Enabling NVMe® I/O Determinism @ Scale

Chris Petersen, Hardware System Technologist
Wei Zhang, Software Engineer
Alexei Naberezhnov, Software Engineer
Facebook
Facebook @ Scale

800 Million 1.3 Billion 2.2 Billion
Industry Trends

Flash and CPU continue to diverge

Flash Capacity Growth
CPU Performance
Industry Trends
NAND Flash and SSDs

Number of NAND flash die and write BW per TB vs. die density

- Number of die per TB
- Seq Write BW/TB

Number of NAND die per 1 TB

Sequential Write BW per TB (MB/s)

NAND Flash die density
Dark Flash

Flash capacity utilization trend vs. target

Growing gap of underutilized Flash (Dark Flash)

Note: Includes 25% generation over generation performance improvements
Flash Workloads @ Facebook

- Read Intensive with bursty writes
- Facebook flash applications are sensitive to read latency
  - Especially read latency outliers
- Multiple, concurrent instances

[Caching]
Sources of Read Latency
External (aka Noisy Neighbors)

Huge latencies result from naively sharing flash.

225ms AVERAGE latency

< 10 ms

- AVERAGE latency
- External (aka Noisy Neighbors)
Sources of Read Latency

Internal (within the application)

- Read latency outliers are caused by “collisions” with
  1. Concurrent flash writes
  2. Flash background operations:
      - Garbage collection
      - Wear leveling
      - Read scrub
      - Block erase
  - Error correction
  - Exception handling (e.g. program/erase failures)

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read time</td>
<td>~60–100us</td>
</tr>
<tr>
<td>Program time</td>
<td>~1–1.5ms</td>
</tr>
<tr>
<td>Erase time</td>
<td>~10–15ms</td>
</tr>
</tbody>
</table>
Solution: NVMe® I/O Determinism

NVMe standards have been ratified!

- NVM Sets and Read Recovery Level (TP 4018a)
- NVMe Predictable Latency Mode (TP 4003a)
- Additional improvements are a work in progress!

https://nvmexpress.org/resources/specifications/
NVMe® I/O Determinism: NVM Sets
NVM Sets

- An abstract allocation of SSD HW resources
- Each set has dedicated NAND resource
- Each set can have dedicated channels, depends on architecture
- Each set carries out its own writes and background operations
- Physically isolated to avoid “Collison” caused by the noisy neighbors
Benefits

- Enables QoS Regions at the SSD level
  - Better support of multi-tenants on an SSD
- Host software can leverage sets as-is
  - Part of the NVMe Standard
  - Sets are exported as namespaces
  - Host OS does NOT need to be sets-aware
Use Cases @ Facebook

App A
App B
App C

Set 1
Set 2
Set 3
Set 4

SSD with NVM Sets

Improved QoS!
Increased Flash Utilization!

Aligns with Facebook’s Disaggregated Flash Strategy!
Evaluation Setup

SSD with 4 Namespaces (No Sets)

Workload Patterns
- All namespaces run the same workloads
- Noisy Neighbors
  - One namespace runs the target workload (NS1)
  - The rest of three namespaces act as noisy neighbors (NS2-4)

VS

SSD with 4 Sets
Neighbors with Same Workloads

read latency averaged across all NS under test

~8X Improvement with Sets

RR @ 4K

70/30 RR/RW @ 4K

NS: Namespace
RR: Random Read
RW: Random Write
Noisy Neighbors

Read Latency (us)

~10X Improvements

- NS: Namespace
- NN: Noisy Neighbor
- RR: Random Read
- RW: Random Write

<table>
<thead>
<tr>
<th>Latency</th>
<th>No Sets</th>
<th>Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>P99</td>
<td>1000</td>
<td>20</td>
</tr>
<tr>
<td>P99.99</td>
<td>2000</td>
<td>20</td>
</tr>
<tr>
<td>P99.9999</td>
<td>3000</td>
<td>20</td>
</tr>
<tr>
<td>P99</td>
<td>4000</td>
<td>20</td>
</tr>
<tr>
<td>P99.99</td>
<td>5000</td>
<td>20</td>
</tr>
<tr>
<td>P99.9999</td>
<td>6000</td>
<td>20</td>
</tr>
</tbody>
</table>
Current Implementation Limitations

