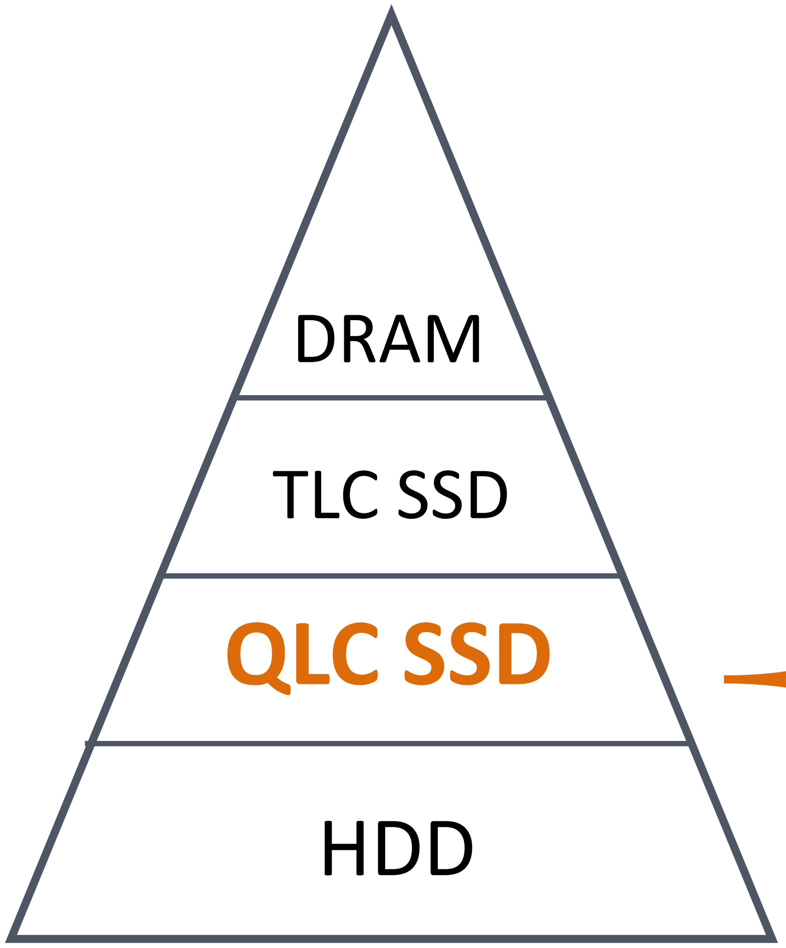
QLC in the Real World

Ross Stenfort, Meta

Agenda

- Why QLC?
- QLC Challenges
  - Performance/ Power
  - Form Factors
  - Management At Scale
- Challenges for the future

WHY QLC?



- Increases Storage Chassis Density
- Increases Device TB / W
- Improves Performance / TB Scaling
  - Power Based Performance Scaling

QLC Creates Storage Tier Between TLC SSD and HDD

Performance / Power

Performance and Power Guidance

What should QLC Performance & Power be?

Performance

Read Bandwidth (MB/s)/ Useable TB	Write Bandwidth (MB/s) / Useable TB
32	0
28.8	0.8
25.6	1.6
22.4	2.4
19.2	3.2
16	4.0
12.8	4.8
9.6	5.6
6.4	6.4
3.2	7.2
0	8

