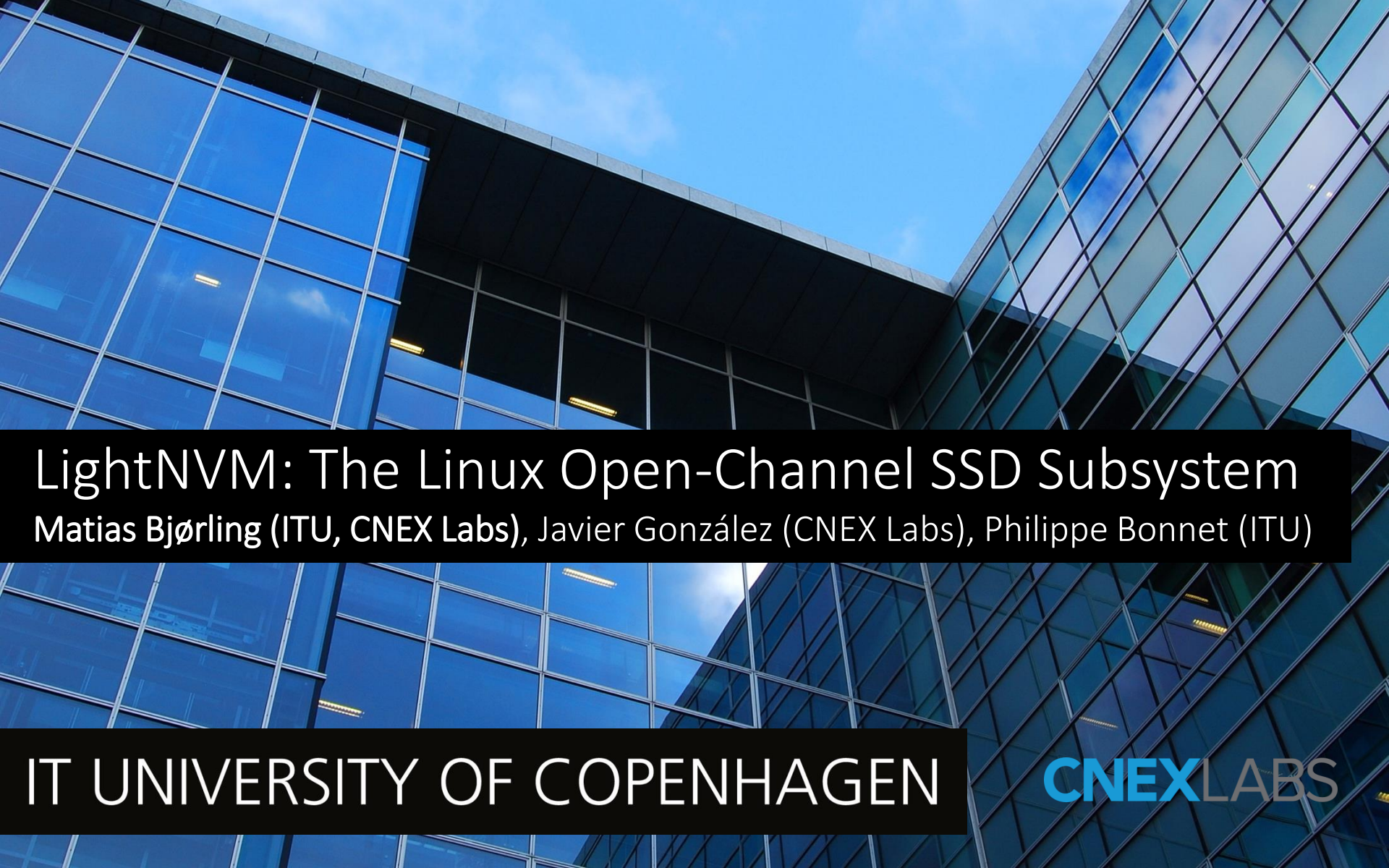


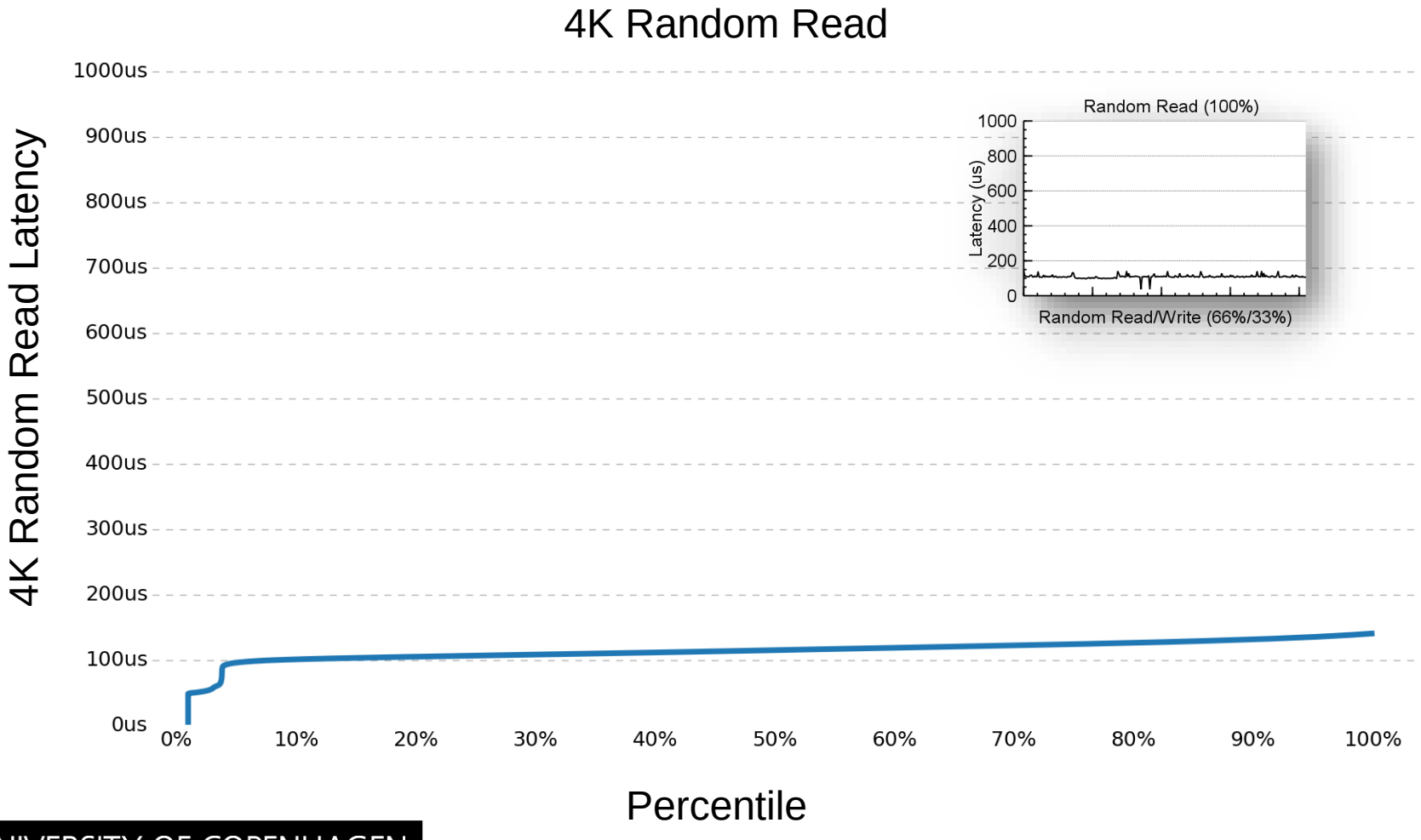
LightNVM: The Linux Open-Channel SSD Subsystem

Matias Bjørling (ITU, CNEX Labs), Javier González (CNEX Labs), Philippe Bonnet (ITU)

