



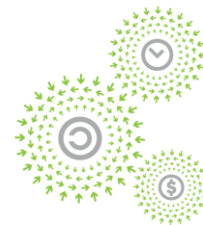
STORAGE

OCP Storage Tech Talk 2024

Ross Stenfort, Meta, Storage Project Co-Lead

Lee Prewitt, Microsoft, Storage Project Co-Lead

Mike Allison, Samsung, Storage Steering Committee Representative



OPEN
PLATINUM™

2024 OCP Storage Tech Talk

May 15, 2024

Thank you to our Sponsors



SAMSUNG





Celebrate Our Wins

2022

NVMe HDD rev. 0.7

[\[spec\]](#)

Microsoft, Seagate, Western Digital

Barton Springs 1S Server

[\[spec\]](#)

Meta

Grand Canyon Storage System

[\[spec\]](#)

Meta

Poseidon E1.S SSD Storage System

[\[spec\]](#)

Samsung and Inspur

Hyperscale NVMe Boot Spec

[\[spec\]](#)

Google and Meta

1st OCP Storage Teck Talk

[\[virtual\]](#)

April 31st

2023

NVMe HDD rev. 1.0

[\[spec\]](#)

Microsoft, Seagate, Western Digital

Datacenter SAS-SATA Device rev. 1.0

[\[spec\]](#)

HPE, Meta, Microsoft

Datacenter NVMe SSD rev. 2.5

[\[spec\]](#)

Dell, Google, HPE, Meta, Microsoft

NVMe Telemetry Scripts

[\[scripts\]](#)

Samsung

2nd OCP Storage Teck Talk

[\[virtual\]](#)

May 16^{sh}

2024

Datacenter SAS-SATA Device rev. 1.X

Coming

HPE, Dell, Meta, Microsoft

Datacenter NVMe SSD rev. 2.6

Coming

Dell, Google, HPE, Meta, Microsoft

EDSFF SSD Reference Design

Coming

ScaleFlux

OCP Marketplace

Today

8 SSDs and Multiple Testing Services

3rd OCP Storage Teck Talk

Today

Storage Tech Talk Agenda

9:00-9:10	Introduction
9:10-9:30	Live Migration
9:30-10:00	Flexible Data Placement (FDP) in the Real World Panel
10:00-10:15	OCP Community Update
10:15-10:30	Device Specification Updates
10:30-10:45	Storage Resiliency leveraging OCP Datacenter NVMe SSD rev. 2.5
10:45-11:05	L.O.C.K.
11:05-11:35	Hyperscale Fleet Learnings
11:35-11:45	Close

Action Packed Line-up of Talks

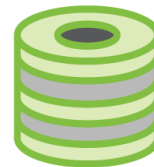
Virtual Event Logistics

- Be present for the great content and engage

- Please type your questions into the chat box
 - Questions will be answered based on time available
 - We will follow up on unanswered questions after the event

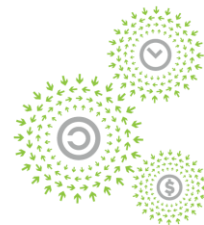
- Recording will be available after the event

NVM Express® Host Managed Live Migration



STORAGE

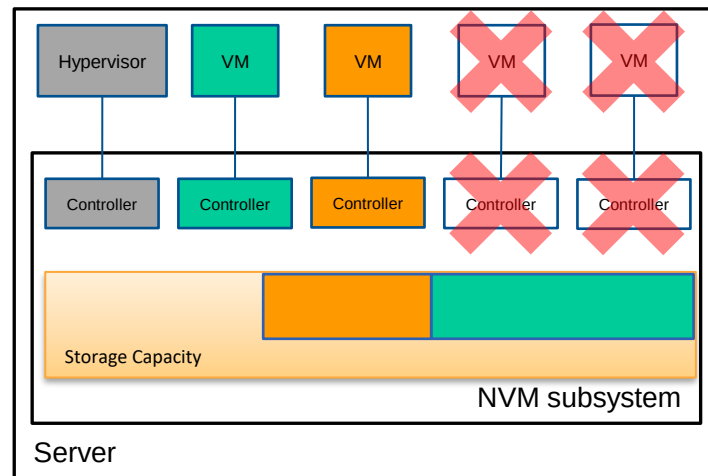
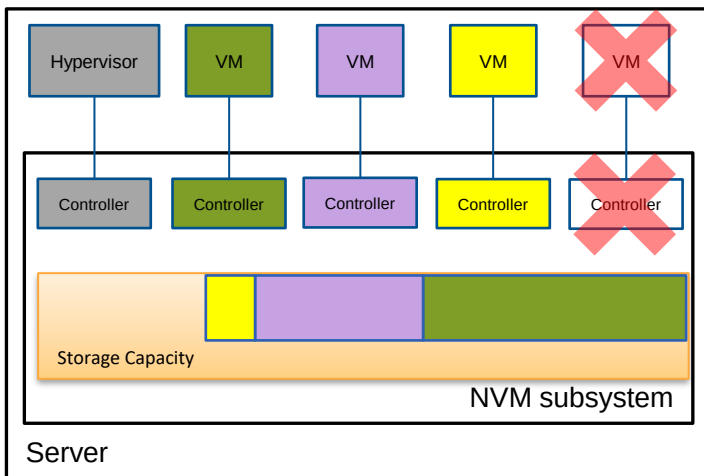
Mike Allison
Sr. Director NAND Product Planning – Standards
Samsung



OPEN
PLATINUM™

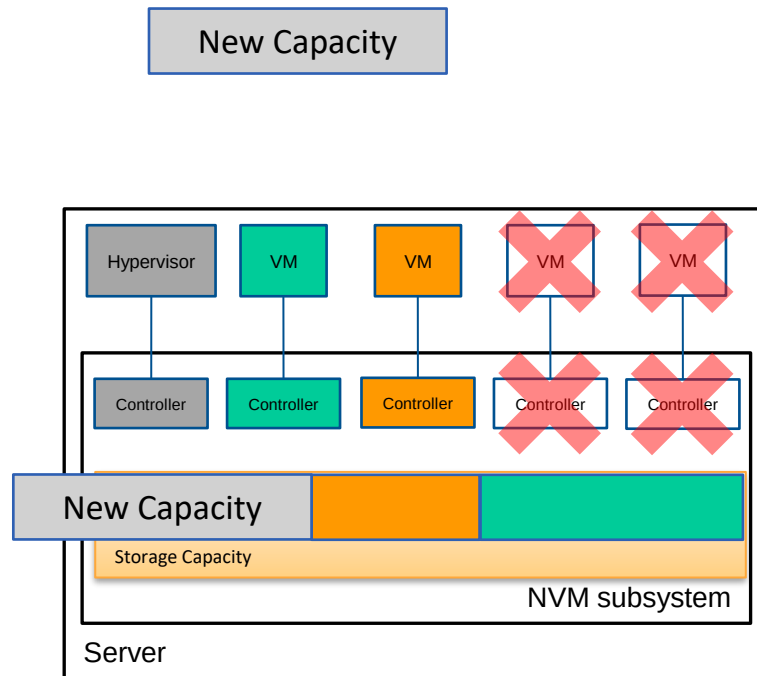
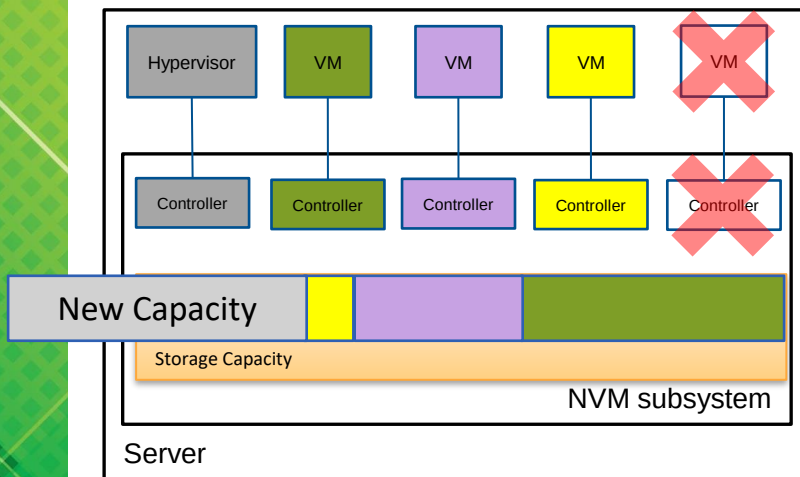
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload



Why Live Migration in NVMe™?

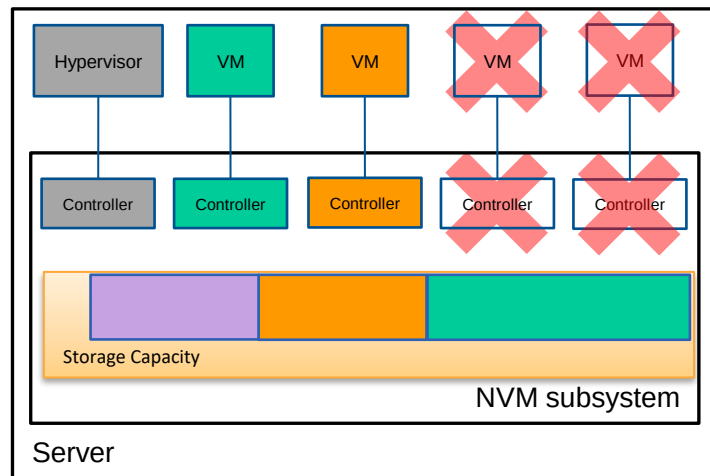
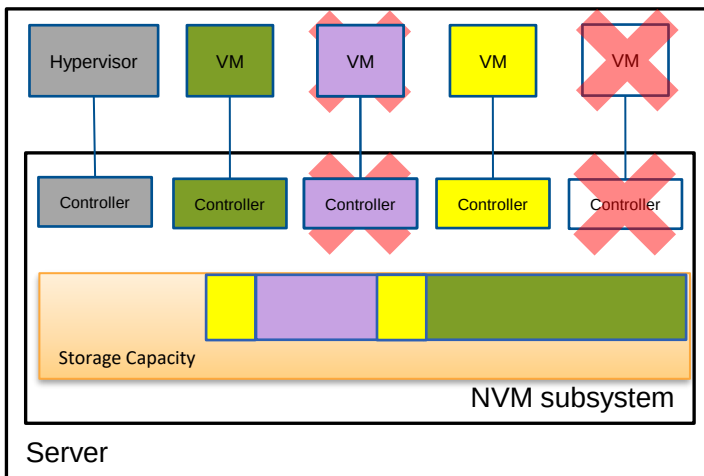
- Key Use Cases:
 - Load balancing for a new workload



Why Live Migration in NVMe™?

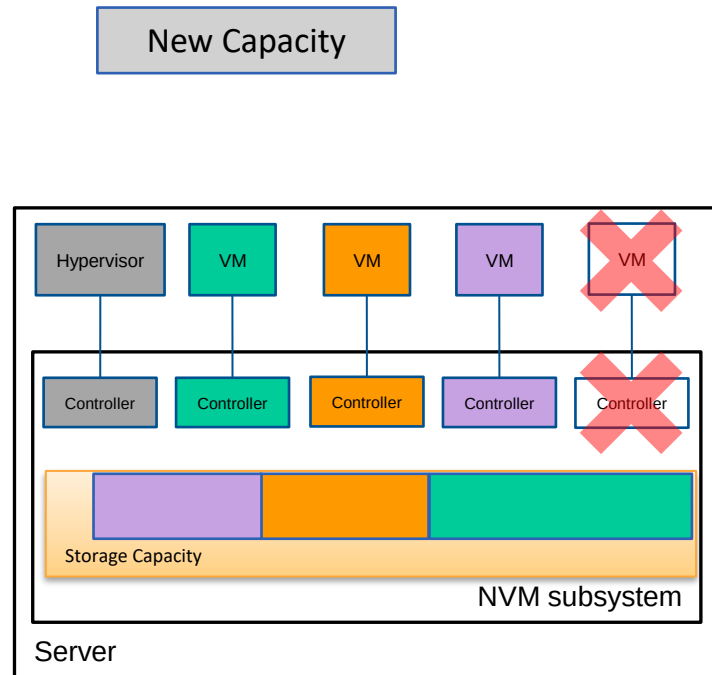
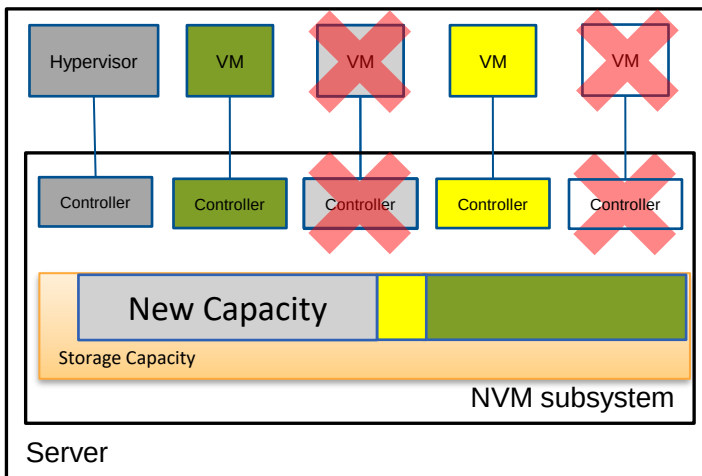
- Key Use Cases:
 - Load balancing for a new workload

New Capacity



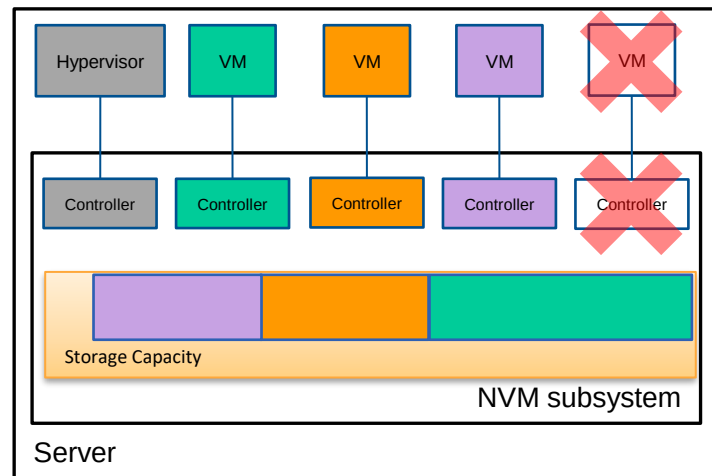
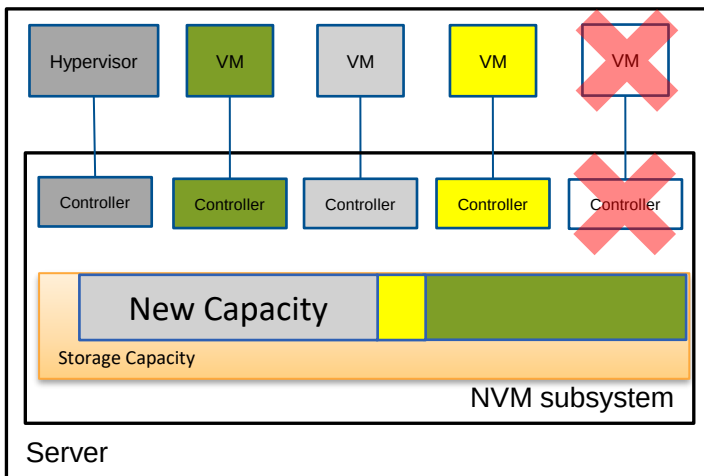
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload



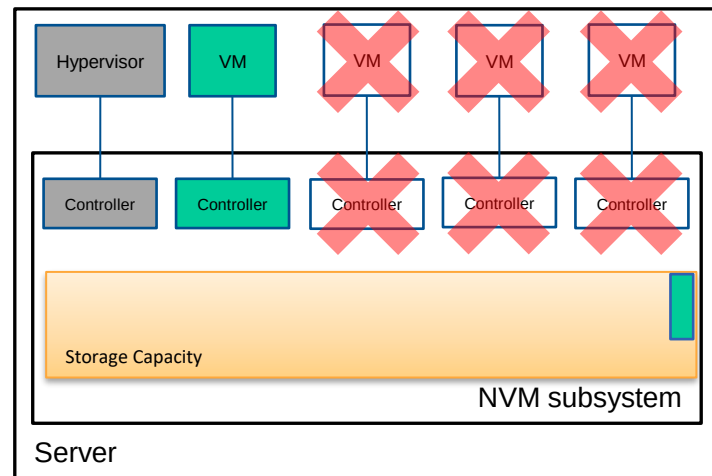
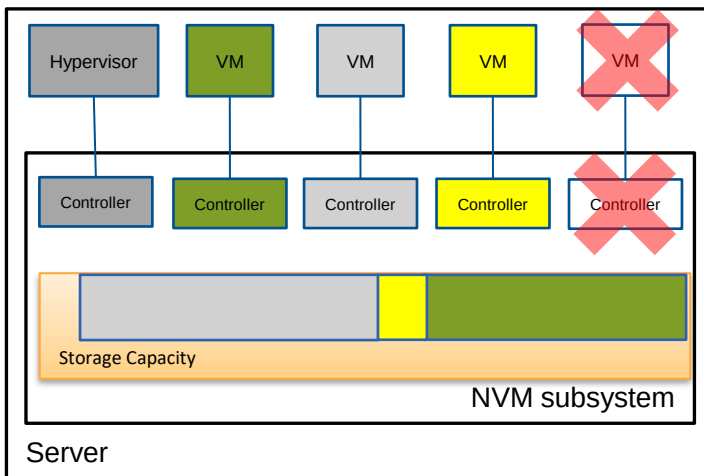
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload



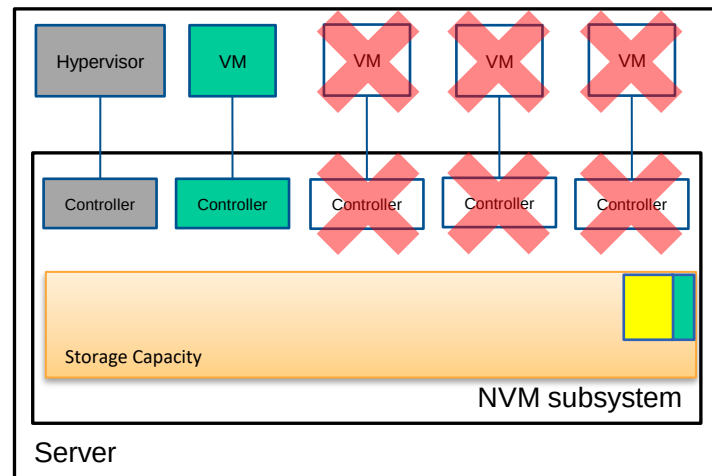
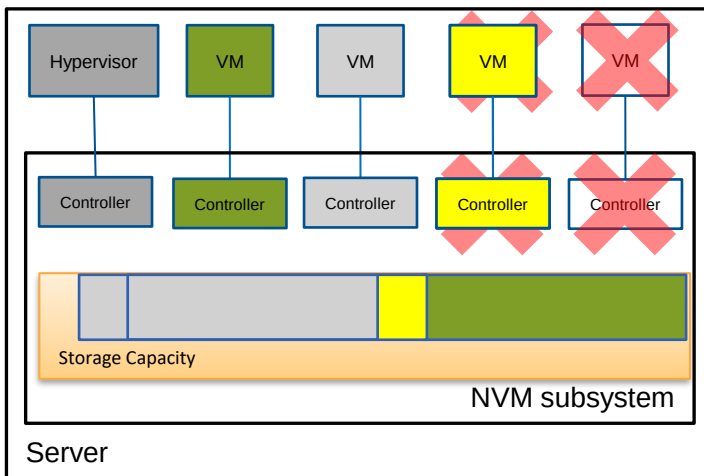
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload
 - Hardware maintenance



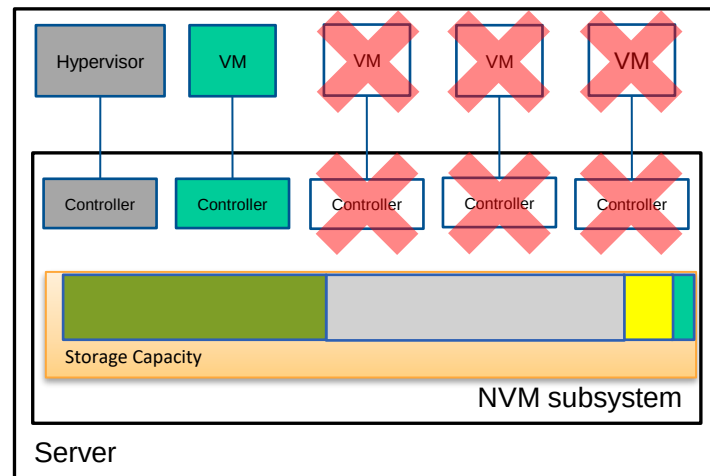
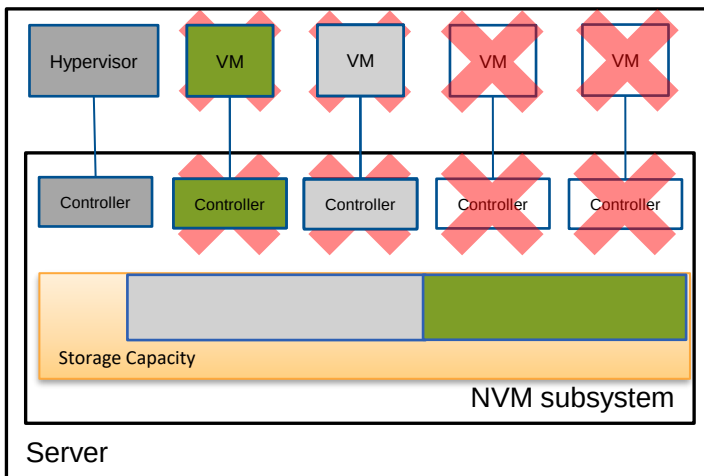
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload
 - Hardware maintenance



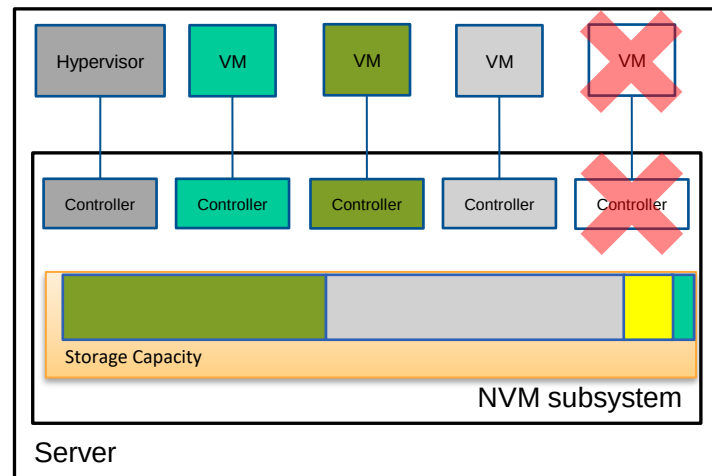
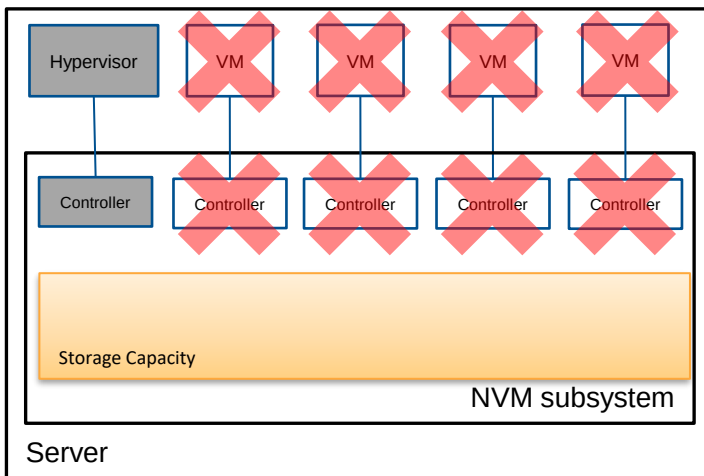
Why Live Migration in NVMe™?

- Key Use Cases:
 - Load balancing for a new workload
 - Hardware maintenance



Why Live Migration in NVMe™?

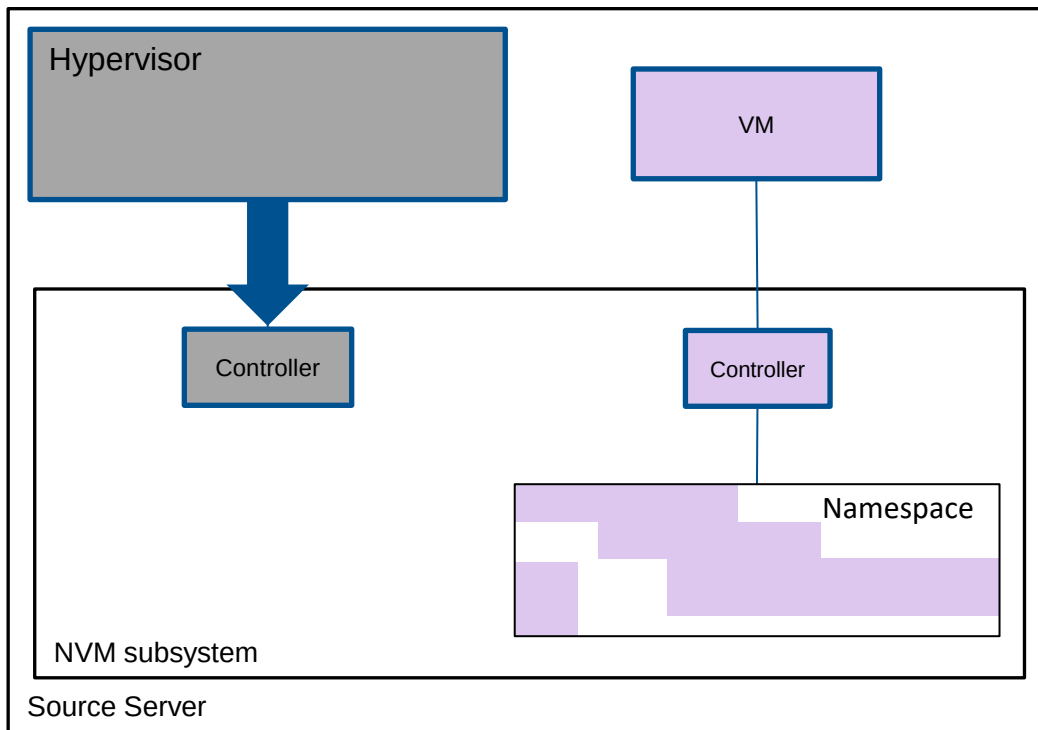
- Key Use Cases:
 - Load balancing for a new workload
 - Hardware maintenance



NVMe™ Host Managed Live Migration

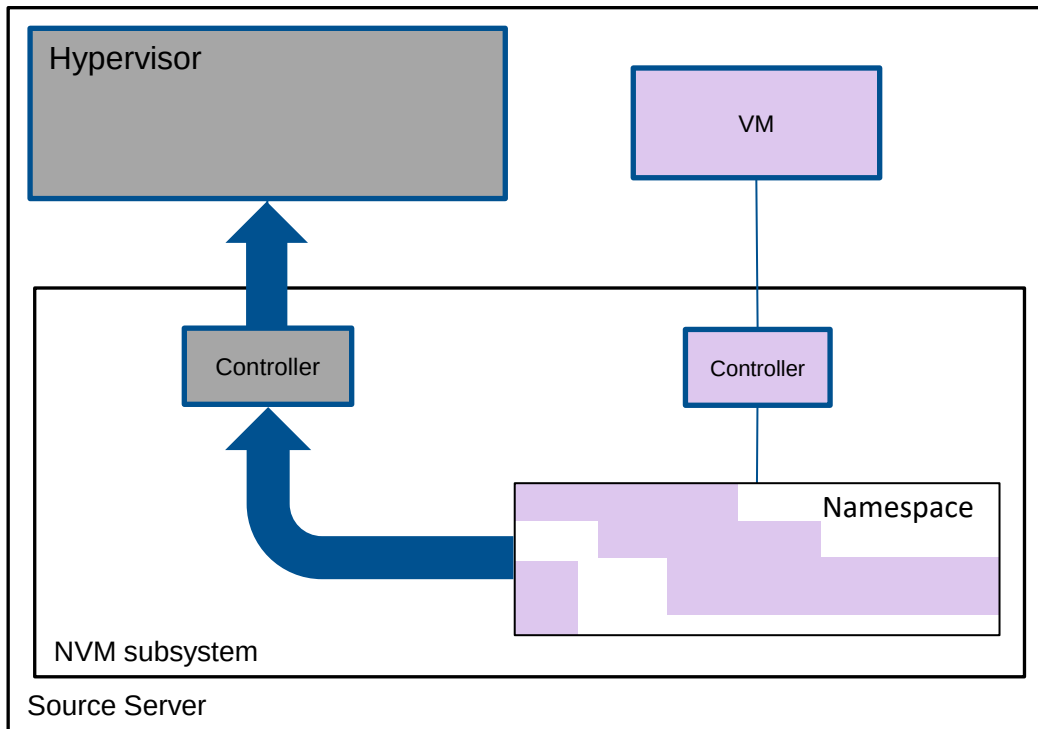
- NVMe Host Managed Live Migration provides services to a Host to:
 - Get the NVMe state from a source controller
 - Put that NVMe state into the target and resume operation
- Source NVM Subsystem
 - Read allocated logical blocks

