Hyperscale Storage Perspectives

Ross Stenfort, Meta

May 2023

Agenda

- SSD Form Factors
- Storage Boxes
- SSD Specification Challenges
- Write Amplification Challenges

SSD Form Factor Challenge/Solution

M.2 Challenges

- Power limited
- Connector challenges
  - Need connector designed for PCIe 5.0 and beyond
- Insufficient NAND placements
- Serviceability
- Security

Market Needs

- PCIe® 5.0 and beyond
- Scalable power, performance, thermal
- Density in 1 OU
- Serviceability
- Security

E1.S Solution

- **Connector designed for PCIe 5.0 and beyond**
- **Scales performance, power and thermal**
- **Supports 1 OU density**
- **Serviceability**
- **Security**

Significant E1.S Growth

TRENDFOCUS Share of Datacenter/ Enterprise PCIe Units*: 2021 **2.4%** → 2027 **40.4%**

* Data excludes SSD consumption where companies buy NAND and build SSDs for internal use.

Real World Hyperscale Systems

4OU Chassis with 48 25mm E1.S SSDs

Up to 768 TB

4OU Chassis with 36 25mm E1.S SSDs

Up to 576 TB

Links to OCP YV3 Contributions:

Yosemite V3: E1.S Faceplate:

https://www.opencompute.org/documents/e1s-faceplate-reference-design-specification-pdf

Yosemite V3: Vernal Falls E1.S 1OU Flash Blade and Expansion Board

https://www.opencompute.org/documents/e1s-expansion-1ou-1s-server-design-specification-pdf

Yosemite V3: Sierra Point E1.S 2OU Flash Blade and Expansion Board

https://www.opencompute.org/documents/e1s-expansion-2ou-1s-server-design-specification-pdf

Yosemite V3 Platform Design

https://www.opencompute.org/documents/ocp-yosemite-v3-platform-design-specification-1v16-pdf

Delta Lake 1S Server Design

https://www.opencompute.org/documents/delta-lake-1s-server-design-specification-1v05-pdf

SSD Specification Challenge

- Customer requirements are confidential
  - Standards have many optional features
  - Real customer requirement is unclear
  - Limited competition
    - Access to specifications are limited based on customer/supplier relationships
- SSD industry highly fragmented with lots of SKUs
  - Many customers ask for similar, but different features
  - SSD Suppliers have finite resources
- 3rd party test providers don't know what customers require

Result

- Product introduction delays
- Lower quality
- Difficult product/feature decisions

SSD Spec Solution: OCP Datacenter NVMe™ SSD Spec

Datacenter NVMe SSD Spec Goals

- Align Hyperscale/OEMs and SSD Vendors
  - Common features results in focused resources, improved speed and quality of results
- Share learnings based on deployments at scale
  - Example: Latency Monitoring
- Provide everything needed to build a Hyperscale / OEM SSD

Datacenter NVMe SSD Spec Coverage

- NVM Express®
- PCI Express®
- Reliability
- Thermal
- Security
- Form Factor
- SMART Logs
- Power
- SMBUS

Open-Source Tooling

- NVMe-CLI/ plugins / OCP

https://github.com/linux-nvme/nvme-cli

2019
Meta / MSFT start development of public SSD Specification

2020
Meta / MSFT contribute NVMe™ Cloud SSD V1.0

2021
Meta / MSFT/ HPE/ DELL EMC contribute Datacenter NVMe™ SSD V2.0

2023
Meta / MSFT/ HPE/ DELL EMC/ Google targeting Datacenter NVMe SSD V2.5 release

OPEN Compute Project
Datacenter NVMe® SSD Specification
Version 2.0 07/30/2021

Link to specification: https://www.opencompute.org/documents/datacenter-nvme-ssd-specification-v2-0r21-pdf

Result:

- **More features, Better quality, and Faster**
- **OCP Datacenter NVMe Specification is an industry collaboration win**

Write Amplification Overview

- What is Write Amplification (WA)?
  - When the host sends write data to the device it is additional data that is written to the media.
  - Write Amplification Factor (WAF) = media written data/ host written data

- WAF = 2.5 Example
  - Host writes 1 MB
  - Device writes 2.5 MB to the media
  - Thus Device
    - Media Writes
      - 1 MB Host Data
      - <u>*Additional 1.5 MB Garbage Collected Data*</u>
    - Extra Media reads to enable host write
      - <u>*1.5+ MB*</u>

Why is Write Amplification Undesirable?

