NVMe Client and Cloud Requirements, and Security

9:45 – 10:50

Features needed for SSD deployments at the client	Gwendal Grignou Lee Prewitt	Software Engineer, Google Principle Program Manager, Microsoft
Features needed for large scale SSD deployments	Lee Prewitt Monish Shah	Principle Program Manager, Microsoft Hardware Engineer, Google
Security Vision and Collaboration with TCG	Jeremy Werner Dave Landsman	VP SSD Marketing and Product Planning, Toshiba Director Standards Group, Western Digital

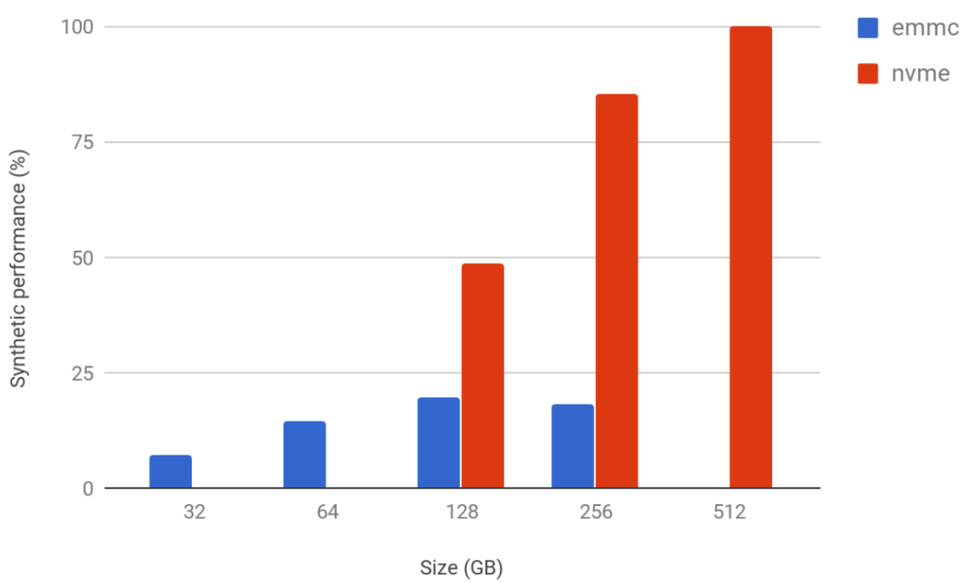
Features needed for SSD deployments at the client

Gwendal Grignou, Software Engineer, Google

Flash Memory Summit

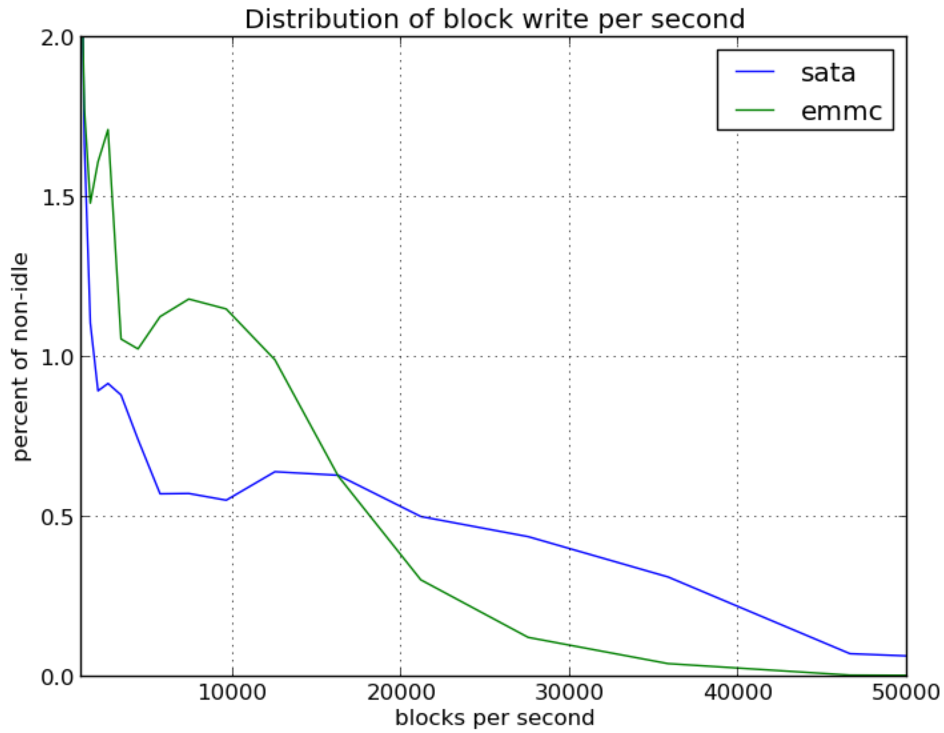
Case for NVMe on client

- Chromebook were not storage intensive
- Changing with Android application support (ARC++)
- Considering model with large storage capacity.



