OPEN.

FOR
BUSINESS.

OCP
SUMMIT

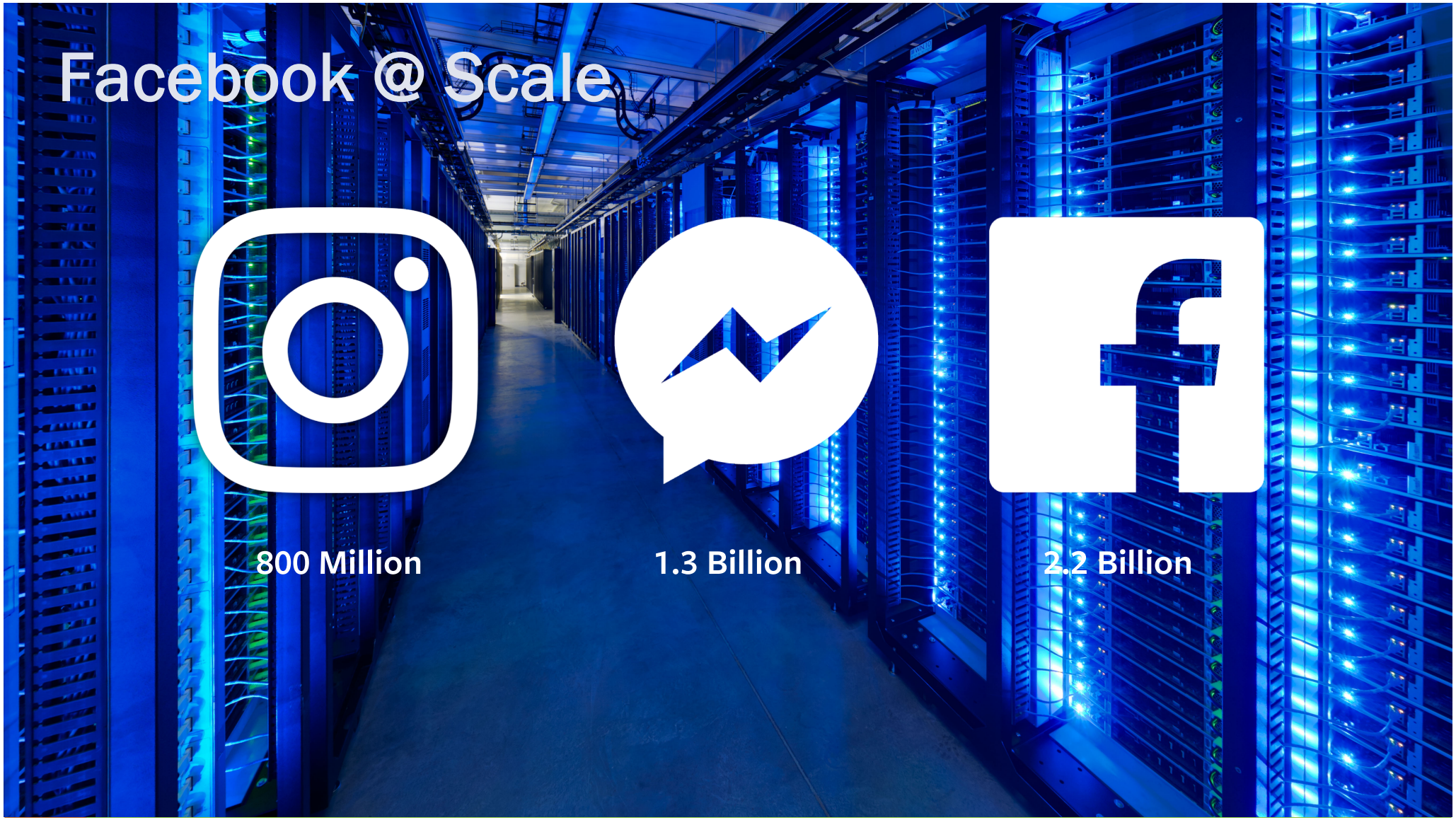
Server & Storage

Achieving Deterministic I/O with Lightning JBOF

Wei Zhang and Alexei Naberezhnov
Software Engineer

Facebook

Facebook @ Scale



Lightning JBOF

- NVMe based JBOF (Just Bunch Of Flash)
- End-to-end PCIe Gen 3 connections
- Support SSD surprise hot plug
- Support various SSD form factors
  - M.2 and U.2

Lightning OCP Specification
http://files.opencompute.org/oc/public.php?service=files&t=6a1da9f48000610fbf003798fd988098

Lightning OCP Market Place
https://www.opencompute.org/products/217/wiwynn-lightning-nvme-jbof-up-to-60-nvme-ssds

