



How Facebook and Microsoft Leverage NVMe™ Cloud Storage

Live Webcast
November 19, 2019
10:00 am PT

Today's Presenters



Ross Stenfort
Facebook



Lee Prewitt
Microsoft



J Metz
Cisco

SNIA-at-a-Glance



185
industry leading
organizations



2,000
active contributing
members



50,000
IT end users & storage
pros worldwide

Learn more: **snia.org/technical**



Technologies We Cover

- ✓ Ethernet
- ✓ iSCSI
- ✓ NVMe-oF
- ✓ InfiniBand
- ✓ Fibre Channel, FCoE
- ✓ Hyperconverged (HCI)
- ✓ Storage protocols (block, file, object)
- ✓ Virtualized storage
- ✓ Software-defined storage

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced in their entirety without modification
 - ◆ The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.



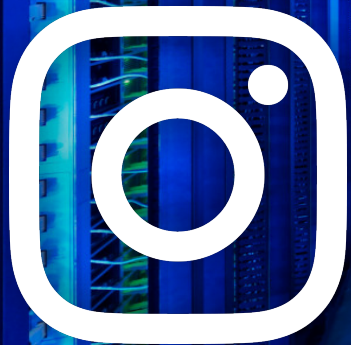
NVMe[™] In The Real World



facebook

**Facebook's mission is to give people the power to build
community and bring the world closer together.**

Facebook @ Scale



1 Billion



1.3 Billion



2.7 Billion

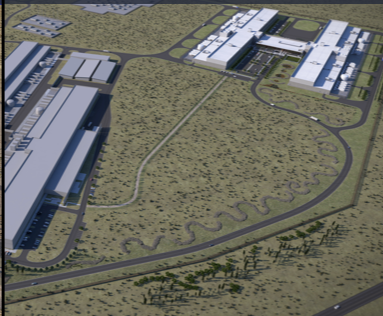
PAPILLION, NE



LOS LUNAS, NM



PRINEVILLE, OR



FOREST CITY, NC



FORT WORTH, TX



NEWTON COUNTY, GA



NEW ALBANY, OH



ODENSE, DENMARK



LULEÅ, SWEDEN



ALTOONA, IA



CLONEE, IRELAND



HENRICO, VA



