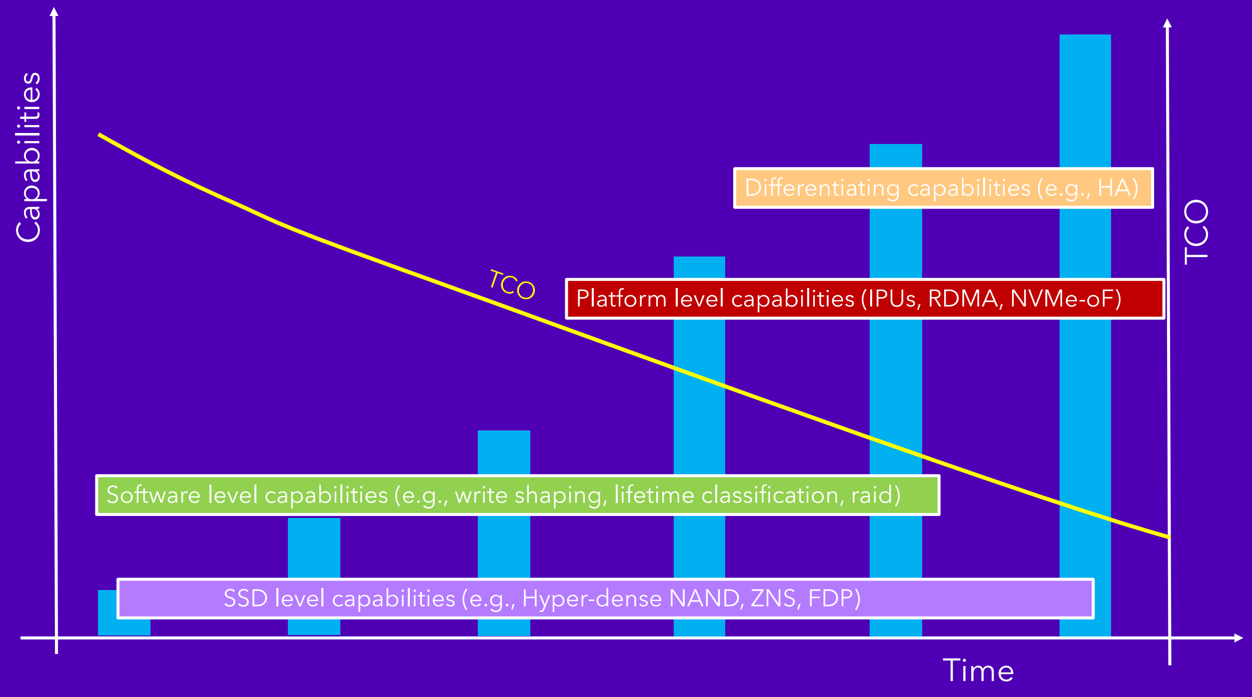
In addition to CSAL, we are taking a community-driven approach to create a Reference Storage Platform (RSP) for everyone

Solution

Reference Storage Platform (RSP)

Why A Reference Storage Platform & Community-driven Approach?

- Open-source building blocks are a complex mix of parts, often challenging to assemble
- This hampers rapid development, assessment, and deployment of storage technologies
- A Reference Storage Platform brings it all together into a turnkey solution
- A community-driven approach enables faster innovation, transparency, and easy evaluation and adoption of "part" technologies inside a unified "whole" solution



The first instantiation of the reference platform already done!

First Reference Storage Platform Instantiation – NVMe-oF Target for Disaggregation

- Thanks to our **Reference Storage Platform (RSP) partners**, we have created the first instantiation of an open-source Reference Storage Platform sandbox
- Reference Storage Platform provides an easy button
    - SPDK NVMe-oF TCP target packaged in a turnkey VM image
    - SPDK NVMe-oF TCP target packaged in a turnkey Container
    - GUI to manage a pool of high-density NAND
    - Reference hardware platform is an off the shelf commodity server from typical OEMs (Dell and Supermicro to start) with Intel CPUs.
    - Getting started guide on spdk.io and Solidigm's website

RSP Partners:

StarWind HYPERCONVERGENCE

intel

SOLIDIGM

Western Digital

SPDK

Next, we used this reference storage platform to illustrate the benefit of our hyper-dense QLC P5336 61.44 TB

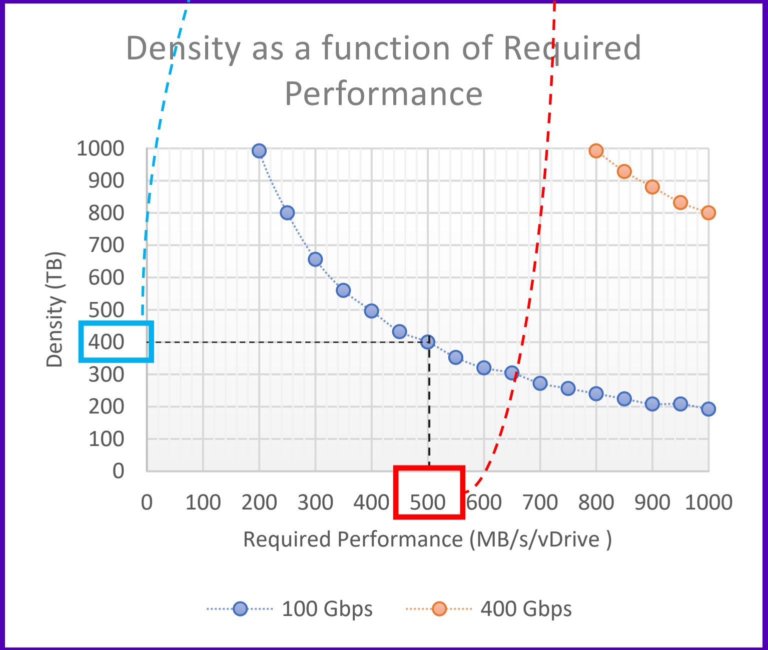
Solution

Performance Results and TCO Benefits

Disaggregated CSAL+SLC+QLC Perf/TCO vs. Incumbent TLC

	10x TLC	3xSLC + 7xQLC	Unit
Total available capacity	138	344	TB
Raw aggregate R/W BW	14	10.3	GB/s
Network Bound	11.4		GB/s
vDrive Capacity	16000		GB
vDrive Count per server	8	21	
Min(Demanded, Delivered) Perf/vDrive	500	500	MiB/s/16TB

Please see Test Configuration #1 under Sources, References and Test configs section on slide 20



Based on Solidigm internal analysis

Config	10x TLC TLC = 15.36TB	3xSLC + 7xQLC SLC = 800GB, QLC = 61.44TB
1x Drive capacity (GB)	15360	61440
Total storage capacity per node (TB)	138	344
Incremental CAPEX for compute (vCPU + host DRAM) and storage (SLC + QLC)	base	+$14K
OPEX per node (for 5 years)	base	base
Data center tax per node (for 5 years)	base	base
Virtual drive capacity (GB)	16000	16000
Virtual drives per node	8	21
% TCO savings per virtual drive (for 5 years)	base	**35%**

