Western Digital®

From Open-Channel SSDs to Zoned Namespaces

Matias Bjørling

Director, Solid-State System Software

23th January 2019

Forward-Looking Statements

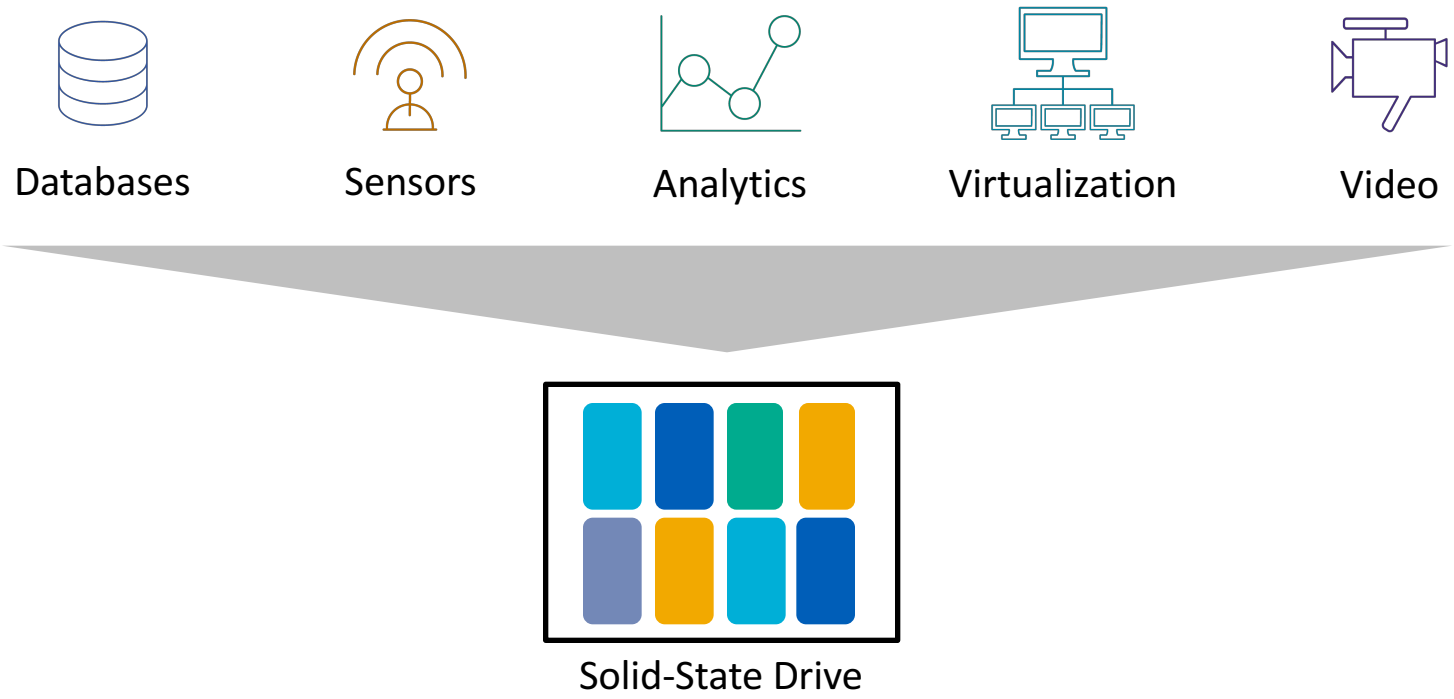
Safe Harbor | Disclaimers

This presentation contains forward-looking statements that involve risks and uncertainties, including, but not limited to, statements regarding our solid-state technologies, product development efforts, software development and potential contributions, growth opportunities, and demand and market trends. Forward-looking statements should not be read as a guarantee of future performance or results, and will not necessarily be accurate indications of the times at, or by, which such performance or results will be achieved, if at all. Forward-looking statements are subject to risks and uncertainties that could cause actual performance or results to differ materially from those expressed in or suggested by the forward-looking statements.

Key risks and uncertainties include volatility in global economic conditions, business conditions and growth in the storage ecosystem, impact of competitive products and pricing, market acceptance and cost of commodity materials and specialized product components, actions by competitors, unexpected advances in competing technologies, difficulties or delays in manufacturing, and other risks and uncertainties listed in the company's filings with the Securities and Exchange Commission (the "SEC") and available on the SEC's website at www.sec.gov, including our most recently filed periodic report, to which your attention is directed. We do not undertake any obligation to publicly update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as required by law.