NVMe™ Host Managed Live Migration



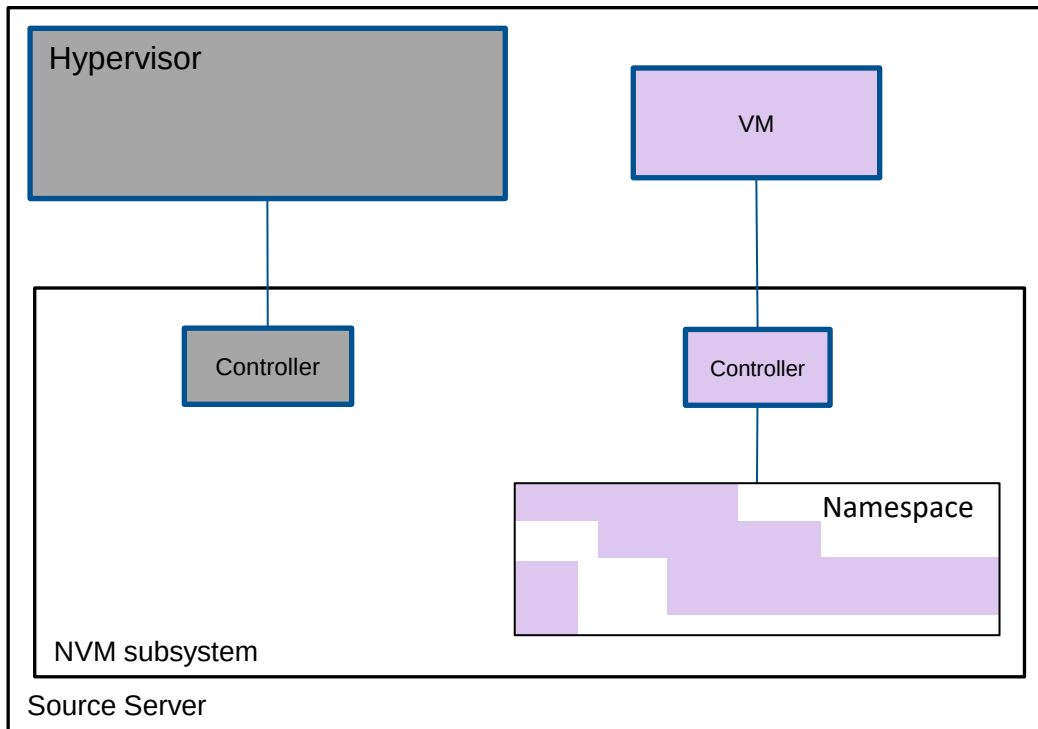
- Source NVM Subsystem
 - Read allocated logical blocks
 - Ratified TP4165 Tracking LBA Allocation with Granularity
 - Allows the host to request allocated LBA reporting

NVMe™ Host Managed Live Migration



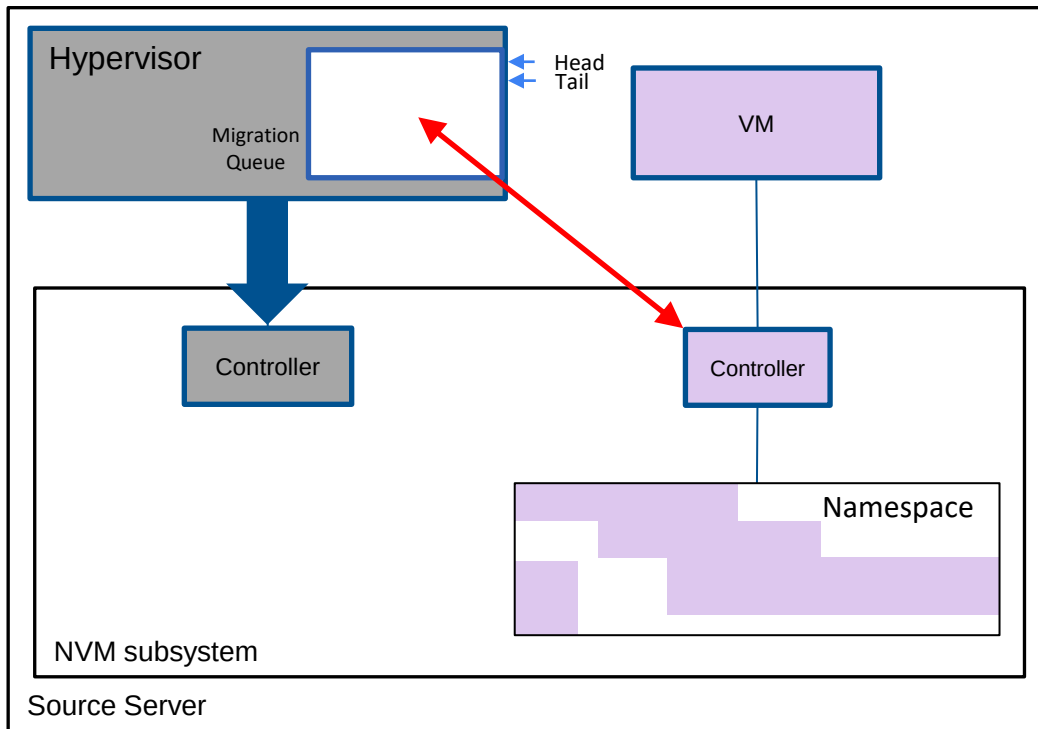
- Source NVM Subsystem
 - Read allocated logical blocks
 - Ratified TP4165 Tracking LBA Allocation with Granularity
 - Allows the host to request allocated LBA reporting
 - Controller reports sequential allocated LBAs on a reported granularity
 - Can be used for other use cases like Snapshots

NVMe™ Host Managed Live Migration



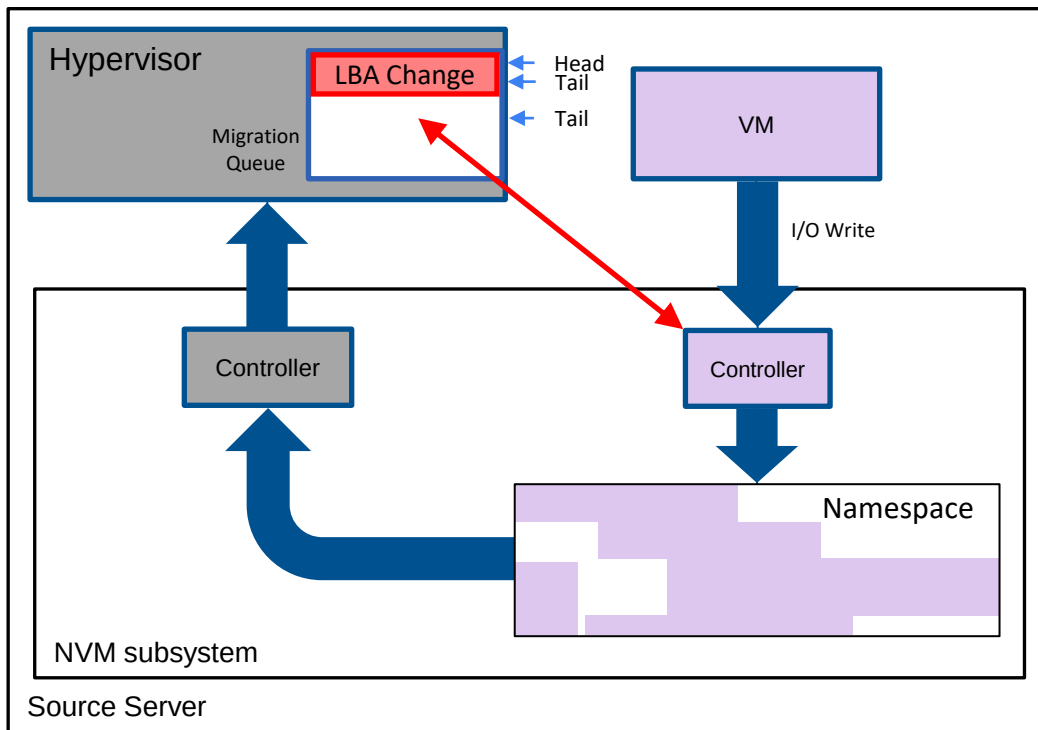
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - For the use case of migrating PCIe SSD namespaces to another PCIe SSDs

NVMe™ Host Managed Live Migration



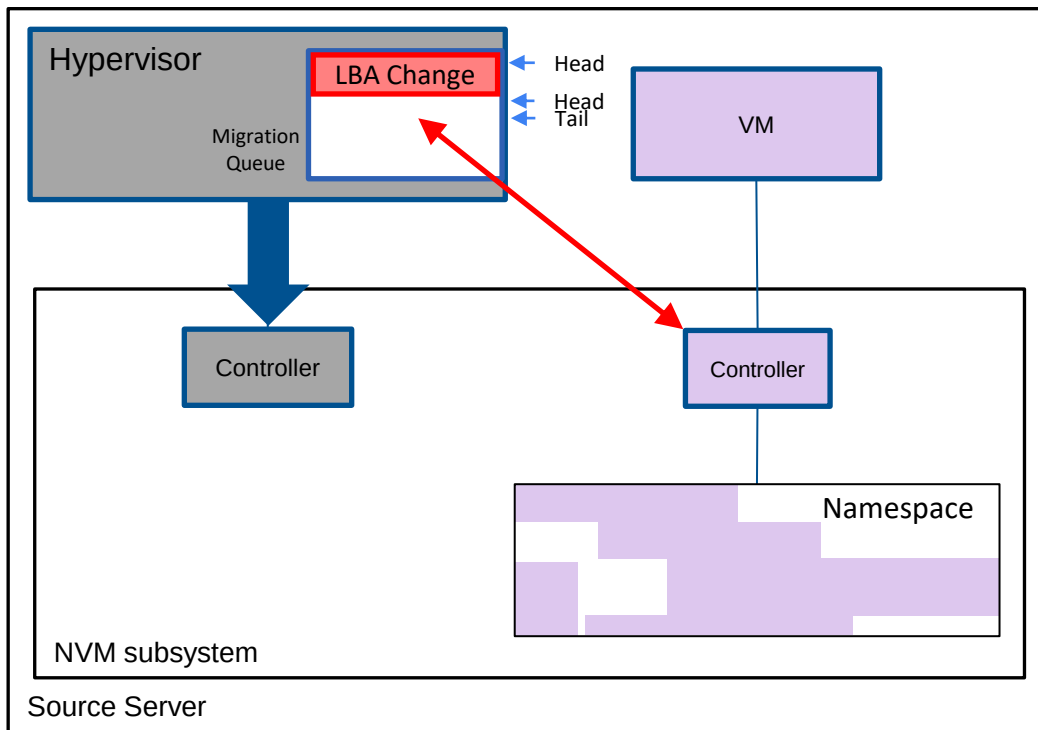
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - Create a Migration Queue
 - Start Logging

NVMe™ Host Managed Live Migration



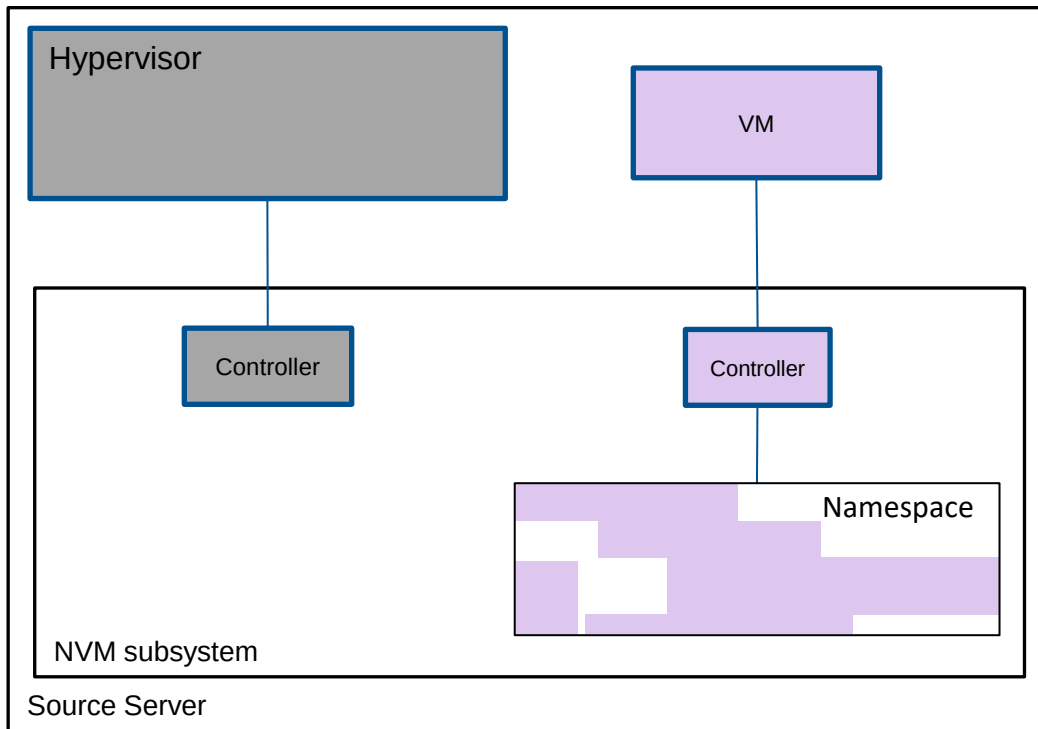
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - Create a Migration Queue
 - Start Logging
 - Post logical block changes

NVMe™ Host Managed Live Migration



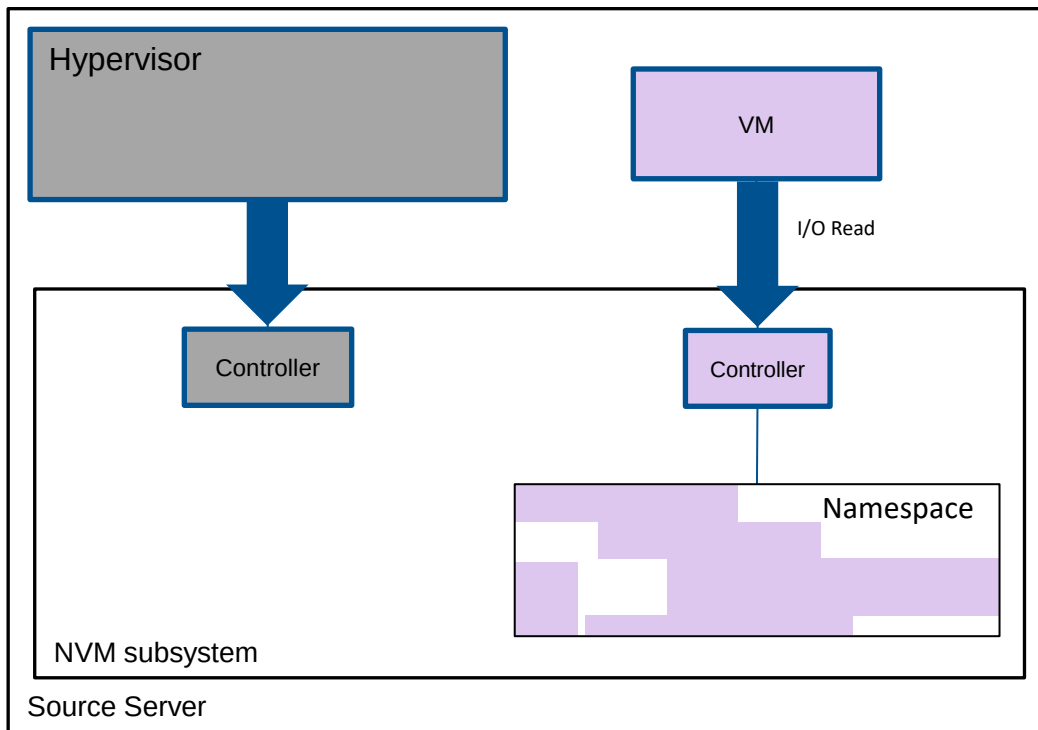
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - Create a Migration Queue
 - Start Logging
 - Post logical block changes
 - Update the Head pointer

NVMe™ Host Managed Live Migration



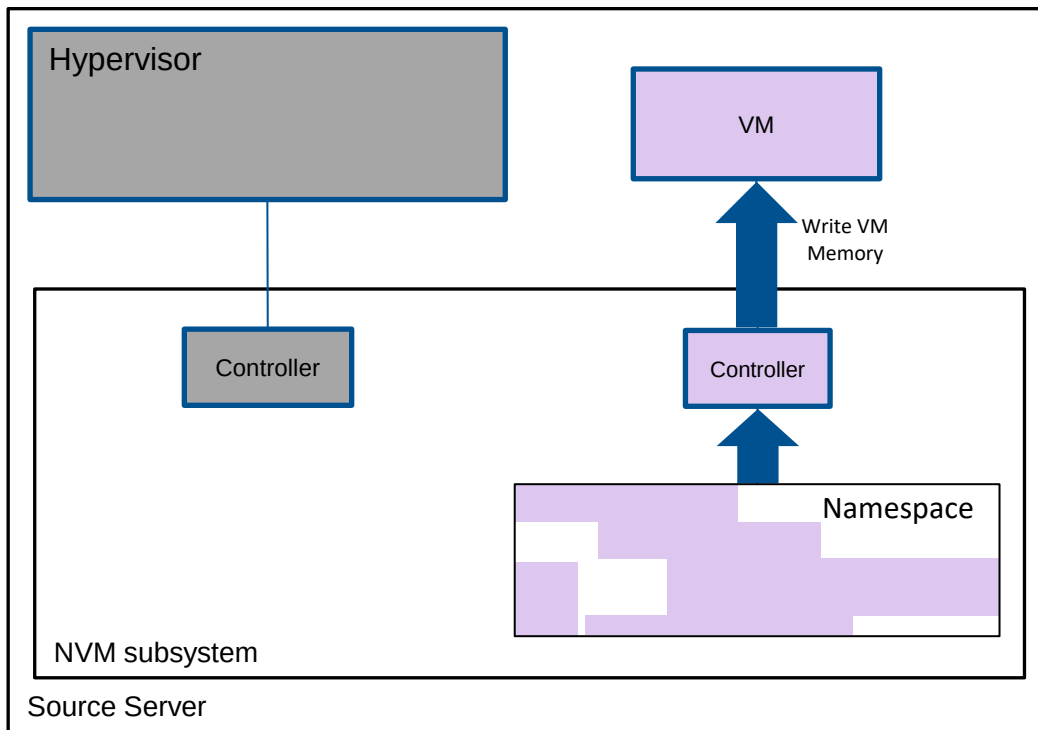
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Use case for migrating DPU controllers using NVMe-oF

NVMe™ Host Managed Live Migration



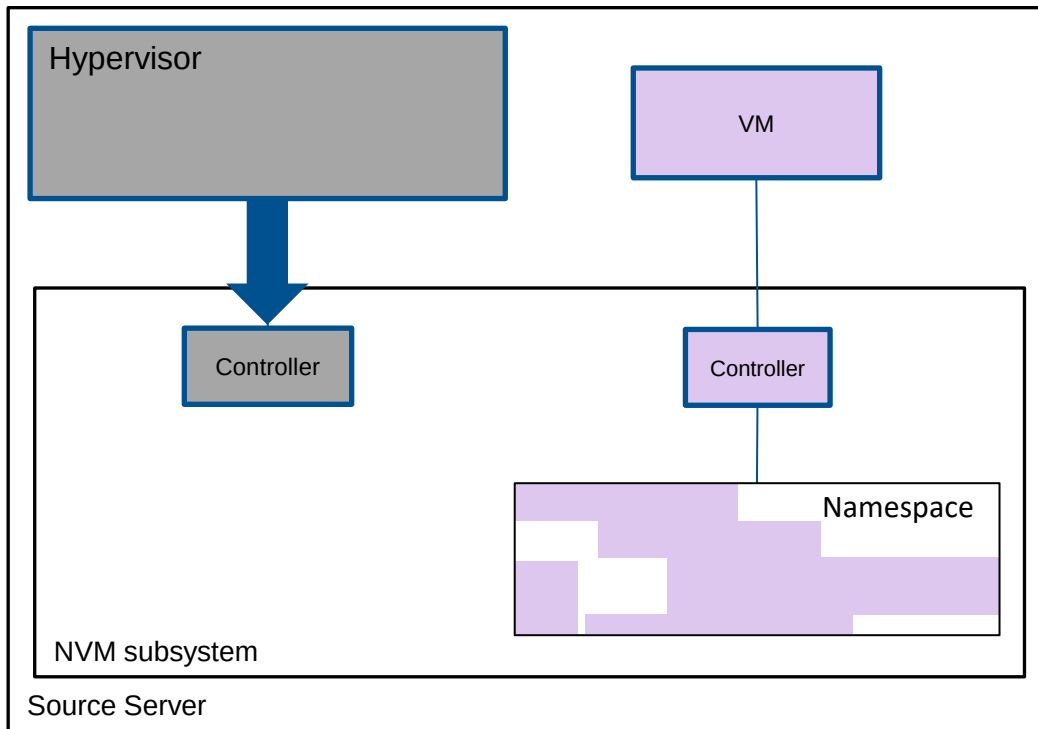
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Start logging and aggregating changes

NVMe™ Host Managed Live Migration



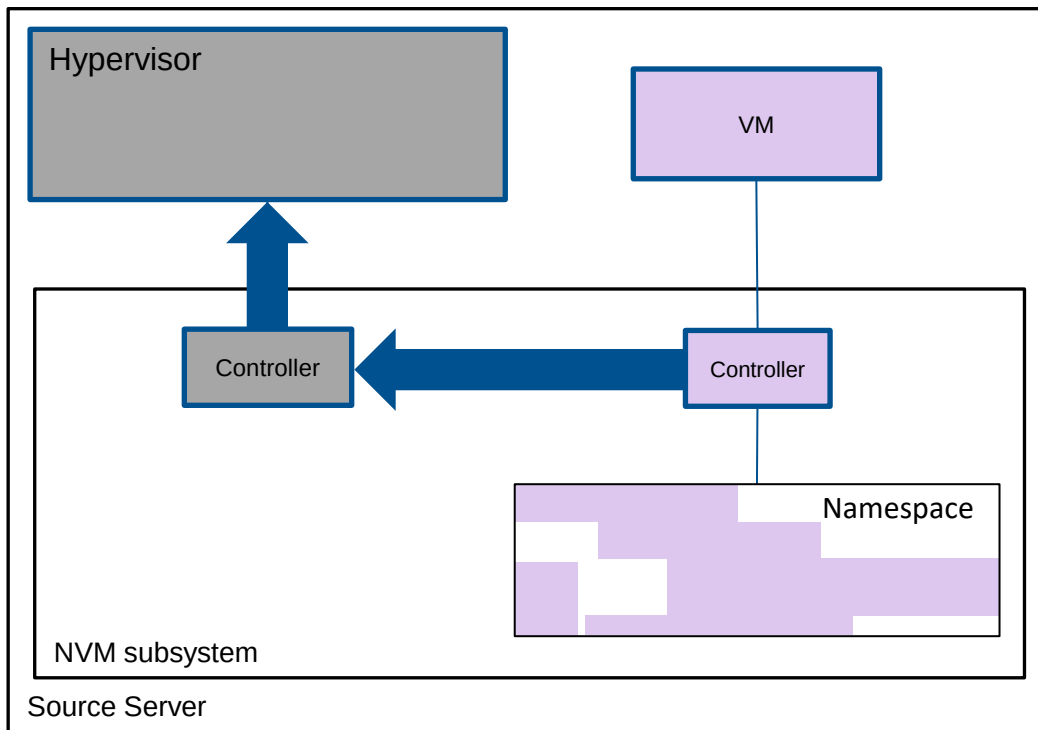
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Start logging and aggregating changes

NVMe™ Host Managed Live Migration



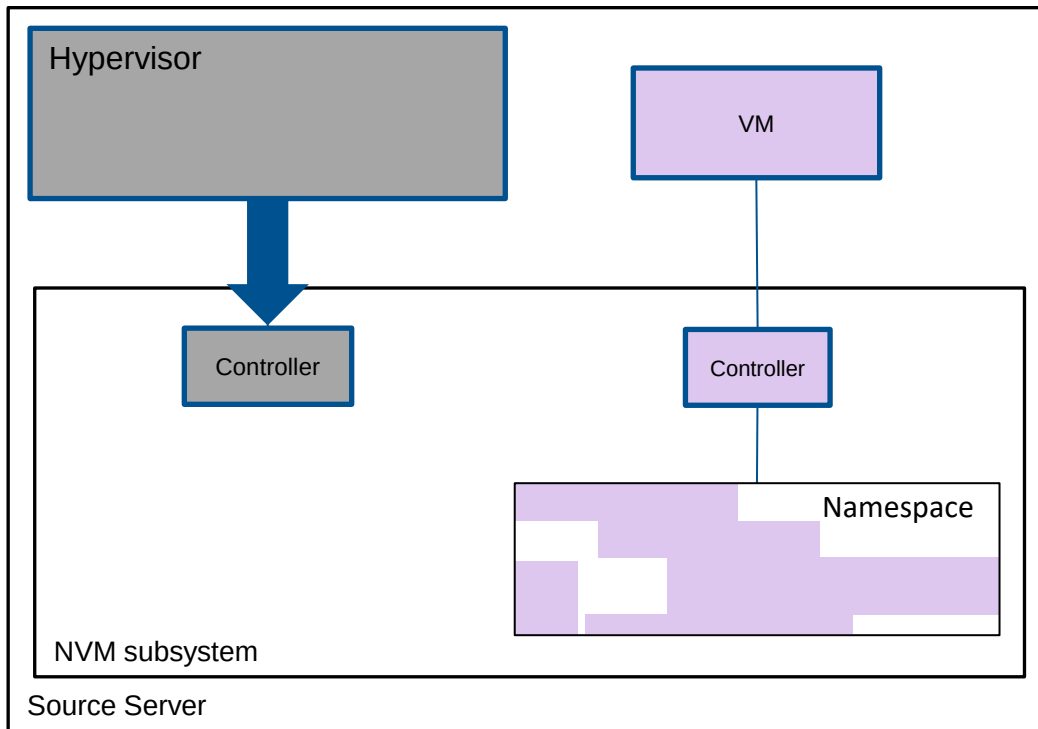
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Start logging and aggregating changes
 - Query aggregated changes

NVMe™ Host Managed Live Migration



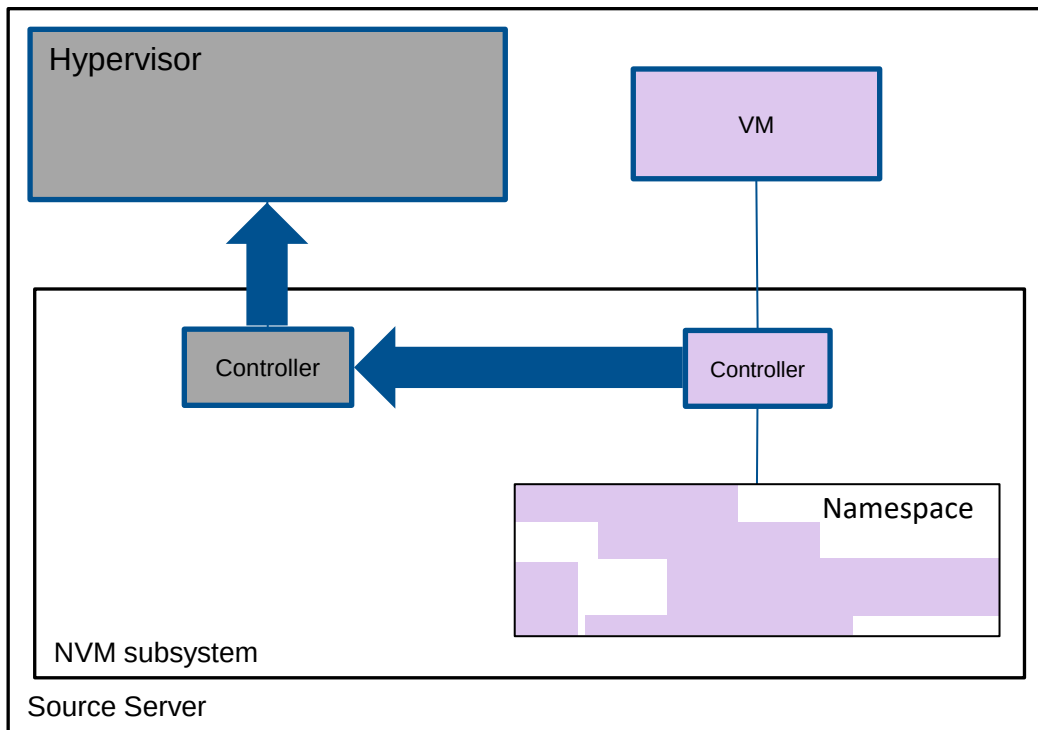
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Start logging and aggregating changes
 - Query aggregated changes

