## 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 |
| Features needed for large scale SSD deployments | Lee Prewitt | Monish Shah | Principle Program Manager, Microsoft |
| 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
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
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
Security: Sanitize

- Sanitize improve security
- When transitioning to developer mode:
  - Crypto-erase if name space is supported
  - Block erase otherwise

/dev/nvme0n1

kernel images / root partitions / other

/dev/nvme0n2

user partition
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
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)

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:

<table>
<thead>
<tr>
<th>Example: NAND flash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read latency w/o blocking</td>
</tr>
<tr>
<td>Read latency w/ blocking behind a write</td>
</tr>
</tbody>
</table>

IOD helps address this:
IOD Concept: NVM Sets

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

<table>
<thead>
<tr>
<th>Config #</th>
<th>NVM Sets provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4 Sets of 1 TB each</td>
</tr>
<tr>
<td>1</td>
<td>8 Sets of 512 GB each</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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
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
No writes to media. No GC or other maintenance. Minimal writes from host. Minimal tail latency on reads.

NDWIN = Non-Deterministic Window
Writes allowed. Host can write; GC and other maintenance allowed. Return to DTWIN when possible.

---

Flash Memory Summit 2017
Santa Clara, CA
Read Recovery Level (RRL)

- RRL: host can trade-off UBER for latency

<table>
<thead>
<tr>
<th>RRL</th>
<th>O/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor specific</td>
<td>Optional</td>
</tr>
<tr>
<td>....</td>
<td>Optional</td>
</tr>
<tr>
<td>Default (normal recovery effort)</td>
<td>Mandatory</td>
</tr>
<tr>
<td>....</td>
<td>Optional</td>
</tr>
<tr>
<td>....</td>
<td>Optional</td>
</tr>
<tr>
<td>Fast Fail (minimal recovery effort)</td>
<td>Mandatory</td>
</tr>
</tbody>
</table>

Better UBER  Better Latency
Opportunity #2: Optimizing memory TCO
Semiconductor geometry scaling is reaching its limit. However, impact on different technologies is variable.

<table>
<thead>
<tr>
<th>DRAM</th>
<th>NAND and new NVMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planar technology, no prospect for monolithic 3D scaling</td>
<td>3D already proven for NAND, expected for PCM and ReRAM</td>
</tr>
<tr>
<td>Limited prospects for cost and capacity scaling</td>
<td>Reasonable prospects scaling in the foreseeable future</td>
</tr>
</tbody>
</table>

Idea: Use NVMs to supplement DRAM.
Caveat: DRAM will have the best performance. Use NVMs for “cold data”.
Implementation: Reinvent Paging

Choosing swap media

<table>
<thead>
<tr>
<th>Media</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDD</td>
<td>~10 ms</td>
</tr>
<tr>
<td>SSD</td>
<td>~100 µs</td>
</tr>
<tr>
<td>New NVM</td>
<td>~10 µs</td>
</tr>
</tbody>
</table>

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?
NVMe device ecosystem has become broad

And...
- NVMe-oF – Non-Pcie fabrics
- NVMe-MI – Out-of-Band Management

NVMe Datacenter to Mobile
Flash Memory Summit

Broad ecosystem means security considerations

**Threat Classes**

- Physical Device Access
  - Device Lost/Stolen
  - Repurposing a device

- Data Access
  - Theft or unauthorized disclosure of data
  - Malicious or criminal change or destruction of data

**Mitigation Strategies**

- Sanitize
- Data-at-Rest Encryption
- Device Locking

Above +

- Media Write Protection
- Data-in-Flight encryption
- E2E Cryptographic Integrity Checks

To left +

- Authentication/Access Control

---

**Threat Classes**

- Physical Device Access
  - Device Lost/Stolen
  - Repurposing a device

- Data Access
  - Theft or unauthorized disclosure of data
  - Malicious or criminal change or destruction of data

**Mitigation Strategies**

- Sanitize
- Data-at-Rest Encryption
- Device Locking

Above +

- Media Write Protection
- Data-in-Flight encryption
- E2E Cryptographic Integrity Checks

To left +

- Authentication/Access Control
TCG/NVMe Work To Date

Opal Family

Opal SSC
- PSID
- Datastore Tables
- Single User Mode
- Block SID Auth
- Config NS Locking

Opalite SSC
- PSID
- Block SID Auth
- Config NS Locking

Pyrite SSC
- Block SID Auth

Enterprise SSC
- PSID
- Block SID Auth
- PSK Secure Messaging
- Locking LBA Range Control

Opal/Enterprise SSCs
- Data-at-rest protection
- Device Locking
- Cryptographic Sanitization

IN DISCUSSION: Solution that supports NVMe and addresses Enterprise use cases using Core Spec v2.01

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

Core Spec v2.01

Core Spec (v1.0 r.09)

Storage Interface Interactions Spec (SIIS)
- SAS, SATA, NVMe, eMMC, UFS
NVMe Native Security Features

1) Security Send-Receive – Protocol Tunneling
   - NVMe Security Receive
   - NVMe Security Send
   - e.g. TCG SWG Protocol

2) Namespace Write Protect
   - Diagram showing transitions between
     - No Write Protect
     - Write Protect
     - Write Protect Until Reset
     - Permanent 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

<table>
<thead>
<tr>
<th>Establish connection</th>
<th>Connect</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Establish secure channel [Channel credential]</td>
<td>Initial command</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authenticated</td>
<td>Initial command</td>
<td>Final response</td>
</tr>
<tr>
<td>[Authentication credential]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authenticated Transmission of Commands</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

- 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
Architected for Performance
Bios TBA
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

Initial state

No Write Protect

Write Protect

Write Protect Until Reset

Permanent Write Protect

NO2

MEK3

Global
LO

LO1

LO2

NS1

MEK1

NS2

MEK2

LBA Range
MEK3

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