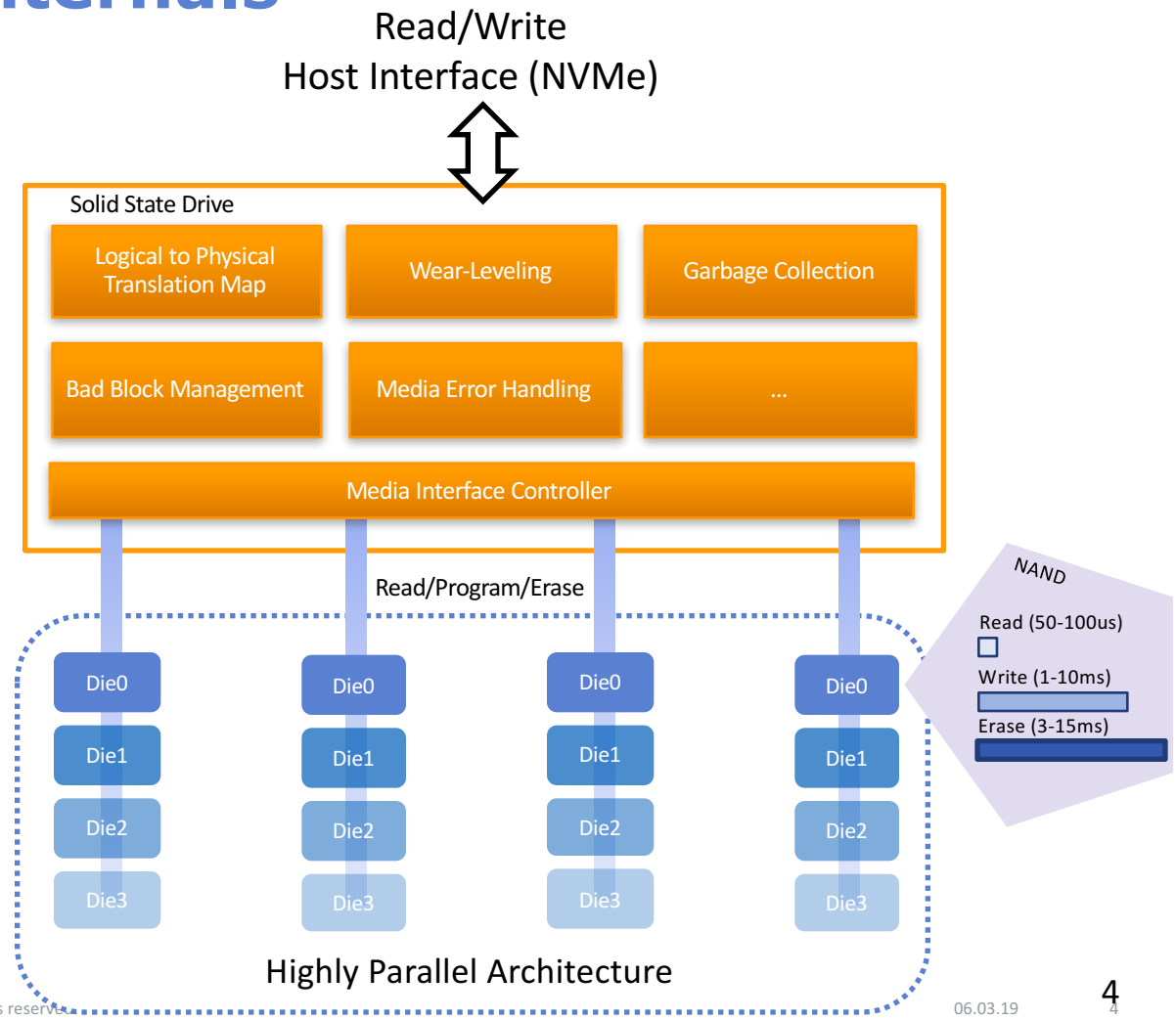
Ubiquitous Workloads

Efficiency of the Cloud requires many different workloads to a single SSD



Solid State Drive Internals

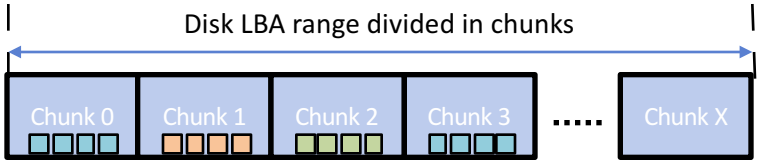
- NAND Read/Program/Erase
- Highly Parallel Architecture
  - Tens of Dies
- NAND Access Latencies
- Translation Layer
  - Logical to Physical Translation Layer
  - Wear-leveling
  - Garbage Collection
  - Bad block management
  - Media error handling
  - Etc.
- Read/Write/Erase -> Read/Write