- NVM Sets is non-trivial to implement with the current SSD architecture
  - Hard to partition all resource in the current generation of controllers
  - Design the NextGen sets-aware SSD controller
- Lack of per-set endurance group info
  - TP 4050: Endurance Group Information Enhancements
- SQs are not associated with Sets
  - TP 4045: SQ Associations
NVMe® I/O Determinism: Predictable Latency Mode
Effects of Internal Activities on Latency

- Internal operations account for most of NAND activity:
  - 7% OP results in 14.3x worst-case WA
  - 28% OP results in 3.6x worst-case WA
- Internal operations usually happen in batches
- Scheduling of internal operations is a black box to the host
Latency Improvement Approaches

- Load limiting (e.g. queue depth, bandwidth)
- Over-provisioning
- Program/erase suspend
- Open-channel
- NVMe Predictable Latency Mode (PLM)
NVMe Predictable Latency Mode (PLM)

- Allows host to decide when internal operations may happen
- Drive encapsulates scheduling algorithm and all media details
- Drive advertises only required details about scheduling capabilities:
  - Estimates of time, # of reads and writes until maintenance is required
NVMe PLM: Contract

**Host agrees:**
- Not to send writes or trims during D-window
- Respect window estimates advertised by the drive

**Drive agrees:**
- Not to do operations unrelated to reads during D-window
- Drive may switch back to ND-window if contract is broken
NVMe PLM: Prototype

Goals:
- Improve consistency of read latency
- Achieve read-only like latency for mixed workloads

Approach:
- Leverage data redundancy & PLM to segregate reads from other operations
NVMe PLM: Write Cache

Prototype uses RAM cache that relies on NVMe meta-data for power fail recovery

Benefits:
- Flexible configuration
- Minimal impact on performance
- No need for additional hardware
- Allows R/W access during recovery*

Recovery:
1. Check boot blocks
2. Clean shutdown?
3. Check data block pairs for version mismatch
NVMe PLM: Kernel support

**MD:**
- New Raid1-PLM personality

**Block Layer:**
- Expose PLM interface
- Expose generic metadata interface

**NVMe Driver:**
- Implement PLM interface
- Add support for generic metadata
- Add support for set associated SQs
NVMe PLM: Test setup

- Same usable capacity, read & write rate
- Random read 4K @ QD8
- 1:2 mix of random and sequential writes 128K @ QD8
- Initialized with 2 passes of mixed writes
NVMe PLM: Test results

RR 4K@QD8 + RW/SW 25%/75% 128K@QD8

- NVMe (baseline)
- NVMe (read-only)
- RAID0 50% OP

Read latency (us)

- 90%
- 99%
- 99.9%
- 99.99%
- 99.999%
- 99.9999%
- 99.99999%

2x
11x
NVMe PLM: Test results

RR 4K@QD8 + RW/SW 25%/75% 128K@QD8

Read latency (us)

- NVMe (baseline)
- NVMe (read-only)
- RAID0 50% OP
- RAID1-PLM

21x
NVMe PLM: Test results

RR 4K@QD8 + RW/SW 25%/75% 128K@QD8

<10%
## NVMe PLM: Proposition

<table>
<thead>
<tr>
<th></th>
<th>RAID1 (2 drives)</th>
<th>RAID5 (4 drives)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Array vs 1 Drive</td>
<td>Utilization</td>
</tr>
<tr>
<td>Capacity</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Write BW</td>
<td>50%</td>
<td>25%</td>
</tr>
<tr>
<td>Read BW</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Extra Hardware</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>Read Latency</td>
<td>read-only like</td>
<td></td>
</tr>
</tbody>
</table>
NVMe PLM: Future work

- Explore other redundancy schemes
- Explore other power fail recovery schemes
- Explore multi-set configurations (requires TP4045)
Thank You!

Check out this I/O Determinism talk too:
The Reality of an NVMe IO Deterministic Drive Using QLC
Steven Wells, Fellow – SSD Data Center Architecture
August 7th, 4:50pm