Case for NVMe on client

- Storage usage is spiky
- NVMe latencies will bring better customer experience

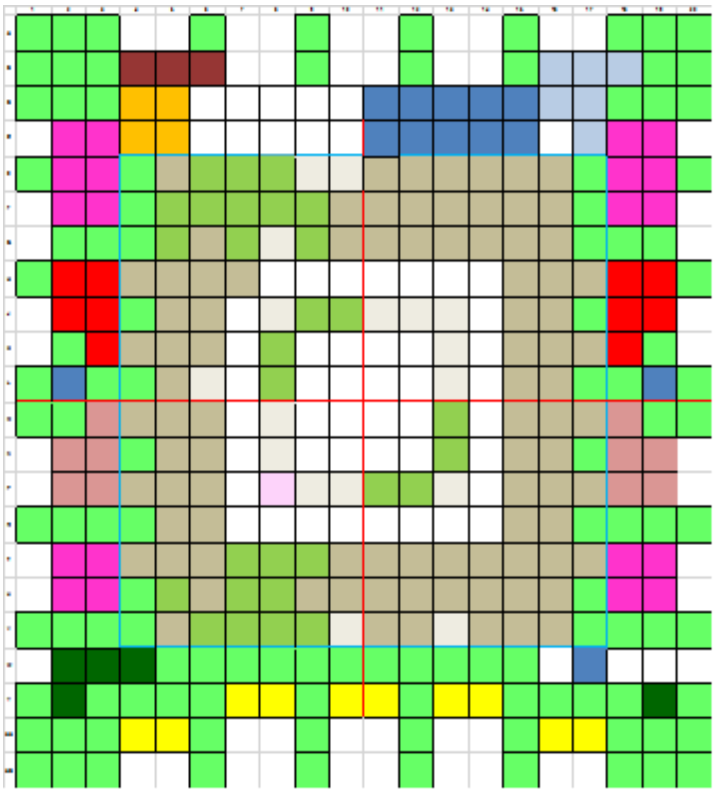


Packaging

- M.2 Connector viable only in some Chromebox
  - Z height
  - real estate on all chromebook motherboard
- 16x20 BGA still too big
- 11.5x13 BGA right format

Flash Memory Summit

11.5x13 BGA



- The same MLB can be stuffed with either NVMe or existing eMMC
- It provides a good transition from eMMC based storage to NVMe storage.
- eMMC and NVMe together provide good **performance, price, feature, and capacity** coverage to meet different customers' needs.

11.5x13 BGA

- 2 PCIe lanes
  - 2 x 8 GT/s: enough 512GB of regular flash
  - Revisit with NextGen Memory
- SPI allow stacking SPI NOR Flash