OPEN ACCEPTED™

Disaggregated Flash @ Facebook

- Lightning enables flash disaggregation at Facebook
  - Flash storage is physically decoupled from the compute node
- Enable applications share flash storage
  - Support more flexible flash allocation
  - Drive up flash overall utilization

Flash Workloads @ Facebook

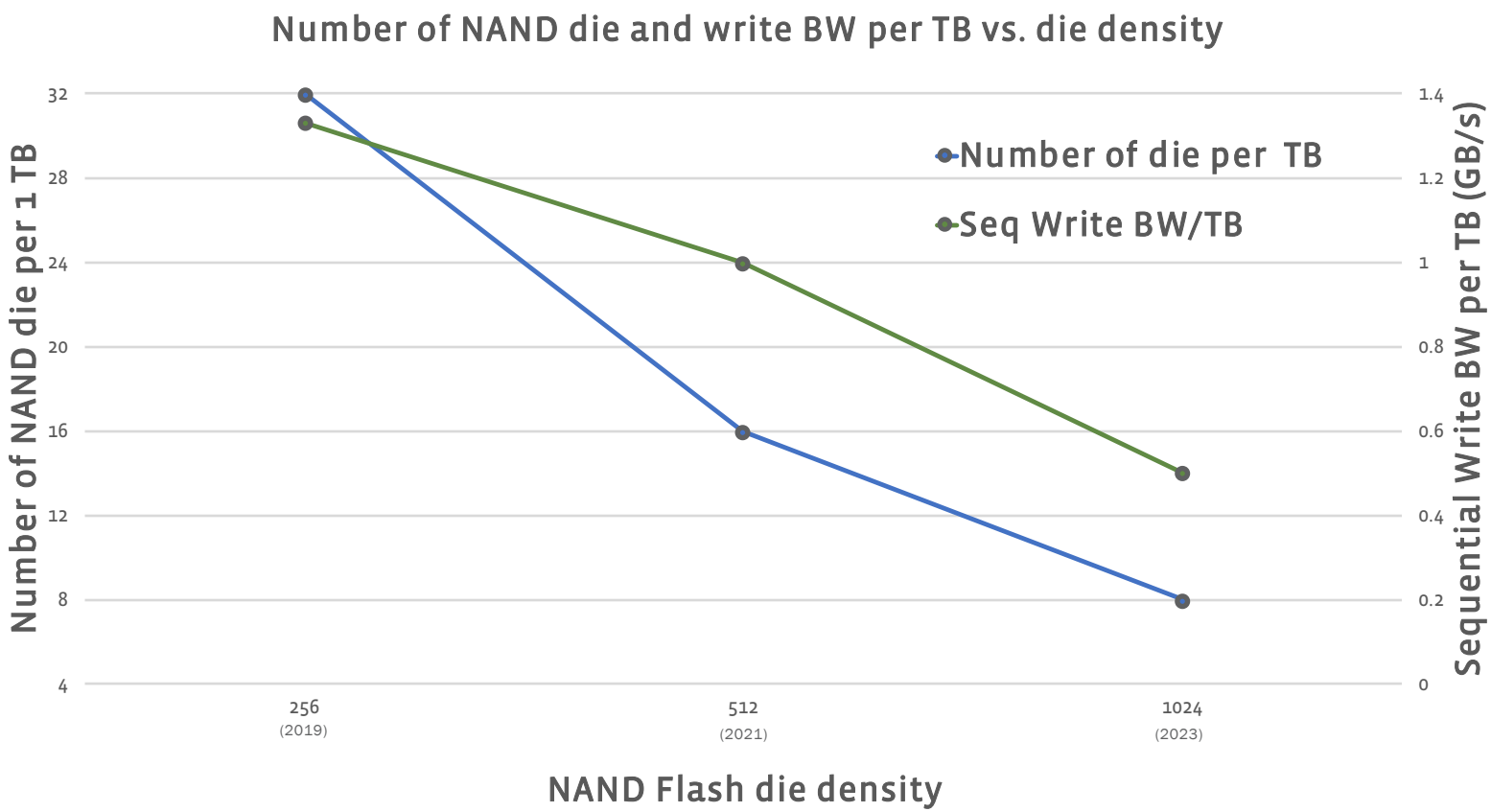
- Read Intensive with bursty writes
- Sensitive to read latency
  - Especially read latency outliers
- Multiple, concurrent instances