Read:
32 (MB/s) / Useable TB

Write:
8 (MB/s) / Usable TB

Power

128 TB: 20W

256 TB: 30W

Form Factors

Background (1 of 2): **EDSFF Comparison: E3, E1**

	E3.L 1T	E1.L 9.5	E3.S 1T	E1.S 9.5
Protocol	NVMe	NVMe	NVMe	NVMe
Transport	PCIe	PCIe	PCIe	PCIe
Connector	SFF-TA-1002	SFF-TA-1002	SFF-TA-1002	SFF-TA-1002
Pinout/electricals	SFF-TA-1009	SFF-TA-1009	SFF-TA-1009	SFF-TA-1009
Number of packages	32-48	32-48	16-32	8-16
Enclosure Length	142.2mm	318.75mm	112.75mm	118.75mm
Enclosure width	76mm	38.4mm	76mm	33.75mm
Enclosure thickness	7.5mm asymmetrical	9.5mm, symmetrical	7.5mm asymmetrical	9.5mm, symmetrical
Connector alignment	27.7mm from Datum	12.415mm from Datum	27.7mm from Datum	12.415mm from Datum
LEDs	Amber/Blue, Green Same PCB side	Amber, Green same PCB side	Amber/Blue, Green Same PCB side	Amber, Green opposite PCB side
Latch/Carrier mount	3 sides, 4 threaded holes	ledge, 2 thru holes	3 sides, 4 threaded holes	ledge, 2 thru holes
EMI/ESD contact point	Side contact pads, mounting holes	Bottom Strike pad, latch area	Side contact pads, mounting holes	Bottom Strike pad, latch area

Unable to scale capacity

Asymmetrical case leads to poor thermals

Unable to fit compute behind storage in chassis

SSD carrier required

Background (2 of 2)

Future of Memory and Storage (FMS) 2024

What form factor is best for High Capacity?

Considerations:

- **Max Capacity Requirement**
  - Number of NAND package placements?
    - NAND Packaging
    - Standard or custom packages
    - 4 Tb die timelines
    - 16 or 32 dies per package
- **Performance Requirement**
  - Power?

Goal:

Industry Unified Form Factor for 128 TB and above.

Concern:

Fragmentation is bad for industry.

Call to Action:

Industry discussion needed to avoid fragmentation.

EDSFF Comparison: E2, E3, E1

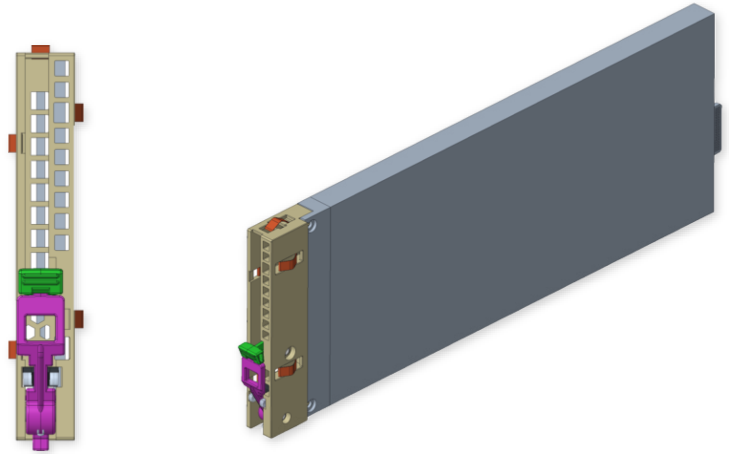
	E2	E3.L 1T	E1.L 9.5	E3.S 1T	E1.S 9.5
Protocol	NVMe	NVMe	NVMe	NVMe	NVMe
Transport	PCIe	PCIe	PCIe	PCIe	PCIe
Connector	SFF-TA-1002	SFF-TA-1002	SFF-TA-1002	SFF-TA-1002	SFF-TA-1002
Pinout/electricals	SFF-TA-1009	SFF-TA-1009	SFF-TA-1009	SFF-TA-1009	SFF-TA-1009
Number of packages	64+	32-48	32-48	16-32	8-16
Enclosure Length	200mm	142.2mm	318.75mm	112.75mm	118.75mm
Enclosure width	76mm	76mm	38.4mm	76mm	33.75mm
Enclosure thickness	9.5mm, symmetrical	7.5mm asymmetrical	9.5mm, symmetrical	7.5mm asymmetrical	9.5mm, symmetrical
Connector alignment	27.7mm from Datum	27.7mm from Datum	12.415mm from Datum	27.7mm from Datum	12.415mm from Datum
LEDs	Amber, Green opposite PCB side	Amber/Blue, Green Same PCB side	Amber, Green same PCB side	Amber/Blue, Green Same PCB side	Amber, Green opposite PCB side
Latch/Carrier mount	ledge, 2 thru holes	3 sides, 4 threaded holes	ledge, 2 thru holes	3 sides, 4 threaded holes	ledge, 2 thru holes
EMI/ESD contact point	Bottom Strike pad, latch area	Side contact pads, mounting holes	Bottom Strike pad, latch area	Side contact pads, mounting holes	Bottom Strike pad, latch area

E2 merges the best attributes and learnings from E3 and E1 to enable a scalable high capacity QLC form factor.