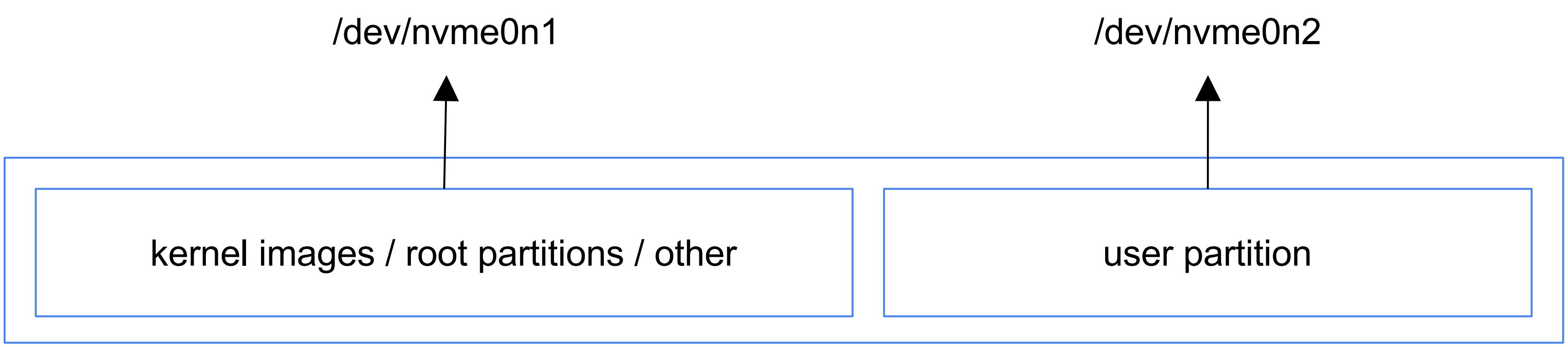
Cost Reduction

- Controller cost has to go down
- Host Memory Buffer (HMB)
  - Support in Linux kernel has been proposed
  - Other options

Flash Memory Summit

Security: Sanitize

- Sanitize improve security
- When transitioning to developer mode
  - crypto-erase if name space is supported
  - block erase otherwise



Conclusion

- NVMe devices in Chromebooks around the corner
  - New usage model requires better storage
- Controller fixed cost is the limitation
  - Coming on larger capacity first
  - Proposed as SATA SSD replacement
  - Replacing eMMC down to 64GB considered

Client & Mobile Needs for NVMe

Lee Prewitt

Principal Program Manager - SFS

Agenda

- Why is client different?
- What NVMe features are required?

Why is the client different?

Design Principles For Client Hardware

- Support a broad set of hardware in smaller and smaller form factors
  Thin and light laptops (2 in1), Phone, Tablet, others
- Reduce BOM cost through integration
  Multiple chips with different functions can be eliminated
- Harden device to security threats
  Malware attacks, DOS attacks, etc.
- Flexible tradeoff between power usage and performance
  Power constraints, thermal constraints, thermal events, race to sleep

What NVMe features are required?

NVMe Optional Features

- When are Optional features not Optional?
  - Required by Windows HLK
  - Needed for smooth interop with Windows

Flash Memory Summit

Client & Mobile Features

- Boot Partitions
- RPMB
- Name Spaces
- HMB
- Drive Telemetry
- Power Management
- Write Protect (targeted for v1.4)

Data Center Needs for NVMe

Lee Prewitt

Principal Program Manager - SFS

Laura Caulfield

Senior Software Engineer - CSI

Agenda

- Why is the Data Center different?
- What NVMe features are required?

Why is the Data Center different?

Design Principles For Cloud Hardware

- Support a broad set of applications on shared hardware
  Azure (>600 services), Bing, Exchange, O365, others

- Scale requires vendor neutrality & supply chain diversity
  Azure operates in 38 regions globally, more than any other cloud provider

- Rapid enablement of new NAND generations
  New NAND every n months, hours to precondition, hundreds of workloads

- Flexible enough for software to evolve faster than hardware
  SSDs rated for 3-5 years, heavy process for FW update, software updated daily

What NVMe features are required?

NVMe Optional Features

- When are Optional features not Optional?
  - Required by Data Center RFP
  - Needed to meet DC use cases

Data Center Features

- Streams
- Fast Fail
- I/O Determinism
- Drive Telemetry

Cloud Requirements for NVMe

Google perspective

Monish Shah

Typical Cloud Application

Application running on globally distributed server farms

Minimize Total Cost of Ownership (TCO)

Internet

Serving millions to 1B+ users

Minimize Latency

Opportunity #1: I/O Determinism

How I/O Determinism helps

- To optimize TCO, large SSDs are often shared between multiple applications
- I/O Determinism (IOD) helps control latency

Note: NVMe Technical Working Group is actively working on I/O Determinism Technical Proposal

Read / Write Interference

- Common element in all NVM technologies:

Example: NAND flash	
Read latency w/o blocking	~100 µs
Read latency w/ blocking behind a write	~2-5 ms

IOD helps address this:

IOD Concept: NVM Sets

- NVM Sets: Provides a mechanism to partition the SSD with performance isolation

Example table of NVM Set configs

Config #	NVM Sets provided
0	4 Sets of 1 TB each
1	8 Sets of 512 GB each
…	…..

All NVM Sets in a config need not be of the same size.