Based on Solidigm internal analysis

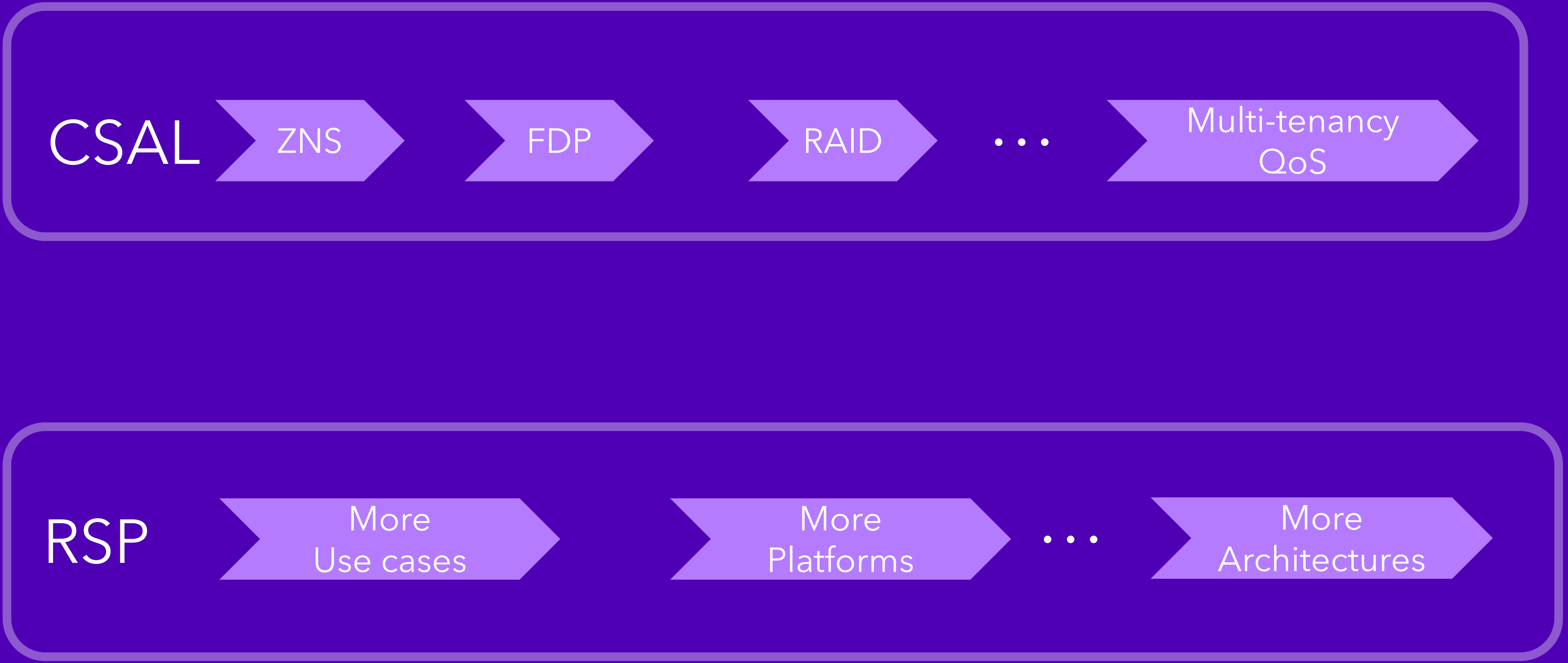
Both TLC and QLC saturate the 80% of the 100Gbps network, but CSAL+SLC+QLC does it with 35% better TCO with D5-P5336

The 35% TCO gain is attributable only to greater density; disaggregation, caching, raid5f, ZNS/FDP, etc. capabilities further the TCO reduction…

Solution

Future and Roadmap

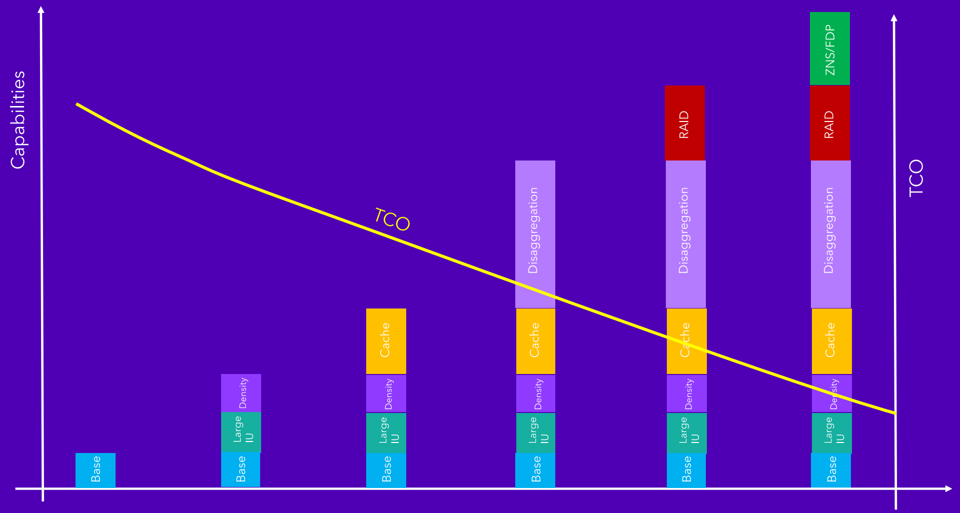
CSAL and Reference Storage Platform (RSP) Add Capabilities Over Time



