A Solid Case for NVMe

Jason Burris – Sony Interactive Entertainment

Ceph Days Silicon Valley – March 25, 2025

A Quick Intro

Jason Burris

Staff SRE – Storage Product Owner

Sony Interactive Entertainment

What we do with Ceph

- Global Platform
  - Object Storage
  - Block Storage
  - Both HDD and NVMe

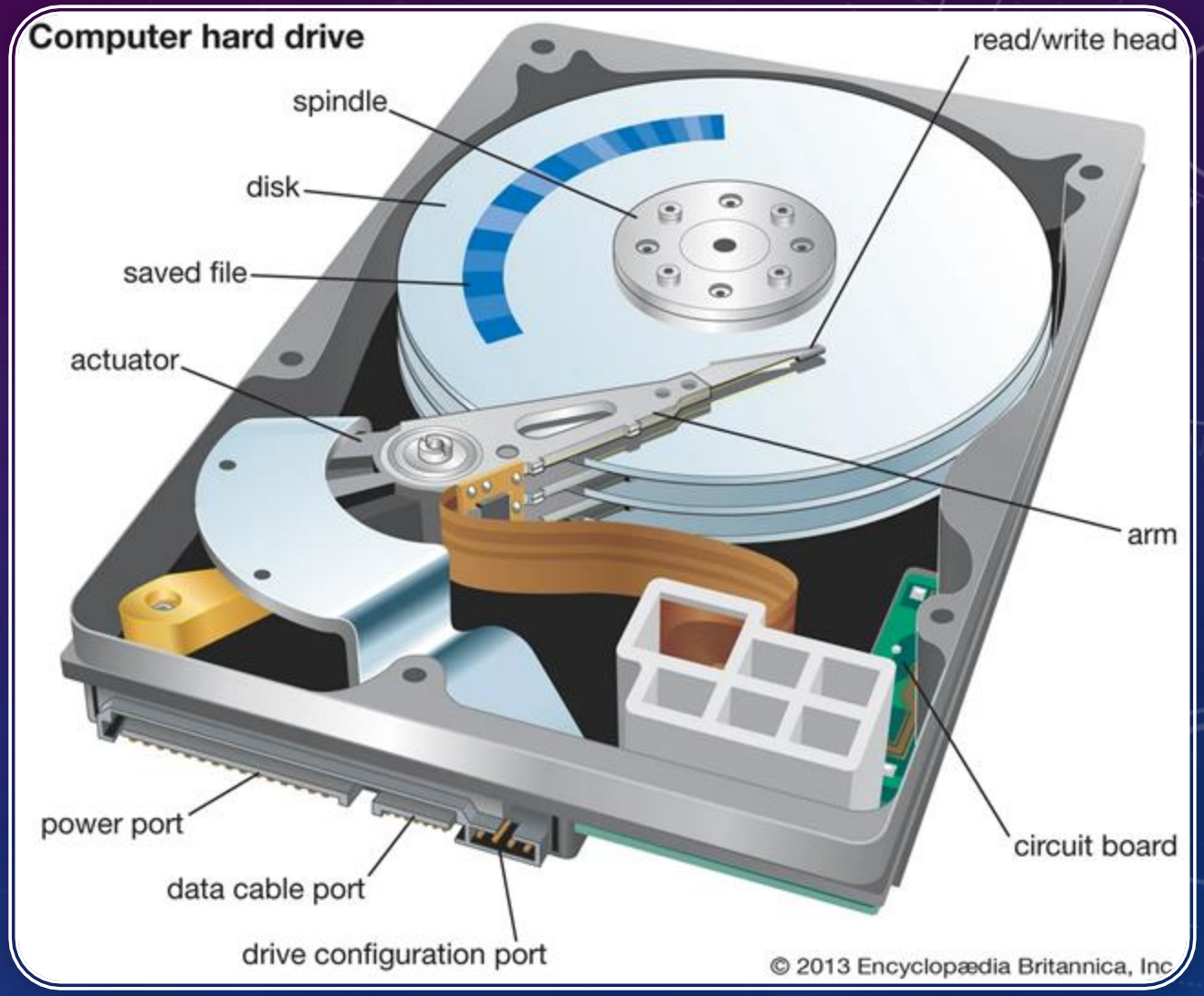
HDD vs NVMe

~~Why should we move from Hard Drives to Solid State?~~

When should we move from Hard Drives to Solid State?