EAGLE MOUNTAIN, UT



HUNTSVILLE, AL



SINGAPORE



2 Million

miles
intra-datacenter fiber

72+

Tb per second
backbone

54

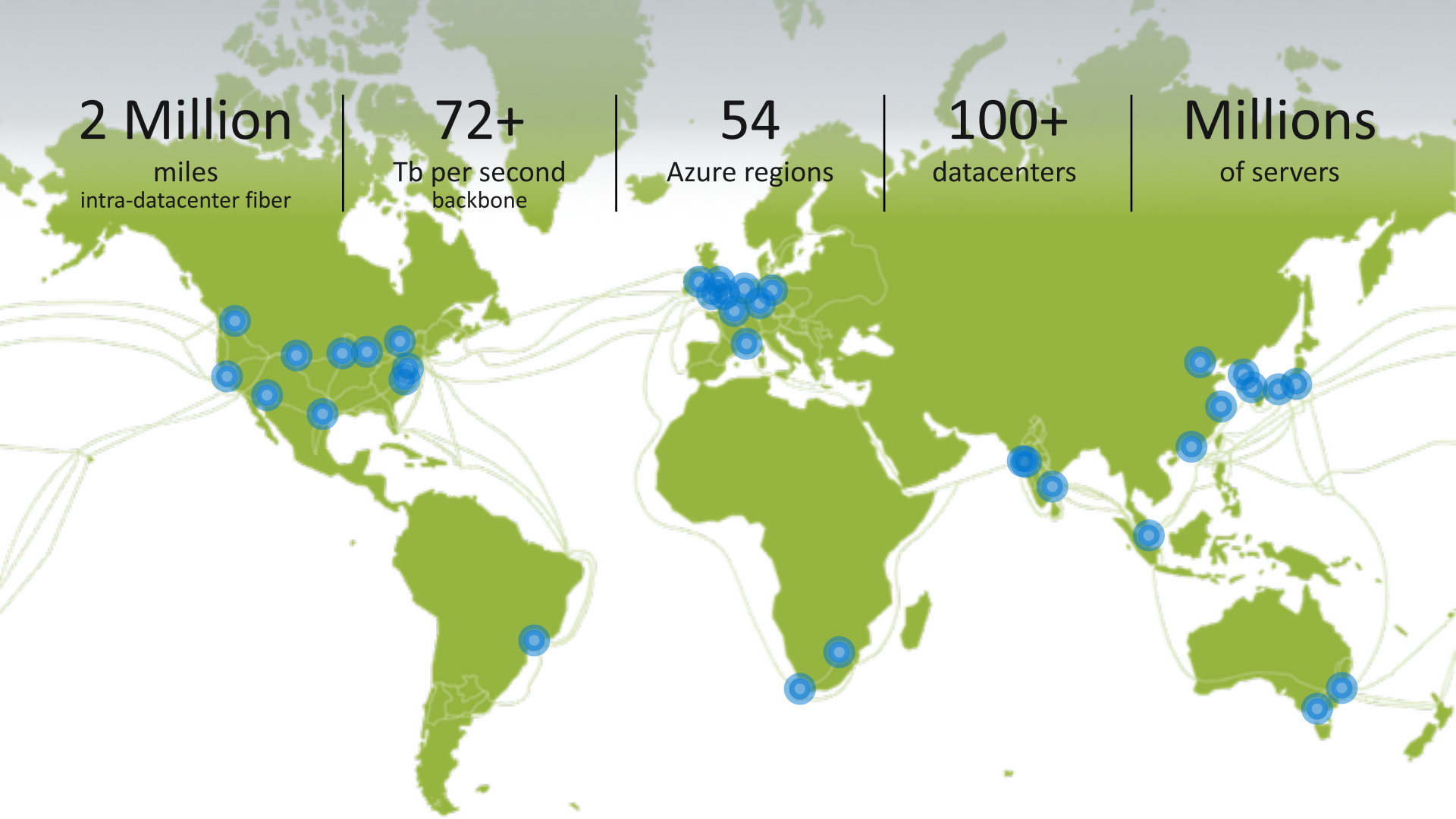
Azure regions

100+

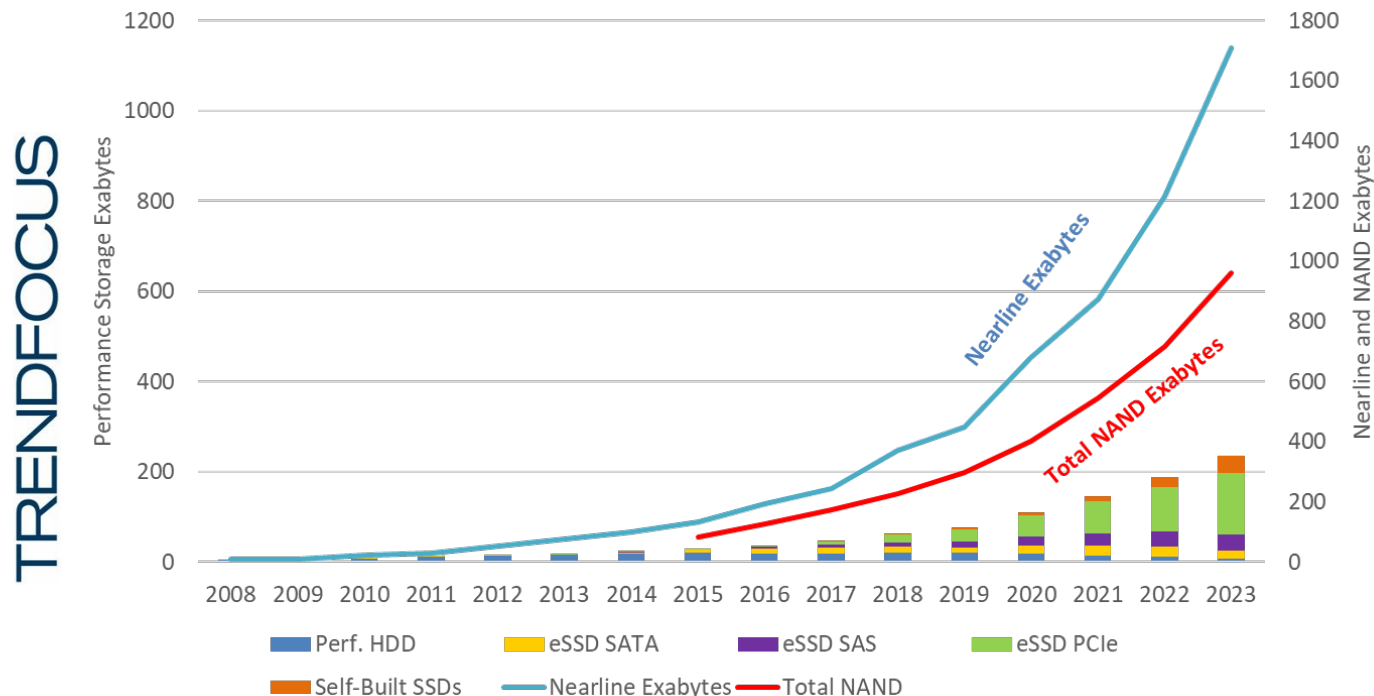
datacenters

Millions

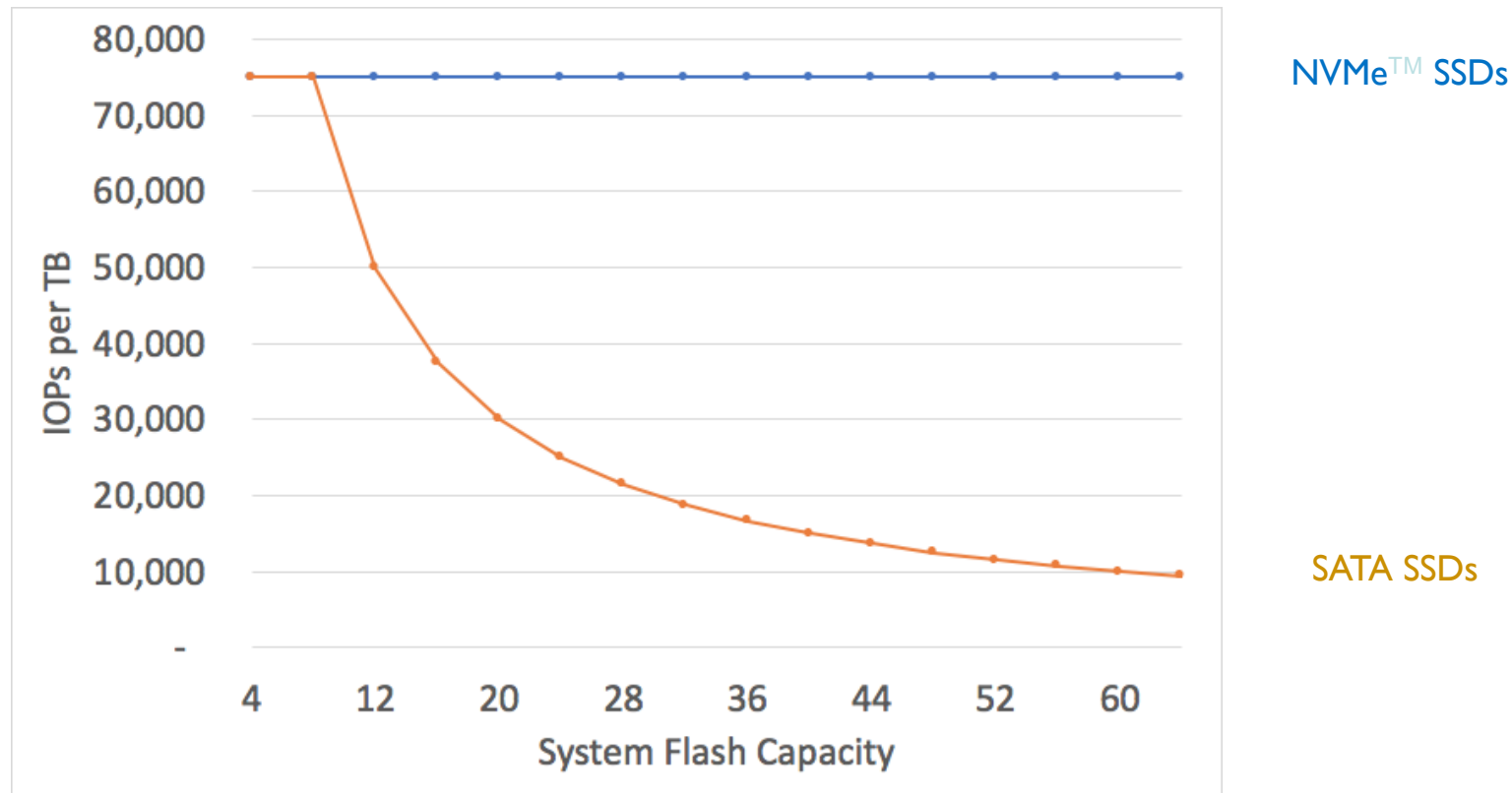
of servers



Why NVMe? - Exploding Storage Growth



Hyperscale Requires IOPS to Scale with Capacity



- Issues at Scale
 - Deallocate
 - Remote debugging
 - Security
 - “Rot in Place”
 - Form factors
- What’s New?



NVMe De-Allocate: Challenges and Improvements

➤ NVMe De-allocate

- ◆ Goal: It's a hint from the system to the SSD that the system is no longer tracking certain LBAs