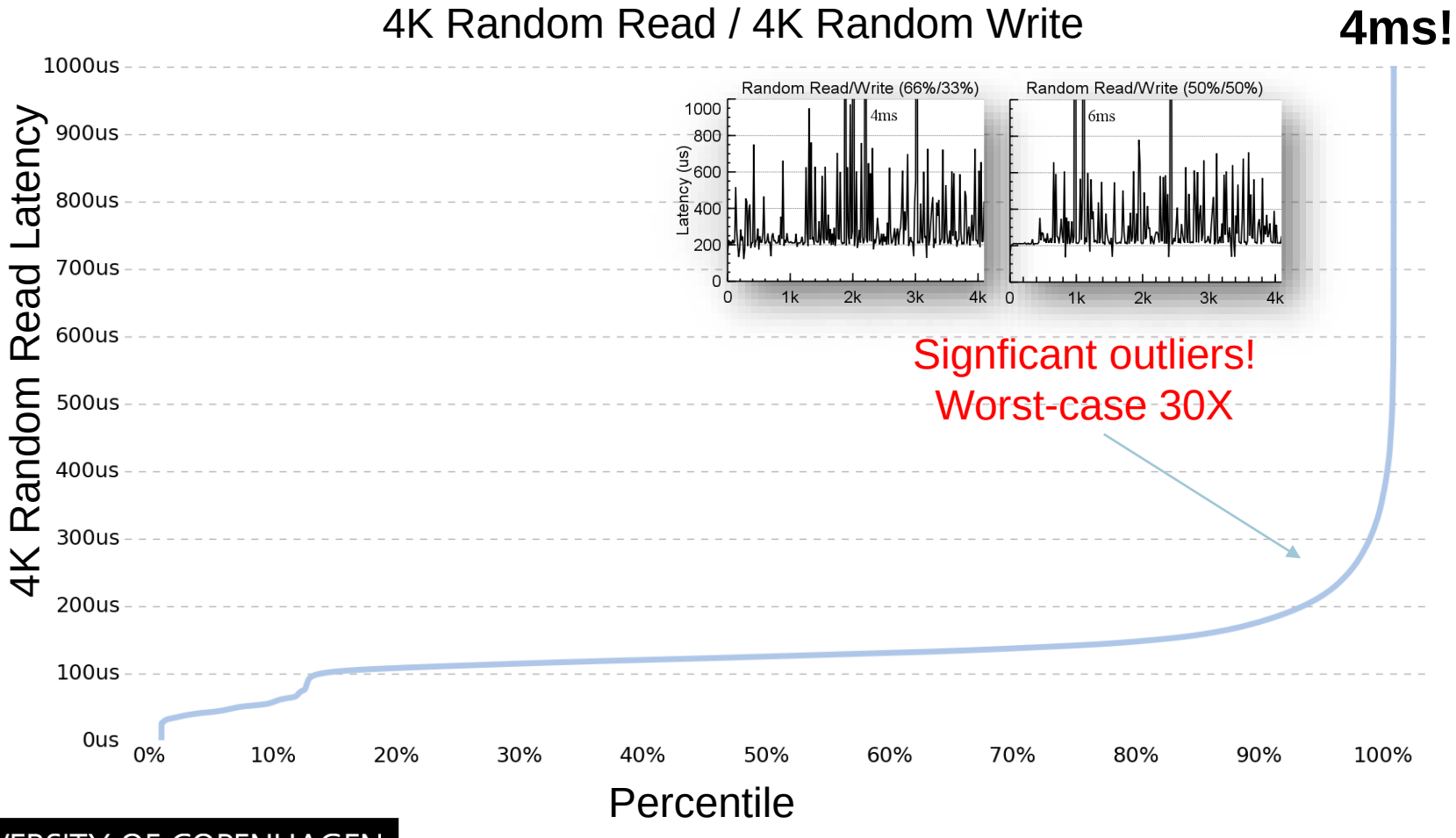
IT UNIVERSITY OF COPENHAGEN

CNEXLABS

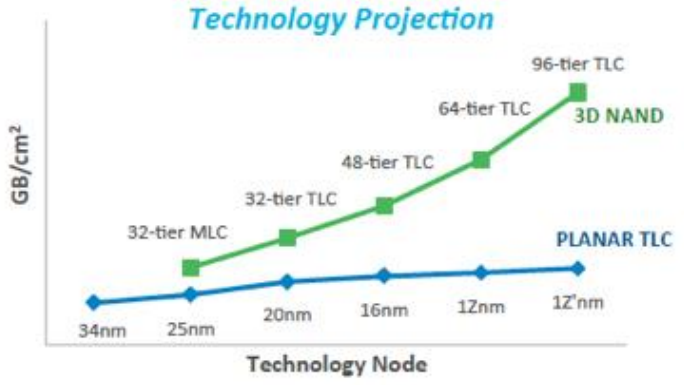
0% Writes - Read Latency

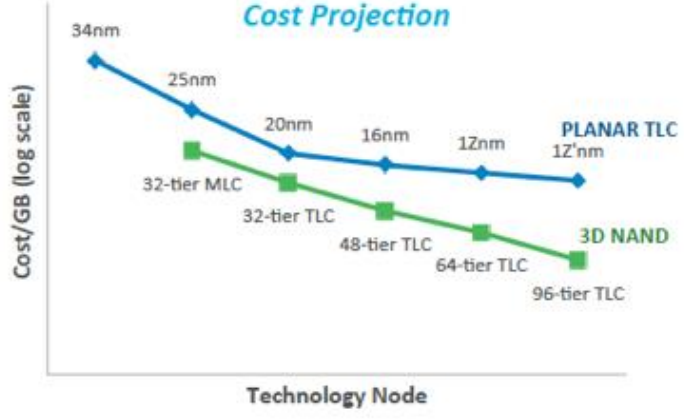


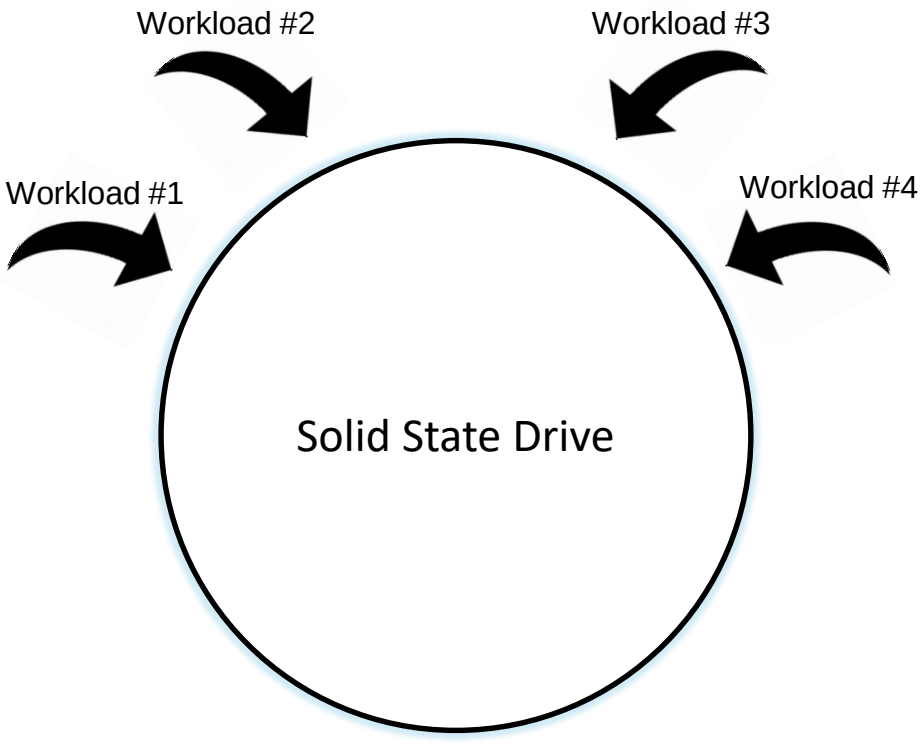
20% Writes - Read Latency