Refresher on HDD

- 100-150 IOPS
- 2ms latency
- Max capacity 36TB
- Depends on mechanical reliability
- ”Spinning Rust”



What is NVMe?

- 400k-1m IOPS
- **80µs/15µs r/w latency**
- Max capacity 16TB/150TB
- Need to consider DWPD

U.3: 7mm and 15mm

E1.S: 5.9mm, 15mm and 25mm

M.2: 22x80mm and 22x110mm

NVMe Variants

QLC

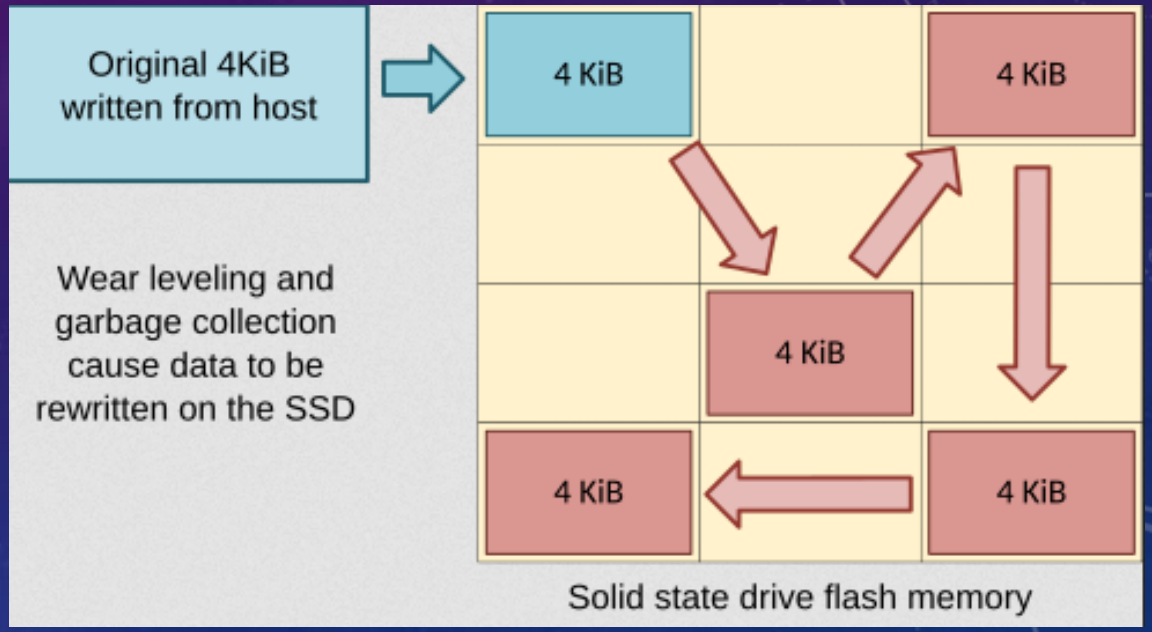
- 16 Charge levels
- Stores 4 bits per data cell
- Capacity 32-150 TB
- DWPD roughly half of TLC (0.5)

TLC

- 8 Charge levels
- Stores 3 bits per data cell
- Capacity 8-16 TB
- DWPD is higher (1.0)

DWPD and Write Amplification

- 1 DWPD = Writing the capacity of the drive
- Write Amplification is the additional data written due to GC outside of the host