NVM Sets: Performance Isolation

- Degree of performance isolation is an implementation choice. Recommendation:
  - Do not allow writes in one NVM Set to block reads in another NVM Set
  - To the extent possible, avoid sharing of internal controller resources
  - PCIe BW is shared: no isolation is possible

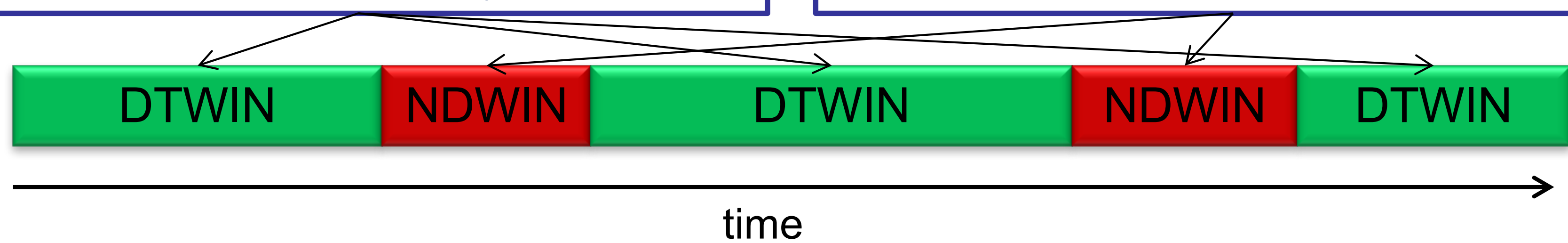
Flash Memory Summit

Predictable Latency Mode

Optional enhancement to NVM Sets, for more advanced applications

- Allows host to control read / write interference within a single NVM Set

DTWIN = Deterministic Window	NDWIN = Non-Deterministic Window
No writes to media. No GC or other maintenance. Minimal writes from host. Minimal tail latency on reads.	Writes allowed. Host can write; GC and other maintenance allowed. Return to DTWIN when possible.



Read Recovery Level (RRL)

- RRL: host can trade-off UBER for latency

Better UBER

RRL	O/M
Vendor specific	Optional
….	Optional
Default (normal recovery effort)	Mandatory
….	Optional
….	Optional
Fast Fail (minimal recovery effort)	Mandatory

Better Latency

Opportunity #2: Optimizing memory TCO

3D NVMs: Skirting End of Moore's Law

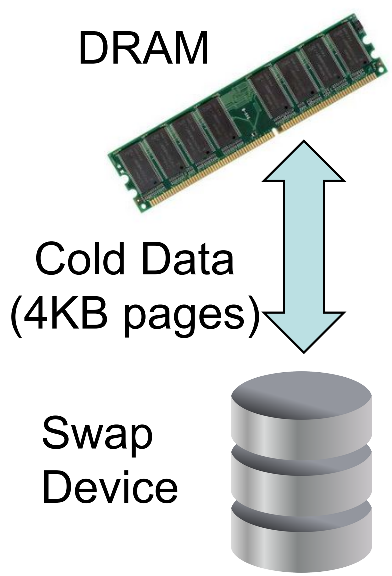
Semiconductor geometry scaling is reaching its limit.
However, impact on different technologies is variable.

DRAM	NAND and new NVMs
Planar technology, no prospect for monolithic 3D scaling	3D already proven for NAND, expected for PCM and ReRAM
Limited prospects for cost and capacity scaling	Reasonable prospects scaling in the foreseeable future

Idea: Use NVMs to supplement DRAM.
Caveat: DRAM will have the best performance. Use NVMs for "cold data".

Implementation: Reinvent Paging

Paging



Choosing swap media

Media	Latency
HDD	~10 ms
SSD	~100 μs
New NVM	~10 μs

Google Experimental Results

Promising results with 10 μs swap device: Negligible application performance hit when paging cold data.

Choice of media: PCM, ReRAM, Low Latency SLC NAND

NVMe optimization: Use Controller Memory Buffers (CMB)

Summary

Opportunities for next gen NVMe SSDs:

1. I/O Determinism:
   - Control latency @ constant TCO
2. Re-inventing paging
   - Reduce DRAM TCO @ constant performance

NVMe,TCG, and security solutions for the NVMe ecosystem

Flash Memory Summit 2017

Dave Landsman – Western Digital
Jeremy Werner – Toshiba
David Black – Dell EMC

Agenda for today

- NVMe and TCG are working together on NVMe security
  - New features developed in Opal family
  - Discussing enabling enterprise capabilities using same core spec as Opal family
- What threats are we trying to address? (and not)
- What's being implemented in NVMe and TCG specs?
- What's next?