NAND Capacity Continues to Grow







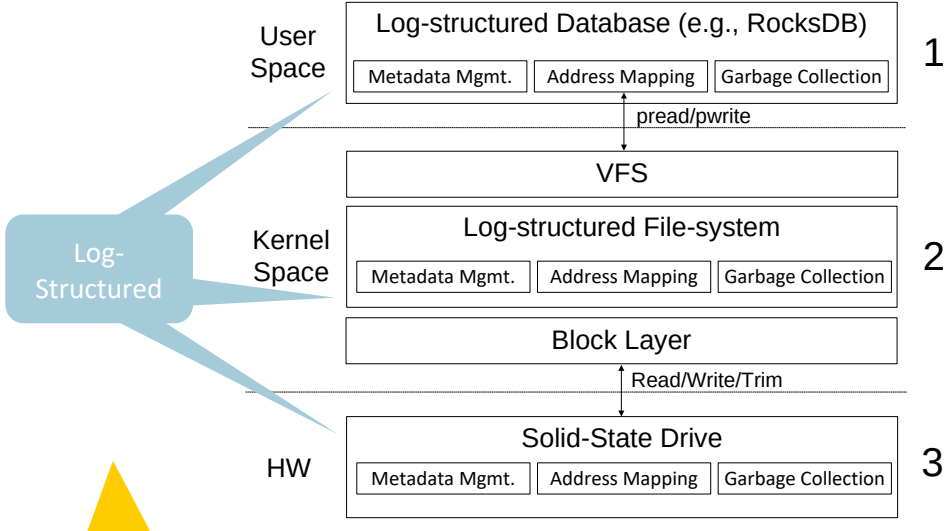
Performance – Endurance – DRAM overheads

Source: William Tidwell -The Harder Alternative – Managing NAND capacity in the 3D age

What contributes to outliers?