NVMe™ Host Managed Live Migration



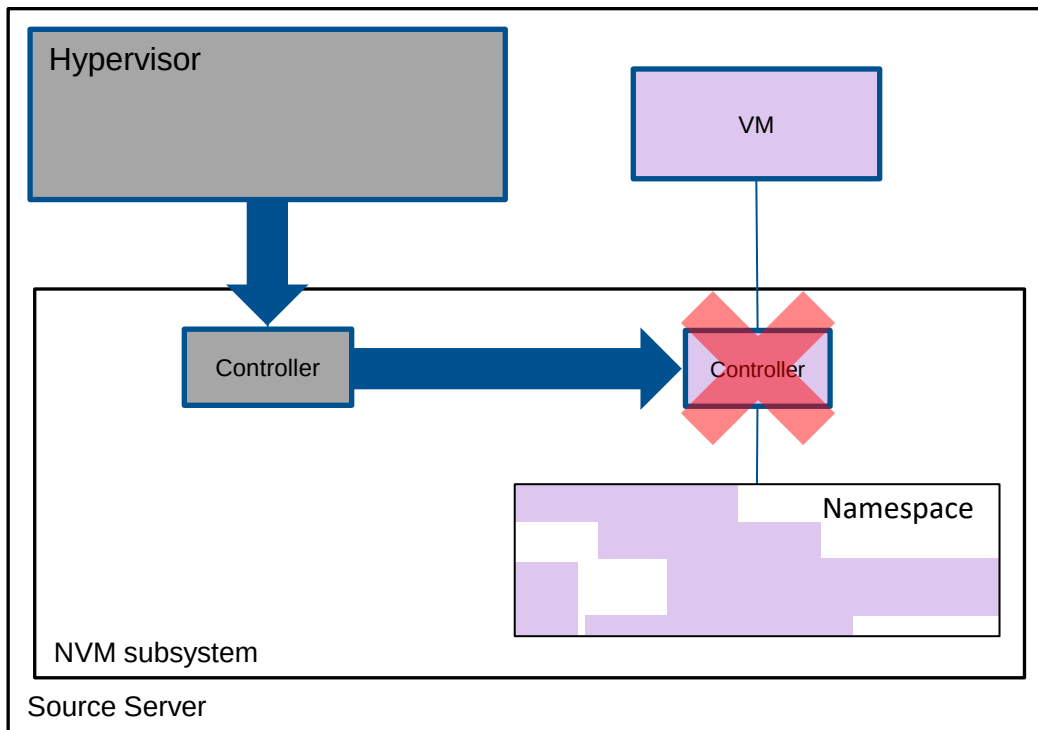
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Read controller state

NVMe™ Host Managed Live Migration



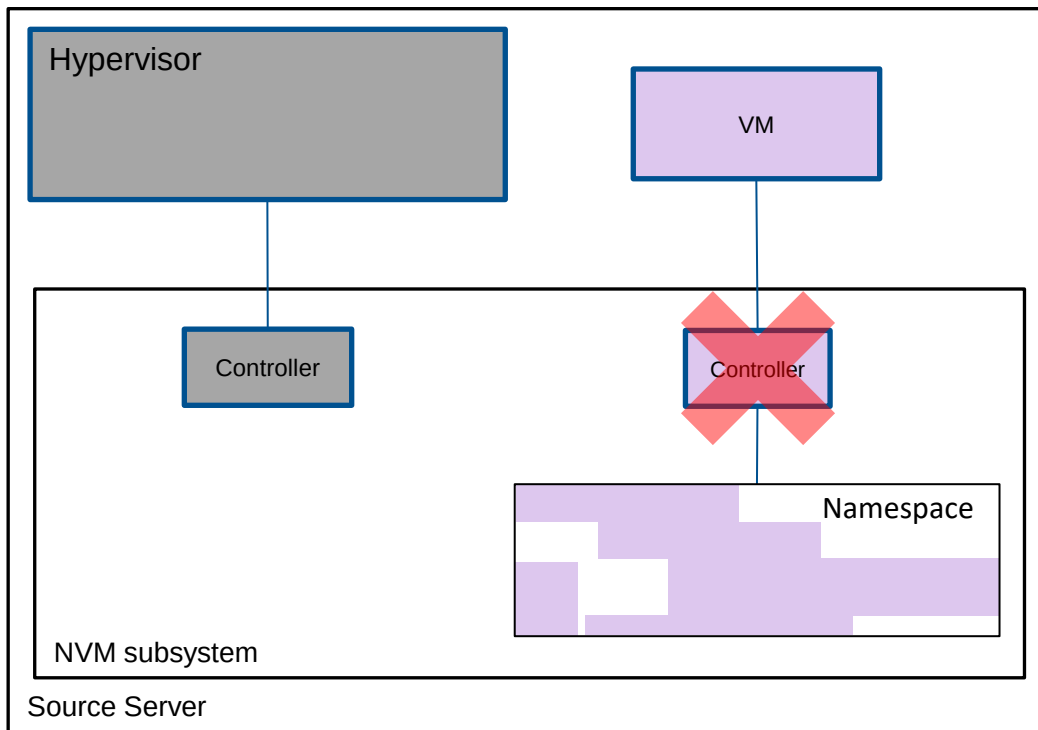
- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Read controller state
 - NVMe state
 - Vendor specific state

NVMe™ Host Managed Live Migration



- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Read controller state
 - Pause controller

NVMe™ Host Managed Live Migration

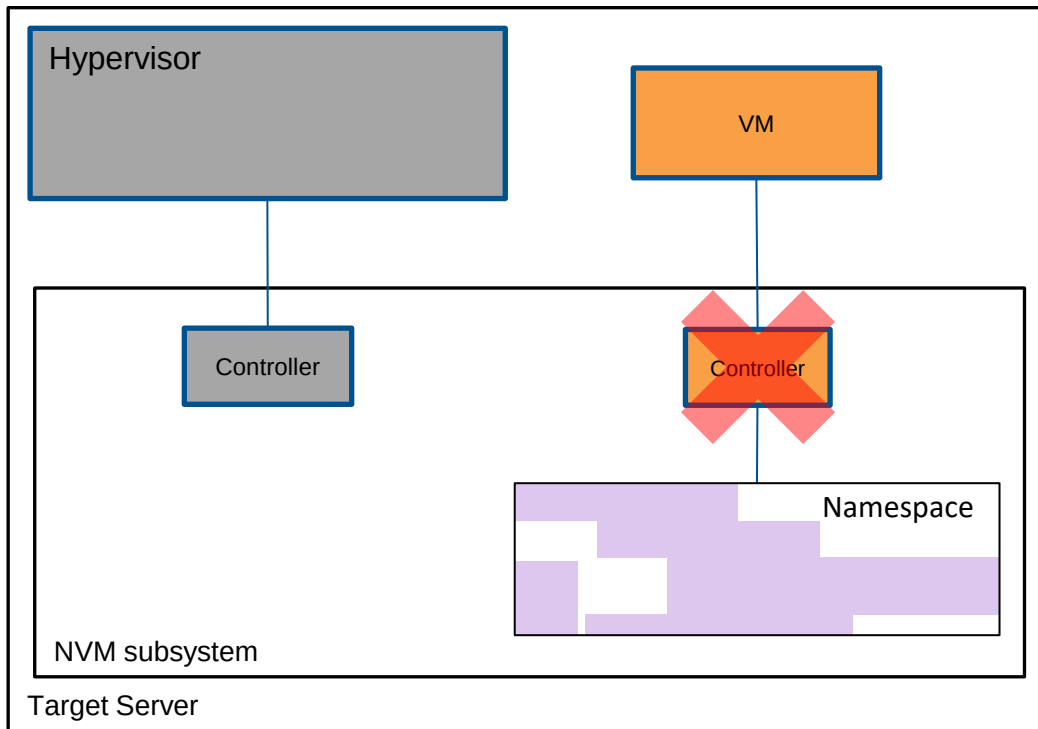


- Source NVM Subsystem
 - Read allocated logical blocks
 - Request logging changes
 - Logical Blocks
 - VM memory
 - Read controller state
 - Pause controller

NVMe™ Host Managed Live Migration

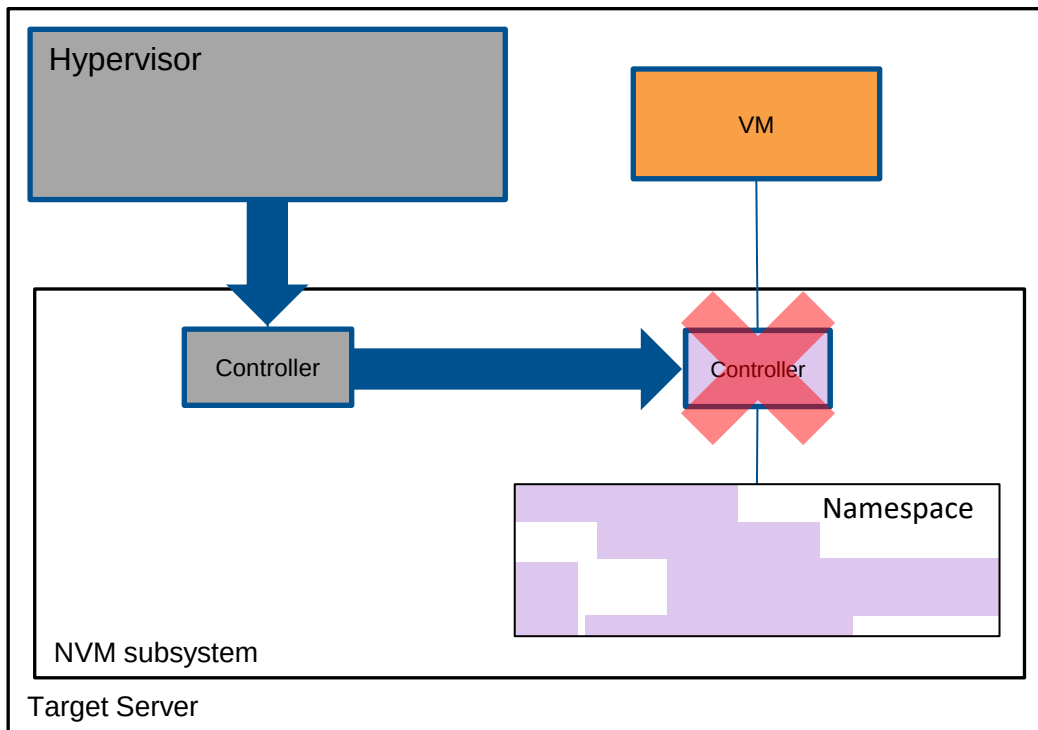
- NVMe Host Managed Live Migration provides services to a Host to:
 - Get the NVMe state from a source controller
 - Put that NVMe state into the target and resume operation
- Target NVM Subsystem
 - [Set the Controller State](#)

NVMe™ Host Managed Live Migration



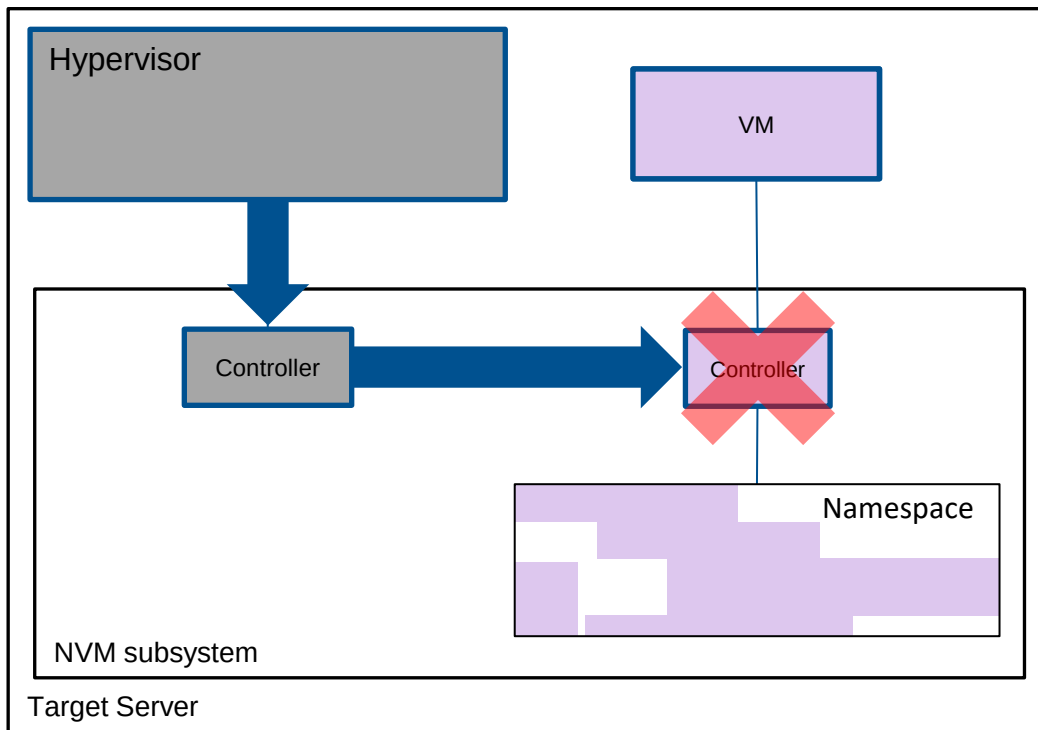
- Target NVM Subsystem
 - Set the Controller State

NVMe™ Host Managed Live Migration



- Target NVM Subsystem
 - Set the Controller State

NVMe™ Host Managed Live Migration



- Target NVM Subsystem
 - Set the Controller State
 - Resume operations on the migrated controller

NVMe™ Host Managed Live Migration

- The reading and writing of namespaces is using existing read and write I/O commands.
- The transfer of namespaces and controller state between servers is outside the scope of the NVMe Express® specifications.



OCP Global Summit 2024

- I am planning on submitting a presentation for OCP Global Summit 2024 to present the ratified version of the specification detailing the full NVMe™ protocol supporting Host Managed Live Migration.

Questions



OPEN
Compute
Project®



STORAGE

Flexible Data Placement Panel

Ross Stenfort, Keith Busch, Meta
Javier Gonzalez, Samsung
Luca Bert, Micron
David Wang, SiliconMotion
Jonmichael Hands, Fadu



OPEN
COMMUNITY®



STORAGE

Open-Source Update

Ross Stenfort, Meta

FDP: Open-Source Status

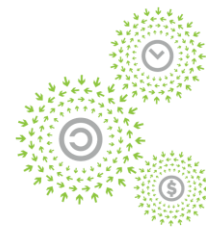
- **Linux Kernel:** Full support with I/O Passthru (**Upstream since 5.19**)
- **xNVMe[®]** : Full support (**Upstream since v0.7**)
- **QEMU:** FDP Emulation (**Upstream since v8.0**)
 - Validation of host stack.
- **Fio:** Supported (**Upstream**)
- **nvme-cli:** Support for FDP commands and log pages (**Upstream**)
- **Cachelib:** Reduce WAF through FDP (**Upstream**)



STORAGE

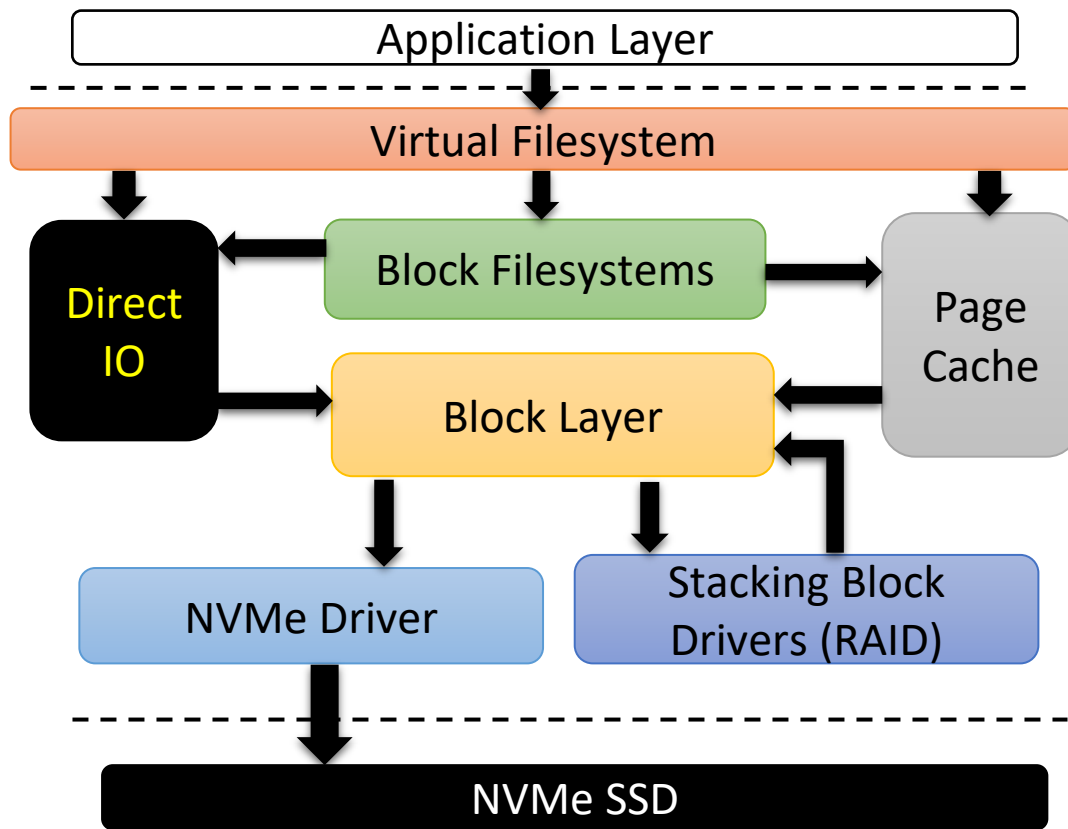
Linux support for FDP

Keith Busch, Software Engineer, Meta



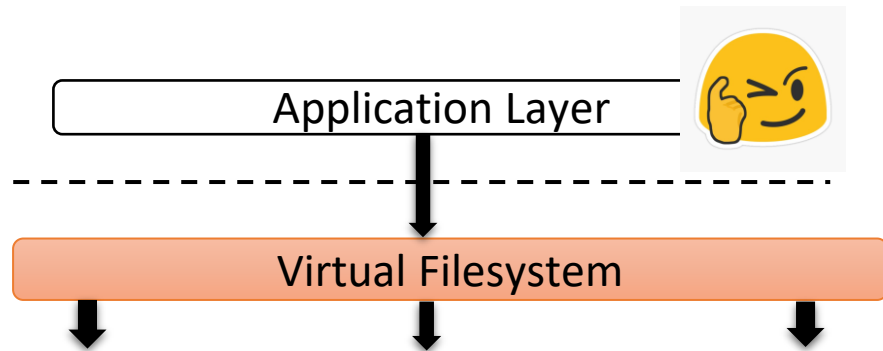
OPEN
PLATINUM™

The Linux Storage Stack



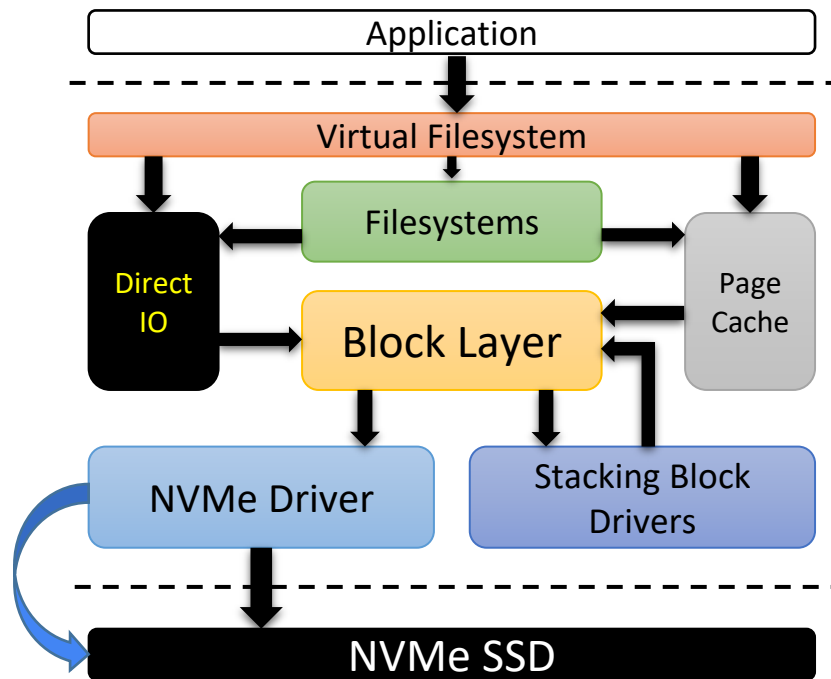
Application Write Hints in Linux

- Applications generically access their storage through the syscall interface provided by the Virtual Filesystem (ex: read(), write())
- FDP requires software provide intelligent write hints
- Applications account for the majority of written data, so it is in the best position to get the most impact
- Problem: there is no generic interface to provide such hints



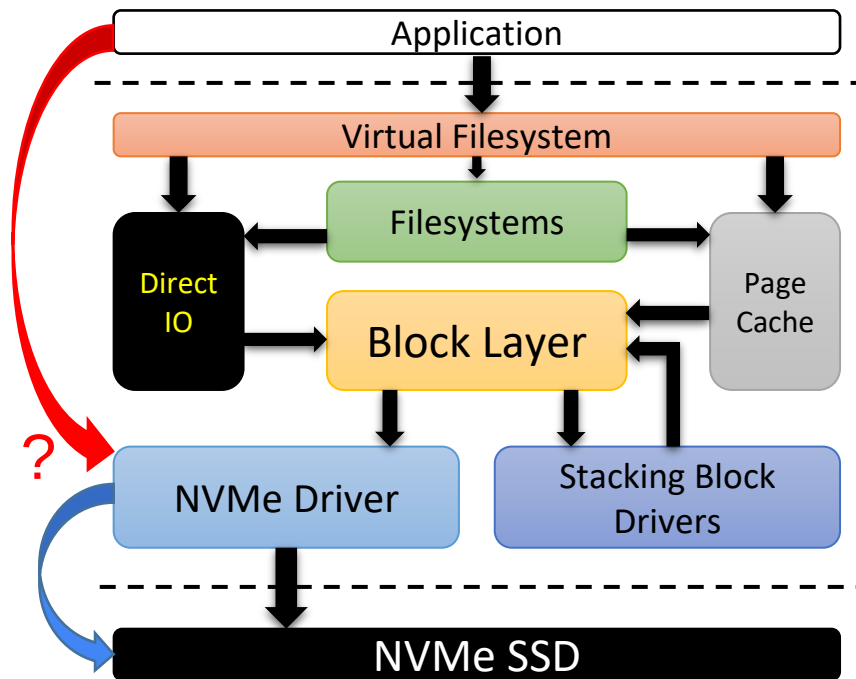
Application Write Hints in Linux

- It's easy enough for the driver to insert an FDP placement handle into the command ...



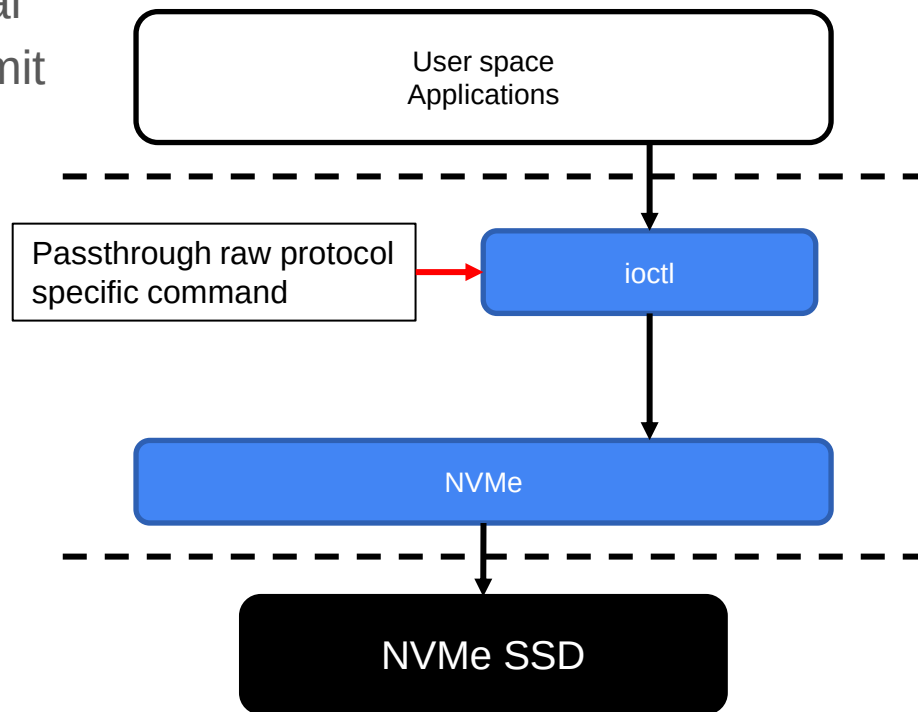
Application Write Hints in Linux

- It's easy enough for the driver to insert a placement handle into the command
- How does the application tell the driver which handle to use?



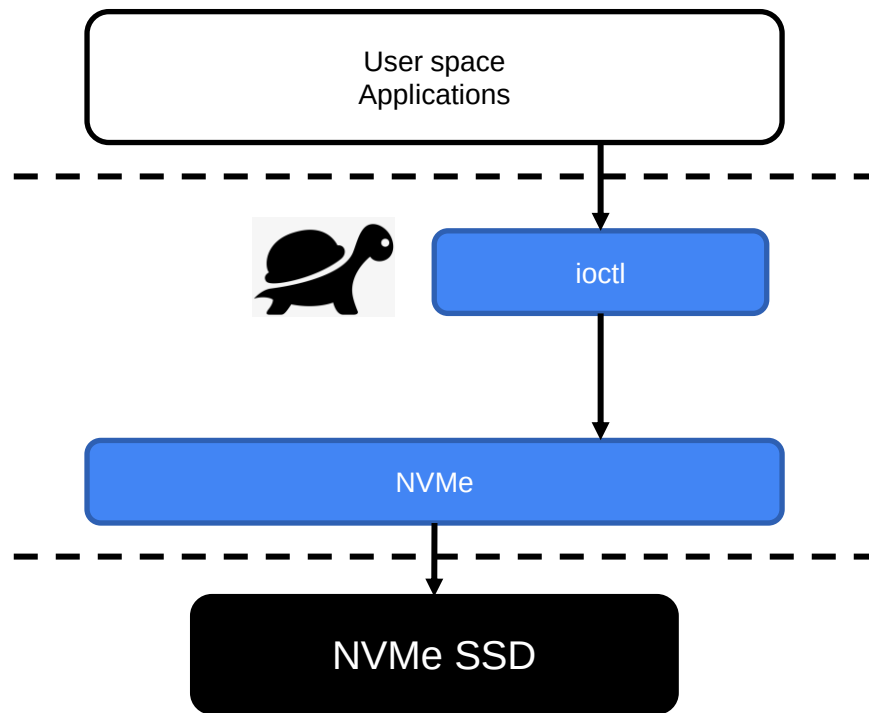
Using FDP on Linux

- The driver provides a special path for applications to submit user defined protocol commands



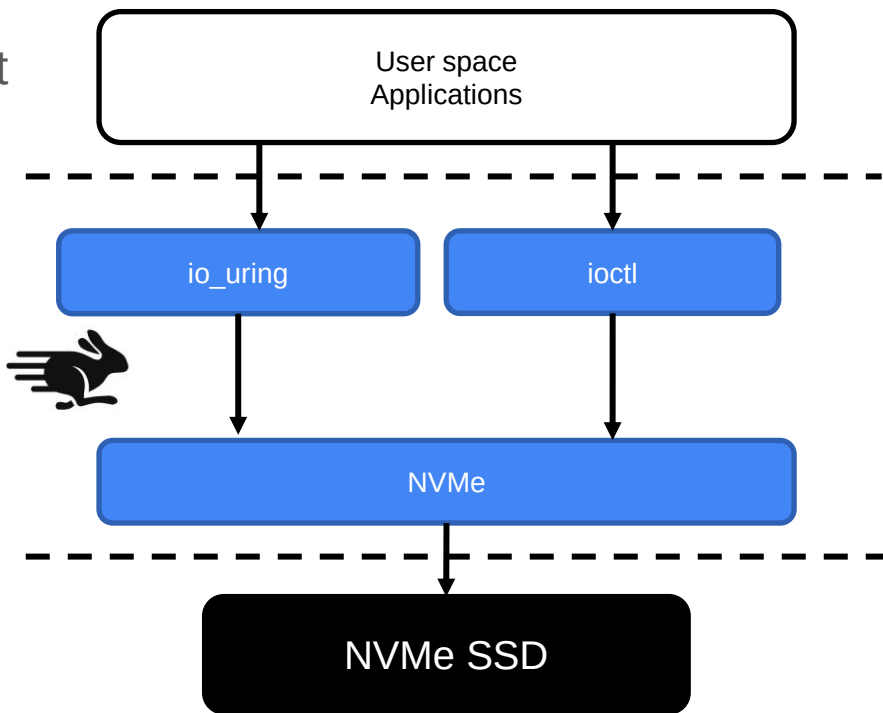
Using FDP on Linux

- The driver provides a special path for applications to submit user defined protocol commands
- But the “ioctl” is not considered a high performance interface



Using FDP on Linux

- The driver provides a special path for applications to submit user defined protocol commands
- But the “ioctl” is not considered a high performance interface
- io_uring to the rescue!