High Capacity Form Factor
Industry Collaboration Result: E2

E2 enables efficient high-capacity flash storage:

Leverages EDSFF including the learnings and best attributes from E1 and E3



Overview

Nand Placements

64+

EDSFF Connector Scaling

· x4 PCI Gen 6 and beyond
· ~80W

Resource Efficient

· Single PCB
· Thermally optimized enclosure

Industry Standard

· SFF-TA-1042 V1.0

Enables

Device Capacity Scaling

· 1 PB and beyond

Dense Chassis

· 40 Devices

Thermal

· Cooling with low airflow above 25W

Performance

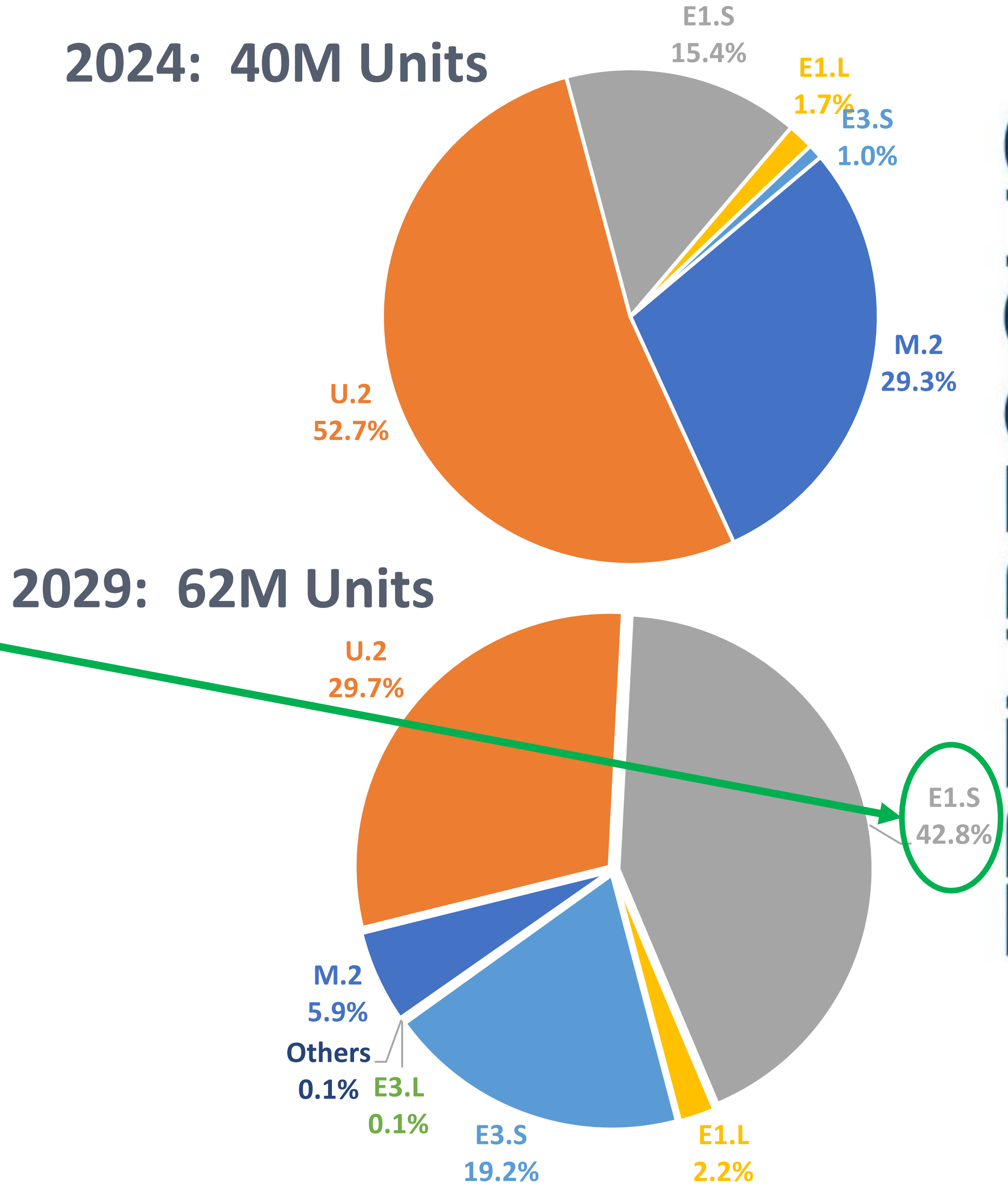
· Scales with power

Serviceability

Form Factor Overview and Market Summary

- E1.S
  - High Density Compute Storage
  - Market Growth Trend
    - Projected Volume 2029
- E2
  - High-Capacity Form Factor
    - High-capacity form factor of the future

E1.S and E2 Compliment each other for Compute and Storage



Management @Scale

OCP Datacenter NVMe® SSD Specification Update

2020

OCP NVMe Cloud SSD Spec v1.0

- Meta
- MSFT

2021

OCP Datacenter NVMe SSD V2.0

- Meta
- MSFT
- Dell
- HPE

2023

OCP Datacenter NVMe SSD V2.5

- Meta
- MSFT
- Dell
- HPE
- Google

2024

OCP Datacenter NVMe SSD V2.6

- Meta
- MSFT
- Dell
- HPE
- Google

2025

Coming Soon

OCP Datacenter NVMe SSD V2.7

- Meta
- MSFT
- Dell
- HPE
- Google

Improvements in V2.7

- **High-Capacity QLC SSD Improvements**
- **Device Self-Measured Power**
- **Additional Improvements to Telemetry / Debug**

OCP Datacenter NVMe SSD Enables:

- More Features, Better Quality and Faster
- Open-Source: OCP NVMe-CLI & Test Cases

Key Features for Managing at Scale (1 of 2)

Improvements Based On Deployment @Scale

- **OCP Health Information Extended Log**
  - Telemetry Metrics based on deployments at scale
- **OCP Latency Monitoring Feature**