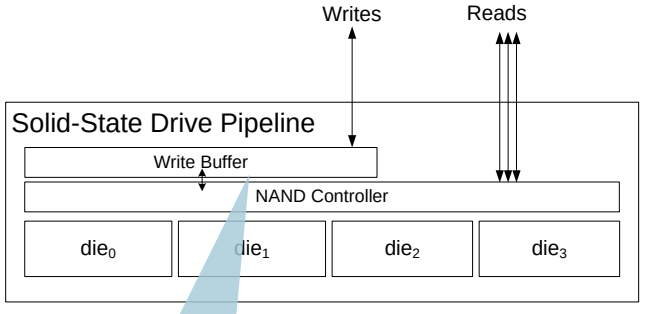
Even if Writes and Reads does not collide from application
Indirection and a **Narrow Storage** interface cause outliers

Host: Log-on-Log



Unable to align data logically
= Write amplification
increase + extra GC

Device: Write Indirection & Unknown State



Drive maps logical data
to the physical location
with **Best Effort**

Host is oblivious to physical
data placement due to
indirection

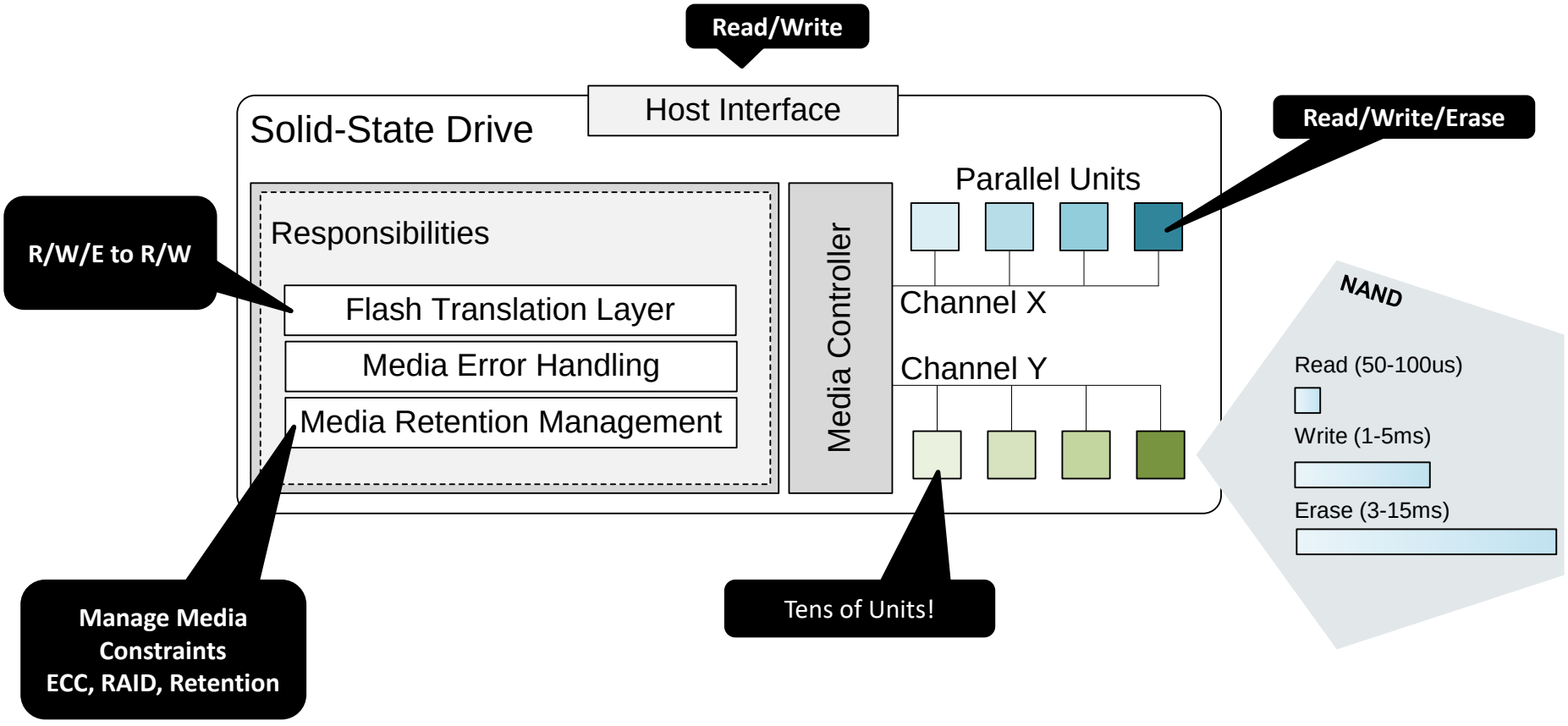
Open-Channel SSDs

I/O Isolation
Provide isolation between tenants by allocating independent parallel units

Predictable Latency
I/Os are synchronous. Access time to parallel units are explicit defined.

Data Placement & I/O Scheduling
Manage the non-volatile memory as a block device, through a file-system or inside your application.