- ◆ Good
 - Reduces Write Amplification
 - Improves performance/endurance
- ◆ Bad
 - Latency spikes due to De-allocate blocking Read/ Write

➤ Old Solution

- ◆ Tune De-Allocate size on a system
- ◆ Problem: The optimized de-allocate size varies based on supplier. Thus which supplier should I optimize for?

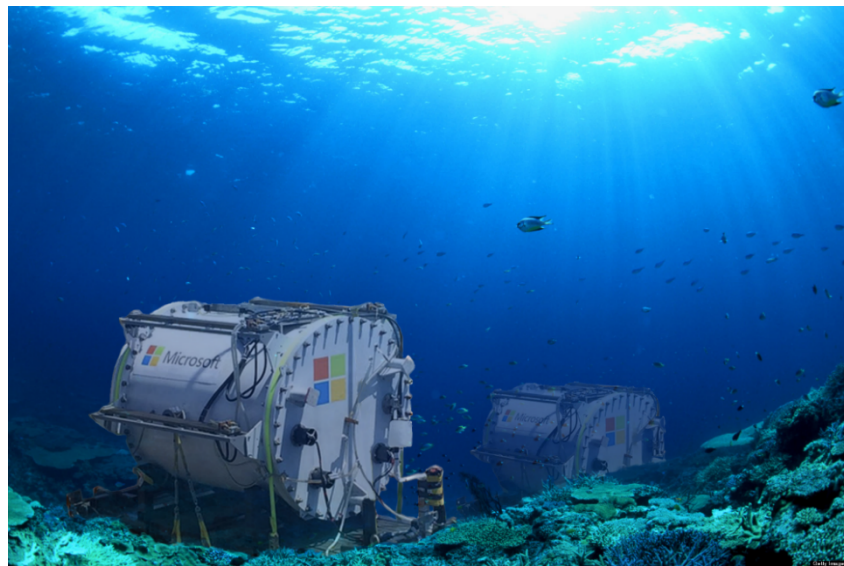
➤ Improved solution

- ◆ NVMe 1.4 allows the SSD to advertise it's preferred De-allocation size
 - If NSFEAT bit 4 = 0x1 then Namespace Preferred Deallocate Granularity (NPDG) is valid
- ◆ This allows systems to be optimized standard mechanisms.