Architectural Decisions

Getting to NVMe has been a journey…

Architectural Decisions

Our first generation

12 HDD + 4 SAS SSD

Using CDG groups 3 HDD per SSD

Architectural Decisions

Our second generation

24 HDD + 2 NVMe

Some using Raid1 some splitting the HDDs into two CDGs

Architectural Decisions

Our third generation

22 NVMe TLC

DB/WAL for each OSD stored directly

on the individual NVMe

Architectural Decisions

All 3 generations are still used today

Benefits of NVMe - Latency

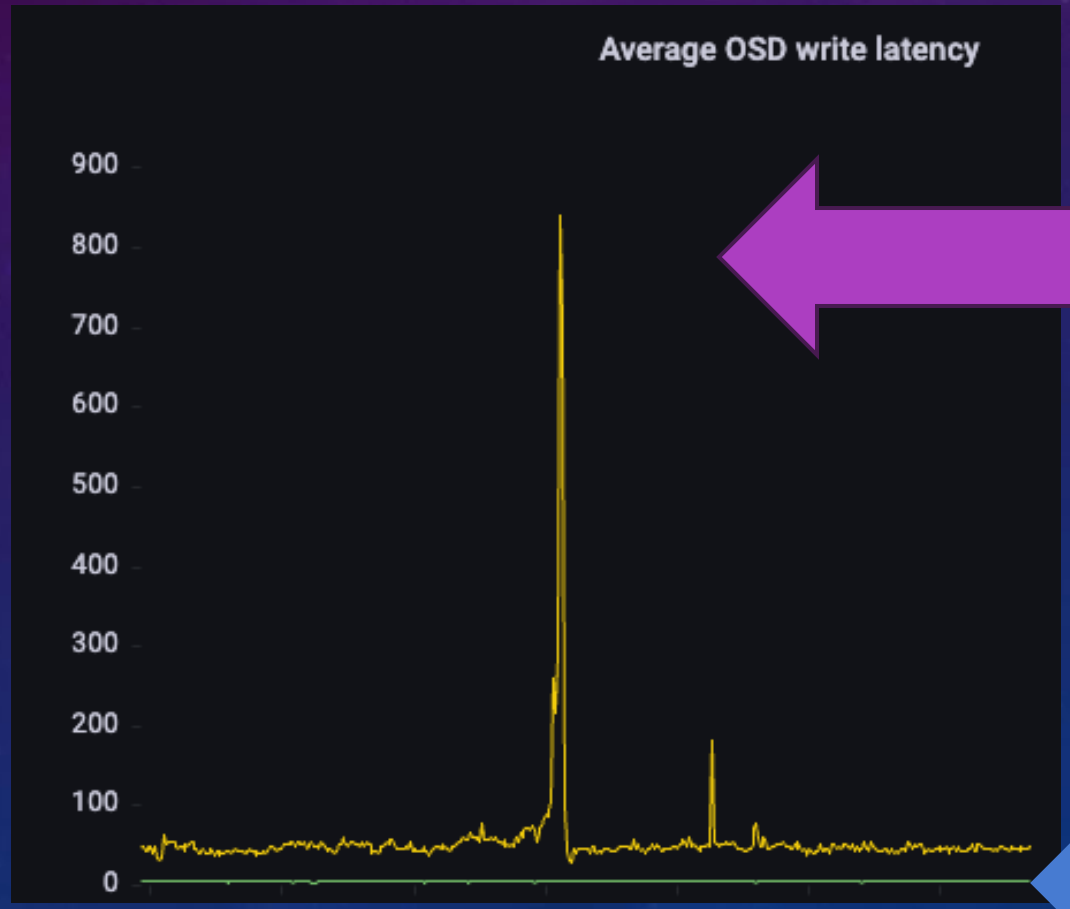
HDD

Avg 50-70 ms

NVMe

Avg 5 ms

Benefits of NVMe



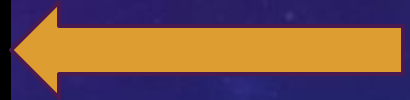
Errors on requests

NVMe latency

NVMe DWPD

8TB Drive

```
Smart Log for NVME device:nvme8n1 namespace-id:ffffffff
critical_warning                    : 0
temperature                         : 33 °C (306 K)
available_spare                     : 100%
available_spare_threshold           : 10%
percentage_used                     : 0%
endurance group critical warning summary: 0
Data Units Read                     : 303701025 (155.49 TB)
Data Units Written                  : 29041006 (14.87 TB)
host_read_commands                  : 1837313729
host_write_commands                 : 522383397
controller_busy_time                : 2316
power_cycles                        : 16
power_on_hours                      : 18727
unsafe_shutdowns                    : 13
media_errors                        : 0
num_err_log_entries                 : 0
Warning Temperature Time            : 0
Critical Composite Temperature Time     : 0
Temperature Sensor 1           : 33 °C (306 K)
Temperature Sensor 2           : 42 °C (315 K)
Thermal Management T1 Trans Count       : 0
Thermal Management T2 Trans Count       : 0
Thermal Management T1 Total Time        : 0
Thermal Management T2 Total Time        : 0
```