  - Isolates, monitors, debugs latency spikes at scale
- **OCP Formatted Telemetry for Human Readable Logs**
  - Customer usable telemetry with improved security
- **Open-Source Tooling: OCP NVMe CLI**
  - Open-Source Tooling
- **DSSD Power State Support**
  - Simplifies power state control

Key Features for Managing at Scale(2 of 2)

- **Hardware Component Log**
  - Hardware manufacturing information is available to customer
- **Device Self Test Improvements**
  - Failing Segment codes is universal rather than supplier/product dependent
- **Device Self-Reported Power**
  - Device Power measurements are simplified in qualification and at scale

Additional QLC Improvements in OCP Datacenter NVMe Spec V2.7

- **New QLC NAND statistics**
  - Total dies on device
  - Bad dies on device
  - Dies taken offline due to predicting die will go bad
  - Bad blocks based on “useable” dies not total dies

- **Capacity Points**
  - Question
    - How much over-provisioning?
    - What should the user capacity be?
  - Answer
    - 128TB raw reports a capacity of: 122.88 TB
    - 256TB raw reports a capacity of: 245.76 TB

FIO/ SPRANDOM

- Problem: 128 TB drive takes ~12 days to precondition
  - At 256 TB these numbers double
- Solution: FIO with Sprandom enables 1 pass (from fresh SSD) to precondition in less than 1 day
- Sprandom does a LBA overlap method of preconditioning in FIO
- Link to sprandom examples: https://github.com/axboe/fio/blob/master/examples/sprandom.fio

Future QLC Challenges as Chassis Capacity Increases

- ❖ E2 solves device capacity scaling
- ❖ Bottlenecks shift with capacity and time
- ❖ Future Challenges for scaling denser capacity
  - CPU Performance
  - NIC Bandwidth
  - TOR Bandwidth

Thank You

Questions?