- Challenge: Hyperscale Requires Debug with no physical access to the SSD.
- Challenge#1: Restricted access for vendor unique tools
- Solution:
 - ◆ NVMeTM CLI – Open source with active industry contribution and updates
 - › <https://github.com/linux-nvme/nvme-cli>
 - ◆ Vendor-unique CLI plugin that pulls and reports the logs in a common format
- Challenge#2: How do I get the debug information needed to resolve the issue
- Solution: Telemetry
 - › This allows SSD providers to get remote debug information to resolve issues
 - › Different data areas allows for different levels of debugging

Issues at Scale – Need for Remote Debugging

- **Timestamp**
 - ◆ Drive events correlated to system (BIOS and OS) events
- **Telemetry**
 - ◆ Host initiated - IO failures
 - ◆ Drive Initiated - Firmware panic?
- **SMART**
 - ◆ Both standard and vendor unique collected once an hour
 - Hey SSD IHVs. How many terabytes would you like to see?
- **Caveat: Any data that leaves the datacenter must be in human readable form!**



Managing at Scale

- **Background:** The amount of data written to a SSD may exceed the endurance of the SSD given the expected lifetime of the SSD. Given a fixed amount of write bandwidth a low the capacity SSD will wear out faster than a higher capacity SSD. Examples of applications where this can ocure are logging and caching.
- **Challenge/ Real World Example:**
 - ◆ Application only needs 256 GB but will use all the SSD capacity
 - ◆ Application write rate is high enough that it will wear out the 256 GB SSD
 - ◆ Application write rate scales per TB: Thus increasing capacity will not keep the SSD from wearing out
- **Solution: Namespace Management**
 - ◆ Allows a 512 GB SSD to be configured as a 256GB SSD with double the endurance of a 256 GB SSD
 - ◆ Thus the application view is a 256GB with double the endurance

- Challenge: How many blocks in my SSD have data and how many do not? If I de-allocate some blocks how many blocks really contain data? What is the effective over provisioning from a performance perspective?
- Solution:
 - › Namespace Utilization (NUSE)
 - › Allows user to determine the number of LBAs that actually contain data.

➤ Security challenges are growing

- ◆ NVM Express supports SECURITY_SEND/ RECEIVE will allow for security protocols to be tunneled into NVM Express
- ◆ There is even an open source tool for NVMe™ Opal security:
 - <https://github.com/Drive-Trust-Alliance/sedutil>
- ◆ Secure Boot is also a common security requirement. This is a process that ensures the firmware running on the device is from the manufacturer and not some other source.

➤ Problem/ Industry call to action:

- ◆ There is no standard way to know if secure boot failed
- ◆ If firmware on a device is compromised, how is this identified vs any other type of failure?



Issues at Scale – Need for Security



eDrive on Windows

- Opal v2 plus IEEE 1667 secure silo

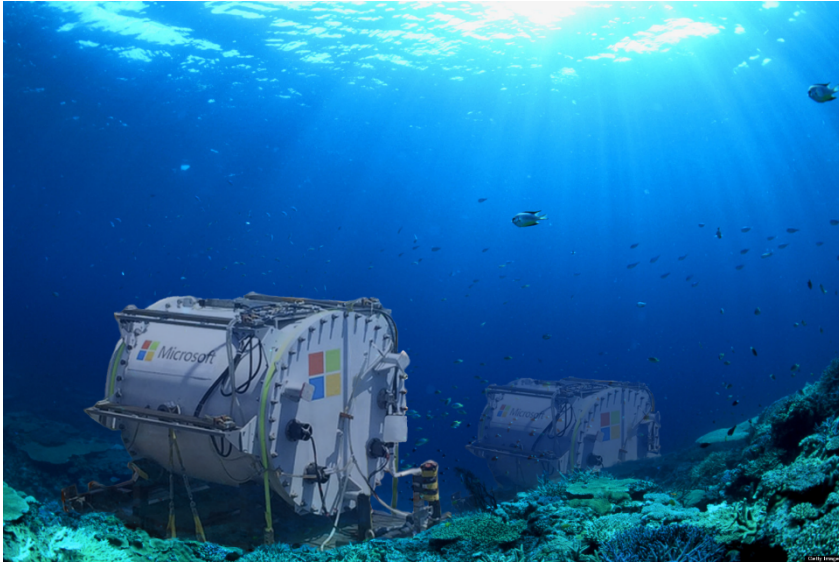
Hardware Root of Trust

- Secure boot
- Signed firmware
- Cerberus

Device Hardening

- Pen and Fuzz testing
- Locking of debug ports and vendor unique commands

Issues at Scale – Need to allow for “Rot in Place”