Current State and Potential Future

- FDP capabilities work well for applications that access their storage directly through `io_uring`
 - Reference implementation is provided in 'fio'
 - Does not work with filesystems or stacking block devices
- On going investigations to make filesystems friendly to write hints from both application and in-kernel use cases
 - Progress is moving cautiously as previous attempts for similar ideas created some maintenance difficulties in the past



STORAGE

FDP Update

Javier Gonzalez, Samsung



OPEN
COMMUNITY®

Introduction: Data Placement

- Data Placement is a prevalent problem across NAND consumers & Industries
 - Impacts: WAF, TCO, predictability (latencies), and overall performance
- Several approaches in the past few years account for innovation in this area
 - Well explored design space facilitates a good understanding of the trade-offs

Block Interface (CNS)

Description & Trade-offs

- Traditional block device
- Most innovation in-device
- Mature host software stack
- WAF ~1: No assurances

Status

- Commonplace for storage devices

Streams / Directives

Description & Trade-offs

- Extension to block device with backwards compatibility
- Use of write hints
- DSM deallocation mechanism
- Minor changes to Host SW
- WAF ~1: Initialize & trust

Status

- Little industry traction

Flexible Data Plac. (FDP)

Description & Trade-offs

- Extension to block device with backwards compatibility
- Capacity-based placement without seq. write requirement
- Use of write tags
- DSM deallocation mechanism
- Minor changes to Host SW
- WAF ~1: Iterative query/check

Status

- Customer-driver technology
- Blooming adoption & Eco

Zoned Namespaces (ZNS)

Description & Trade-offs

- Departs from block device
- No backwards compatibility
- LBA-based placement with strict seq. write requirement
- Explicit host deallocation & state machine management
- Major changes to Host SW
- WAF ~1: assured

Status

- Vendor-driven technology
- Fragmentation across vendors
- Tech validation for +3 years

Open-Channel SSDs

Description & Trade-offs

- Full host-based FTL
- Most innovation in host
- Drastic changes to Host SW
- WAF ~1: Assured

Status

- Not standardized
- Dropped by the industry



Introduction: Data Placement Model

- The goal is to compare Data Placement (DP) technologies across these 4 metrics
 - WAF (+Space Amp), Performance (BW & latency), Device Utilization, and Engineering Effort
 - TCO is a function of all these metrics

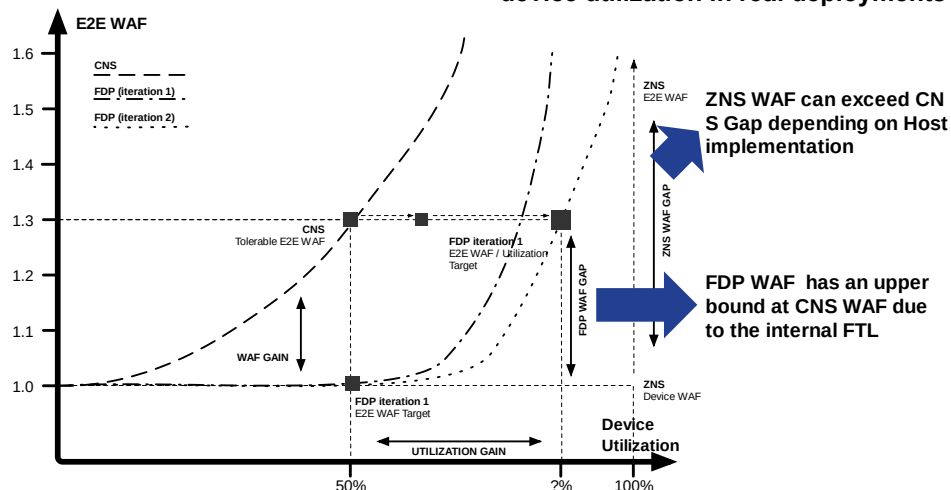
Technology Viewpoint

- How is each customer dealing with the trade-off across these metrics?
- How much can we push device utilization at acceptable metrics?
- Is 80% of the benefit enough? Other %?

Business Viewpoint

- Which DP tech. allows customers to meet their requirements at the *lowest* TCO?
- What is the trade-off between the initial investment and the expected benefits on a full deployment?

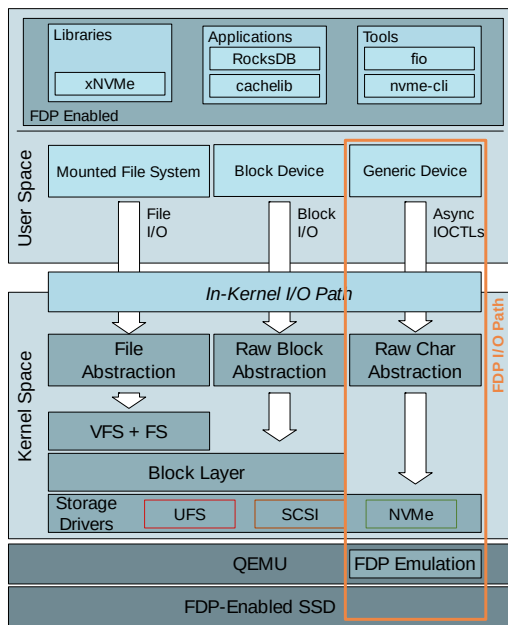
Experimental Space



Ecosystem: Linux Software Stack

- FDP enablement easy to integrate in existing storage stacks
 - Hint-based, Implicit (no write errors), Backwards compatibility,

Upstream Linux I/O Paths



Ecosystem Ready

Support in Linux Kernel through I/O Passthru

- Use in-kernel NVMe device directly from user-space
- Similar to SPDK, without the need to detach the PCIe device from the kernel
- Upstream since 5.13
- Increasing number of adopters in hyperscalers and enterprise
- Scalability through io_uring and higher performance than block I/O

Application enablement through xNVMe

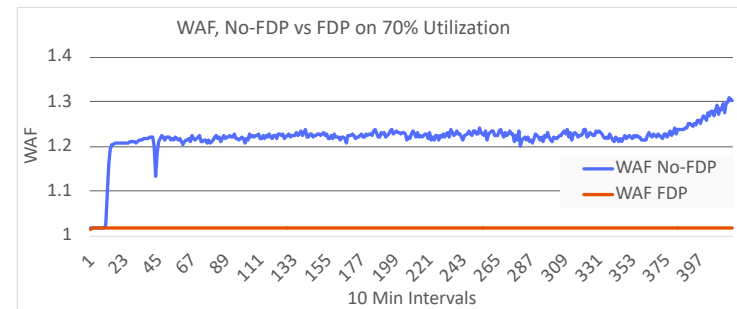
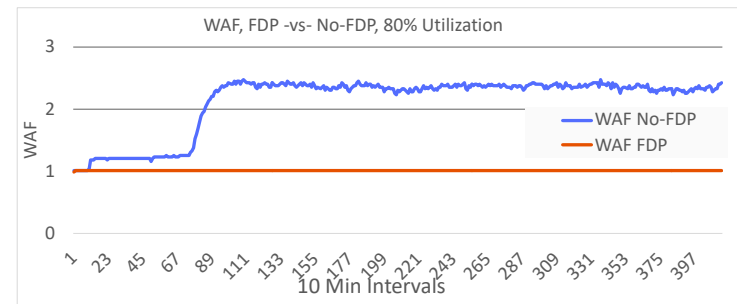
- Easy application integration through different I/O Paths (block, I/O Passthru, SPDK)
- Upstream support in SPDK (native & xNVMe) to enable use of SPDK bdev
- Samsung backed, community driven
- Cachelib: Development upstream with community
- RocksDB: POC Implementation

Tools

- Support in nvme-cli and fio
- Emulation in QEMU since v8.0

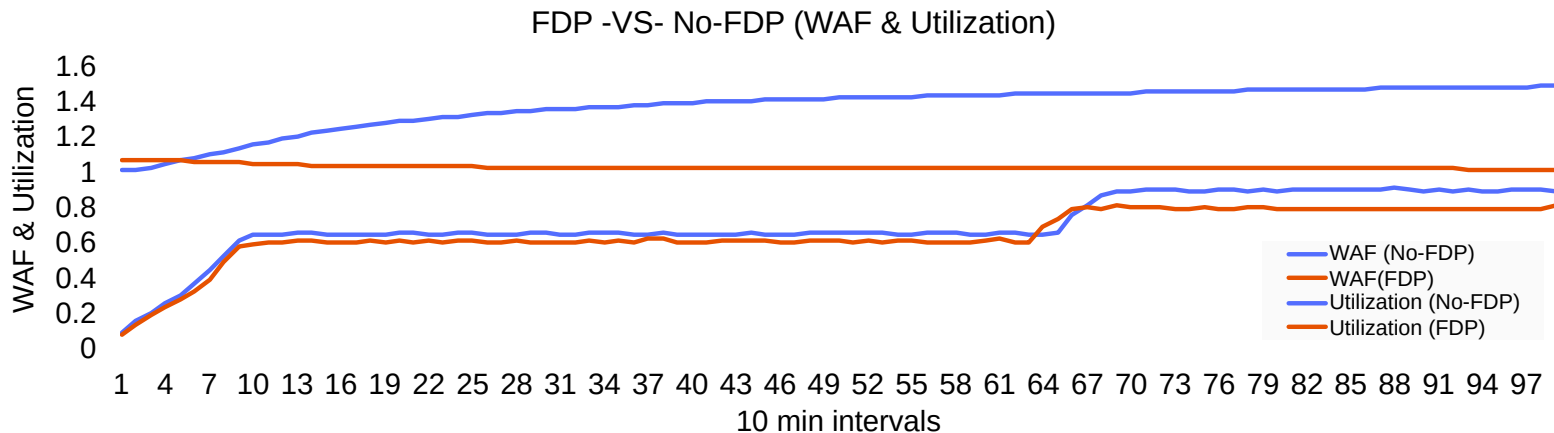
Ecosystem: Cachelib Case Study

- Significant improvement to Cachelib WAF without major changes
 - No impact on other metrics
 - Throughput, latency, cache hit-rate
 - Cachelib code in process to merge in under 5 months
 - PR: <https://github.com/facebook/CacheLib/pull/247>
 - Merged
 - WAF for 80% and 70% utilization
 - WAF ≈ 1 when FDP enabled
 - Even for high utilization (80%)
 - Throughput maintained
 - Future
 - Further segregation of Large Items
 - Generalize FDP library?



Ecosystem: RocksDB Case Study

- WAF = ~1 when FDP enabled
 - Even for high utilization (~80%)
- Experimental
- Testbed for FDP and other Data Placement approaches



Larger Indirection Units (IU) Ecosystem and Relationship with FDP for Storage



STORAGE

Luca Bert, Distinguished Member SSD Architecture,
Micron technology



OPEN
COMMUNITY®

Problem statement

- IU = General term to define a FTL mapping unit larger than LBA size
- Large IU is necessary to support large capacity SSD
 - DRAM size to keep 4K IU maps becoming prohibitive
 - 16KB is used as reference size for large SSD but others can be considered
- Main concern is around induced Write Amplification (WAF) due to unsized/ unaligned Writes
 - $WAF_{Total} = WAF_{App} * WAF_{SSD} * WAF_{IU}$
 - WAF_{IU} Is the multiplicative factor induced by Large IU
 - $1 \leq WAF_{IU} \leq 4$ for 16KB IU
 - Perception is that 16K IU will result in $WAF_{IU} = 4$ and thus 4x worse endurance

- We need real life data to support/ challenge above statement
- We need to validate how, in Real Life Apps, FDP may help reducing it further

Real life data of WAF_{IU} from sample Benchmarks

$$WAF_{Total} = WAF_{App} * WAF_{SSD} * WAF_{IU}$$

$1 \leq WAF_{IU} \leq 4$ for 16KB IU – The lower the better

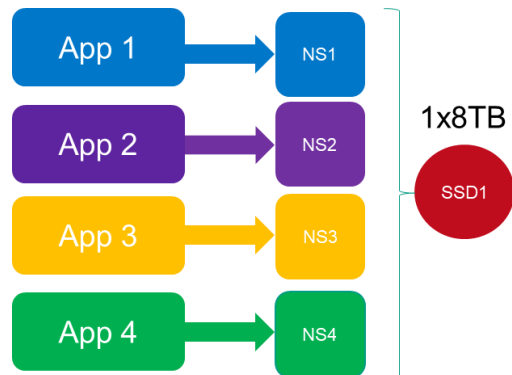
Application	WR count%	Bucketized Writes (by IO count)										Avg Size Wr (KB)	Measured 16K WAF _{IU}
		0-4KB+	4096	8192	16384	32768	65536	131072	262144	524288	1048576		
Expected based on 4KB RW		0	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4	4.00
TPC-H/ 512GB/ TempDB files=8/ XFS - TempDB nvme2n1	59.7%	0.0%	0.0%	0.1%	0.0%	0.0%	99.9%	0.0%	0.0%	0.0%	0.0%	64	1.22
TPC-H/ 2.3TB/ TempDB files=8/XFS- TempDB nvme2n1	54.5%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.00
TPC-H/ 512GB/ TempDB files=384/XFS- TempDB nvme2n1	62.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.02
TPC-H/ 2.3TB/ TempDB files=384/XFS- TempDB nvme2n1	69.6%	0.0%	0.1%	0.0%	0.0%	0.0%	99.9%	0.0%	0.0%	0.0%	0.0%	64	1.00
1363-A: YCSB on RocksDB - Workload A	4.8%	0.0%	2.1%	2.1%	2.1%	2.1%	2.1%	2.1%	25.0%	54.2%	8.3%	432	1.01
1363-B: YCSB on RocksDB - Workload B	1.1%	0.0%	36.8%	19.7%	10.4%	5.4%	3.2%	1.7%	14.8%	8.0%	0.0%	89	1.03
1363-F: YCSB on RocksDB - Workload F	4.5%	0.0%	2.2%	2.2%	2.2%	2.2%	0.0%	2.2%	24.4%	55.6%	8.9%	442	1.01
1413-00: Cassandra/XFS YCSB 512GB Load	100.0%	0.0%	0.2%	32.9%	0.1%	0.0%	0.0%	0.0%	64.7%	0.6%	1.1%	183	1.03
1413-01: Cassandra/XFS YCSB 512GB Workload A	4.2%	0.0%	1.2%	27.6%	0.2%	0.2%	0.2%	0.5%	49.4%	8.6%	8.2%	257	1.03
1413-02: Cassandra/XFS YCSB 512GB Workload B	3.2%	0.0%	4.3%	26.9%	1.2%	0.6%	0.6%	0.9%	48.5%	7.7%	4.6%	215	1.03
1413-04: Cassandra/XFS YCSB 512 GB Workload F	3.9%	0.0%	19.8%	23.2%	0.8%	0.5%	0.5%	0.8%	41.0%	6.6%	3.8%	182	1.03
1413-EXT4-01: Cassandra YCSB/ EXT4 128 GB Workload A	3.3%	0.0%	1.3%	17.4%	6.3%	1.2%	1.0%	1.6%	25.9%	34.2%	11.2%	361	1.02
1413-EXT4-02: Cassandra YCSB/ EXT4 128 GB Workload B	2.5%	0.0%	8.9%	21.8%	4.8%	0.4%	0.4%	0.8%	30.2%	6.0%	17.3%	290	1.03
1413-EXT4-04: Cassandra YCSB/ EXT4 128 GB Workload F	0.7%	0.0%	3.2%	23.3%	4.3%	0.7%	0.4%	0.8%	44.4%	15.3%	6.5%	263	1.02
1453-02dc Ceph RadosBench- Both data and metadata	68.8%	0.0%	37.5%	13.1%	1.8%	0.2%	46.9%	0.0%	0.2%	0.0%	0.0%	33	1.18
1453-b7ca Ceph RadosBench- Data nvme0n1	53.8%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.12
1453-b7ca Ceph RadosBench- Metadata nvme7n1	99.6%	0.0%	69.9%	25.5%	3.7%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	6	2.53

Green = low %; Red = high %; Rest is gradient colors

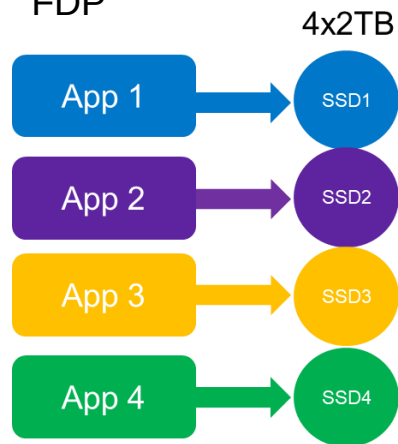
FDP Simple simulation case

- Use reference benchmarks to test Write Amplification improvements in Conventional and FDP configurations
 - Focus on Persistently Isolated FDP
- Run with FIO:
 - App 1: SeqWr 4KB 64QD
 - App 2: SeqWr 16KB 32QD
 - App 3: SeqWr 64KB 16QD
 - App 4: SeqWr 256KB 8QD
- Run with JEDEC JESD219A.01
- Run with Aerospike
 - 4DB 2TB each
 - 32K, 64K, 128K and 256K write sizes
 - Benchmark: YCSB WL A (50% R, 50% W, uniform Distribution)

Conventional NS



FDP



Results overview

- All benchmarks benefit of FDP-aware configurations
- Real Life benchmarks are 37%-41% but given WAF baseline is 1, improvement is much higher
- Real Life benchmarks (Aerospoke and, in lesser degree, FIO) benefit much more than purely synthetic (JESD)
 - We cannot rely only on benchmarks who do not follow realistic behaviors (i.e. JESD219A)
 - There is an opportunity to improve existing WAF characterization methods

	Write Amplification		
	Conv SSD	FDP SSD	%
FIO	1.63	1.02	-37%
JESD219A.01	3.95	3.8	-4%
Aerospoke	1.84	1.09	-41%

Summary

- Very large SSD will not have any major impact in WAF due to:
 - Nature of benchmarks that has an inherent larger granularity and alignment
 - When FDP is supported, this will enable data segregation keeping data of different nature in dedicated Regions further reducing WAF impact
- Large IU and FDP are complimentary technologies in enabling very large SSD used in Storage Systems
- New test tools and metrics may become necessary to validate these new data layout

Overcoming Challenges in Enabling FDP in High Capacity QLC SSDs



STORAGE

David Wang, SSD FW Architect,
Silicon Motion



OPEN
COMMUNITY®

The content of this document including, but not limited to, concepts, ideas, figures and architectures is furnished for informational use only, is subject to change without notice, and should not be construed as a commitment by Silicon Motion Inc. and its affiliates. Silicon Motion Inc. assumes no responsibility or liability for any errors or inaccuracies that may appear in the informational content contained in this document.

Nothing in these materials is an offer to sell any of the components or devices referenced herein. Silicon Motion Inc. may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Silicon Motion, Inc., the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property. © 2023 Silicon Motion Inc. or its affiliates. All Rights Reserved. Silicon Motion, the Silicon Motion logo, MonTitan, the MonTitan logo are trademarks or registered trademarks of Silicon Motion Inc.

Challenges in enabling FDP in QLC

Flexible Data Placement (FDP) is ideal for accelerating QLC adoption in Data Center/Primary Storage Applications enabling \$/GB SSDs

- Dramatic reduction in WAF, improving DWPD
- Maximizes performance, while minimizing negative impact of low-PE QLC media

WAF	IOPS (4K RW on 4K IU)	DWPD (based on QLC PE 1.5K)
WAF=5.5* (RndWr)	120K IOPS	0.18
WAF=3 *	230K IOPS	0.32
WAF=1.5 ** (FDP)	450K IOPS	0.64
WAF = 1* (SeqWr)	725K IOPS	0.96
* Measured data based on 16TB PCIe Gen5 QLC SSD; NAND specification PE=1500		
** Estimated data based on ~1.5 WAF enabled with FDP		

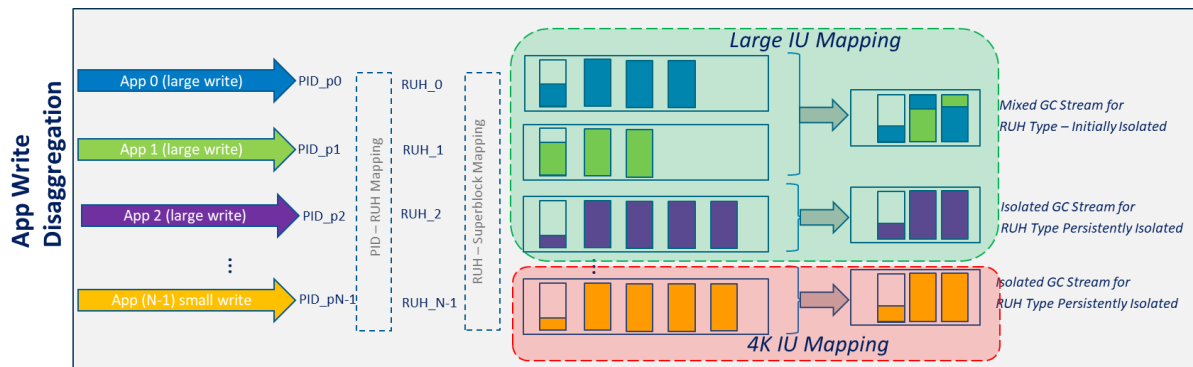
We see a wide range of design challenges in high capacity (32TB-128TB) QLC-based SSDs:

- Problem1: How to reduce FTL DRAM footprint to save BOM cost?
- Problem2: How to manage the limited controller write resource for up to 32 FDP RUHs?

Design for reducing DRAM footprint

- Design 1: use configurable Indirection Unit (IU) in QLC SSD and 16K IU for large size drive
- Design 2: Controller uses HW assisted bit-pack L2P entry for further saving L2P Table size. E.g., using 33bit entry (instead of 40bit) to address 8G IUs can further reduce L2P size up to 17.5%
- Design 3: to reduce WAF caused by small write on large IU, support both RUHs with 4K IU and RUHs with 16K IU. App can place the small writes to the RUH with 4K IU.

	4K IU		8K IU		16K IU	
SSD Size TB	PPA bits	L2P Size (GiB)	PPA bits	L2P Size (GiB)	PPA bits	L2P Size (GiB)
16	33	15.01	32	7.28	32	3.64
32	34	30.92	33	15.01	32	7.28
64	35	63.66	34	30.92	33	15.01
128	36	130.97	35	63.66	34	30.92



Design for multiple RU Handles

Due to QLC's coarse-fine program scheme with N WL prog interval, each RUH (for full size RU of one SSD) will take a few hundred MBs of write buffer. In multiple RUH config, the required write buffer could easily surpass what PLP circuit can protect during surprise power loss.

- Design1: Use the write flow with SLC copy-back for multiple RUHs QLC.
 - Support 8/16/32 RUHs with RAID die-failure protection
 - SLC write buffer size is manageable and can be protected by Supercap
- Design2: Use PerformaShape™ engine to regulate per RUH Write throughput and appropriate write buffer allocation

Diagram A - Host Write Flow

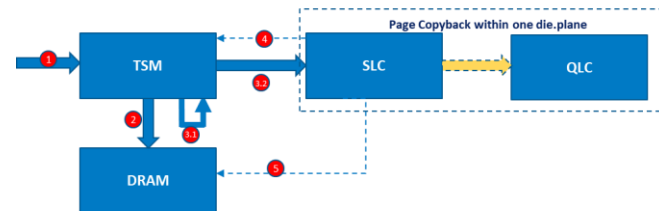
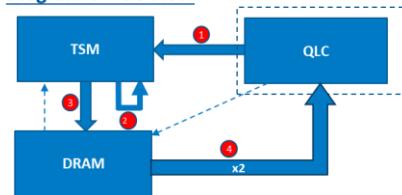


Diagram B - GC Flow



Summary

- FDP brings down WAF and makes high density SSD with low-PE QLC NAND viable. 16TB QLC SSD can achieve 1800-2900MB/s Write and DWPD 0.6+ on FDP-aware applications.
- High capacity QLC SSD requires more design optimizations on DRAM reduction and write resource management for up to 32 RUHs. We share design methods to tackle these challenges:
 - Reduce DRAM with large IU and compact L2P entry,
 - Support co-exist RUHs with 4K IU and RUHs with 16K IU,
 - QLC Write flow with SLC copyback, and performance and QoS control among multiple RUHs.



Maximizing Storage Efficiency and TCO with Flexible Data Placement Technology



STORAGE

Jonmichael Hands, Sr. Director Product Planning, Fadu

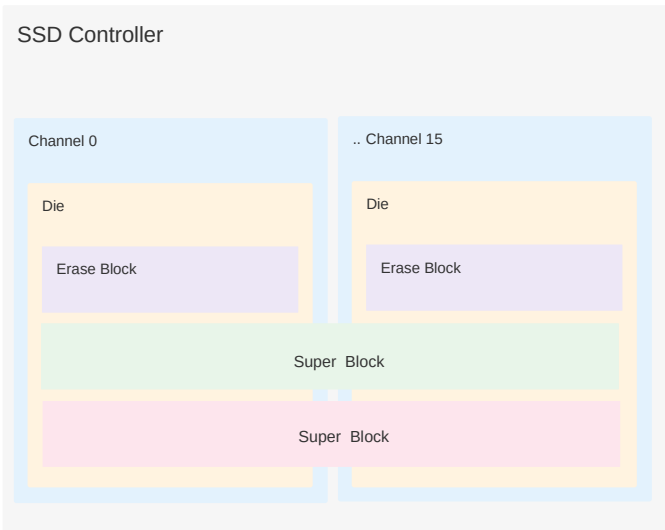


OPEN
COMMUNITY®

What is Flexible Data Placement?

- Improved application performance and extended SSD endurance - reduce SSD write amplification (WAF)
- Software intelligently places data on SSD that is friendly with how flash works (blocks, channels, etc.)
- Improved isolation leads to better quality of service
- Fully backwards compatible, namespace support
- Scale SSD capacity with performance for compute workloads
- Lower overprovisioning → higher device utilization → lower TCO
- Lower WAF → higher endurance → lower TCO

VMs, containers, apps



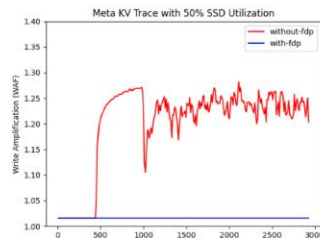
FDP Customer Use case

- Developed by Google, Meta, SSD vendors
- Multi-tenant, multiple namespaces, multiple applications, VMs, containers, microservices, log structured DB, caching
- Presentation at [OCP](#) detailing customer use cases

Cachelib: Case Study

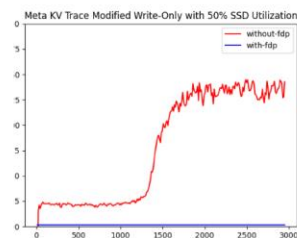
❖ Significant improvement to Cachelib WAF without major changes

- No impact on other metrics: Throughput, latency, cache hit-rate
- Cachelib code in process to merge in *under 5 months*

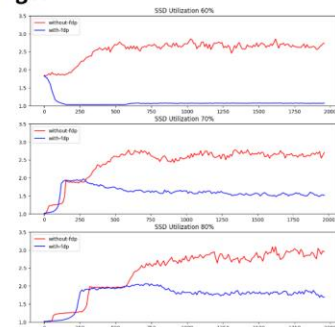


1. Raw Meta read-heavy workload

https://cachelib.org/docs/cachelib-1.0.0-rc1-features.html#fb_hww_eval#running-ca

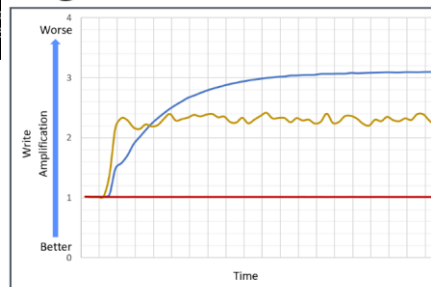


2. Altered write-heavy workload



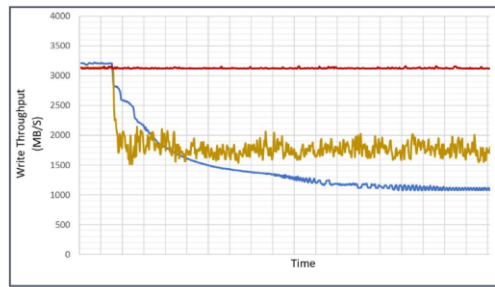
3. Increased utilization on Meta write-heavy workload

Log Structured Workload Pattern



Standard SSD Workload Summary:

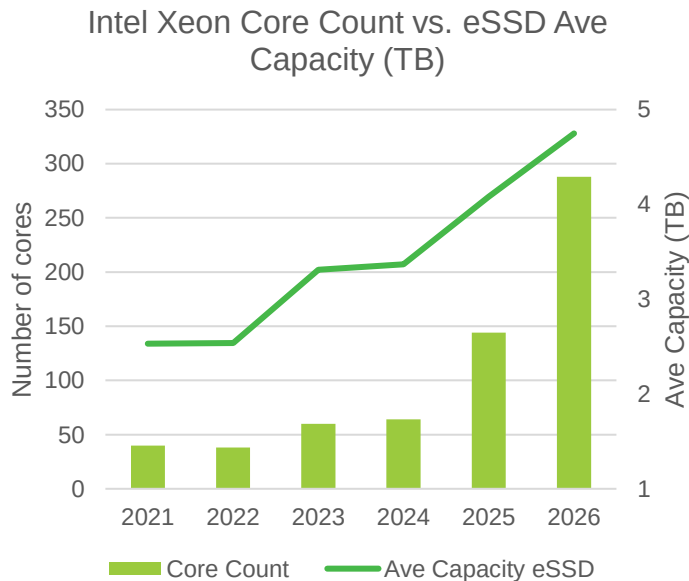
- Blue: 64KB Random Write
- Yellow: Log Structured 8 Writers 64KB
- Red: Log Structured 8 Writers 64KB with FDP



Example Workload Summary:

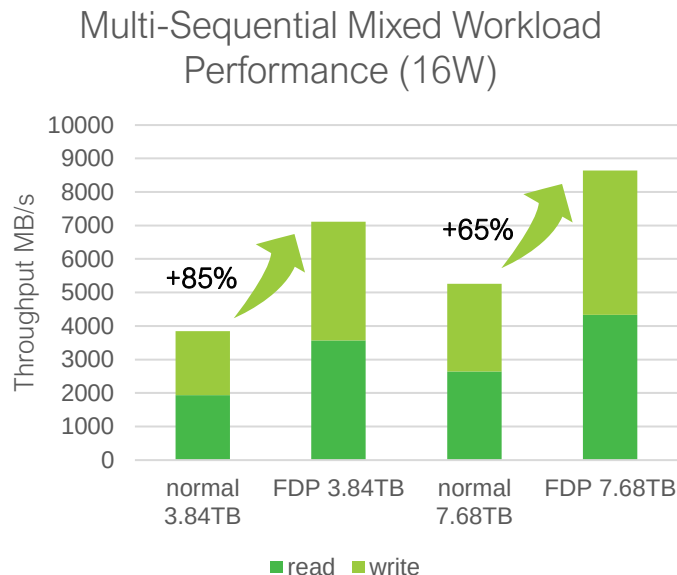
- FDP Reduces Wear: ~2-3X
 - SSD lasts 2-3X longer
- FDP Improved Write Throughput: ~2-3X
 - Improved Performance and Quality Of Service