- Write Amplification results in additional:
  - Media Reads/ Writes affecting performance/ QOS
  - Flash media writes causing non-host induced media wear
  - Additional power needed to perform the additional reads/writes
- Random Write example:

Write Amplification Factor	Write Performance
1	Random Write = Sequential Write
5	Random Write = 20% Sequential Write

Write Amplification Improvements History

Write Amplification Improvement Timeline:

~1991

NAND Based SSDs
Solution #1: Overprovisioning

~2007/2008

Host provides SSD LBA Hints
Solution#2 TRIM/Deallocate

2022

Host provided data placement hints
Solution #3 Flexible Data Placement
(TP4146)

- How did Flexible Data Placement come about?
  - Google Write Amplification Investigation Result
    - Data placement on media is key
    - SMART FTL Proposal
  - Meta Write Amplification Investigation Result
    - Data placement on media is key
    - Direct Placement Mode Proposal
  - Google & Meta merged their independent learnings into Flexible Data Placement (FDP) merging the best features of each proposal to enable best industry solution
- What is the status of this in NVM Express?
  - TP4146 is fully ratified
  - Link: https://nvmexpress.org/wp-content/uploads/NVM-Express-2.0-Ratified-TPs_12122022.zip

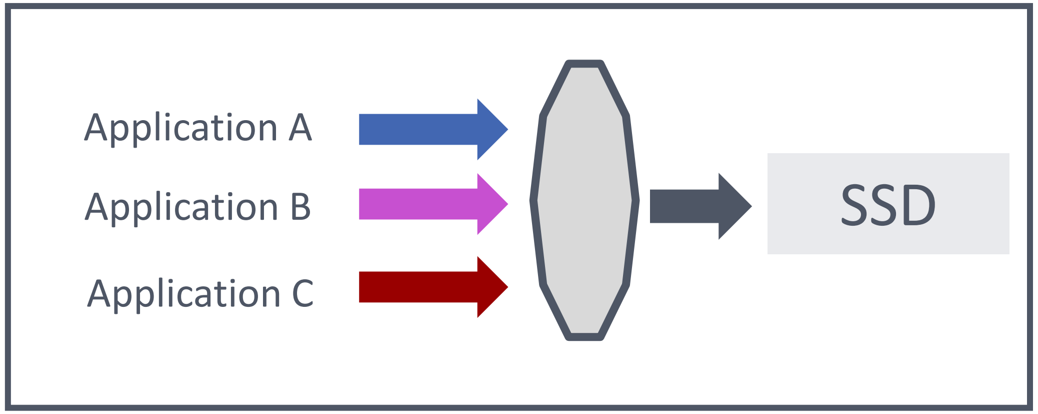
Flexible Data Placement (FDP) Overview

- Enables host to provide hint where to place data
  - Virtual handle/pointer
- Device changes:
  - Places data in super block based on a host hint rather than choosing it's own super block.
  - Advertises size of super block
- What functionality does not change
  - Read
  - Write (Optional media placement hint added)
  - Deallocate/TRIM
  - Security
- Backwards compatibility
  - FDP may be enabled/disabled on standard devices
  - Applications are not required to understand FDP to benefit
  - Applications which understand FDP have increased benefits

FDP Use Case Example: Disaggregated Storage

- Multi-user/ Multi-workload/ Disaggregated Storage

- Today's Challenges
  - Application's Data is Mixed
  - Device performance is unstable
    - Never reaches "steady state" due to mixed workloads
  - Overprovisioning is increased until Write Amplification (WA) is low enough and performance appears stable
  - Workload changes causes process above to repeat



Flexible Data Placement (FDP) Use Case Example: Disaggregated Storage

Results:

Today’s Method:
All Media blocks must be garbage collected resulting in a WAF ~3.

FDP Method:
Only single media block erased resulting in WAF = ~1

FDP Open-Source Activities

- **Goal: Support FDP through a full upstream I/O Path**
- Current Support:
  - **Linux Kernel**: Full support through 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. No simulation (e.g., WAF, performance)
  - **Fio**: Basic support for RU and RUH (Upstream)
    - Working on improving generic trim in io_uring (Ongoing)
  - **nvme-cli**: Support for FDP commands and log pages (Upstream)
  - **Cachelib**: Ongoing effort to reduce WAF through FDP (Ongoing)

Resources

- OCP Storage Project Link: https://www.opencompute.org/projects/storage
    - Meeting calendar with dial in information
- OCP Contribution database:

  https://www.opencompute.org/contributions
- OCP Referenced Contributions:
    - Yosemite V3: E1.S Faceplate https://www.opencompute.org/documents/e1s-faceplate-reference-design-specification-pdf
    - Yosemite V3: Vernal Falls E1.S 1OU Flash Blade and Expansion Board: https://www.opencompute.org/documents/e1s-expansion-1ou-1s-server-design-specification-pdf
    - Yosemite V3: Sierra Point E1.S 2OU Flash Bald and Expansion Board: https://www.opencompute.org/documents/e1s-expansion-2ou-1s-server-design-specification-pdf
    - Yosemite V3 Platform Design: https://www.opencompute.org/documents/ocp-yosemite-v3-platform-design-specification-1v16-pdf
    - Delta Lake 1S Server Design: https://www.opencompute.org/documents/delta-lake-1s-server-design-specification-1v05-pdf
    - Datacenter NVMe SSD Specification V2.0: https://www.opencompute.org/documents/datacenter-nvme-ssd-specification-v2-0r21-pdf

Thank You