Use the Endurance and Performance metrics for auto tiering

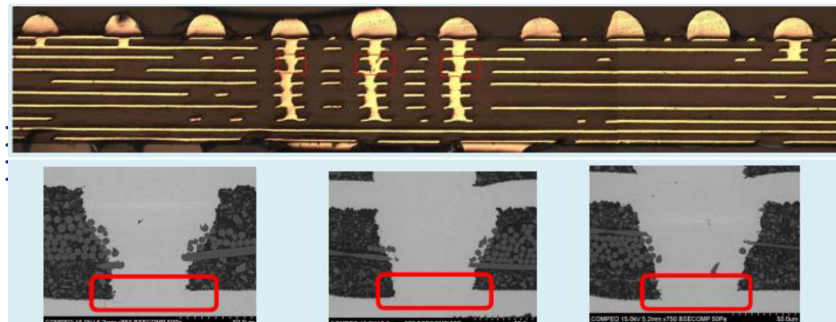
- Allows for fitting the workload to the device
- Allows for the ability to adjust the temperature of the data over time
- Allow for 5 to 7-year device service life

Zoned Name Spaces for QLC

- Reduce WAF due to large sequential writes
- Reduce DRAM due to large indirection unit
- Reduce overprovisioning due to minimal garbage collection

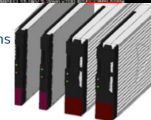
Issues at Scale – Form Factors

- m.2 has run its course
 - ◆ Power and thermal constraints
 - ◆ Fragile PCB and connector
 - ◆ Not hot-swappable
- E1.L and E1.S are here to replace it
 - ◆ Built from the ground up for datacenter use cases
- Good news is that they support NVMe too!



E3 (SFF-TA-1008)

- Ultra high-performance applications
- (104.9/142.2) x 78mm
- Supports up to 70W
- Up to 48 Standard NAND sites



- [Same](#) Connector, SFF-TA-1002
- [Same](#) Pinout and Functions
- [Different](#) Usages
- [Same](#) Expectations!

What's new in storage with Microsoft & Facebook?

Microsoft/Facebook are merging their SSD drive requirements into a single document

What is the Problem?

- Hyperscale providers have features that are needed to manage SSDs at scale but are not sharing these features with the rest of the industry
- Many features are common across providers. Each implemented slightly differently
- Cloud SSD consumers have confidential SSD specifications which doesn't encourage industry collaboration and discussion
- There is no public document on what a Cloud SSD should be
- SSD industry fragmentation due to lots of different skews that are "similar" but different
- 3rd Party compliance suites don't know what features cloud providers care about since they do not know what features Cloud consumers use.

- Microsoft/Facebook have merged their SSD drive requirements into a single document
 - **Microsoft/Facebook would like to contribute this document to OCP**
- Benefits:
 - Allows the market to understand what features Microsoft/Facebook need to manage a SSD at scale
 - Allows the market to understand and use the SSD's Microsoft/Facebook are using
 - Reduces SSD market fragmentation
 - Enables open source tools like NVMe CLI to manage the SSD
 - Allows 3rd parties to focus their test/validation efforts
- Summary
 - Benefits both system makers and SSD providers

Questions?

After This Webcast

- Please rate this webcast and provide us with feedback
- This webcast and a PDF of the slides will be posted to the SNIA Networking Storage Forum (NSF) website and available on-demand at www.snia.org/forums/nsf/knowledge/webcasts
- A full Q&A from this webcast, including answers to questions we couldn't get to today, will be posted to the SNIA-NSF blog: sniansfblog.org
- Follow us on Twitter @SNIANSF

Thank you