Sources of SSD Read Latency

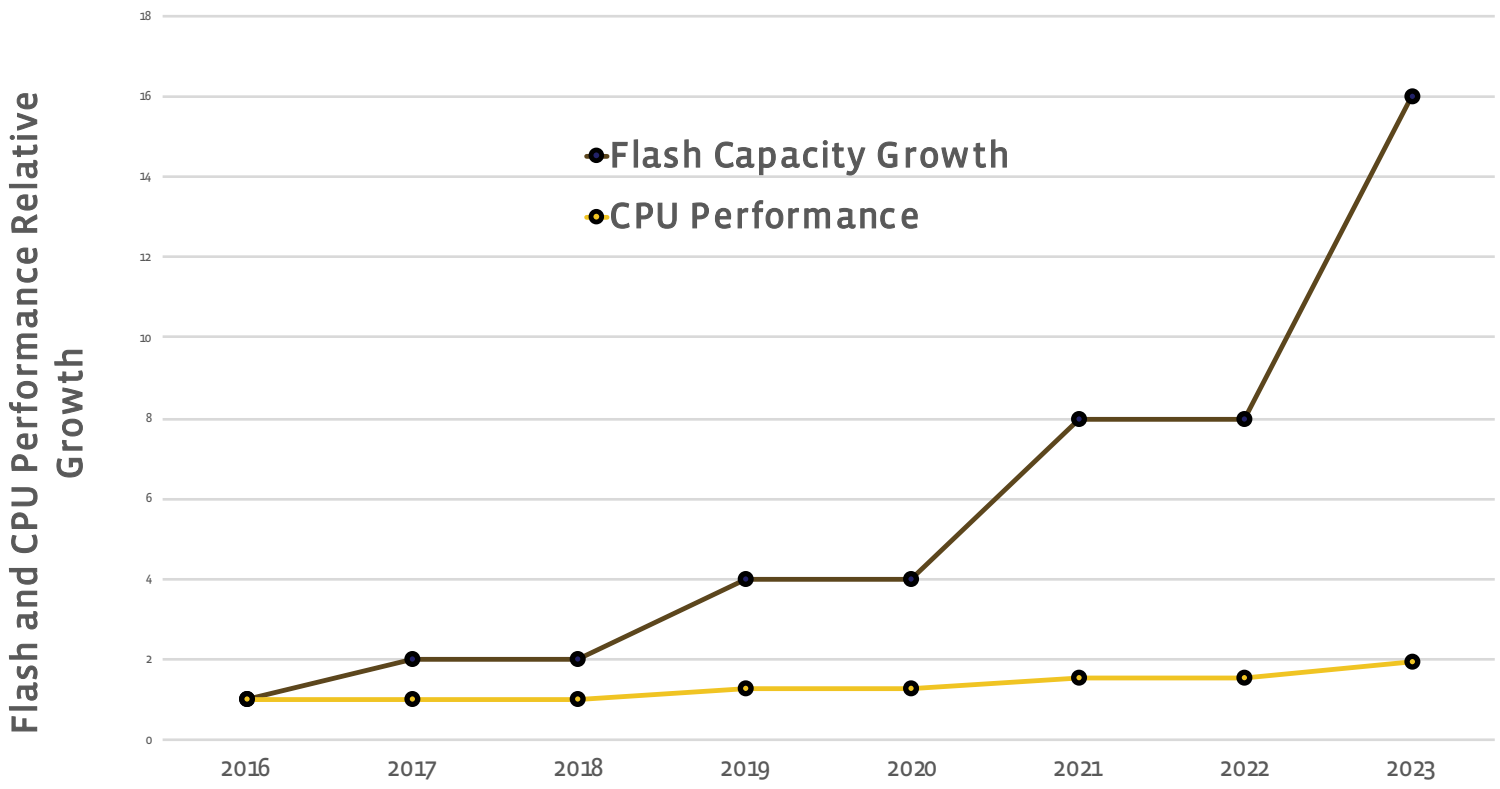
- Read latency outliers are caused by “collisions” with
  - Concurrent flash writes
  - Flash background operations:
    - Garbage collection
    - Wear leveling
    - Read scrub
    - Block erase
- Error correction
- Exception handling (e.g. program/erase failures)

Read time	~60– 100us
Program time	~1 – 1.5ms
Erase time	~10 – 15ms

Flash Industry Trends



Flash Industry Trends



Solution: NVMe® I/O Determinism

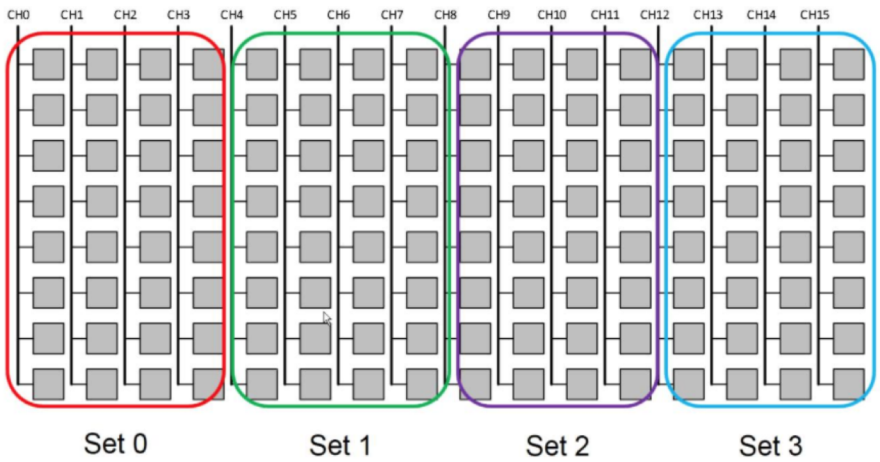
NVMe standards have been ratified!