Solid-State Drives



Rebalance the Storage Interface

Expose device parallelism

- Parallel units (LUNs) are exposed as independent units to the host.
- Can be a logical or a physical representation.
- Explicit performance characteristics.

Log-Structured Storage

- Exposes storage as chunks that must be written sequentially.
- Similar to the HDD Shingled Magnetic Recording (SMR) interface.
- No need for internal garbage collection by the device.

Integrate with file-systems and databases, and can also implement I/O determinism, streams, barriers, and other new data management schemes without changing device firmware.

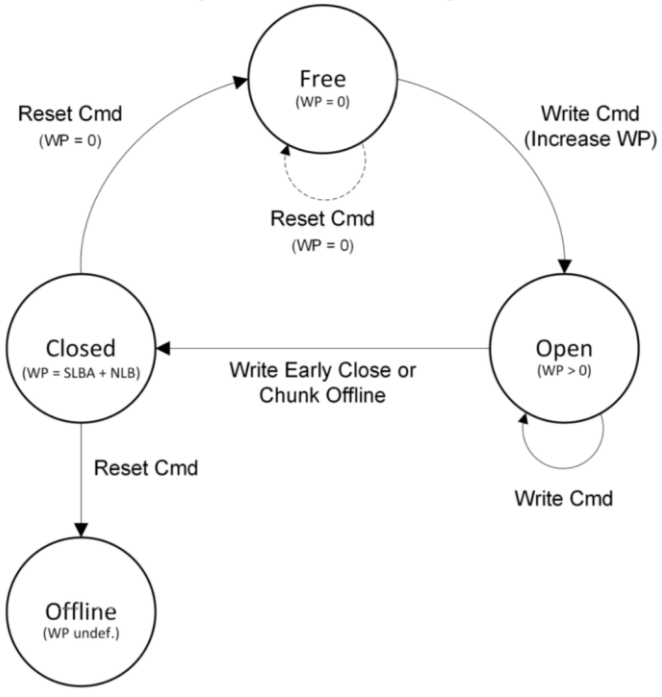
Specification

Device model

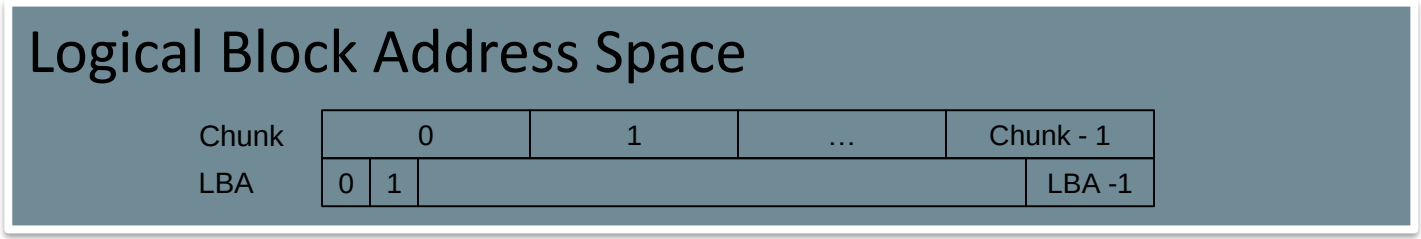
- Defines parallel units and how they are laid out in the LBA address space.
- Defines chunks. Each chunk is a range of LBAs where writes must be sequential. To write again, a chunk must be reset.
  - A chunk can be in one of four states (free/open/closed/offline)
  - If a chunk is open, there is a write pointer associated.
  - The model is media-agnostic.

Geometry and I/O Commands

- Read/Write/Reset – Scalars and Vectors



Drive Model - Chunks



Reads

Logical block granularity

For example 4KB

Write

Min. Write size granularity

Synchronous – May fail – An error marks write bad, not whole SSD

Reset

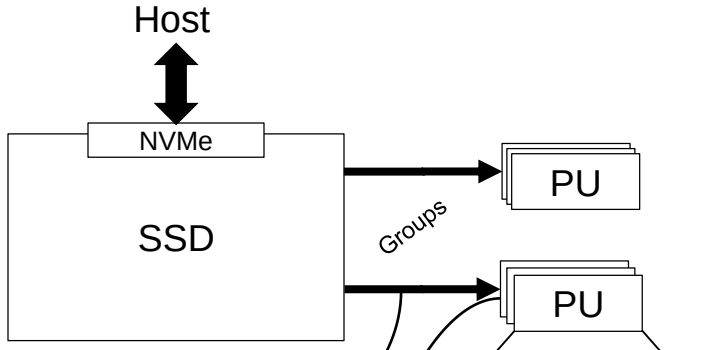
Chunk granularity

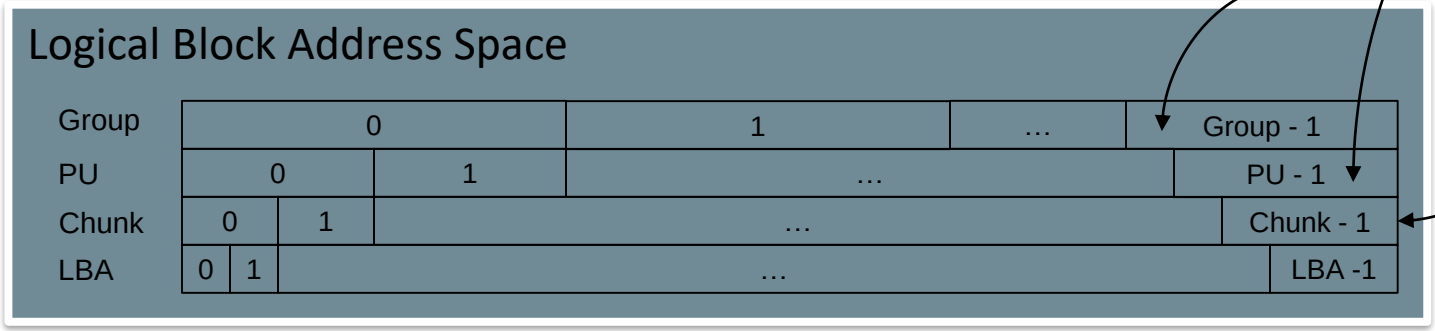
Synchronous – May fail – An error only marks chunk bad, and not whole SSD