FDP Improves Mixed Workload Performance



CPU core count driving need for higher capacity SSD

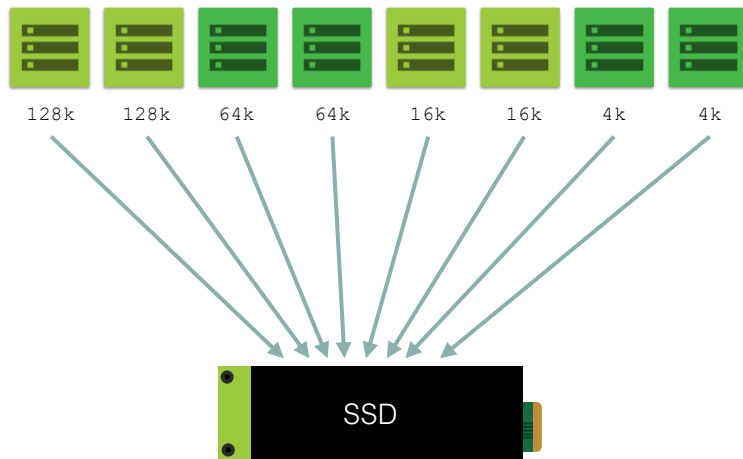
Source: ark.intel.com, Forward Insights Q'24



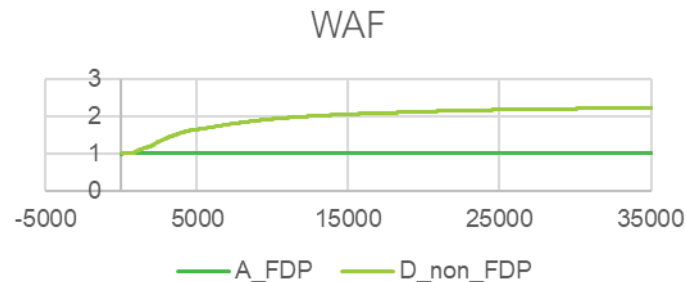
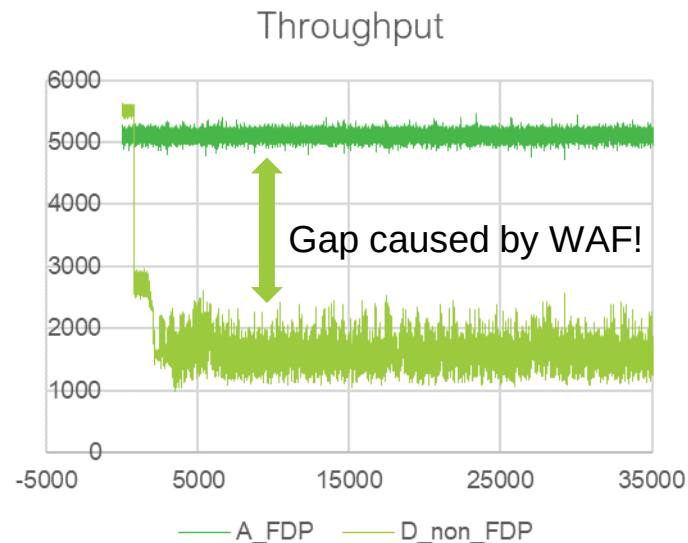
Drives designed for multi-tenant and cloud-native applications

Source: Measured by Fadu. 13th Gen Intel(R) Core(TM) i5-13600K, ASRock Z790, Ubuntu 22.04.2 LTS, kernel 6.2.10, fio-3.35

FDP Improves Write Bandwidth



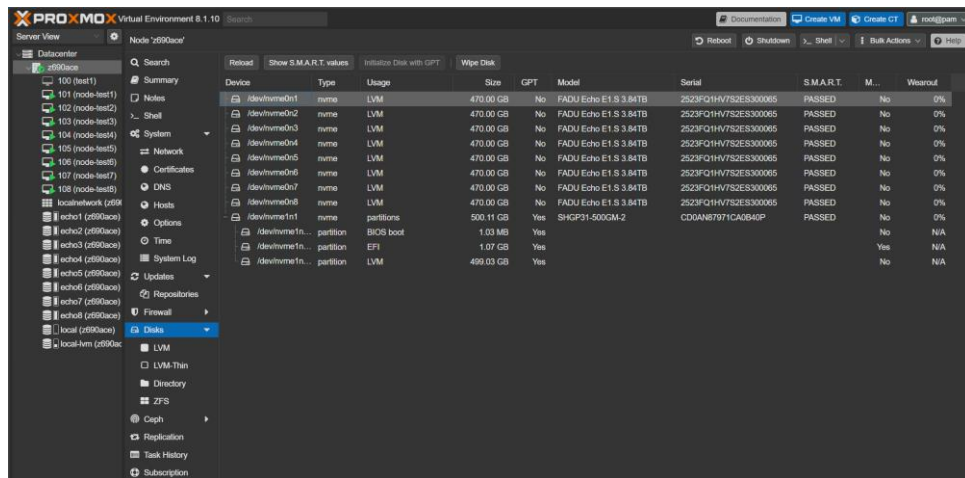
8 different write threads of different block sizes
Achieve WAF=1, full NAND bandwidth with FDP



Multi-Namespaces FDP (no SRIOV)

- With FDP enabled, each namespace gets its own RUH (up to 8 on this drive)
- No software development changes required
- Create an LVM per namespace and attach to VM

Mixed	Normal	FDP
Bandwidth (MB/s)	3080	4116
Consistency	62%	96%
WAF	2.5	2



```
for i in {1..8}; do nvme delete-ns -n $i /dev/nvme0; done
for i in {1..8}; do nvme create-ns -S 470G -C 470G -f 0 /dev/nvme0; done
for i in {1..8}; do nvme attach-ns /dev/nvme0 -n $i -c 1; done
```

Questions



OPEN
Compute
Project®

OCP Foundation Updates



OPEN
Compute Project®

Rob Coyle - CTPM

rob@opencompute.org

Michael Schill - Director of Community

michael@opencompute.org

Steve Helvie - VP Emerging Markets

steve@opencompute.org



OPEN
Compute
Project®

Community-driven hyperscale innovation for all.

Contribution Process Updates

- Updates to Contribution area of opencompute.org
- Review of Process
- Errata Process Review

Contribution Process Page Update



[About](#) [Marketplace](#) [Solution Providers](#) [Contributions](#) [Projects](#) [Events](#) [Membership](#) [Blog](#)

Welcome

[Home](#) > [Contributions](#) > Welcome

Welcome to the Open Compute Project contribution process!

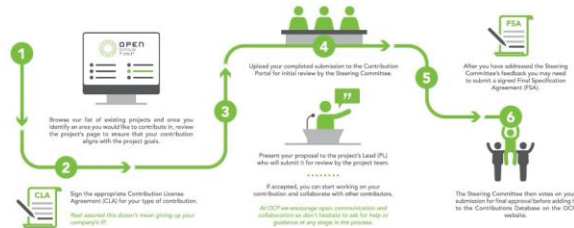
We are excited to have you join our [Community of contributors](#) who are passionate about driving innovation and collaboration in the data center industry. We aim to design and share efficient, scalable, sustainable hardware and software solutions that empower businesses and organizations worldwide. Working together can create more innovative, efficient, and sustainable data centers that benefit everyone. So whether you are an engineer, developer, or IT professional, we invite you to join us in developing cutting-edge solutions that advance the world of data center technology.

[Contribution Database](#) [How To Contribute](#) [Contribution Agreements](#) [Contributor Portal](#)

There are several ways you can contribute:



Want to become an OCP Contributor?



BUILD UP YOUR EXPERTISE, SHARE YOUR EXPERIENCE, GIVE BACK TO THE COMMUNITY

<https://www.opencompute.org/contributions/welcome>



Contribution Process Review

Want to become an OCP Contributor?



There are several ways you can contribute:



Specs



Design Files



Product Recognition



Facility Recognition



Reference Architecture



Software Contributions



Embedded Software



Case Studies



White Papers

BUILD UP YOUR EXPERTISE, SHARE YOUR EXPERIENCE, GIVE BACK TO THE COMMUNITY

<https://www.opencompute.org/contributions/how-to-contribute>

Contribution Agreements

 Rob Coyle [Contact](#)

 [About](#) [Marketplace](#) [Solution Providers](#) [Contributions](#) [Projects](#) [Events](#) [Membership](#) [Blog](#)

Contribution Agreements

- [FAQs about Licenses and OCP's Legal Process](#)
- [How to Contribute](#)
- Please contact contributors@opencompute.org if you would like to initiate any joint development activity

Templates		Licenses
Hardware Specifications 	• OCP Specification All-in-One (v1.0.0) [click to make a copy] • 2023 OCP Base Specification [click to make a copy] • 2023 OCP Design Specification [click to make a copy] • 2023 OCP Product Specification [click to make a copy]	• Modified OWF-CLA-1.0 (As of August 16, 2021) [Click here to sign the agreement] • Modified OWFa1.0 Final Specification Agreement (FSA) (As of August 16, 2021) [Click here to sign the agreement]
Specification Errata Template 	• Specification Errata Template (v.1.0.0)	• <i>(Follows / Same as Original Contribution)</i>
Hardware Design Packages	• Design Package Guidelines	• Modified OWF-CLA-1.0 (As of August 16, 2021) [Click here to sign the agreement]

<https://www.opencompute.org/contributions/agreements>

Errata



An errata document provides corrections or clarifications to errors and ambiguities found in a published contribution.

The scope of an errata document typically includes:

- Listing the specific contribution name, number, and date of the document being corrected
- Identifying the errors, omissions, and ambiguities in the standard requiring clarification
- Providing the correct information or additional details needed to resolve the issues

The goal of a standards errata document is to efficiently communicate needed technical corrections without reissuing the entire standard. Well-defined errata help ensure consistent understanding and implementation of the standard.

OCP Community Updates

- Governance Update - Now [OCP Operating Structure and Policies Document](#)
- Planned changes to OCP CLA/FSA process
- Addition of Feedback Agreement for Contributors
- Regional Summit Recap
- 2024 OCP Global Summit Update



OCP Operating Structure and Policies Document

Changes:

- Updates made to volunteer leadership requirements, criteria & expectations
- OCP Contribution Process - streamlined
- New programs (OCP S.A.F.E., Experience Center) introduced and defined.

Planned changes to OCP CLA/FSA process

Changes:

- Reverting to 0.9 version of OWF CLA/FSA in order to allow for easier contribution, less restrictions for contributors.
- Once approved by OCP BoD, 30 notification period will occur for OCP community to review.

Addition of Feedback Agreement for Contributors

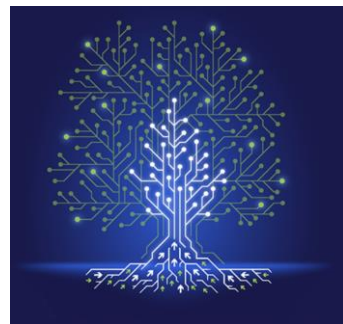
Problem Statement:

- There exists a need for a process by which non-contributors may provide technical feedback to contributors while not committing to join the contribution/sign a full CLA.
- The OCP BoD legal committee, along with the help of Samsung and others is close to ratifying an agreement which will address this problem.

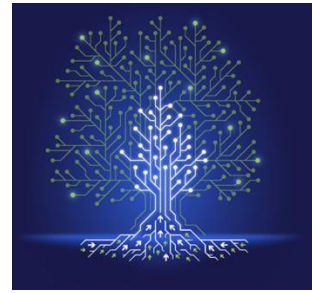


Regional Summit Recap

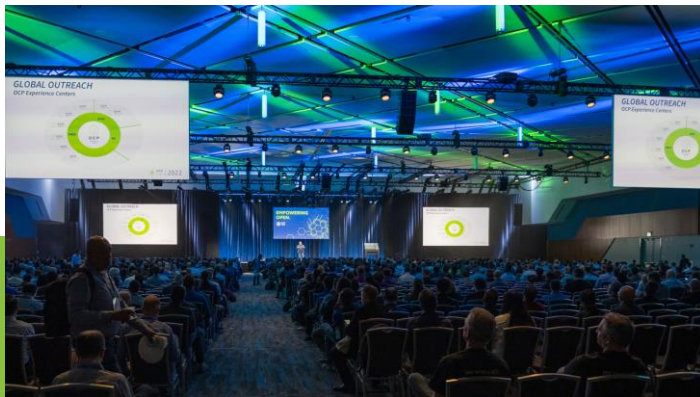
- Slides/Recordings:
<https://www.opencompute.org/events/past-events/2024-emea-regional-summit>
- ~730 attendees (up from ~560 in Prague in 2023)
- 2025 OCP Regional Summit will be held in Dublin, IE



2024 OCP Global Summit Update



- San Jose, CA. October 15th–17th, 2024
- Registration starts in June! Speakers receive a free pass..
- Call for Presentations launching soon...



OCP Security Appraisal Framework & Enablement

OCP S.A.F.E. Program Overview

Microsoft | Tech Community Community Hubs ▾ Blogs Events ▾ Microsoft Learn ▾ Lounge

[Home](#) > [Azure](#) > [Azure Infrastructure Blog](#) > OCP-SAFE, a systematic hardware security appraisal framework

[Back to Blog](#)

[< Newer Article](#)

[Older Article >](#)

OCP-SAFE, a systematic hardware security appraisal framework



By [Bryan Kelly \(NCSA\)](#)

Published Oct 17 2023 08:00 AM

👁 5,485 Views



OCP
S.A.F.E.

Azure Blog [Link](#)



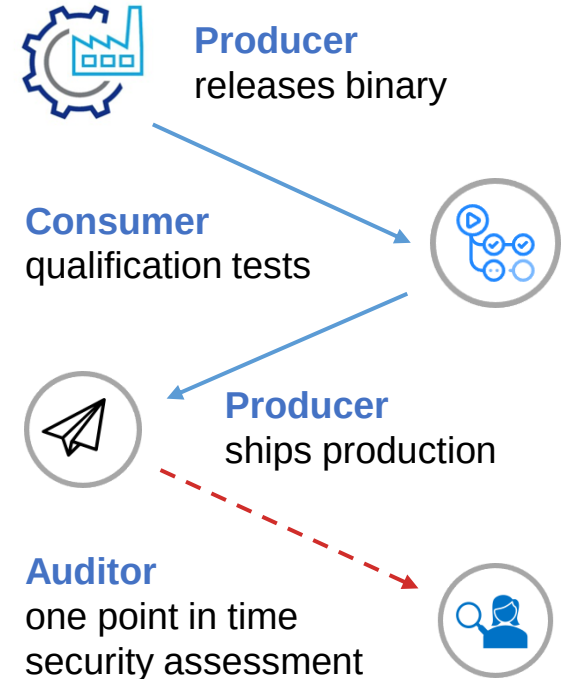
Challenges with Today's Model

Security reviews are **one point in time only**

Vulnerabilities often found **late in production lifecycle**

Assessment results **often exclusive** to large consumers

Multiple audits performed by different consumers
without added value



Key Components to Hardware Integrity



Security Assessment Report provides assurance the firmware reported by the device is trustworthy



Reference Integrity Manifest provides reference information for firmware reported by device



Software Bill of Materials inventory of components and dependencies



Hardware **Root of Trust** provides cryptographic endorsement of all firmware and config is running on a device



OPEN
Compute
Project®

Community-driven hyperscale innovation for all.

OCP S.A.F.E. Model

Systematic **appraisal** framework

Not bound to device “Profiles”,
not exclusive to crypto boundaries

Assessment focused on
OCP S.A.F.E. assurance

Removes redundancy of duplicate audits



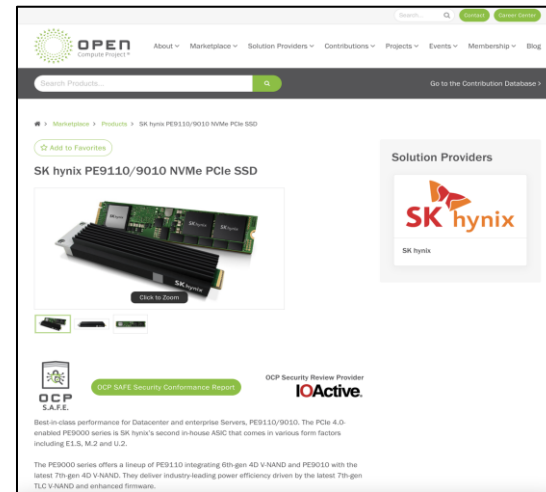
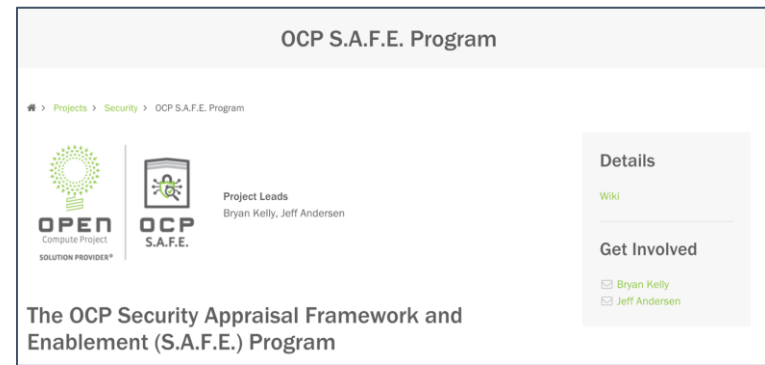
OCP S.A.F.E. Details

The program is under the leadership of the OCP Security Project

There is a Security Review Provider (SRP) Checklist that an auditor fills out to participate

There is a Device Checklist that a supplier fills out to have a review completed

When a review is complete, a badge appears with the device in the OCP Marketplace



Learn more at <https://www.opencompute.org/blog/open-compute-project-tackles-data-center-hardware-and-firmware-security>:

- Visit the OCP S.A.F.E. Project Page (<https://www.opencompute.org/projects/ocp-safe-program>)
- Visit the OCP S.A.F.E. GitHub (<https://github.com/opencomputeproject/OCP-Security-SAFE>)
- Visit the OCP Marketplace and view the OCP S.A.F.E. approved products (<https://www.opencompute.org/products>)
- For questions about the S.A.F.E. program please contact:
 - Steve Helvie (steve.helvie@opencompute.org)



OCP
S.A.F.E.

OCP Marketplace



Integrated Solutions



Facilities & Services



Telecom & Edge



Products



OPEN
Compute Project
SOLUTION PROVIDER®



OPEN
Compute Project
COLO SOLUTION PROVIDER™



OCP
ACCEPTED™



OCP
INSPIRED™



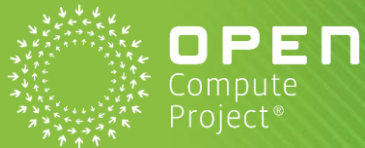
OCP
READY™



OCP
Experience
Center



OCP
S.A.F.E.



OPEN
Compute
Project®

Community-driven hyperscale innovation for all.

SAMSUNG



KIOXIA



riscure

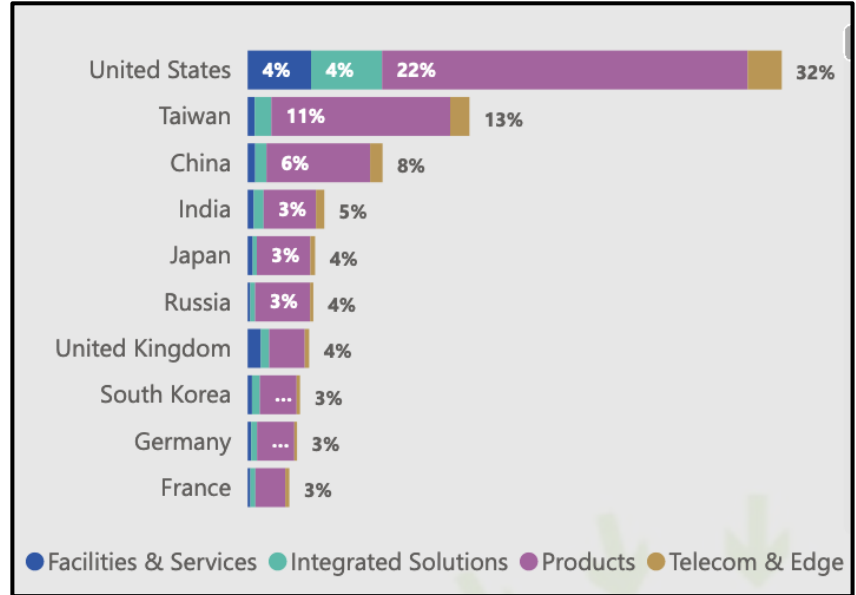
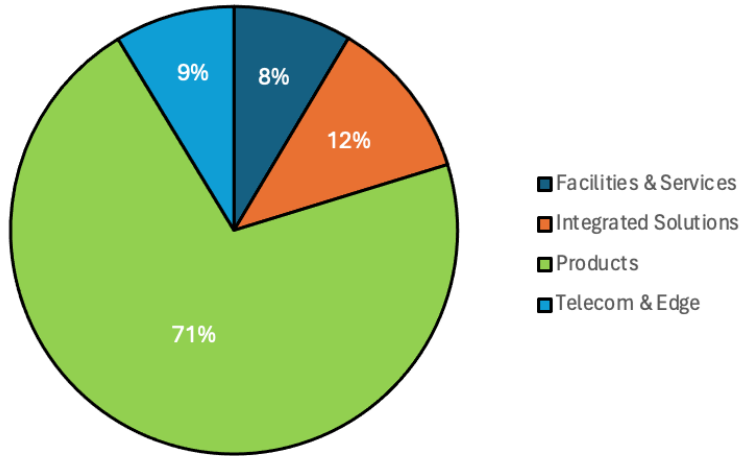
Driving your security forward



Marketplace Summary

275 listings on the Marketplace
> 60 members are now part of the SP program
Averaging = 13,340 views per month

Page Views



Stand-alone Hardware ▾

- ☐ Firmware (1)
 - ☐ OCP S.A.F.E. (1)
- ☐ Storage (14)
 - ☒ SSD (8)
 - ☐ Storage Systems (Server & JBOD) (3)
 - ☐ Storage Testing Services (3)

Firmware ▾

- ☐ All Open Source Firmware (1)
- ☐ OCP S.A.F.E. (1)

Software ▾

Provider ▾

- ☐ Kioxia (2)
- ☐ Micron Technology (3)
- ☐ SKhynix (1)
- ☐ Samsung Electronics (2)

[Clear refinements](#)

Results: All [☆ Favorites](#)

Sort by: [Relevance](#) ▾

Samsung PM9A3 NVMe PCIe SSD

High Performance The Samsung PM9A3 NVMe SSD supports PCI Express Gen 4, breaking through the read performance bottleneck experienced b...

Solution Provider: Samsung Electronics

Model #: Samsung PM9A3 NVMe PCIe SSD

[Associated Approved OCP Contributions](#)



SK hynix PE9110/9010 NVMe PCIe SSD

OCP SAFE Security Conformance Report OCP Security Review Provider...

Solution Provider: SKhynix

Model #: SK hynix PE9110/9010 NVMe PCIe SSD



[Associated Approved OCP Contributions](#)



Micron XTR NVMe PCIe SSD

The Micron XTR SSD is engineered to deliver extreme levels of endurance for write-intensive workloads. The Micron XTR was purpose-built for data ...

Solution Provider: Micron Technology

Model #: Micron XTR NVMe PCIe SSD

[Associated Approved OCP Contributions](#)



KIOXIA XD7P Series - 15mm with Heatsink

KIOXIA XD7P Series E1.S SSDs are designed to the Enterprise and Datacenter Standard Form Factor (EDSFF) E1.S specification to address t...

Solution Provider: Kioxia

Model #: KIOXIA XD7P Series - 15mm with Heatsink

[Associated Approved OCP Contributions](#)



Micron 7450 NVMe PCIe SSD

The Micron 7450 NVMe SSD enables storage for cloud-scale and enterprise data centers to support a variety of workloads. The Micron 7...

Solution Provider: Micron Technology

Model #: Micron 7450 NVMe PCIe SSD

[Associated Approved OCP Contributions](#)



KIOXIA XD7P Series - 9.5 mm

KIOXIA XD7P Series E1.S SSDs are designed to the Enterprise and Datacenter Standard Form Factor (EDSFF) E1.S specification to address t...

Solution Provider: Kioxia

Model #: KIOXIA XD7P Series - 9.5 mm

[Associated Approved OCP Contributions](#)



Samsung PM9D3a PCIe Gen5 NVMe SSD

Storage works wonders byte after byte. It's the power of PCIe 5.0 Data is growing exponentially and storage must adapt . It's a trend that is deman...

Solution Provider: Samsung Electronics

Model #: Samsung PM9D3a PCIe Gen5 NVMe SSD

[Associated Approved OCP Contributions](#)



Micron 7500 NVMe PCIe SSD

The Micron 7500 SSD is the world's first mainstream data center SSD with 200+ layer NAND. Micron's industry-first 232-layer NAND enables class-1...

Solution Provider: Micron Technology

Model #: Micron 7500 NVMe PCIe SSD

[Associated Approved OCP Contributions](#)



SAMSUNG



KIOXIA

[Link](#)

Marketplace Listings – Testing Services

Stand-alone Hardware ▾

- ☐ Storage (14)
- ☐ SSD (8)
- ☐ Storage Systems (Server & JBOD) (3)
- ☒ Storage Testing Services (3)

Software ▾

Firmware ▾

Provider ▾

- ☐ Teledyne LeCroy (2)
- ☐ UNH-IOL (1)

Clear refinements

Results: All ☆ Favorites

Sort by: Relevance ▾

UNH - IOL - OCP NVMe® SSD Testing Services

The UNH-IOL is the sole provider of NVMe Express® conformance and interoperability testing that is required for products to be certified for the ...

Solution Provider: UNH-IOL

Model #: UNH - IOL - OCP NVMe® SSD Testing Services

[Specifications](#)



Teledyne LeCroy - OakGate SSD Test Solutions

Teledyne LeCroy is a leading provider of test, validation, and benchmarking test platforms to the storage industry. Ability to run the Teledyne LeCroy ...

Solution Provider: Teledyne LeCroy

Model #: Teledyne LeCroy - OakGate SSD Test Solutions

[Specifications](#)



Teledyne LeCroy - OCP Compliance Test Suites 1.0a and 2.0

Teledyne LeCroy is a leading provider of test, validation, and benchmarking test platforms to the storage industry. Based on OCP v1.0a and 2.0 Com...

Solution Provider: Teledyne LeCroy

Model #: Teledyne LeCroy - OCP Compliance Test Suites 1.0a and 2.0

[Specifications](#)



OCP Experience Centers - Storage

OCP Experience Center (Hwaseong, Republic of Korea) - Hosted by Samsung

Samsung Electronics is proud to present OCP Experience Center within Samsung Memory Research Center (SMRC) located in Hwaseong, the he...

Solution Provider: Samsung Electronics

Model #: OCP Experience Center - (Hwaseong, Korea)



OCP Experience Center (Austin, TX) - Hosted by Teledyne LeCroy

With state-of-the-art test facilities located around the world and industry expertise Austin Labs takes advantage of the wide array of Teledyne LeCr...

Solution Provider: Teledyne LeCroy

Model #: OCP Experience Center (Austin, TX) - Hosted by Teledyne LeCroy



OCP Experience Center for NVMe Testing - UNH-IOL (Durham, NH) - Hosted by Univ. of New Hampshire

The University of New Hampshire Interoperability Laboratory (UNH-IOL) is pleased to offer OCP NVMe 1.0 & 2.0 NVMe testing against target device...

Solution Provider: UNH-IOL

Model #: OCP Experience Center for NVMe Testing - UNH-IOL (Durham, NH)
- Hosted by Univ. of New Hampshire



OCP
Experience
Center

SAMSUNG



University of New Hampshire
**InterOperability
Laboratory**



TELEDYNE LECROY
Everywhere you look™

[Link](#)

Storage System Listings

Stand-alone Hardware ▾

- ☐ Storage (14)
- ☐ SSD (8)
- ☒ Storage Systems (Server & JBOD) (3)
- ☐ Storage Testing Services (3)

Software ▾

Firmware ▾

Provider ▾

- ☐ Wiwynn (3)

Clear refinements



Results: All ☆ Favorites

Sort by: Relevance ▾

Wiwynn High Density Storage System - SV7000G4

The SV7000G4-L configuration effectively manages all 72 x 3.5" hard drives in a micro server setup. The SV7000G4/ST7000G4 series is equi...

Solution Provider: Wiwynn
Model #: Wiwynn - SV7000G4

[Specifications](#)



Wiwynn High Density Storage System - SV7000G4-L

SV7000G4-L is the one micro server configuration that manages all 72 x 3.5" hard drives. SV7000G4/ST7000G4 series uses SAS-4 24Gbps IOC...

Solution Provider: Wiwynn
Model #: Wiwynn - SV7000G4-L

[Specifications](#)



Wiwynn High Density Storage System - ST7000G4

The SV7000G4-L micro server configuration effectively manages all 72 x 3.5" hard drives. The SV7000G4/ST7000G4 series incorporates SAS-4 ...

Solution Provider: Wiwynn
Model #: Wiwynn - ST7000G4

[Specifications](#)



Call to Action

Please reach out to steve.helvie@opencompute.org to learn more about the Marketplace and OCP S.A.F.E.