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

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

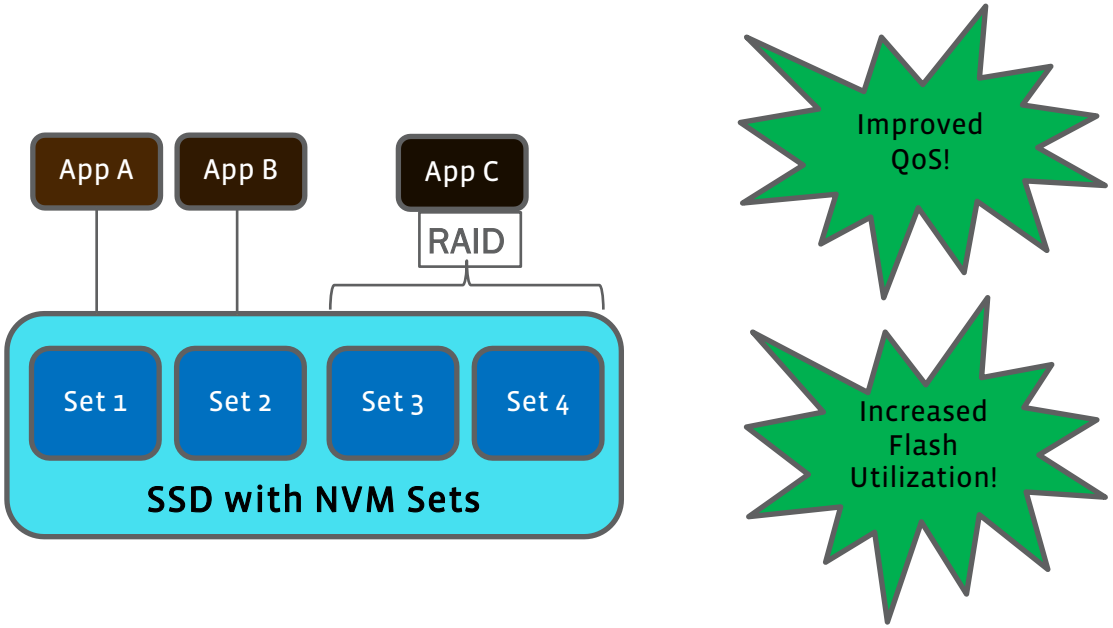


Stripe *within* channels like this:

Benefits

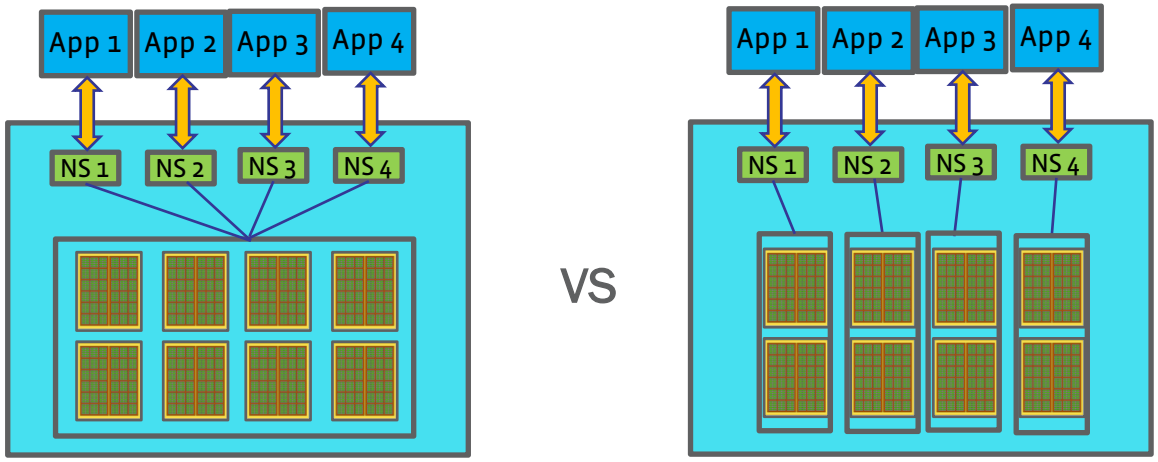
- Enables QoS Regions at the SSD level
  - Better support of multi-tenants on an SSD
- Host software can leverage sets as-is
  - Part of the NVMe Standard
  - Sets are exported as namespaces
  - Host OS does NOT need to be sets-aware

Use Cases @ Facebook



Aligns with Facebook's Disaggregated Flash Strategy!