Summary and Call to Action

Summary:

- Solidigm's D7-**P5810** provides optimal balance of cost, performance, and endurance to replace P5800X Optane SSDs and enable O+Q use cases
- CSAL has the necessary host-based FTL abstractions you need for the emerging SSD interfaces and high-density NAND
- We have provided a turnkey open-source disaggregated NVMe-oF TCP target as our first instantiation of the Reference Storage Platform
- CSAL and Reference Storage Platform provide an easy way to adopt new technologies
- Continuing technology additions drive TCO lower



Call to Action:

- Develop CSAL with us in the SPDK open source
- Use the NVMe-oF target we've provided and see the great things you can do with it.
- Add your own capabilities and create more reference implementations
- Let's grow this community for the benefit of the entire storage industry!

We invite you to play with us in the sandbox!

Please take a moment to rate this session.

Your feedback is important to us.

Sources and References

Back up content supporting the main presentation

Sources and References

1. CSAL whitepaper – A Media-Aware Cloud Storage Acceleration Layer (CSAL)
2. CSAL+ZNS presentation – Zoned Storage in the Cloud
3. CSAL solution brief – A CSAL-Based Reference Storage Platform
4. Reference Storage Platform – Main Download Page
5. CSAL SDC 2022 Presentation – Enabling Unprecedented Perf and Capacity with Optane and QLC Flash

Author Contact: kapil.karkra@solidigm.com

Test Configuration #1

Storage Server – SuperMicro SYS-220U-TNR System Configuration	
BIOS Version	1.4b
OS	Fedora 37 (Server Edition)
Kernel	6.3.8-100.fc37.x86_64
CPU Model	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz
NUMA Node(s)	2
DRAM Installed	756GB (16x16GB DDR4 3200MT/s [3200MT/s])
Huge Pages Size	2048 kB
NIC Summary	Ethernet Controller X710 for 10GBASE-T, Ethernet Controller X710 for 10GBASE-T
Drive Summary	**3x SLC+ 7x QLC**: SLC is Solidigm's first generation SLC for cache device; QLC is a P5336 D5-P5336 61TB
SPDK	22.09
CSAL	1.0
FIO	3.29

Test Configuration #1: Example FIO job file

```
[global]
ioengine=spdk_bdev
spdk_json_conf=${FTL_JSON_CONF}
filename=${FTL_BDEV_NAME}
# SPDK cores, FTL core mask should avoid core 0
spdk_core_mask=${SPDK_CORE_MASK}
# CPUS allowed fio threads cannot interleave with SPDK cores
cpus_allowed=12
cpus_allowed_policy=split
direct=1
thread=1
buffered=0
time_based
norandommap=1
randrepeat=0
scramble_buffers=1
rw=randrw

[POR]
bs=4k
rwmixread=70
numjobs=1
iodepth=128
runtime=3600s
time_based=1
```

Test Configuration #2

Storage Server – SuperMicro SYS-220U-TNR System Configuration	
BIOS Version	1.4b
OS	Fedora 37 (Server Edition)
Kernel	6.3.8-100.fc37.x86_64
CPU Model	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz
NUMA Node(s)	2
DRAM Installed	756GB (16x16GB DDR4 3200MT/s [3200MT/s])
Huge Pages Size	2048 kB
NIC Summary	Ethernet Controller X710 for 10GBASE-T, Ethernet Controller X710 for 10GBASE-T
Drive Summary	1. **TLC** is a Solidigm TLC SSD D7-P5520 15.36 TB 2. **1x SLC+ 1x QLC:** SLC is Solidigm's first generation SLC for cache device; QLC is a P5336 D5-P5336 61TB
SPDK	22.09
CSAL	1.0
FIO	3.29