Drive Model - Organization

Parallelism across

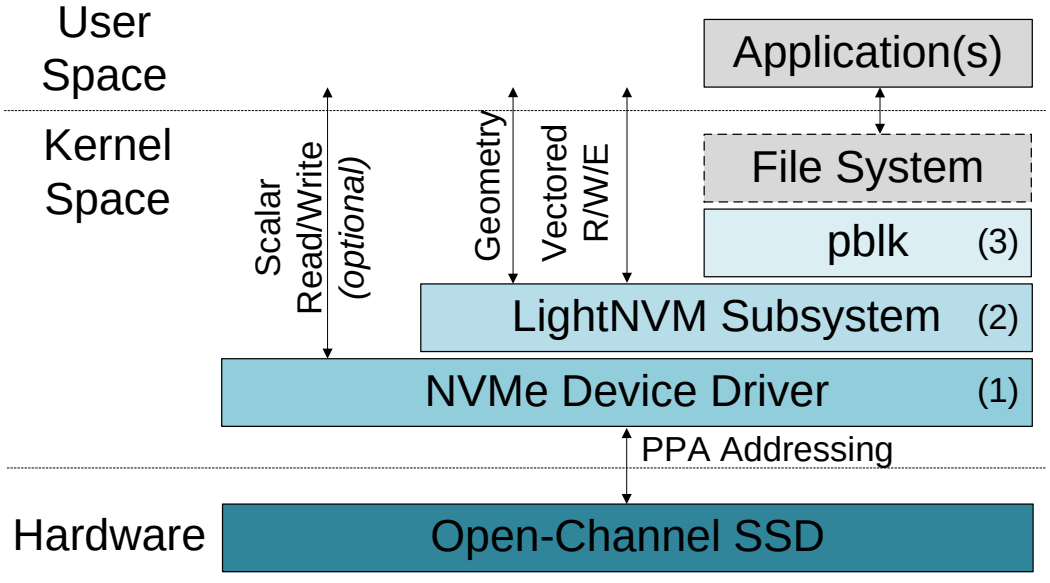
- Groups (Shared bus)
- Parallel Units (LUNs)





LightNVM Subsystem Architecture

1. NVMe Device Driver
   - Detection of OCSSD
   - Implements specification
2. LightNVM Subsystem
   - Generic layer
   - Core functionality
   - Target management
3. High-level I/O Interfaces
   - Block device using a target
   - Application integration with liblightnvm
   - File-systems, ...



pblk - Host-side Flash Translation Layer

Mapping table

- Logical block granularity

Write buffering

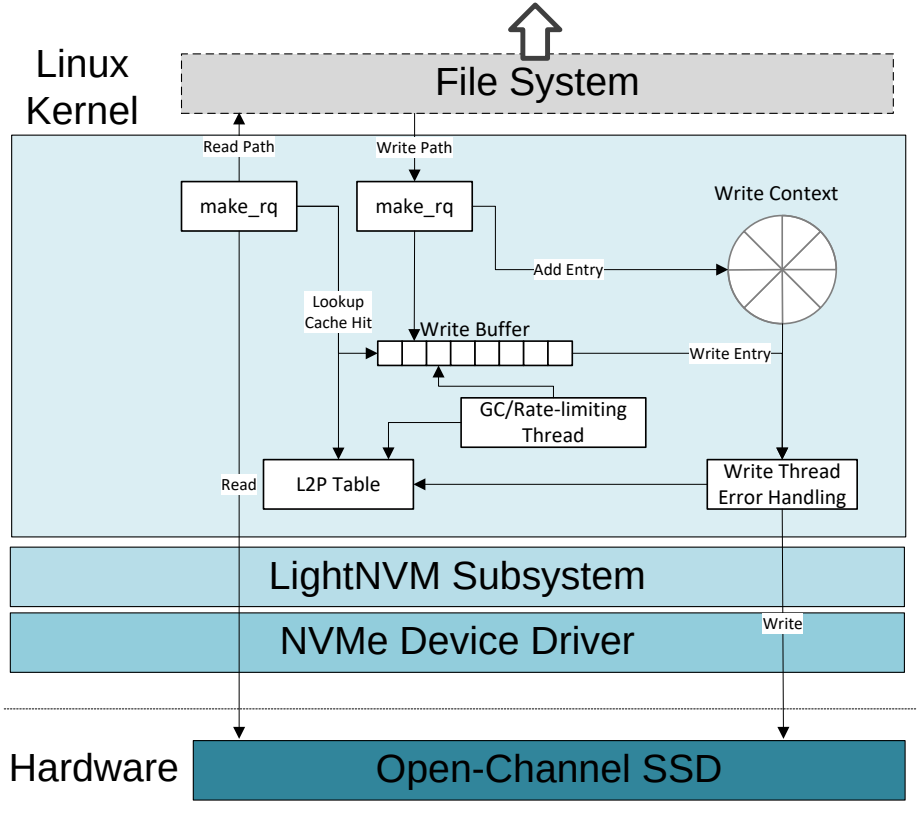
- Lockless circular buffer
- Multiple producers
- Single consumer (Write Thread)