Flash Memory Summit

NVMe device ecosystem has become broad

And…

- NVMe-oF – Non-PCIe fabrics
- NVMe-MI – Out-of-Band Management

NVMe
Datacenter to Mobile

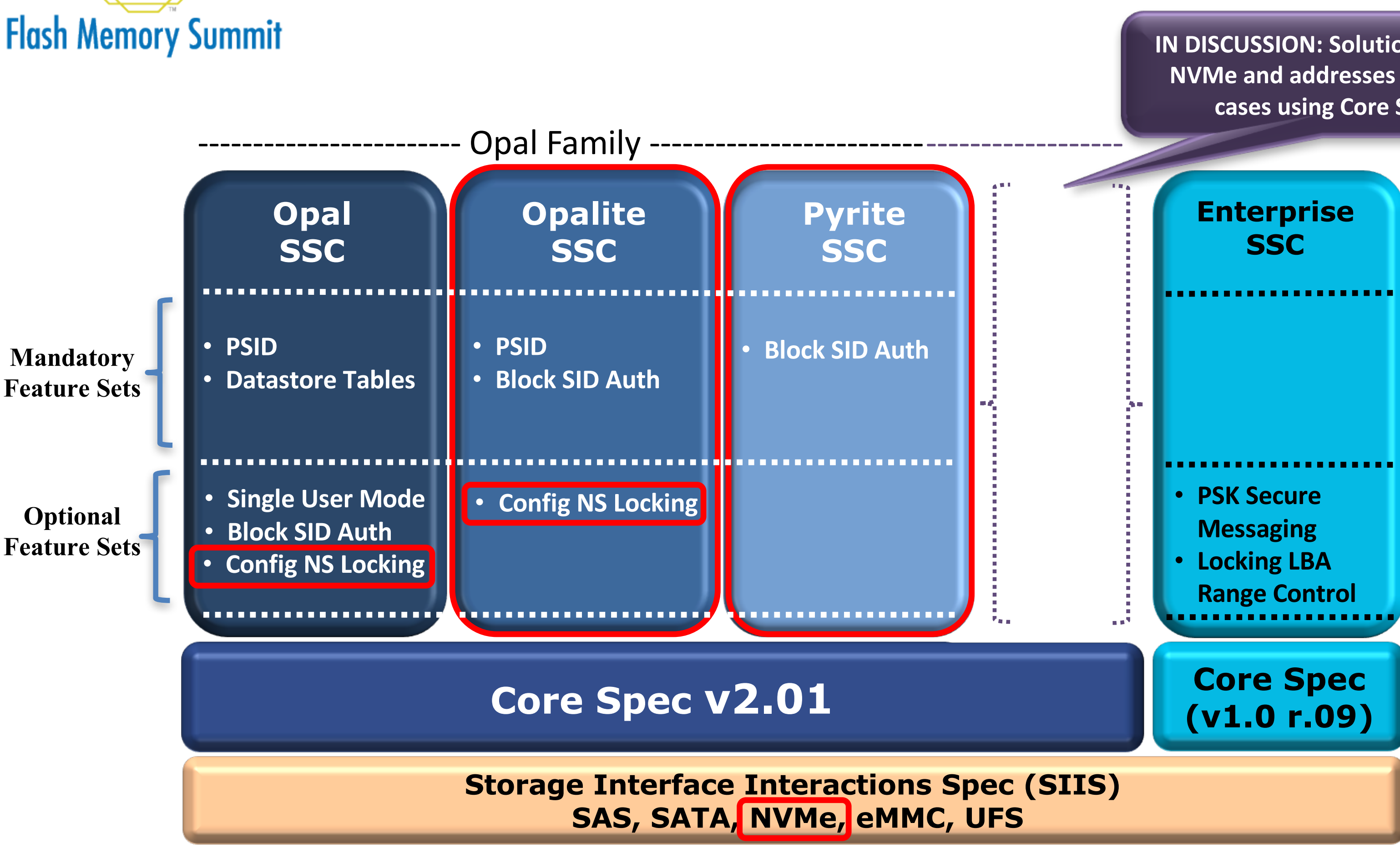
Broad ecosystem means security considerations