If you are a device manufacturer, consider listing your products and solutions on the OCP Marketplace

Visit one of the OCP Experience Centers and utilize testing services

Questions



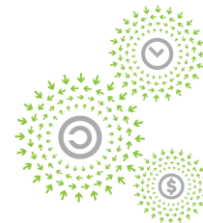
OPEN
Compute
Project®



STORAGE

Datacenter Specifications Updates

Lee Prewitt, Director Cloud Hardware, Microsoft



OPEN
PLATINUM™



Datacenter NVMe SSD Specification v2.6: Updates



OPEN
Compute Project

Datacenter NVMe® SSD Specification

Author: Ross Stenfort, Meta

Author: Lee Prewitt, Microsoft

Author: Paul Kaler, HPE

Author: David Black, Austin Bolen, Dell Technologies

Author: Chris Sabol, Charles Kunzman, Google

Expanded Error Injection

ERRIE-3	3:2	Error Injection Type	2	Error Injection type definition:	
				Value	Description
				0000h	Reserved.
				0001h	CPU/Controller Hang
				0002h	NAND Hang
				0003h	PLP Defect
				0004h	Logical Firmware Error
				0005h	DRAM Corruption Critical Path
				0006h	DRAM Corruption Non-Critical Path
				0007h	NAND Corruption
				0008h	SRAM Corruption
				0009h	HW Malfunction
				000Ah	No More NAND Spares Available
				000Bh	Incomplete Shutdown
				000Ch	Metadata Corruption
				000Dh	Critical Garbage Collection
				000Eh	Latency Spike
				000Fh	I/O command failure
				0010h	I/O command timeout
				0011h	Admin command failure
				0012h	Admin command timeout
				0013h	Thermal Throttle Engaged
				0014h	Thermal Throttle Disengaged
				0015h	Critical Temperature Event
				0016h - FFFFh	Reserved. Shall be cleared to zero.

Telemetry

- **32K default size for Data Area 1**

Telemetry Data Area	Purpose	Profile 0 Data Area Size	Latency Impact to IO
1	Periodic logging for monitoring trends/problems	32768 1638 4 bytes	< 1 ms typical, < 5 ms max

- **Virtual FIFO support added enabling greater than 16 FIFOs in Data Area 1 and 2**

Watch Dog Timer

Default time period for watchdog timer of < 2 seconds

CRASH-5

The device shall have a mechanism (e.g., watchdog timer) to allow recovery from hangs of the firmware. [The timeout interval of this mechanism shall be less than 2 seconds.](#)

Require most panic recoveries to be through NVMe Controller Reset

EREC-2	2	Panic Reset Action	1	Bit field indicating potential reset actions that may need to be taken. If no reset action is needed, do not set any of the bits. More than 1 bit can be set, and it is up to the host to decide the sequence of action(s) to take. Instead of asserting and requiring the host to perform an NVMe Controller Reset to clear the assert condition, the device should reset itself using a methodology similar to what is used to perform a firmware activation without reset (e.g., pause command processing, set the Processing Paused bit to 1b, save runtime state, re-initialize firmware, restore runtime state, resuming processing, clear the Processing Paused bit to 0b). The device shall increment the Panic Count field (see EREC-16) and shall log this event in the telemetry logs and include a vendor specific panic code in the Error ID (see TELRI-1) whether or not it can autonomously recover. If autonomous recovery is not possible, then the assert condition shall be reported to the host (see EREC-5). The strongly preferred order of device advertised recovery is bit 0, 1, 2, 5, 3 then 4 which is based on least impactful to most impactful from a host perspective.												
				<table><tr><th>Bit</th><th>Bit Description</th></tr><tr><td>7:6</td><td>Reserved. Shall be cleared to zero.</td></tr><tr><td>5</td><td>PCIe Conventional Hot Reset.</td></tr><tr><td>4</td><td>Main Power Cycle.</td></tr><tr><td>3</td><td>PERST#.</td></tr><tr><td>2</td><td>PCIe Function Level Reset.</td></tr><tr><td>1</td><td>NVM Subsystem Reset.</td></tr><tr><td>0</td><td>NVMe Controller Reset.</td></tr></table>	Bit	Bit Description	7:6	Reserved. Shall be cleared to zero.	5	PCIe Conventional Hot Reset.	4	Main Power Cycle.	3	PERST#.	2	PCIe Function Level Reset.
Bit	Bit Description															
7:6	Reserved. Shall be cleared to zero.															
5	PCIe Conventional Hot Reset.															
4	Main Power Cycle.															
3	PERST#.															
2	PCIe Function Level Reset.															
1	NVM Subsystem Reset.															
0	NVMe Controller Reset.															

Security

- Require OCP S.A.F.E. compliant firmware audits

SEC-15	A firmware audit of each version of the firmware shall be performed using OCP SAFE .
--------	--

Additional Thoughts:

Add workload tracking to Telemetry?

- Needs more discussion
 - What needs to be recorded?
 - How many commands deep?



Datacenter SAS-SATA Specification v1.x: Update

Authors: Microsoft, Meta, HPE, Dell



OPEN
Compute Project

Datacenter SAS-SATA Device Specification

Version 1.5r7 (05022024)

Author: Ross Stenfort, Meta
Author: Lee Prewitt, Microsoft
Author: Paul Kaler, HPE
Author: David Black, Dell Technologies

Telemetry

- Port over the format and contents of the NVMe Telemetry from 2.5
 - Enables human readable debug logs
- Rework transport mechanisms to fit with SAS and SATA protocols
- Add HDD specific debug information

TCG Configuration and History Logs

- Port over the format and contents of the NVMe TCG Configuration Log from 2.5
- Add TCG History events

Security

- Update the Security section to be on par with OCP Datacenter NVMe SSD V2.6

Questions



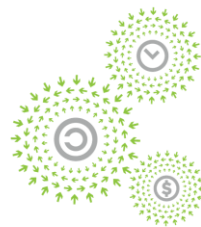
OPEN
Compute
Project®

Enhancing SSD Resiliency in OCP Storage: A V2.5 DSSD Specification Overview



STORAGE

Steven Wells, Fellow Storage Solutions Architecture,
Micron Technology



OPEN
PLATINUM™

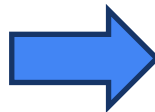
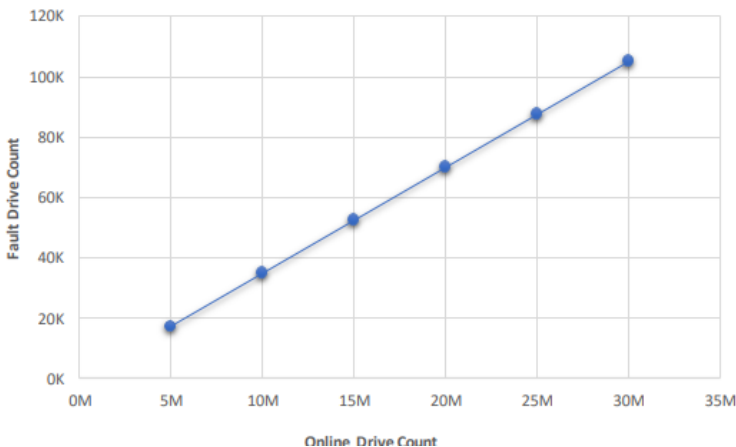


Overview

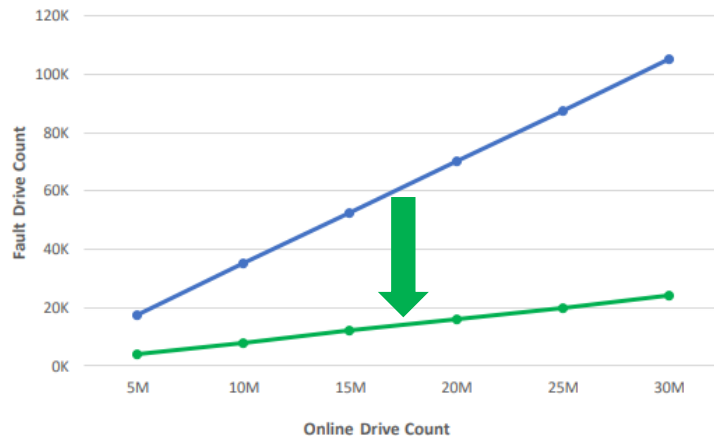
- Defining Resiliency
- Resiliency North Star
- V2.5 Spec Enhancements

Storage Resiliency

Standard SSDs Annualized Fault Drive Count
Projection with MTBF of 2.5M hour

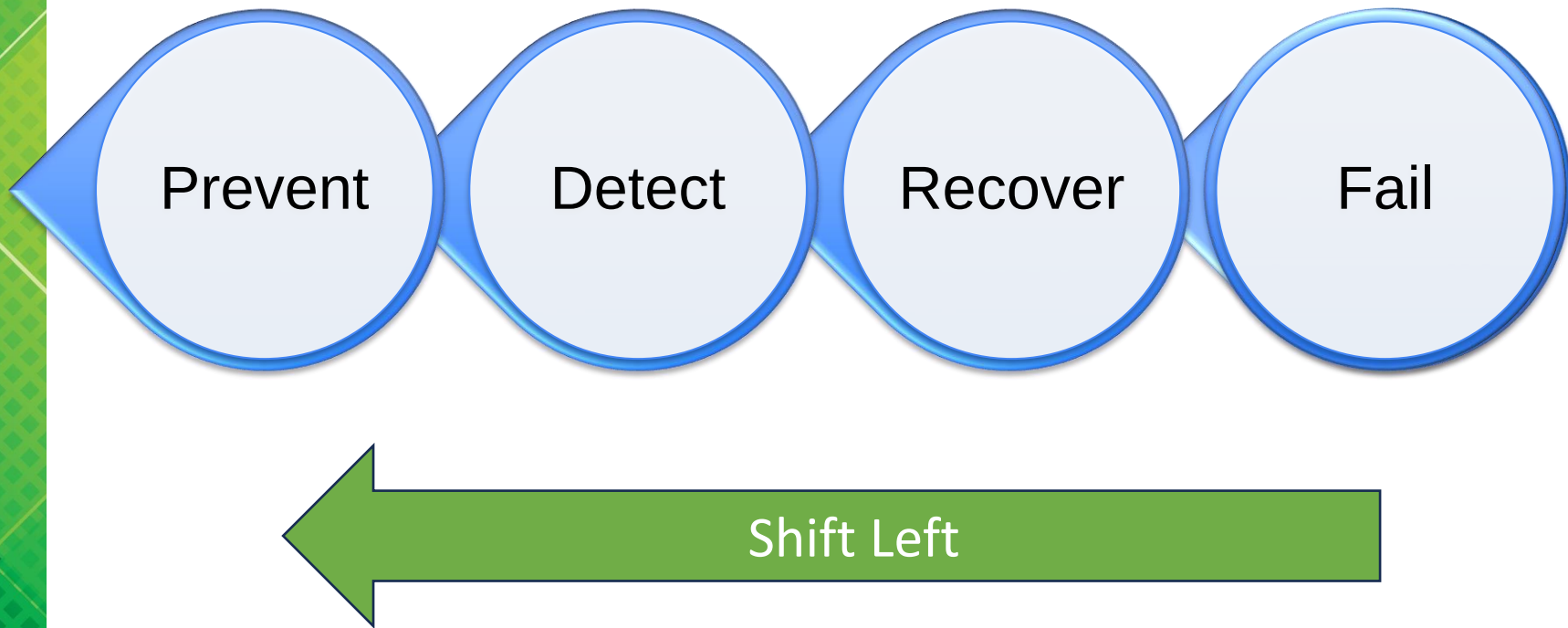


Goal



[Reference: "Vertically Integrated High Resilience SSDs Designed for Cloud Computing", Ayberk Ozturk, FMS 2023](#)

Storage Resiliency



OCP DSSD Resiliency North Star

→ Alert and Continue



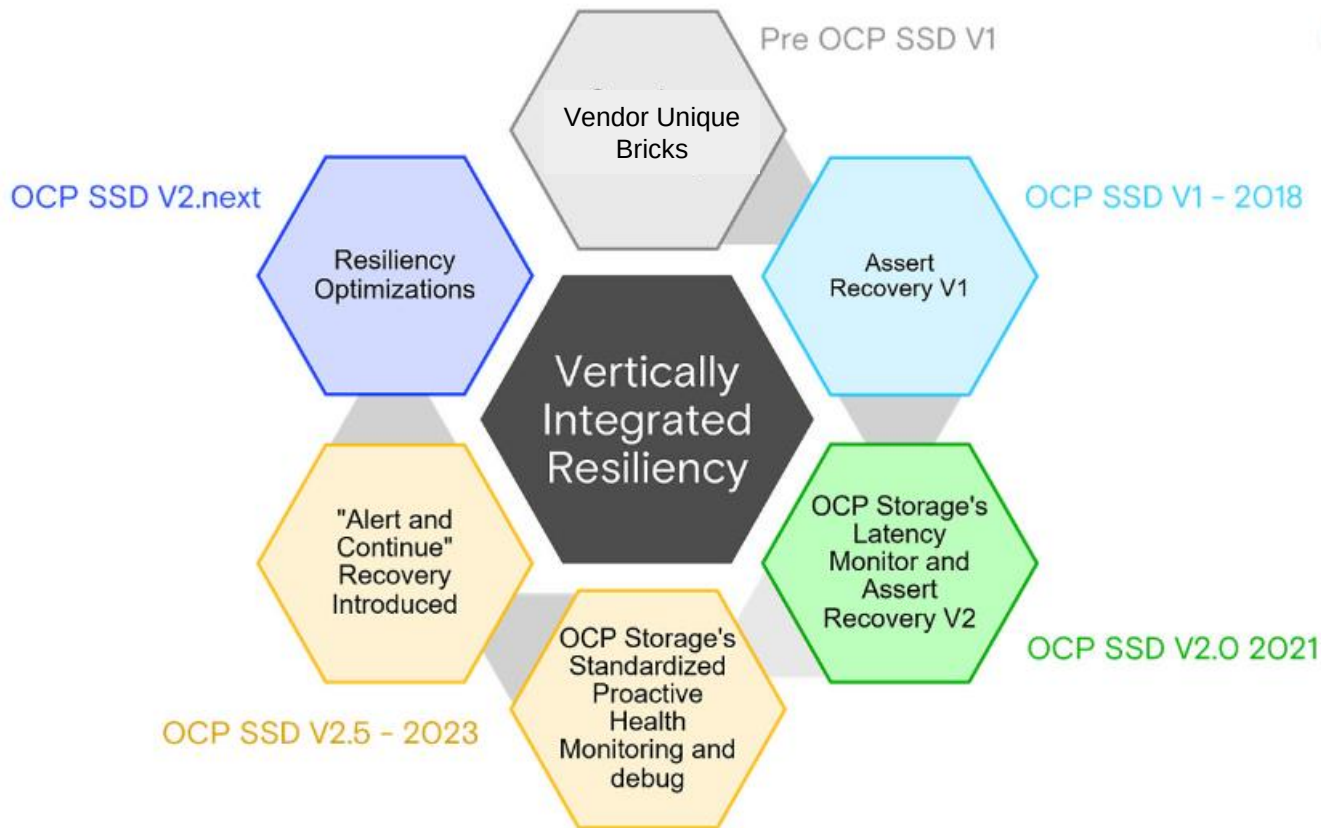
Rather than panic or require a reset as part of recovery, alert the host and continue operation.



The Alerted host can collaboratively work with vendor to evaluate standardized telemetry to determine root of problem and appropriate next steps.

Where appropriate – Panic otherwise

OCP SSD Spec – Resiliency Shift Left in the History of the OCP Specification



Key DSSD V2.5 Resiliency Features Deep Dive

Check Engine Light for OCP DSSD → Alert and Continue

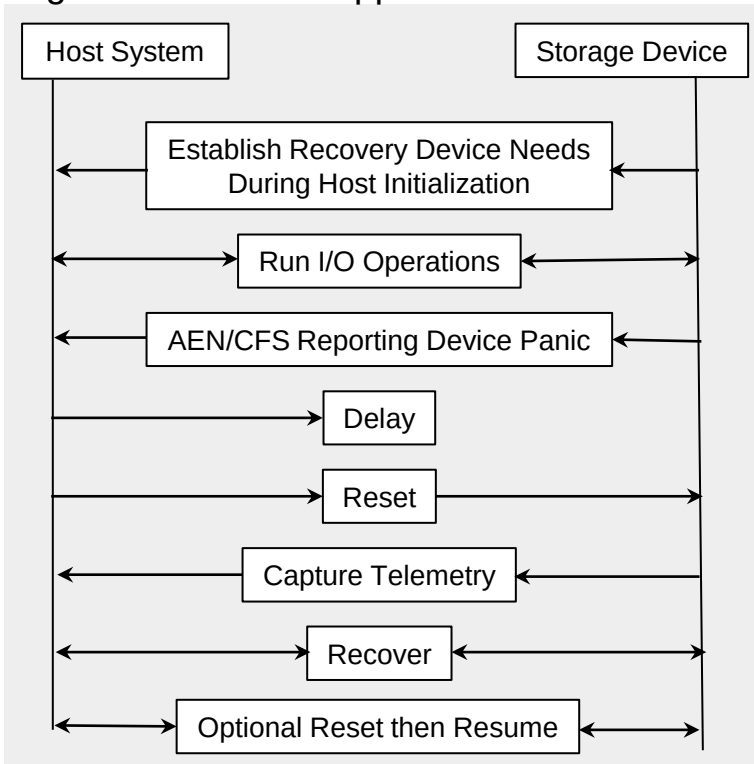
EREC-2	2	Panic Reset Action	4 Bit field indicating potential reset actions that may need to be taken. If no reset action is needed, do not set any of the bits. More than 1 bit can be set, and it is up to the host to decide the sequence of action(s) to take. 1 Instead of asserting and requiring the host to perform an NVMe Controller Reset to clear the assert condition, the device should reset itself using a methodology similar to what is used to perform a firmware activation without reset (e.g., pause command processing, set the Processing Paused bit to 1b, save runtime state, re-initialize firmware, restore runtime state, resuming processing, clear the Processing Paused bit to 0b). The device shall log this event in the telemetry logs and include a vendor specific panic code in the Error ID (see TELRI-1) whether or not it can autonomously recover. 2 If autonomous recovery is not possible, then the assert condition shall be reported to the host (see EREC-5).
TELRI-1	63:0	64	Err 3 The device shall set a unique Error ID defining the persistent or transient error at the time of log capture.

- 1 Device **should** reset itself similar to FW update.
 - No host reset action required
 - Utilize CSTS.PP
- 2 Device **shall** log this event in telemetry logs
- 3 Including a VS panic code in Reason Identifier
- 4 “Alert and Continue”. No reset action is an option

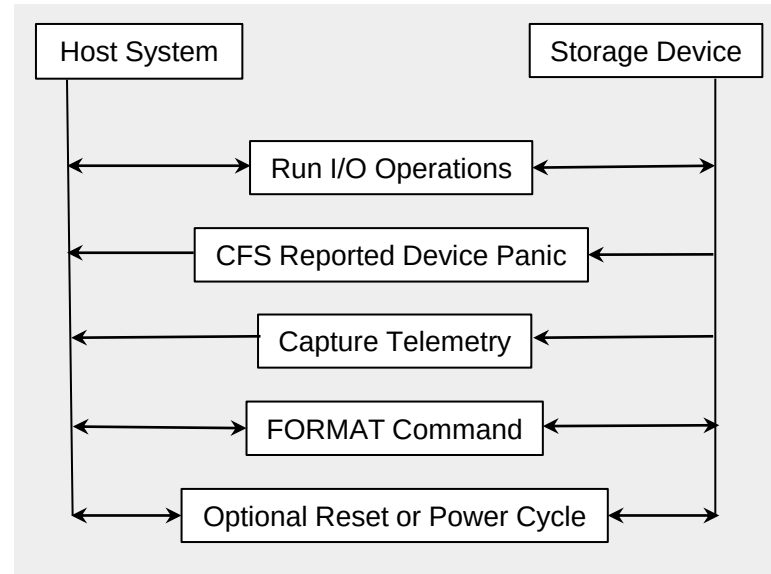
Panic Workflow

Requirement ID	Description
EREC-FUNC-1	The device shall make every attempt to transparently recover from errors and panics without invoking any of the Error Recovery workflows described below.

Log Identifier C1h Supported Workflow:



Otherwise:



Summary & Call to Action

Summary



- Alert through Reason Identifier (TELRI-1...6)
- Guidance on panic recovery (EREC-2)
- Recovery AEN with no recovery resets or actions. (EREC-2)



- OCP 2.5 Standardized Telemetry (Section 4.9)

Call to Action

- SSD Developers
 - Shift Left for Enhanced Device Resiliency
 - Implement Alert vs Panic
 - VU Standardized Telemetry
- SSD Hosts and Linux Foundation
 - Implement host elements to support a resilient system as necessary
- OCP Author Contributors
 - Continue to embrace “Alert and Continue” model similar to check engine light in ongoing specifications. Add clarity and guidance.

Questions



OPEN
Compute
Project®

OCP L.O.C.K.

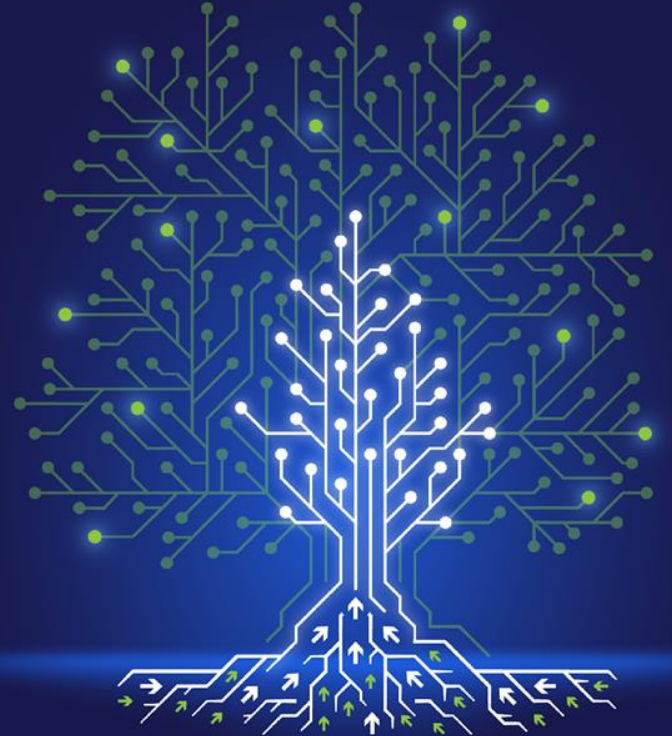
Layered
Open-source
Cryptographic
Key-management

Scaling Innovation Through Collaboration



OCP
REGIONAL
SUMMIT

24-25 APRIL 2024
LISBON, PORTUGAL

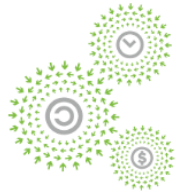




SECURITY

OCP L.O.C.K.

Jeff Andersen, Staff Software Engineer, Google
Eric Eilertson, Principal Engineer, Microsoft



OPEN
PLATINUM®



Who we are

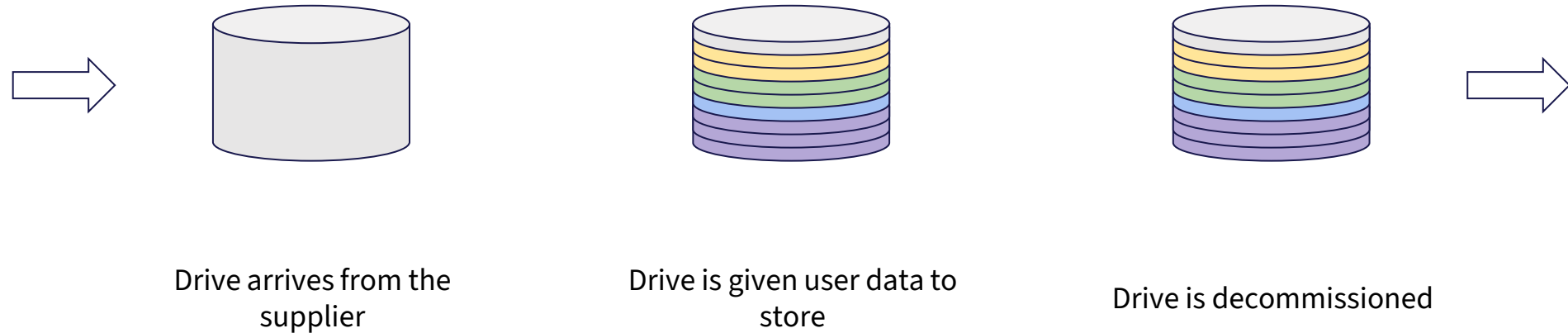


SAMSUNG

KIOXIA

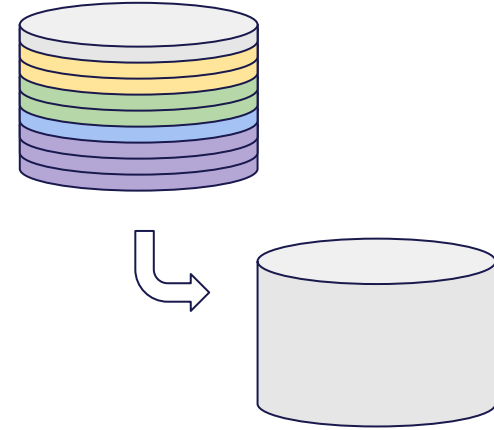


Life of a data center storage device



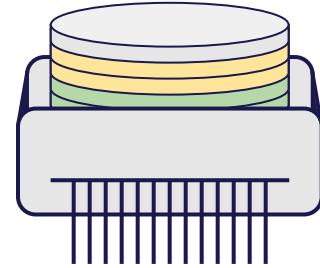
Decommissioning drives

- The physical drive is leaving the data center
- **User data cannot be permitted to escape**



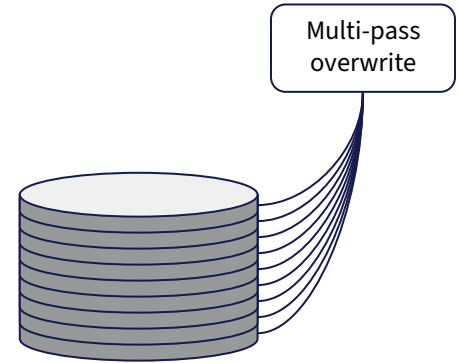
Default policy: destroy the drive

- Safest way to ensure bits on the drive don't escape
- Produces significant e-waste
- Impacts bottomline of drive owner
 - Inhibits second-hand markets



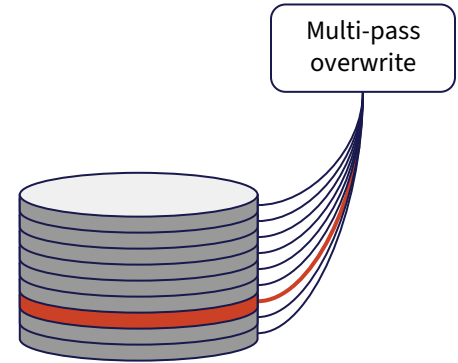
One technique: overwrite

- Write over every piece of data held within the drive
- Every portion of the drive must be overwritten, before the drive is allowed to leave in one piece



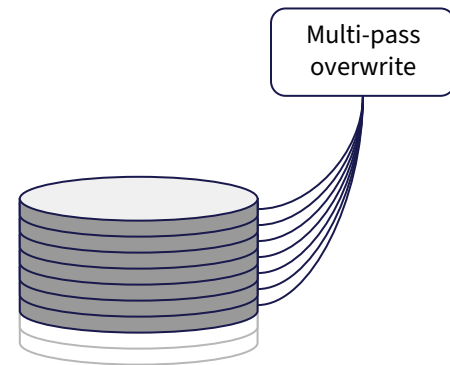
Problem: drive failure

- If *any* portion of the drive cannot be overwritten, erasure fails and the drive must be destroyed
- Ergo, we still destroy a lot more disks than we'd like



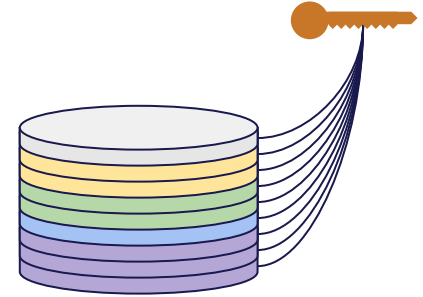
Problem: NVMe page management

- On NVMe drives, bad pages are hidden from the host
- The host cannot even address such pages
- Hidden pages may have user data



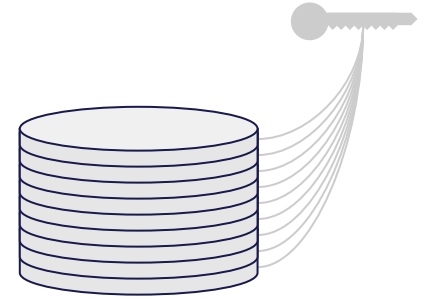
Solution: drive encryption

- Ensure all data on the drive is encrypted to a key



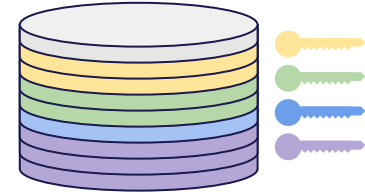
Solution: drive encryption

- Ensure all data on the drive is encrypted to a key
- Forget the key



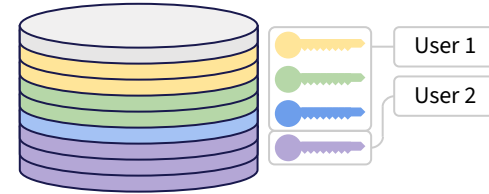
NVMe self-encrypting drives

- The drive manages encryption keys
- Allows granular mapping of keys to address ranges