Evaluation Setup



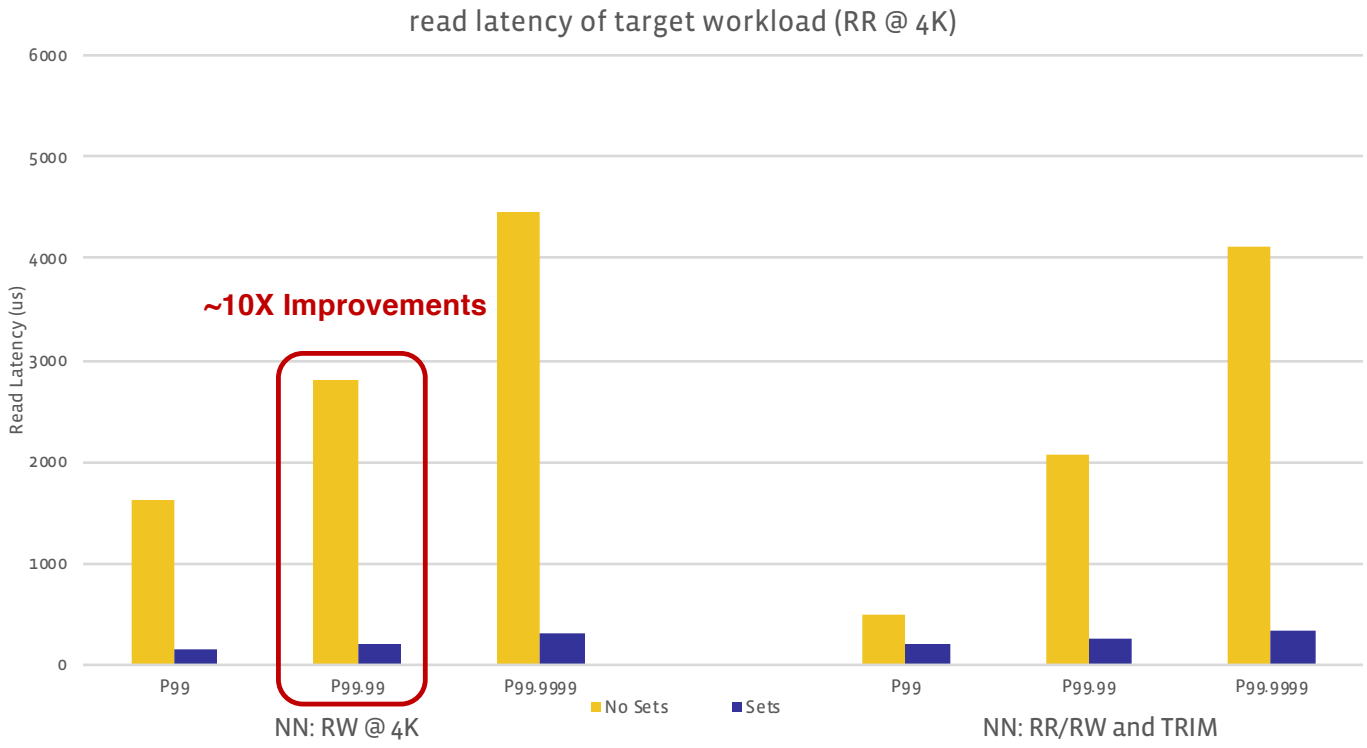
SSD with 4 Namespaces (No Sets)

SSD with 4 Sets

Noisy Neighbor Workload Patterns

- One namespace runs the target workload (NS1)
- The rest of three namespaces act as noisy neighbors (NS2-4)

Silencing the Noisy Neighbors



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

NVMe® I/O Determinism Predictable Latency Mode

Effects of Internal Activities on Latency

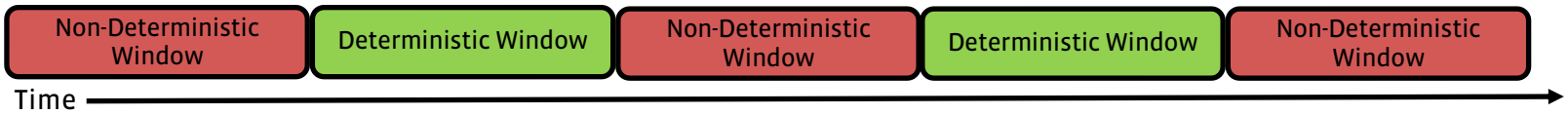
- Internal operations account for most of NAND activity:
  - 7% OP results in 14.3x worst-case WA
  - 28% OP results in 3.6x worst-case WA
- Internal operations usually happen in batches
- Scheduling of internal operations is a black box to the host