Error Handling

- Device write/reset errors

Garbage Collection

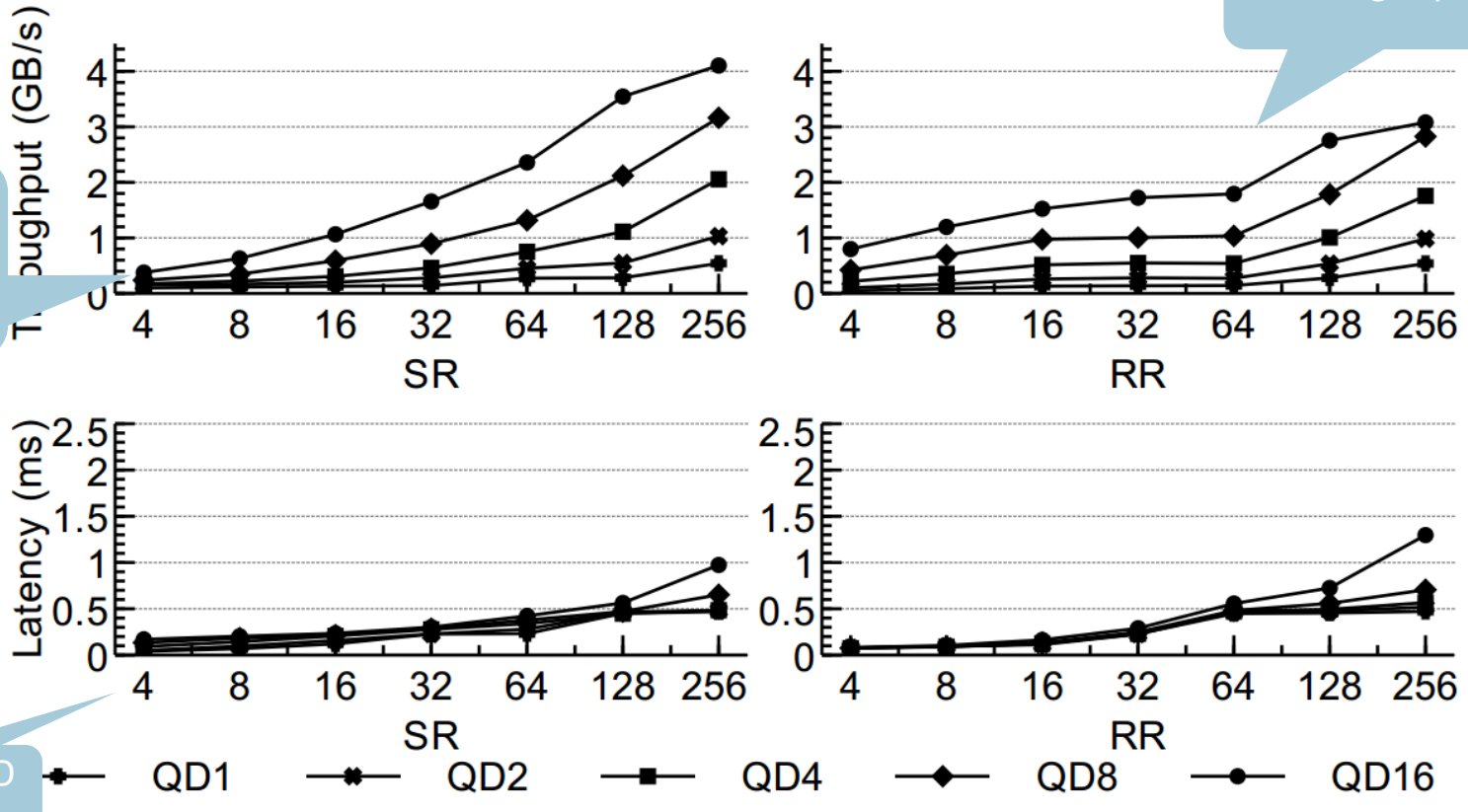
- Refresh data
- Rewrite chunks



Experimentation

- **Drive**
  - CNEX Labs Open-Channel SSD
  - NVMe, Gen3x8, 2TB MLC NAND
  - Implements Open-Channel 1.2 specification
- Parallelism
  - 16 channels
  - 8 parallel units per channel (Total: 128 PUs)
- Parallel unit characteristic
  - Min. Write size: 16K + 64B OOB
  - Chunks: 1,067, Chunk size: 16MB
- Throughput per parallel unit:
  - Write: 47MB/s
  - Read: 108MB/s (4K), 280MB/s (64K)