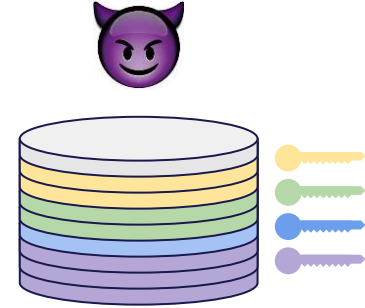
NVMe self-encrypting drives

- The drive manages encryption keys
- Allows granular mapping of keys to address ranges
- Allows granular mapping of keys to users



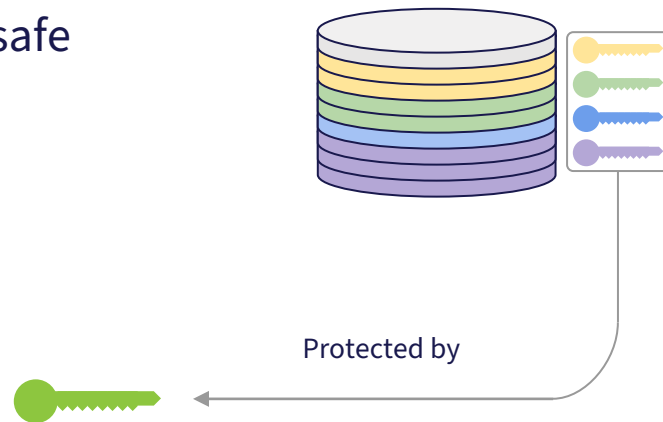
Risk: drive theft

- Keys must be erased before the drive leaves the DC
- If the drive is stolen, the keys survive
- A determined adversary may obtain user data



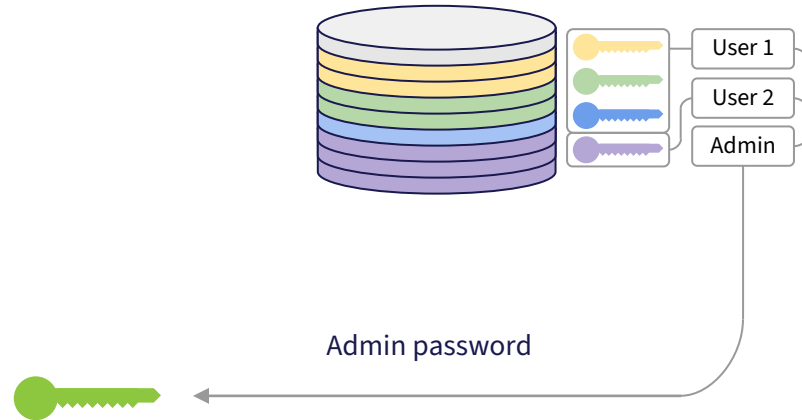
Mitigation: key material held outside the drive

- All media keys protected with a secret the drive does not have
- If the drive is stolen, the 'root secret' remains safe
- By extension, all media keys remain safe

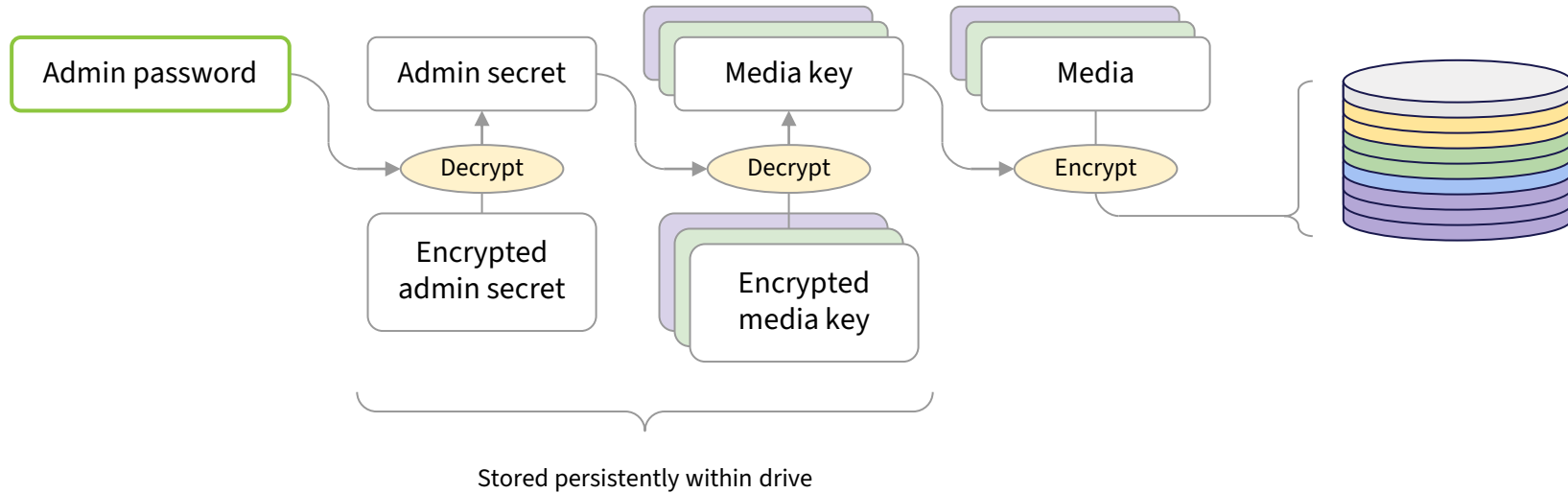


How in practice: admin credentials

- Set up a strong admin password
- Hold the password off-drive, such as in a TPM
- Rely on the drive to transitively protect all media keys with the admin password

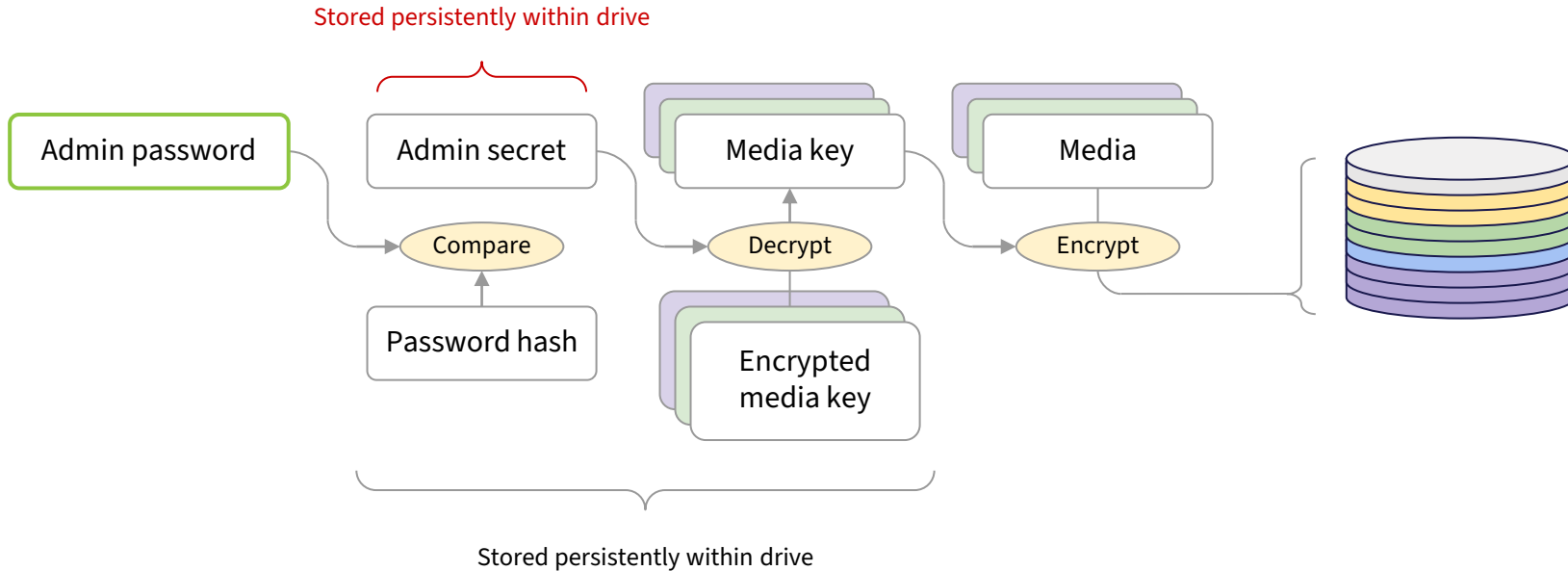


A working implementation



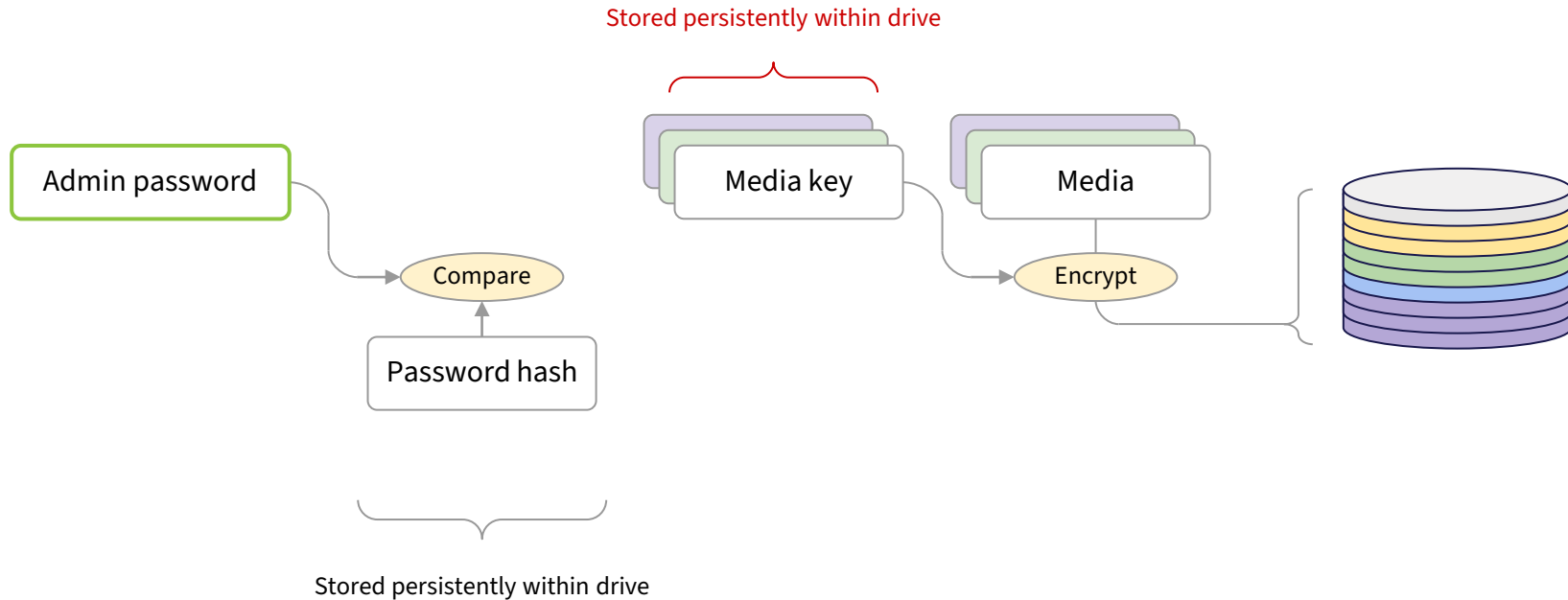
Over-simplified diagram

Broken implementation #1



Over-simplified diagram

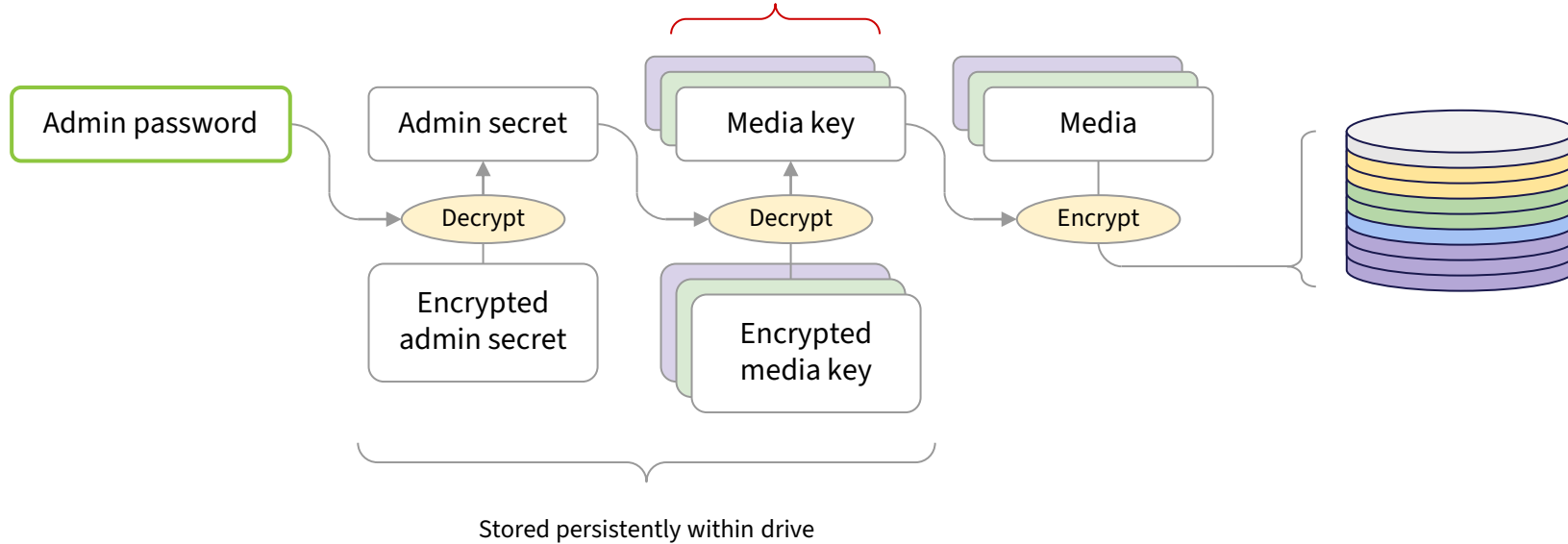
Broken implementation #2



Over-simplified diagram

Broken implementation #3

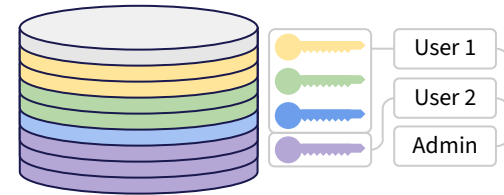
Extractable via external interfaces (JTAG, UART, PCIe, etc.) or glitch attacks



Over-simplified diagram

Overall problem

- Storage key management is critical to get right
- Threat model is significant
 - Drive theft, supplier infiltration, hardware attacks
- Implementations vary in quality
- Auditing implementations is a chore
 - Post-deployment fixes are herculean



Recall: Caliptra

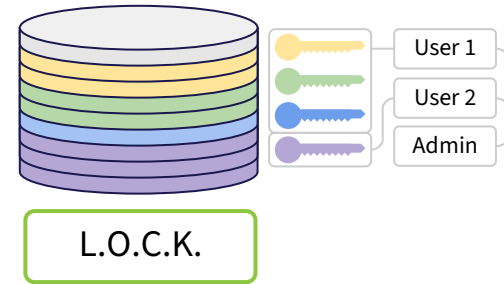


- Silicon roots of trust are critical components in data center hardware
- Caliptra is an OCP specification for an **internal root of trust IP block for SoCs**
- An open source implementation has been delivered at CHIPS Alliance
 - Ensures consistency, transparency, openness and reusability
- At this level, security should be boring



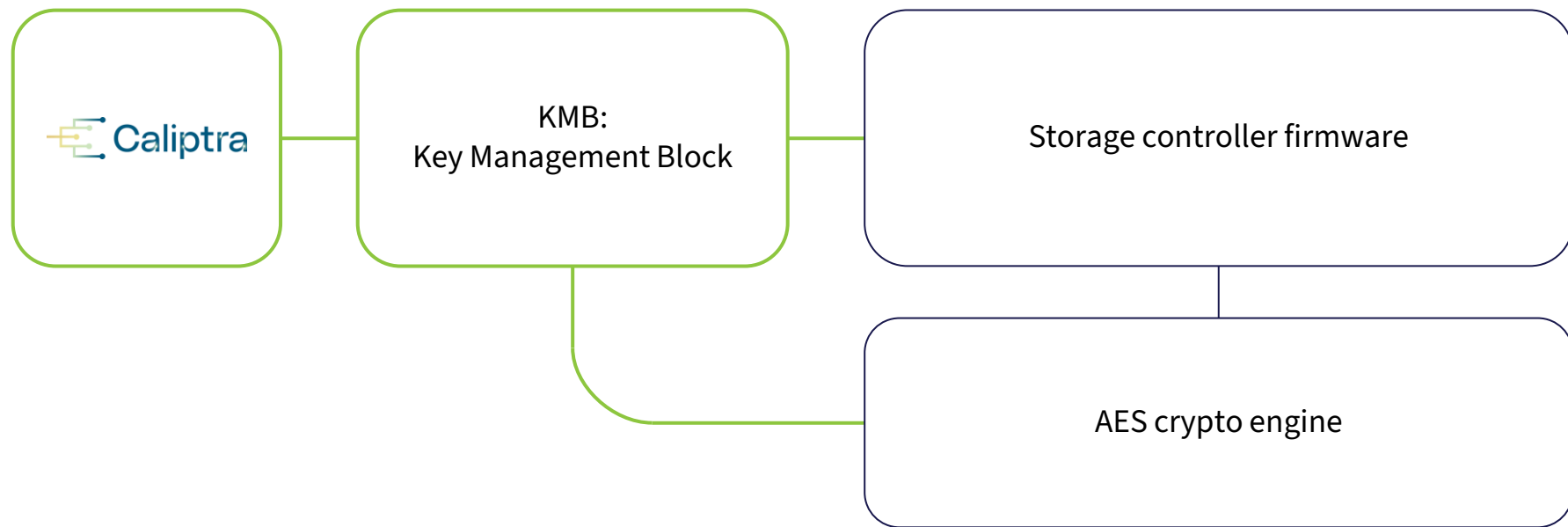
Introducing: OCP L.O.C.K.

- A project to deliver an open implementation at CHIPS Alliance, leveraging and following Caliptra
- Scoped specifically to storage devices
- Provides key management services to the drive and host, utilizing services from Caliptra



Layered
Open-source
Cryptographic
Key-management

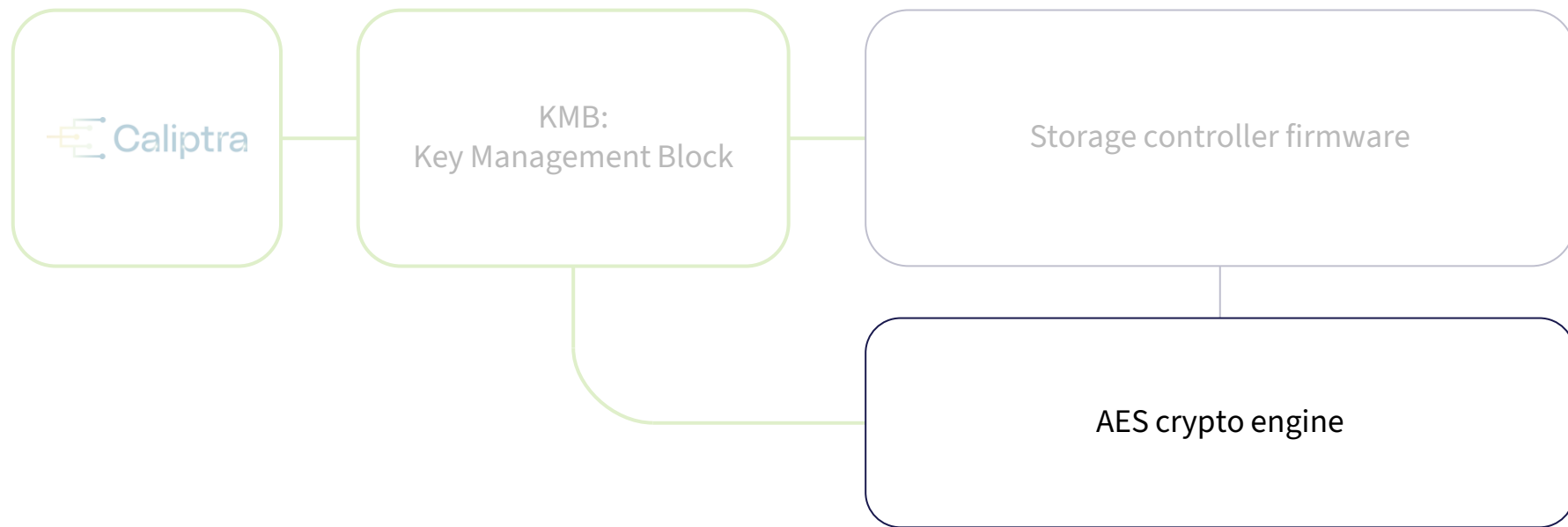
Components



Components

AES crypto engine (existing)

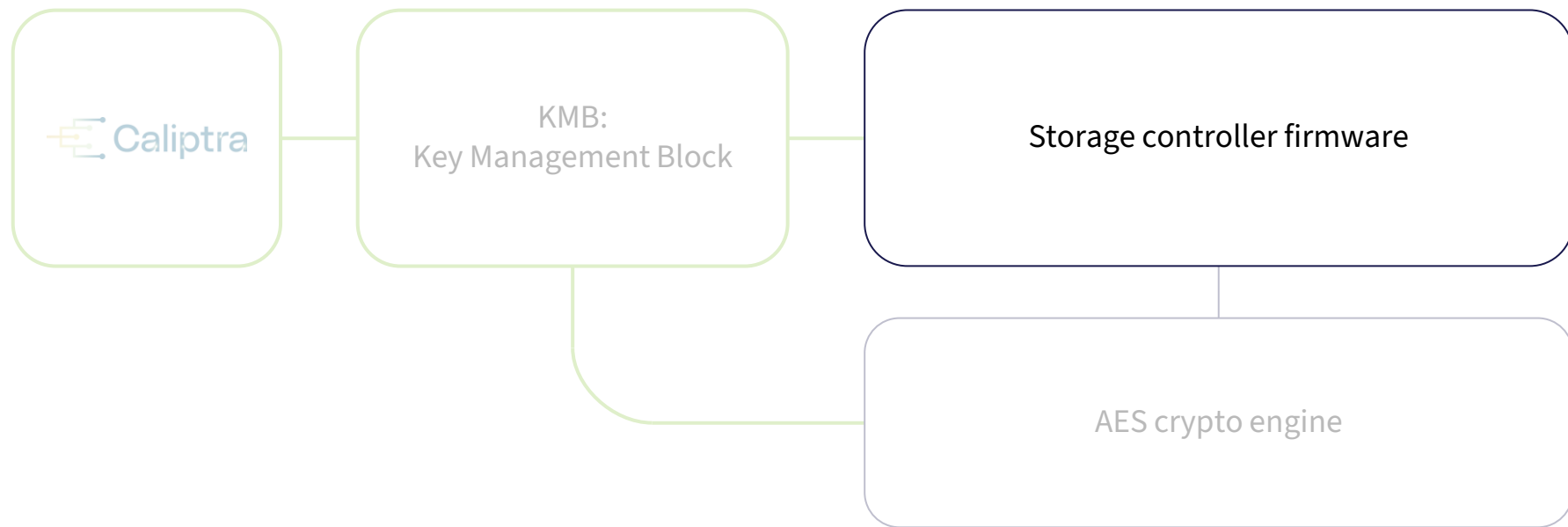
Performs line-rate encryption of data as it enters and exits the storage device



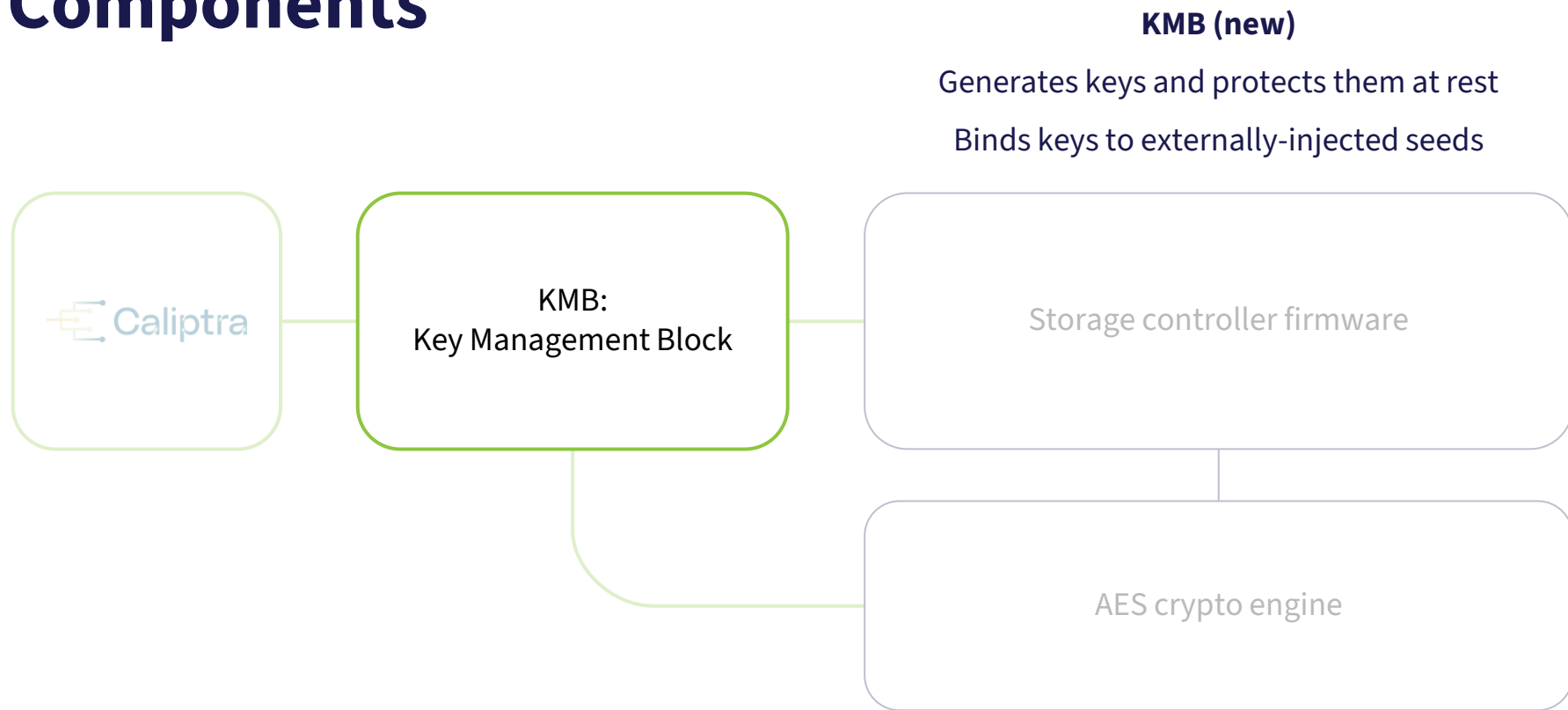
Components

Controller firmware (existing)