Test Configuration #2: Example FIO job file

```
[global]
ioengine=spdk_bdev
spdk_json_conf=${FTL_JSON_CONF}
filename=${FTL_BDEV_NAME}
# SPDK cores, FTL core mask should avoid core 0
spdk_core_mask=${SPDK_CORE_MASK}
# CPUS allowed fio threads cannot interleave with SPDK cores
cpus_allowed=12
cpus_allowed_policy=split
direct=1
thread=1
buffered=0
time_based
norandommap=1
randrepeat=0
scramble_buffers=1
rw=randrw

[POR]
bs=64k
numjobs=1
rw=randwrite
random_distribution=zipf:1.2
iodepth=128
runtime=3600s
time_based=1
```

Test Configuration #3

Storage Server – SuperMicro SYS-220U-TNR System Configuration	
BIOS Version	1.4b
OS	Fedora 37 (Server Edition)
Kernel	6.3.8-100.fc37.x86_64
CPU Model	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz
NUMA Node(s)	2
DRAM Installed	756GB (16x16GB DDR4 3200MT/s [3200MT/s])
Huge Pages Size	2048 kB
NIC Summary	Ethernet Controller X710 for 10GBASE-T, Ethernet Controller X710 for 10GBASE-T
Drive Summary	CSAL ZNS POC branch (not yet upstream) with P5800X + **WDC ZN540 ZNS TLC SSD**
SPDK	22.09
CSAL	1.0
FIO	3.29

Test Configuration #3: Example FIO job file

```
[global]
ioengine=spdk_bdev
spdk_json_conf=${FTL_JSON_CONF}
filename=${FTL_BDEV_NAME}
# SPDK cores, FTL core mask should avoid core 0
spdk_core_mask=${SPDK_CORE_MASK}
# CPUS allowed fio threads cannot interleave with SPDK cores
cpus_allowed=12
cpus_allowed_policy=split
direct=1
thread=1
buffered=0
norandommap=1
randrepeat=0
scramble_buffers=1
```

```
[WRITE_SEQ]
bs=4k
numjobs=1
rw=write
iodepth=128
size=100%
exitall

[WRITE_RAND]
bs=4k
numjobs=1
rw=randwrite
iodepth=128
runtime=1000d

[WRITE_ZIPF_0_8]
bs=4k
numjobs=1
rw=randwrite
random_distribution=zipf:0.8
iodepth=128
runtime=1000d
time_based=1

[WRITE_ZIPF_1_2]
bs=4k
numjobs=1
rw=randwrite
random_distribution=zipf:1.2
iodepth=128
runtime=1000d
time_based=1
```

Test Configuration #4

Storage Server – SuperMicro SYS-220U-TNR System Configuration	
BIOS Version	1.4b
OS	Fedora 37 (Server Edition)
Kernel	6.3.8-100.fc37.x86_64
CPU Model	Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz
NUMA Node(s)	2
DRAM Installed	756GB (16x16GB DDR4 3200MT/s [3200MT/s])
Huge Pages Size	2048 kB
NIC Summary	Ethernet Controller X710 for 10GBASE-T, Ethernet Controller X710 for 10GBASE-T
Drive Summary	**1x SLC+ 1x QLC:** SLC P5810 800GB SLC for cache device; QLC is a P5316 D5-P5316
SPDK	22.09
CSAL	1.0
FIO	3.29

Test Configuration #2: Example FIO job file

```
global]
ioengine=spdk_bdev
spdk_json_conf=${FTL_JSON_CONF}
filename=${FTL_BDEV_NAME}
# SPDK cores, FTL core mask should avoid core 0
spdk_core_mask=${SPDK_CORE_MASK}
# CPUS allowed fio threads cannot interleave with SPDK core
s
cpus_allowed=12
cpus_allowed_policy=split
direct=1
thread=1
buffered=0
time_based
norandommap=1
randrepeat=0
scramble_buffers=1

[POR]
bs=4k
numjobs=1
rw=randwrite
iodepth=128
runtime=3600s
time_based=1
```