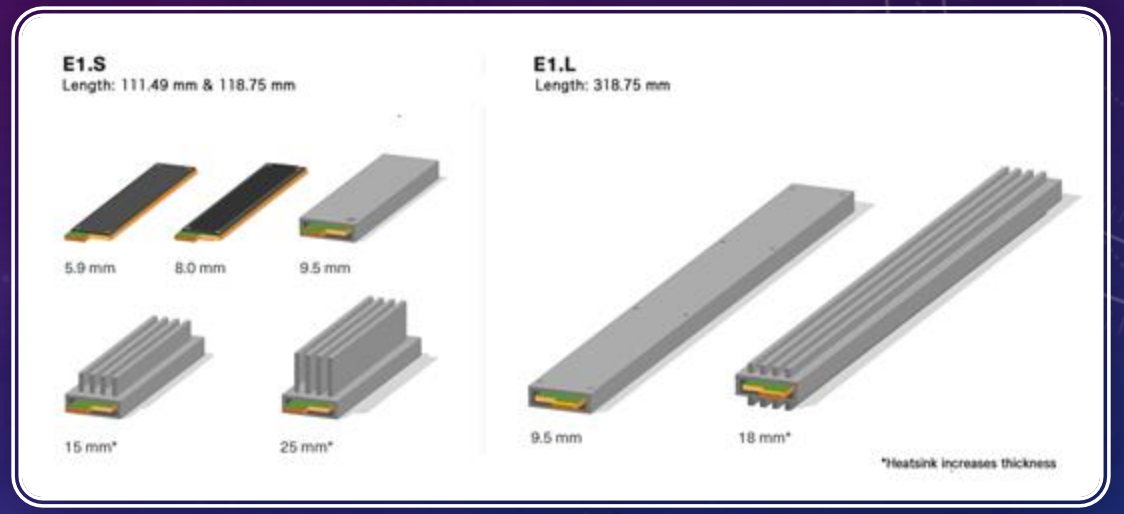
2x of the capacity

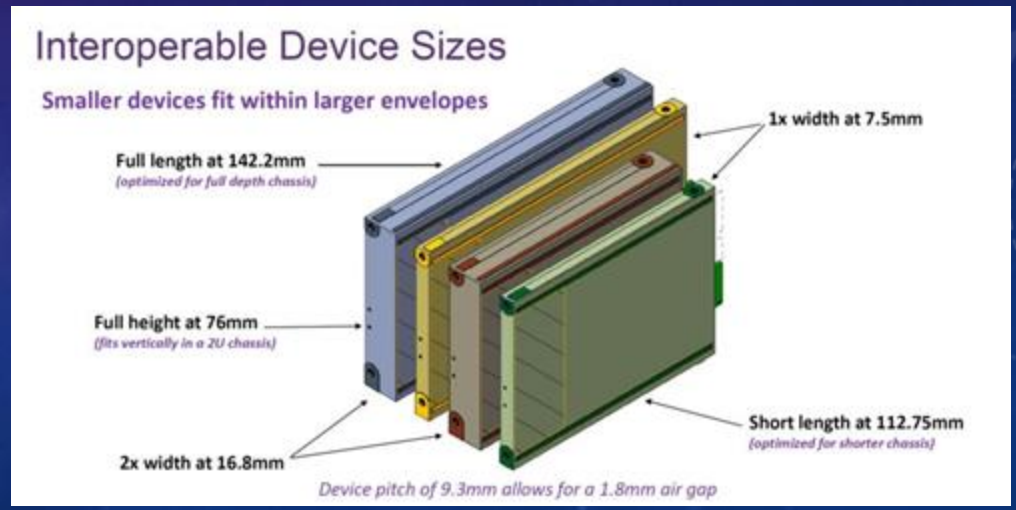
780 Days of use

15 TB / 780 Days = 0.02 DWPD

How drives are changing

- EDSFF Drive Format
  - These are emerging and the adoption rate is in flux
  - M.2 -> E1.S
  - U.2/3 -> E3.S





Looking to the future

- ZNS
  - Zoned Name Space for reducing write amplification
- Ceph Crimson
  - Rewrite of the OSD for highspeed drives
- Ceph Seastore
  - Store data more symmetrically

Things to consider on your journey

- How much overhead and impact are HDDs costing you?
- Do your workloads fit better on TLC or QLC?
- Could adopting NVMe improve your operational model?
- What is the ROI for making the switch to NVMe?
- Are you ready for the future?

Thank you!