Base Performance – Throughput + Latency

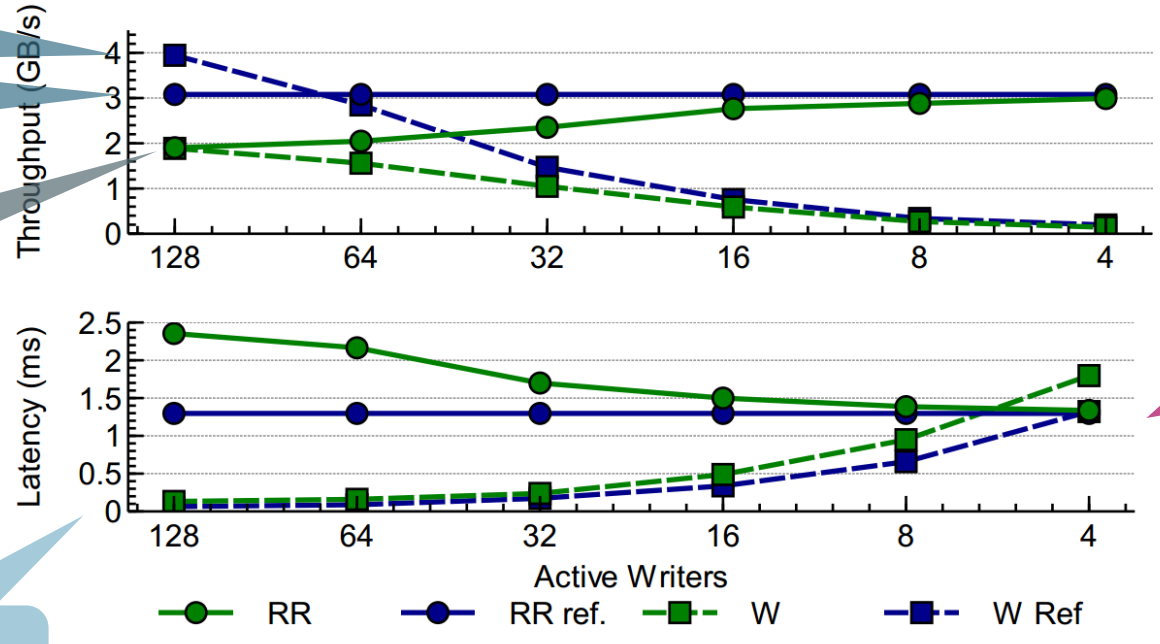


Limit # of Active Writers

Limit number of writers to improve read latency

Single Read or Write Perf.

Mixed Read/Write



Write Perf. at 200MB/s

Write latency increases, and read latency reduces

256K Write QD1
256K Read QD16

A priori knowledge of workload. Write 200MB/s

Multi-Tenant Workloads

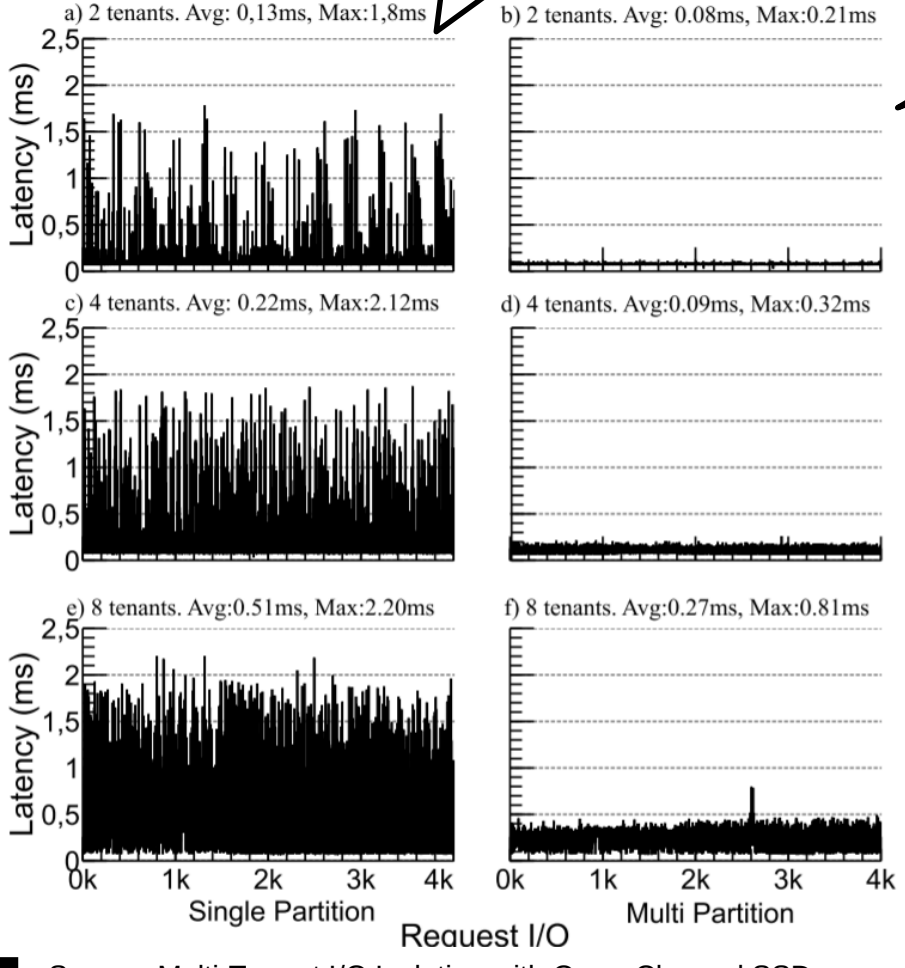
NVMe SSD

OCSSD

2 Tenants (1W/1R)

4 Tenants (3W/1R)

8 Tenants (7W/1R)



Source: Multi-Tenant I/O Isolation with Open-Channel SSDs, Javier González and Matias Bjørling, NVMW '17

Lessons Learned

1. **Warranty to end-users** – Users has direct access to media.
2. **Media characterization is complex** and performed for each type of NAND memory – Abstract the media to a ”clean” interface.
3. **Write buffering** – For MLC/TLC media, write buffering is required. Decide if in host or in device.
4. **Application-agnostic wear leveling is mandatory** – Enable statistics for host to make appropriate decisions.

Conclusion

Contributions

- New storage interface between host and drive.
- The Linux kernel LightNVM subsystem.
- pblk: A host-side Flash Translation Layer for Open-Channel SSDs.
- Demonstration of an Open-Channel SSD.

LightNVM

- Initial release of subsystem with Linux kernel 4.4 (January 2016).
- User-space library (liblightnvm) support upstream in Linux kernel 4.11 (April 2017).
- pblk available in Linux kernel 4.12 (July 2017).
- Open-Channel SSD 2.0 specification released (January 2018) and support available from Linux kernel 4.17 (May 2018).


Thank You