Solid-State Drives and Non-Volatile Media

Solid-State Drive

- Responsibilities
  - Flash Translation Layer
  - Media Error Handling
  - Media Retention Management

Media Controller

Parallel Units
- Channel X
- Channel Y

Host Interface

Read/Write

Tens of Parallel Units!

Transform R/W/E to R/W

Manage Media Constraints
  - ECC, RAID, Retention

Read/Write/Erase

NAND

- Read (50-100us)
- Write (1-10ms)
- Erase (3-15ms)
Mixed Workloads

- **0% writes and latency is consistent**

- **20% writes makes big impact on read latency**

- **50% writes can make SSDs as slow as spinning drives**

- Larger outliers on increased writes
Indirection and Read/Write I/O Interface

Even if Writes and Reads does not collide from application Indirection and loss of information due to the narrow Read/Write I/O interface

Solid-State Drive Pipeline

- Write Buffer
- NAND Controller
- die₀
- die₁
- die₂
- die₃

Writes decoupled from Reads

Data placement + Buffering = Best Effort

SSD state is hidden due to the narrow I/O Interface
There is a need for a Storage Interface that provides

- I/O Predictability
- I/O Isolation
- Reduce write-amplication by tighter integration
- Host-controlled data placement and I/O scheduling
Introduction

1. Physical Page Addressing (PPA) for Open-Channel SSDs
2. The LightNVM Subsystem
3. pblk: A host-side Flash Translation Layer for Open-Channel SSDs
4. Demonstrate I/O Predictability and I/O Isolation using this interface
Physical Page Addressing (PPA) Interface

- Expose geometry
  - Logical/Physical geometry
  - Performance
  - Controller functionalities

- Hierarchical Address Space
  - Encode geometry into the address space

- Vector I/Os
  - Read/Write/Erase

# Channels, # Parallel Units, # Chunk, Chunk Size, Min. Write size, Optimal Write size, ...
Up to the SSD vendor

Encode parallel units into the address space
1. NVMe Device Driver
   - Detection of OCSSD
   - Implements PPA interface

2. LightNVM Subsystem
   - Generic layer
   - Core functionality
   - Target management (e.g., pblk)

3. High-level I/O Interface
   - Block device using pblk
   - Application integration with liblightnvm
Host-side Flash Translation Layer - pblk

- Mapping table
  - Sector-granularity
- Write buffering
  - Lockless circular buffer
  - Multiple producers
  - Single consumer (Write Thread)
- Error Handling
  - Media write/erase errors
- Garbage Collection
  - Rewrite blocks
- Recovery of metadata
Benchmarks

- CNEX Labs Open-Channel SSD
  - NVMe
  - PCIe Gen3x8
  - 2TB MLC NAND

- Geometry
  - 16 channels
  - 8 PUs per channel (Total: 128 PUs)

- Parallel Unit Characteristics
  - Read Size: 4K
  - Write size: 16K + 64B user OOB
  - Chunks: 1.067, Chunk Size: 16MB

- Performance:
  - Write: Single PU 47MB/s
  - Read: Single 108MB/s, 280MB/s (64K)

- Limit # Active Parallel Write Units
- Predictable Latency
- Multi-tenancy using I/O Isolation
Limit # Active Writers

- A priori knowledge of workload. E.g., limit to 400MB/s Write
- Limit number of Active PU Writers, and achieve better read latency

256K Write QD1
256K Read QD16

Single Read and Write Perf.

Mixed Read/Write

Write throughput 400MB/s

Write latency increases, and read latency reduces
Predictable Latency

- 4K reads during 64K concurrent writes
- Consistent low latency at 99.99, 99.999, 99.9999
Multi-Tenant Workloads

2 Tenants (1W/1R)

4 Tenants (3W/1R)

8 Tenants (7W/1R)

NVMe SSD

OCSSD
Conclusion

- New interface that provides
  - I/O Predictability
  - I/O Isolation
  - Puts the host in front seat of data placement and I/O scheduling

- PPA Specification is open and available for implementors

- Active community using OCSSDs both for production and research
  - Multiple drives in development within SSD vendors
  - Multiple papers already on Open-Channel SSDs that shows how this interface can improve workloads

- Fundamental building blocks are available:
  - Initial release in Linux kernel 4.4.
  - User-space library (liblightnvm) support with Linux kernel 4.11.
  - Pblk will be upstream with Linux kernel 4.12.

- The right time to dive into Open-Channel SSDs
  - More information available at: http://lightnvm.io
CNEX Labs, Inc.

Teaming with NAND Flash manufacturers and industry leaders in storage and networking to deliver the next big innovation for solid-state-storage.