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

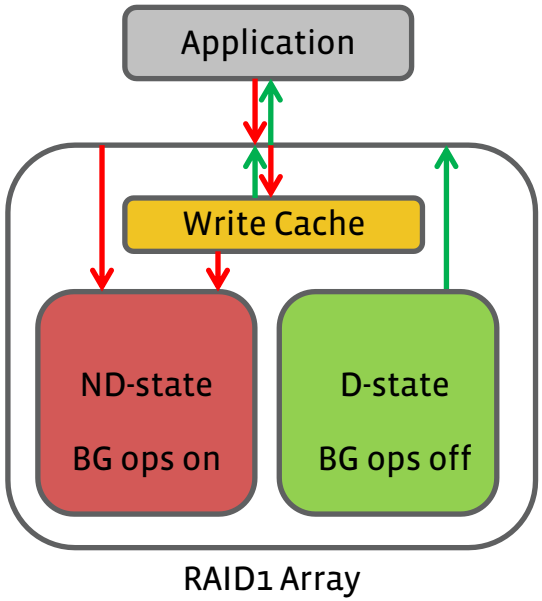
NVMe PLM: Prototype

Goals:

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

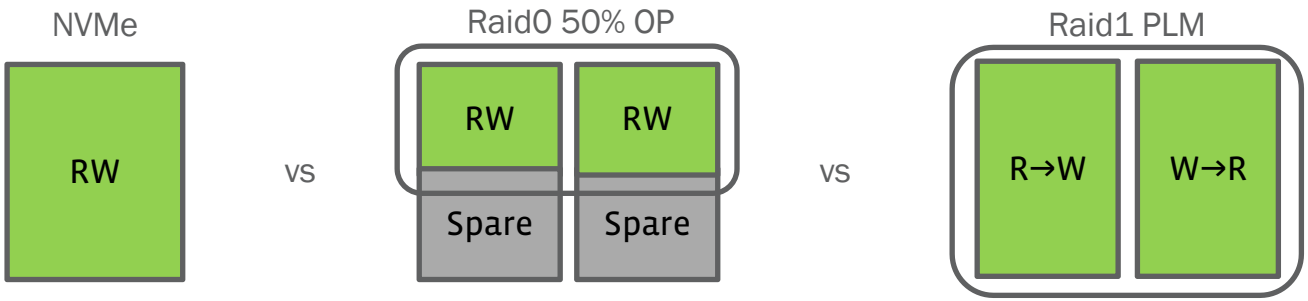
Approach:

- Leverage data redundancy & PLM to segregate reads from other operations

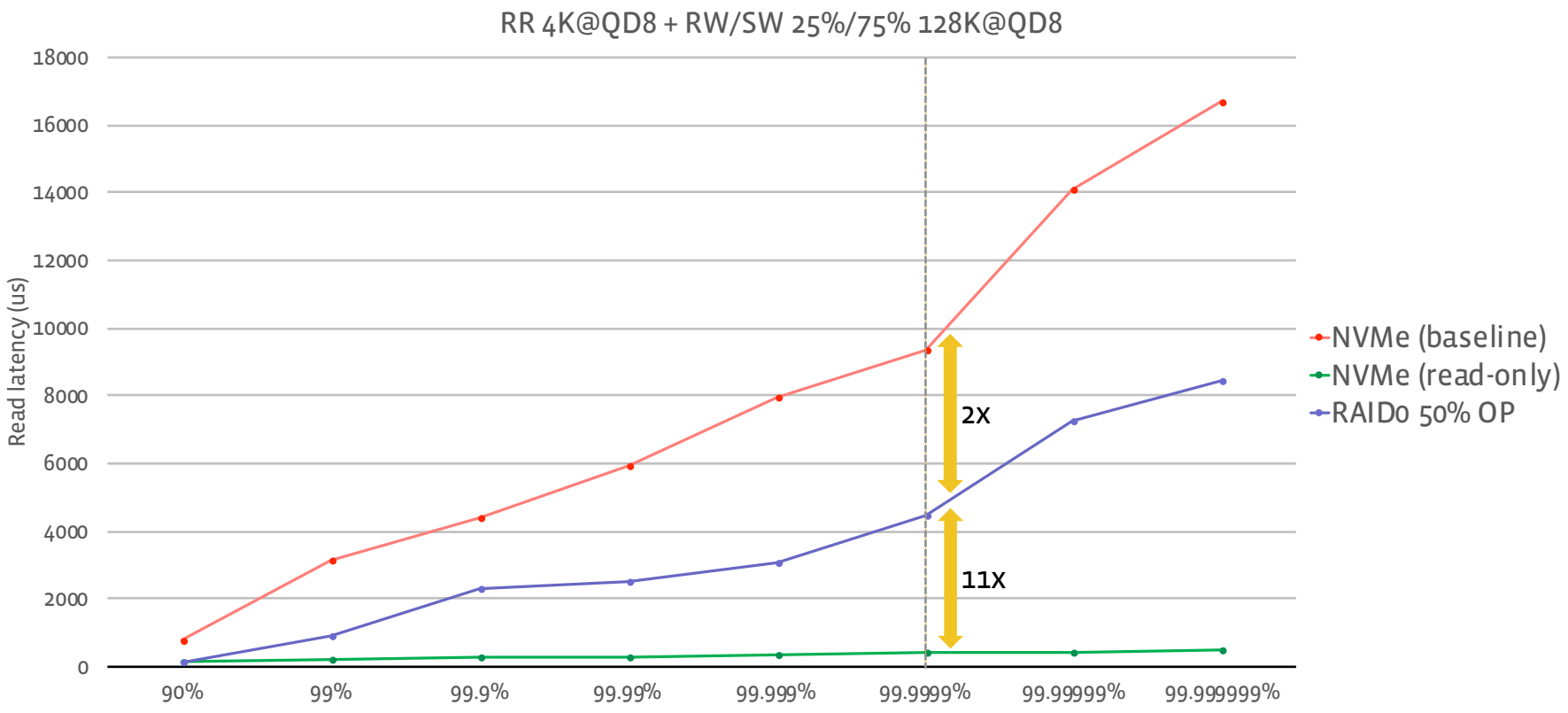


NVMe PLM: Test setup

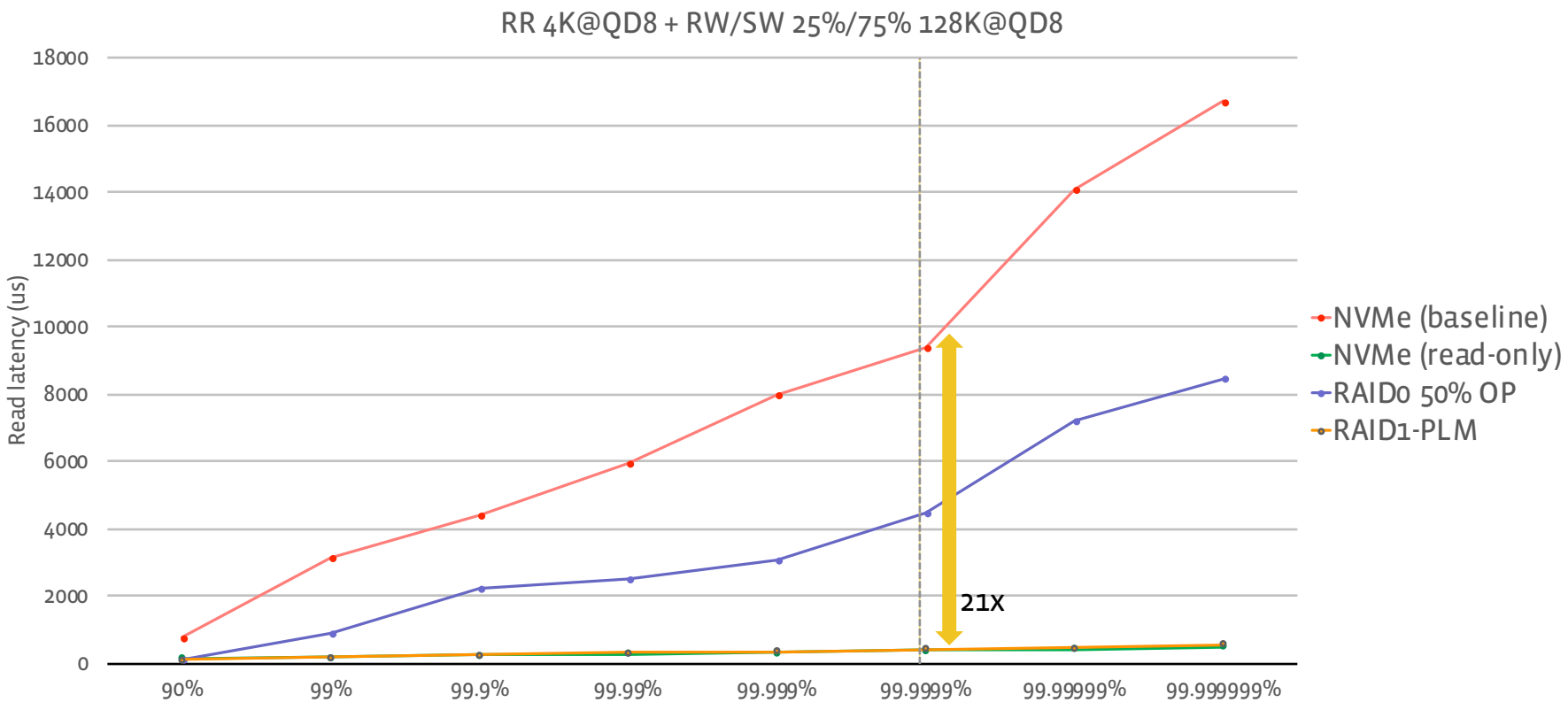
- Same usable capacity, read & write rate
- Random read 4K @ QD8
- 1:2 mix of random and sequential writes 128K @ QD8
- Initialized with 2 passes of mixed writes



