How Facebook and Microsoft Successfully Leverage NVMe™ Cloud Storage

Sponsored by NVM Express™ organization, the owner of NVMe™, NVMe-oF™ and NVMe-MI™ standards
Speakers

Ross Stenfort

Lee Prewitt

Facebook

Microsoft
NVMe™ In The Real World

Ross Stenfort, Hardware System Engineer
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
Hyperscale Requires IOPS to Scale with Capacity

![Graph showing the relationship between System Flash Capacity and IOPS per TB for NVMe™ SSDs and SATA SSDs.](image-url)
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 Perferred Deallocate Granularity (NPDG) is valid
  - This allows systems to be optimized standard mechanisms.
Managing at Scale (1 of 3)

- Challenge: Hyperscale Requires Debug with no physical access to the SSD.

- Challenge#1: Restricted access for vendor unique tools

  Solution:
  - NVMe™ CLI – Open source with active industry contribution and updates
    - [https://github.com/linux-nvme/nvme-cli](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
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 occur 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
Managing at Scale (3 of 3)

- 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.
Industry Challenge

➢ Security challenges are growing
  • NVM Express™ supports SECURITY_SEND/RECEIVE will allows 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 manufacture 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?
NVMe™ at Hyper-Scale

Lee Prewitt, Principle Hardware Program Manager
Azure CSI - Microsoft
• Azure at a Glance
• Why NVMe™?
• Issues at Scale
  – Form factors
  – Need to allow for “rot in place”
  – Need for remote debugging
  – Need for security
Why NVMe™? - Exploding Storage Growth

![Graph showing the growth of storage exabytes from 2008 to 2023. The graph illustrates the performance of different storage types including Perf. HDD, eSSD SATA, eSSD SAS, eSSD PCIe, Self-Built SSDs, Nearline Exabytes, and Total NAND exabytes. The trend shows a significant increase in storage capacity over the years, with a sharp rise in the last few years, indicating the expanding need for storage solutions.](image-url)
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!
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 - 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 over provisioning due to minimal garbage collection
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!
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
Questions?
nvm EXPRESS®
Architected for Performance