Flash Memory Summit

Threat Classes	Mitigation Strategies	
• Physical Device Access – Device Lost/Stolen – Repurposing a device	• Sanitize NVM Express, Trusted Computing Group • Data-at-Rest Encryption Trusted Computing Group • Device Locking Trusted Computing Group	To left + • Authentication/ Access Control NVM Express, Trusted Computing Group
• Data Access – Theft or unauthorized disclosure of data – Malicious or criminal change or destruction of data	Above + • Media Write Protection NVM Express, Trusted Computing Group • Data-in-Flight encryption • E2E Cryptographic Integrity Checks	

Flash Memory Summit

TCG/NVMe Work To Date



Opal/Enterprise SSCs

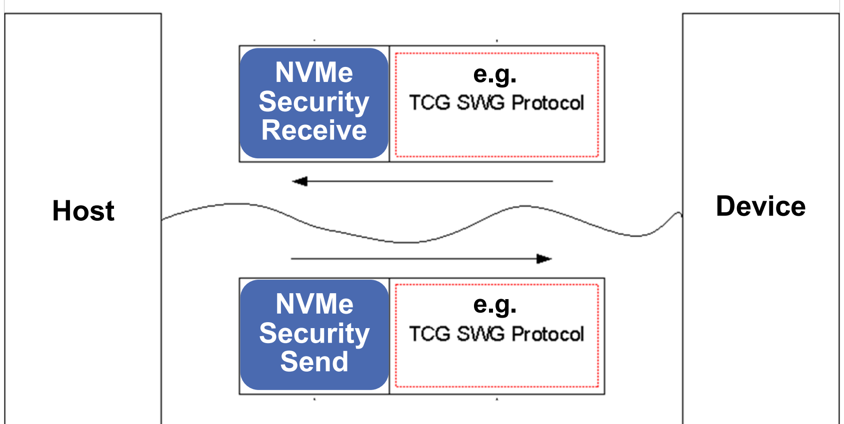
- Data-at-rest protection
- Device Locking
- Cryptographic Sanitization

NVMe/TCG Work

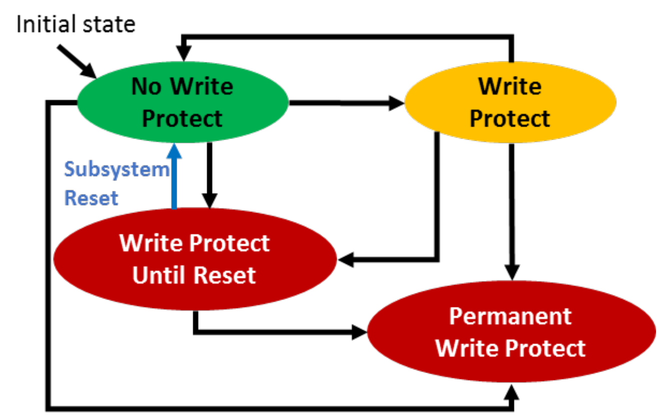
- SIIS binding for NVMe
- Opalite
- Pyrite (device lock only)
- Configurable Namespace Locking