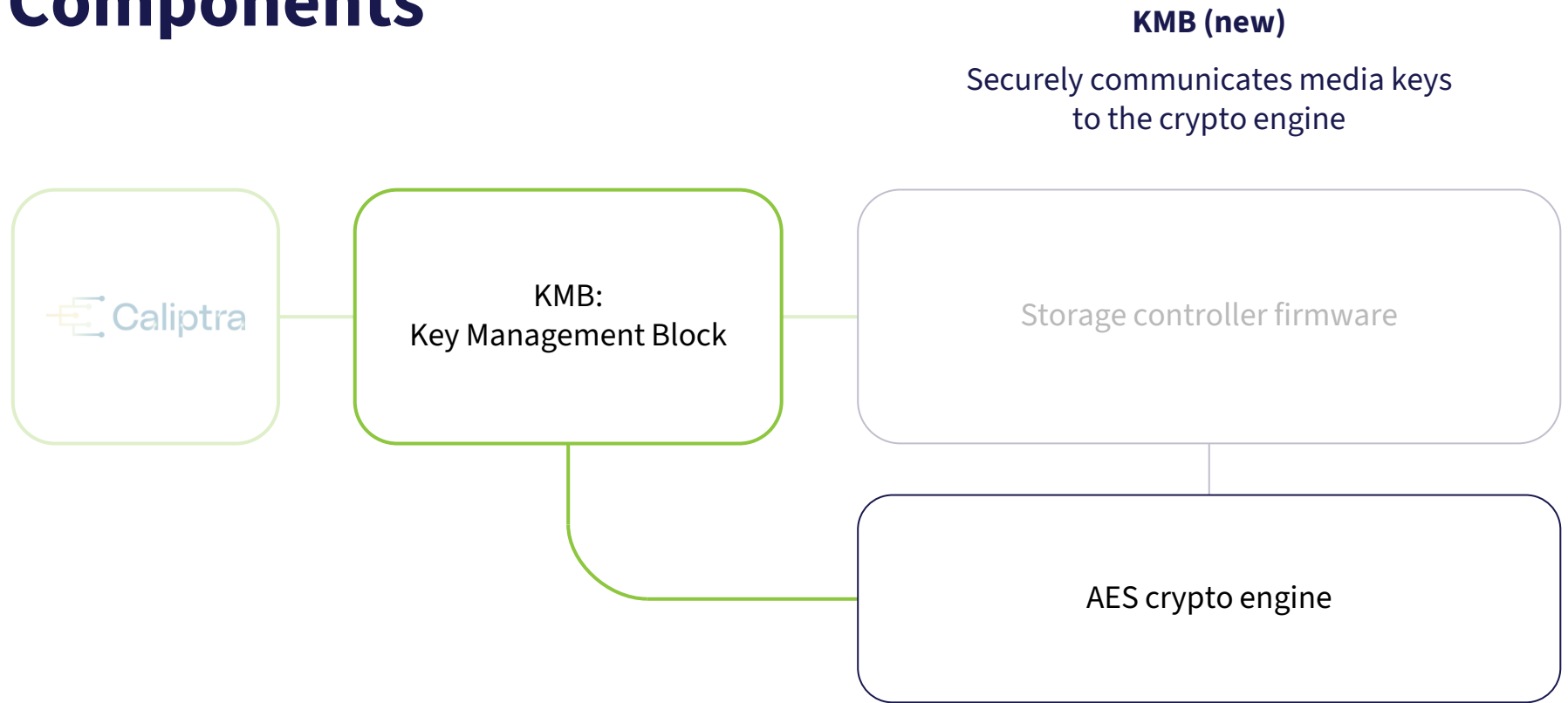
Manages users and wrapped keys



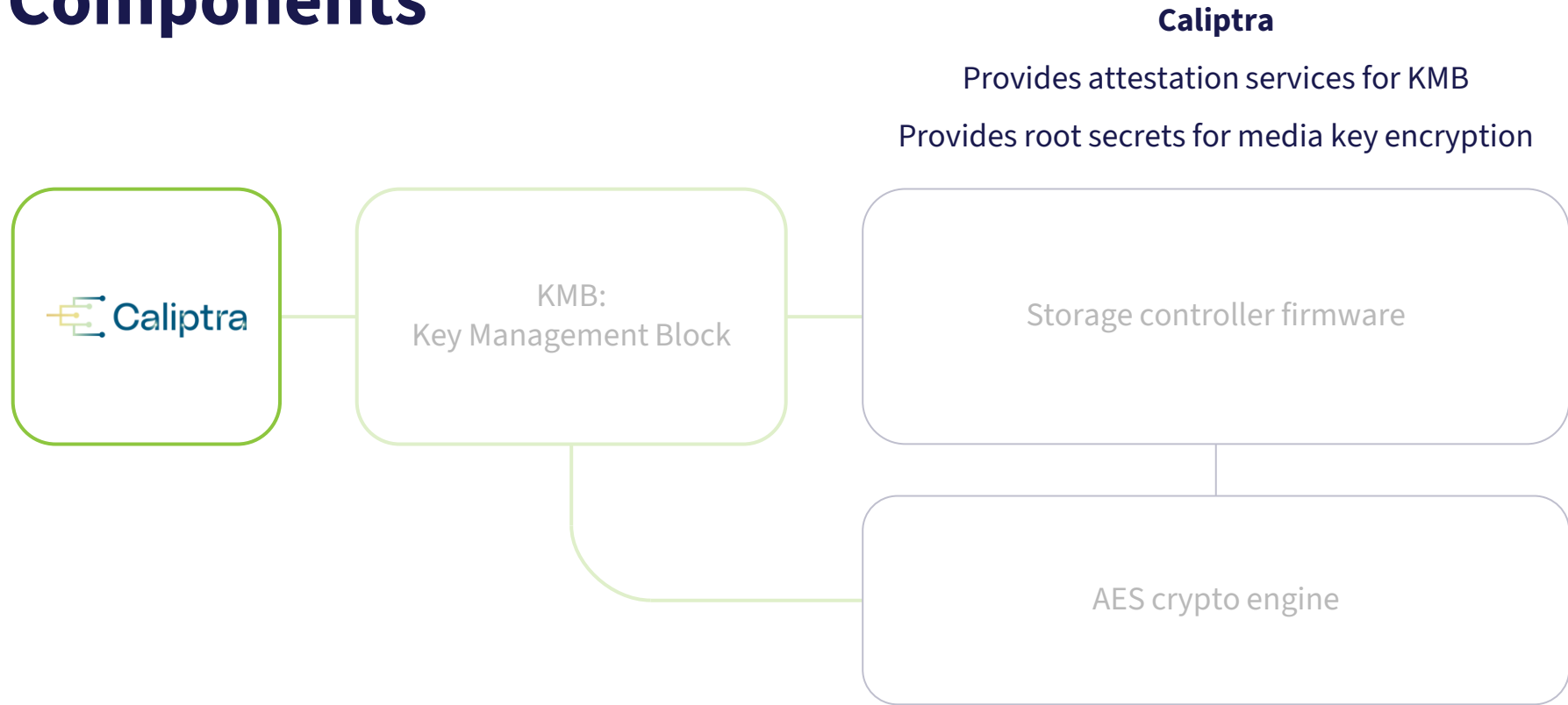
Components



Components



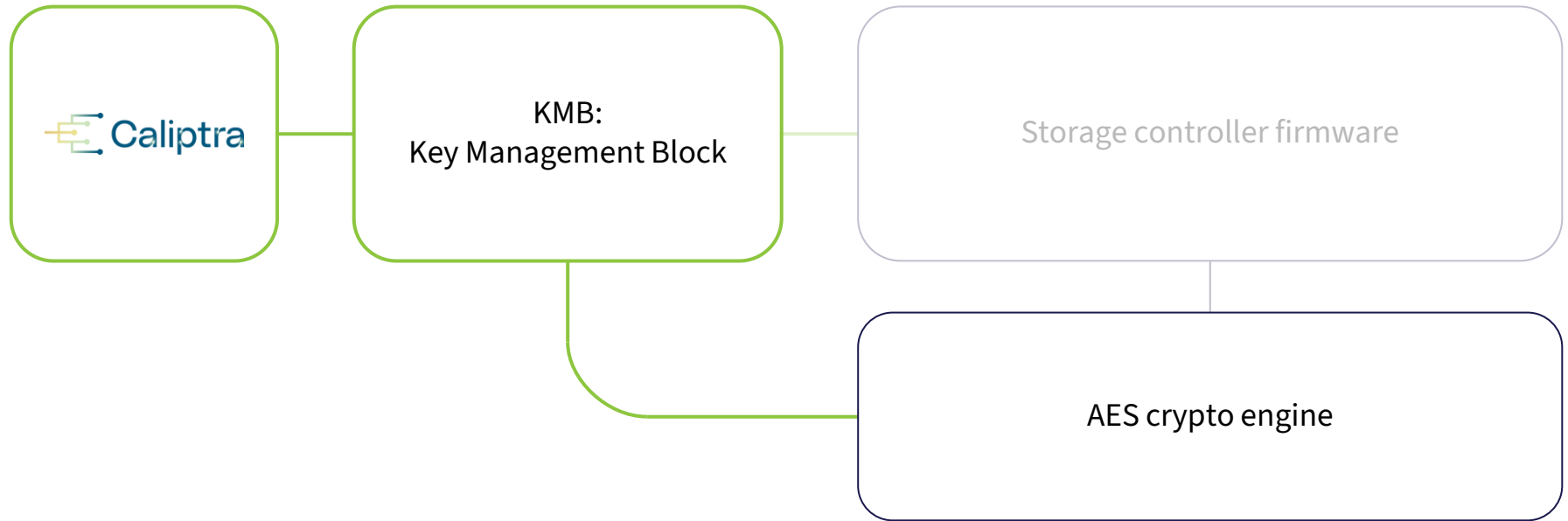
Components



Components

Trust boundary

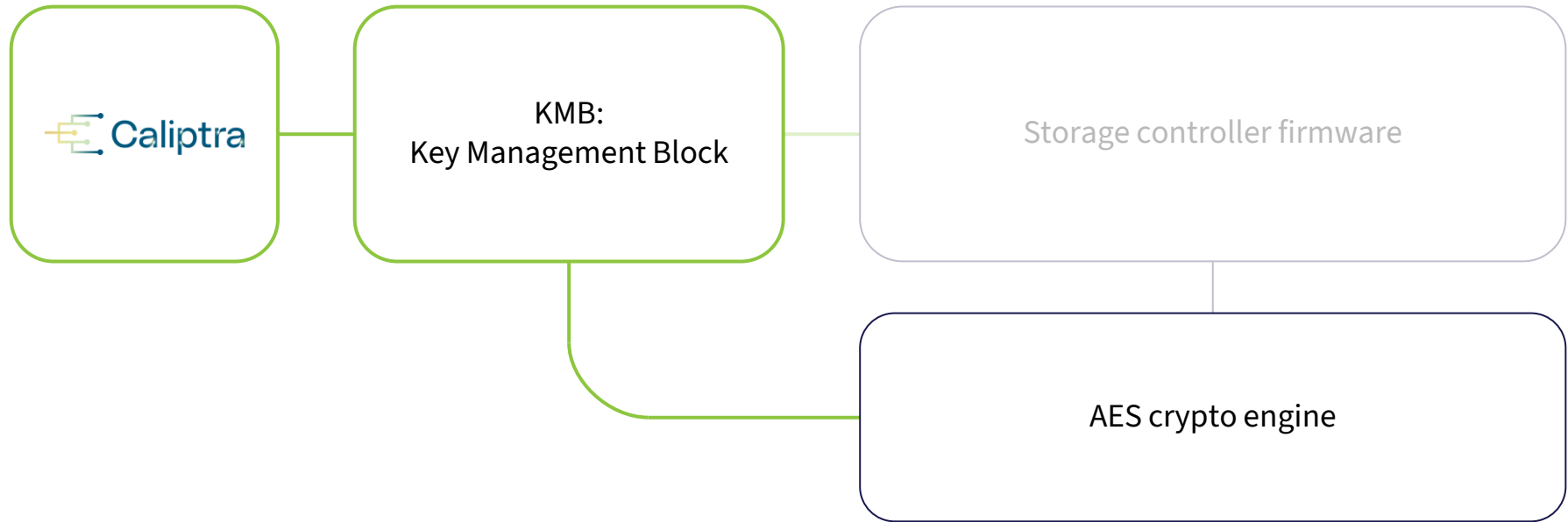
Caliptra + KMB removes system management and control interfaces from the data-at-rest TCB



Components

Trust boundary

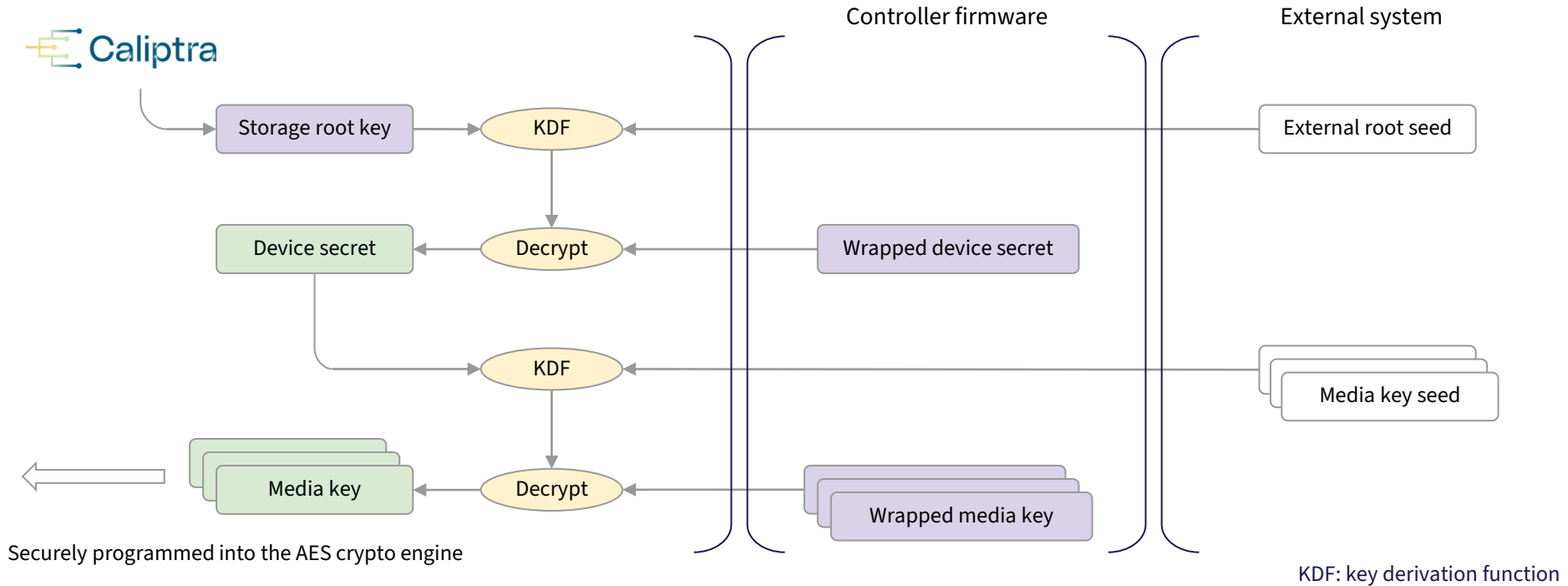
L.O.C.K. enables I/O path innovation, while maintaining a common, minimal TCB



KMB key hierarchy

Generated secret,
visible only to KMB

Persistently-stored data

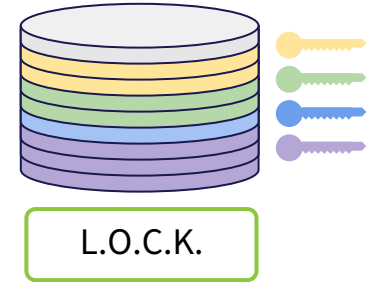


Summary and Call to Action

- L.O.C.K. will deliver a common IP block for storage devices
- L.O.C.K. ensures secure management of media keys

Call to Action:

- Look for the 0.5 spec later this summer
- Join CHIPS Alliance if interested in collaborating on the implementation



Questions



OPEN
Compute
Project®



STORAGE

Hyperscale Fleet Learnings

Vineet Parekh, Meta
Chris Sabol, Google
Lee Prewitt, Microsoft



STORAGE

Open Source Testing

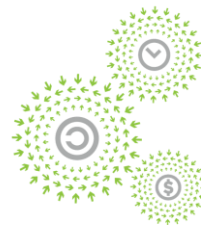
OCP Storage Test Repository

Vineet Parekh, Meta

Ross Stenfort, Meta

Jan Seidel, Meta

Adrian Enache, Meta



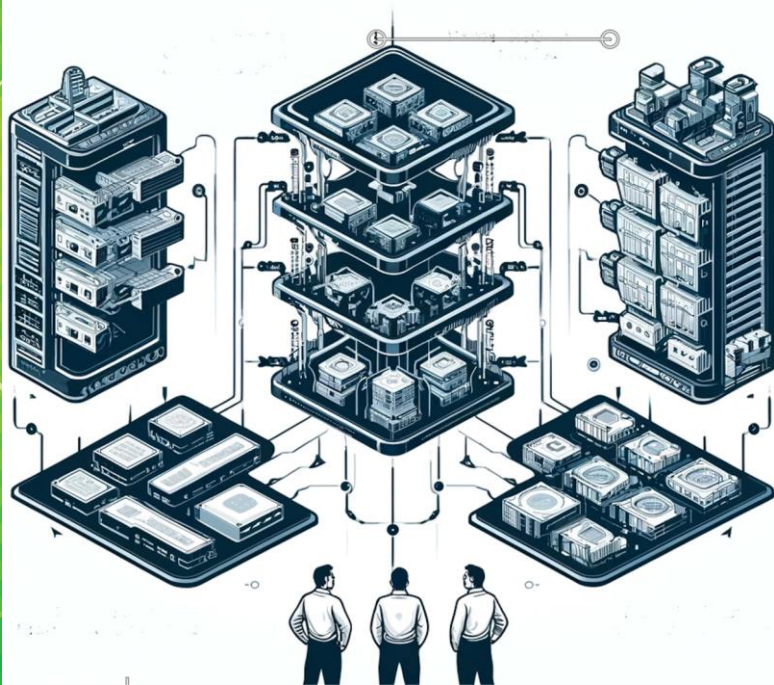
OPEN
PLATINUM™

Imagine a World....

Where open sourcing tests becomes an innovation
cornerstone for seamless and efficient qualifications



The Current State of SSD Qualification



Diverse Approaches: Each hyperscaler, including Meta with its Hi5 tool, employs unique automation frameworks tailored to their specific operational needs.

Vendor Integration: Vendors face significant challenges adapting these frameworks, requiring extensive setup and support, often resulting in inefficiencies in deployment and testing processes.

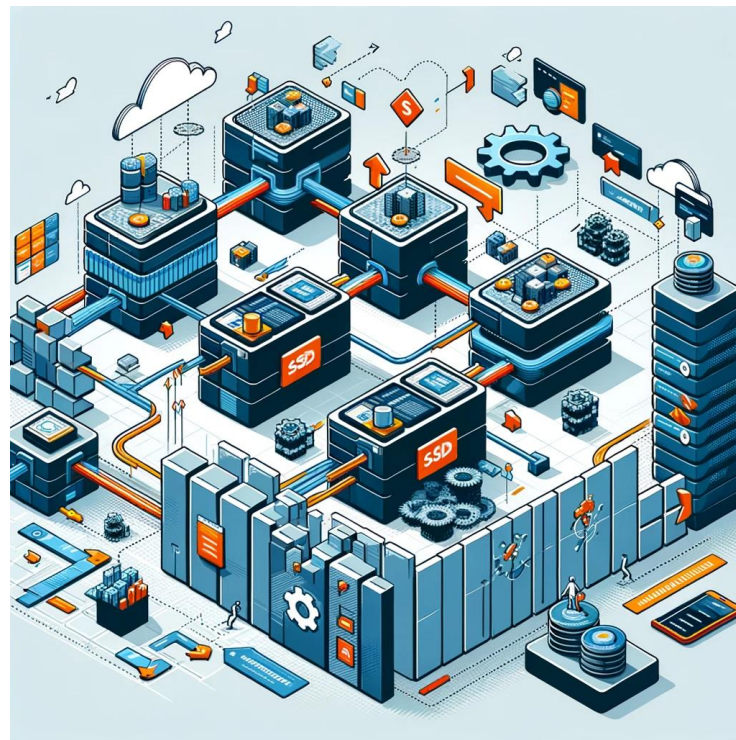
Resource Allocation: The necessity for each hyperscaler to maintain individual support teams for framework implementation and troubleshooting showcases the resource-intensive nature of current practices.

Inefficiency and Fragmentation

Barrier to Efficiency: Fragmentation caused by individualized testing frameworks creates significant barriers, preventing seamless collaboration and swift problem resolution.

Cost of Complexity: Each custom solution not only drains hyperscalers' resources but also burdens vendors, leading to increased operational costs and slower innovation cycles.

Catalyst for Change: These inefficiencies highlight the urgent need for a unified approach that simplifies the qualification process and reduces the heavy reliance on extensive support structures.



Meta Solution: Open Sourcing Tests

Open-Source Collaboration: Meta proposes the adoption of an open-source testing strategy that SSD Customers, SSD Suppliers and the industry in general can contribute to and utilize.

Reduced Complexity and Costs: By harmonizing the testing procedures, we reduce the need for multiple custom solutions. This simplifies vendor integration, decreases the demand for extensive support, and cuts operational costs.

Accelerated Innovation: Efficient qualification will lead to faster time to market.





The Future Has Arrived....

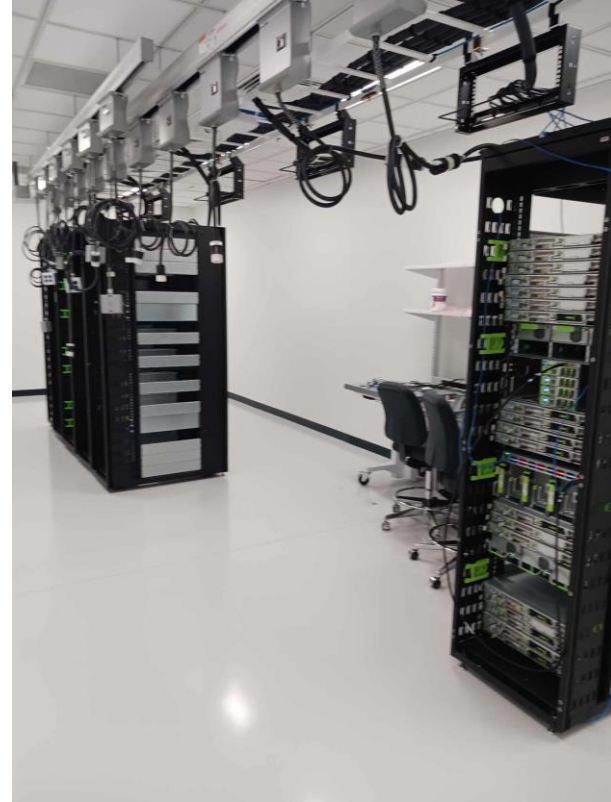
Meta test framework and storage qualification tests are now available at OCP Github Repository!

Meta OCP Framework - <https://github.com/opencomputeproject/ocp-diag-autoval>

Meta OCP Storage Tests - <https://github.com/opencomputeproject/ocp-diag-autoval-ssd>

OCP Tests and Validation - <https://github.com/opencomputeproject/ocp-diag-core>
(ongoing OCP effort)

This repository has been validated at a Meta OCP Community Lab on a OCP Server



OCPT Test Repository @ Github

opencomputeproject / ocp-diag-autoval Public

<> Code Issues Pull requests Actions Projects Security Insights Meta Internal

main 1 Branch 0 Tags

Go to file

Code

mimir-d	add pyre ignore's to be fixed later	d7f93a9 · 4 minutes ago · 6 Commits
src/autoval	add pyre ignore's to be fixed later	4 minutes ago
.gitignore	add pyproject.toml	2 days ago
LICENCE.md	fix readme for usage instructions	4 minutes ago
README.md	fix readme for usage instructions	4 minutes ago
pyproject.toml	add autoval_test_runner script	2 days ago

README MIT license

What is AutoVal?

AutoVal is a test runner framework. The goal of AutoVal is to provide validation functionality and helpful utilities for the hardware validation tests. AutoVal supports running tests on single or multiple hosts.

Key framework components

opencomputeproject / ocp-diag-autoval-ssd Public

<> Code Issues Pull requests Actions Projects Security Insights Meta Internal

main 1 Branch 0 Tags

Go to file

Code

mimir-d	fix readme for usage instructions	469c846 · 1 minute ago · 4 Commits
src/autoval_ssd	import codebase from internal	16 hours ago
.gitignore	add pyproject.toml	16 hours ago
LICENCE.md	fix readme for usage instructions	1 minute ago
README.md	fix readme for usage instructions	1 minute ago
gluster_configuration.md	fix readme for usage instructions	1 minute ago
pyproject.toml	add pyproject.toml	16 hours ago
rpm_repo_hosting.md	fix readme for usage instructions	1 minute ago

README MIT license

ocp-diag-autoval-ssd

ocp-diag-autoval-ssd is a collection of SSD tests using the ocp-diag-autoval test framework.

Overview

- More qualification tests to be added in the coming days
- Autoval will be compatible with OCPTV specification in near future.

Impact and Future Directions

Industry-wide Benefits: Implementing this open-source testing will not only streamline the qualification process but also enhance the overall quality and reliability across the industry.

Scalability and Flexibility: The open-source nature allows for continuous improvement and adaptation, ensuring the framework remains robust against future technological advancements.

Collaborative Success: Encouraging collaboration among hyperscalers, OEMs, vendors, and storage developers fosters an efficient community-focused approach to solving industry challenges.



Call To Action

- ❖ **Invitation to Collaborate:** We invite industry players to join us in this initiative. Your expertise and insights are crucial to creating a more efficient and innovative future.
- ❖ **How to Get Involved:**
 - Utilize open-source test cases across the industry
 - Meta will provide a more detailed presentation and discussion in a future monthly storage meeting
 - Participate in OCP Test And Diagnostics effort ([link](#))
 - Additional innovative ideas? Reach out to Vineet Parekh or Ross Stenfort

Join Us In Open Diagnostics Innovation,
where open sourcing tests becomes an innovation cornerstone
for seamless and efficient qualifications

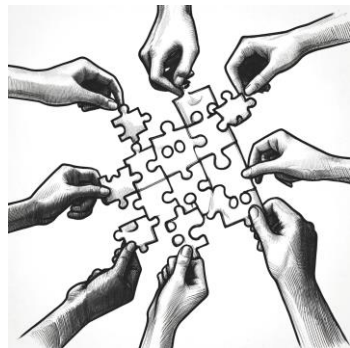
Hyperscale Fleet Learnings



STORAGE

Chris Sabol, Storage Architect, Google

Google High Level Thoughts



OCP Datacenter NVMe Specification

Helpful to have clear and aligned requirements



The Foundation is Security

Our users trust us with their data, we must honor that



Qualifying + Maintaining Storage at Scale is Hard

Need more efficiency in testing and maintenance

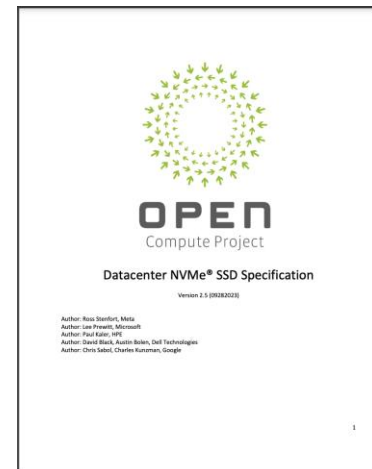
OCP Datacenter NVMe Specification

Challenge: Disparate Customer Asks for NVMe SSDs

- Hard for vendors to determine what is most important to support

Solution: **OCP Datacenter NVMe Specification**

- Focus on one high volume and well validated solution
- Great value in the spec overall, and specifically on:
 - Flexible Data Placement
 - Telemetry
 - Common Opal functionality requirements



Please support OCP DC NVMe revision 2.5

Safeguarding User Data & Trust

User Trust is Paramount

- Security elements are opaque
- An exploit on one SSD results in assumption of **vulnerability of all** SSDs

Approach to Solve: Open Source Hardware

- **Remove opacity** and lift all boats
- Step 1: Audit drives with [OCP S.A.F.E](#)
- Step 2: Integrate [Caliptra](#) as hardened Root of Trust
- Step 3: Standardize NVMe Key Mgmt with [OCP L.O.C.K.](#)
- Step 4: Integrate OCP L.O.C.K.



Securely erasing user data is the foundation for circularity

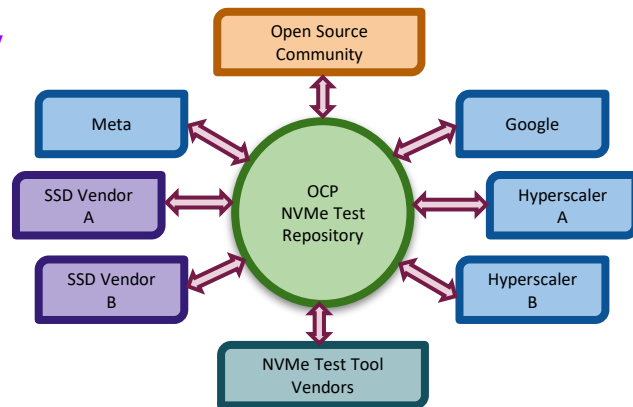
Qualifying + Maintaining Storage at Scale

Needs Shared in 2023:

- Industry needs more efficiency in testing (e.g. open source tests)
- Industry needs to improve debug at scale as well (e.g. telemetry)

Result in 2024: **Open Source OCP Test Repository**

- Broadens visibility into Google test cases
- Allows break fix iteration earlier in development cycles
 - Easier debug; predictable quals.
 - Both firmware & silicon
- Open source enables expertise outside Google



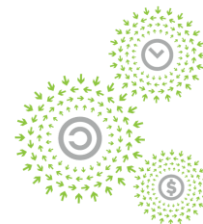
Google finalizing first 3 test cases; stay tuned :-)



STORAGE

Hyperscaler Perspectives

Lee Prewitt, Director Cloud Hardware, Microsoft



OPEN
PLATINUM™

Security

- Security is table stakes!
- Now up to 120 requirements in Section 12 – Security in OCP 2.6
- S.A.F.E.
- FIPS 140-3
- Caliptra
- L.O.C.K.

Resiliency

- Million to one chances happen 9 times out of 10 at scale
- Increasing core counts and device size increases blast radius
- Simple asserts don't cut it anymore
- Drives must be able to recover in the least impactful way
- North star is to recover under the hood and then tell us about it with Controller Initiated Telemetry

Debuggability

- We try and shift left as much testing and validation as possible. But...
- Impossible to cover all corner cases in CDL
- OCP 2.5 Telemetry can greatly decrease the debug turnaround time
- So important that we are pushing to have Telemetry backported to in-fleet devices

Questions



OPEN
Compute
Project®



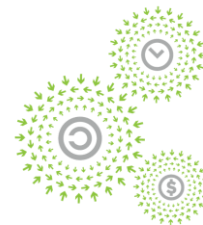
STORAGE

OCP Storage Tech Talk 2024 - Closing

Ross Stenfort, Meta, Storage Project Co-Lead

Lee Prewitt, Microsoft, Storage Project Co-Lead

Mike Allison, Samsung, Storage Steering Committee Representative



OPEN
PLATINUM™

Highlights

- Contributions: Great Progress
- OCP Marketplace:
 - Significant number of new SSDs
 - Significant number of test providers



Summary:

Collaboration on Meaningful Topics
Results In Meaningful Products

Thank You to our Speakers

Vineet Parekh

Mike Allison

Keith Busch

Javier Gonzalez

Lee Prewitt

Luca Bert

David Wang

Jonmichael Hands

Michael Schill

Rob Coyle

Steve Helvie

Jeff Anderson

Steve Wells

Eric Eilertson

Ross Stenfort

Chris Sabol

2024 OCP Storage Tech Talk

May 15, 2024

Thank you to our Sponsors



SAMSUNG



Call to Action

- ❖ Opportunities
 - Present Ideas in monthly meeting
 - Contribution to OCP
 - Participate in OCP Marketplace

Collaboration on Real World Problems

Leads to Great Ideas

Resulting Great Products

Thank You!

Thank You



OPEN
Compute
Project®