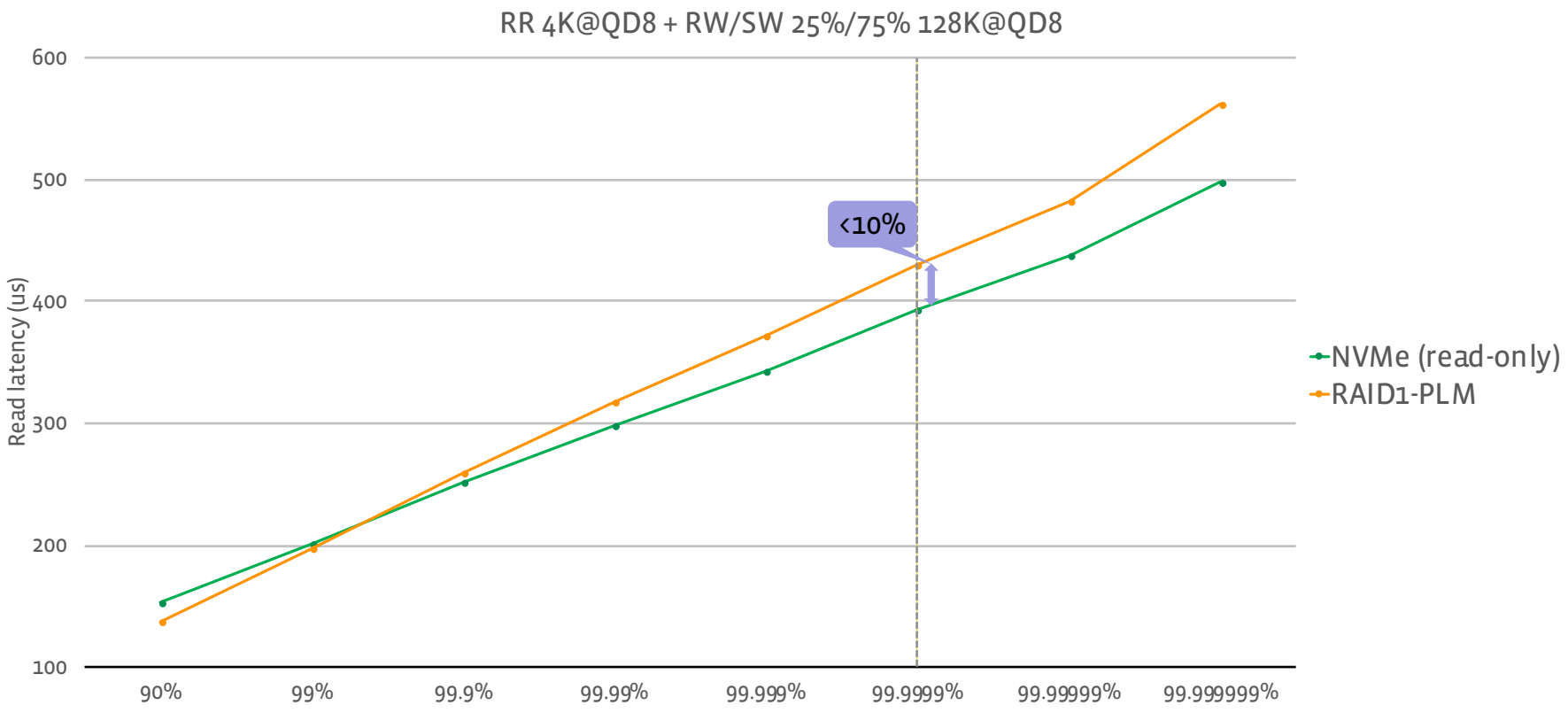
NVMe PLM: Test Results



NVMe PLM: Test Results



NVMe PLM: Test Results



NVMe PLM: Future work

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

OCP SUMMIT

OPEN. FOR BUSINESS.