NVMe Native Security Features

1) Security Send-Receive – Protocol Tunneling



2) Namespace Write Protect



3) Sanitize Device

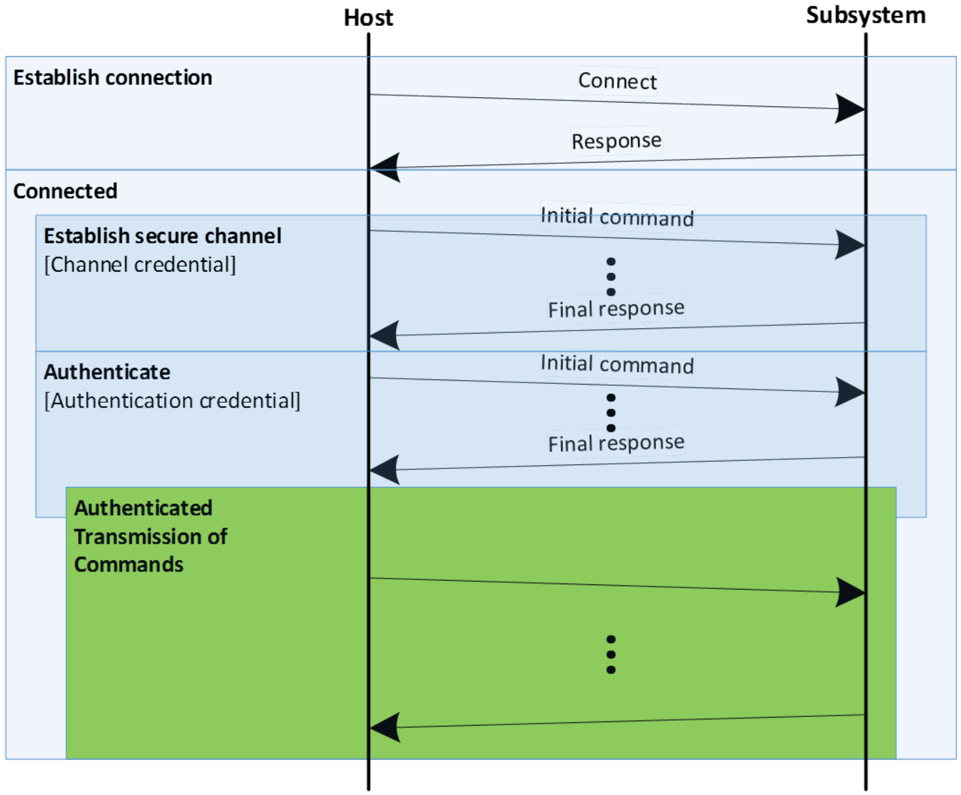
- Make all user data inaccessible through the interface
- Crypto Erase, Block Erase and Overwrite methods supported

4) RPMB (Authentication)

- Private area in non-user part of device, allowing only authenticated access
- Used for Boot and Namespace Write Protection; could be used for other use cases

Other Being Discussed

NVMe-oF In-band Authentication



Summary

Flash Memory Summit

- NVMe and TCG have worked together on baseline security features
  - Data-at-Rest Encryption
  - Device Locking
  - Sanitize
  - Media Write Protection
  - Authentication features
- NVMe and TCG continue working together
  - Data-at-Rest solution for NVMe that addresses Enterprise SSC use cases using Core Spec v2.01
  - NVMe-oF Session Authentication

Get involved at NVMe and TCG

THANK YOU

nvm EXPRESS®

Architected for Performance

Bios TBA

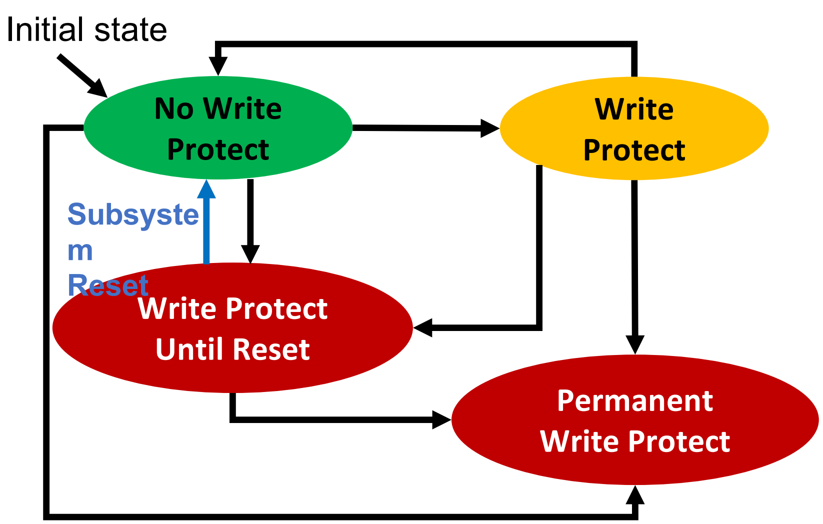
BACKUP

NVMe and TCG Media Write Protection

Flash Memory Summit

NVMe Namespace Write Protect

- Set Features command sets Write Protect states
- State applies to whole namespace
- Use of Write Protect Until Reset and Permanent Write Protect may enabled/disabled using RPMB



TCG Configurable Namespace Locking

- By default, all namespaces controlled by Global Locking Object
- Namespace may be given a separate Locking Object
- Range of LBAs may be given a separate Locking Object
- Authentication handled by Opal xyz