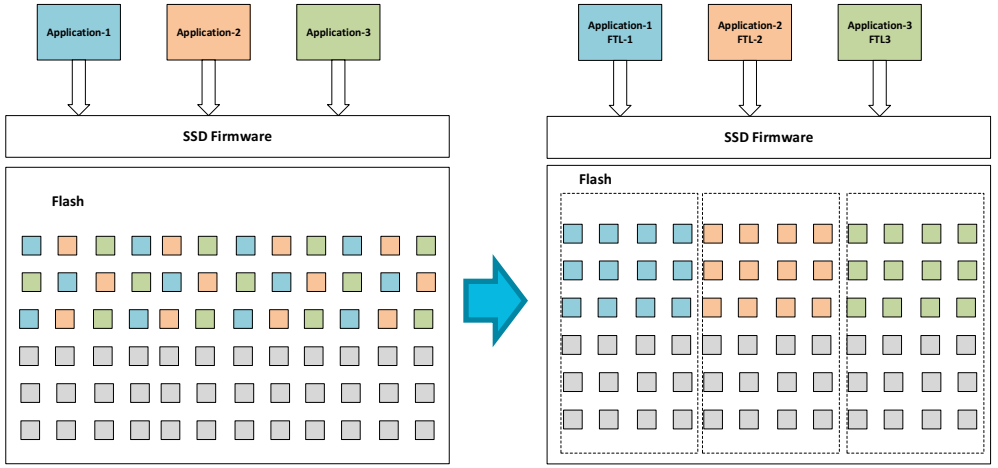
Open-Channel Concepts

Chunks & Parallel Units

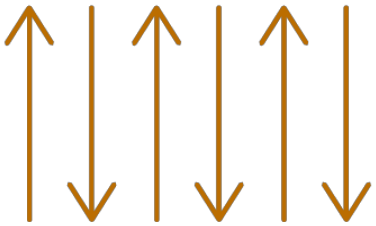
- Chunks
  - Write sequentially within an LBA range
  - Requires reset for rewrites
  - Borrows from HDD's SMR specification (ZAC/ZBC)
  - Optimized for SSD physical constraints
    - Align writes to media layout



- Parallel Units
  - Host can direct I/Os to separate workloads
  - Stripes across single or multiple dies.
  - The parallel units inherits the throughput and latency characteristics of the underlying media
  - Similar concept to I/O Determinism in NVMe

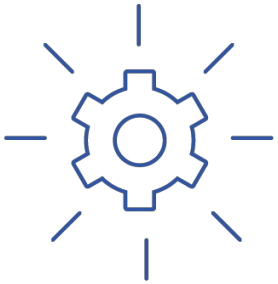


Open-Channel SSDs



I/O Isolation

Predictable Latency



Data Placement & I/O Scheduling

Open-Channel SSDs

Interface Adoption

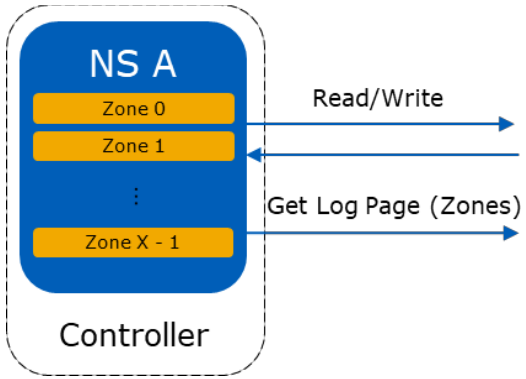
Open-Channel SSD architectures are being adopted by major hyper-scalers

Concepts ready to be part of the NVMe standard specification

NVMe Approach

Drive features into NVMe which address key OCSSD use cases

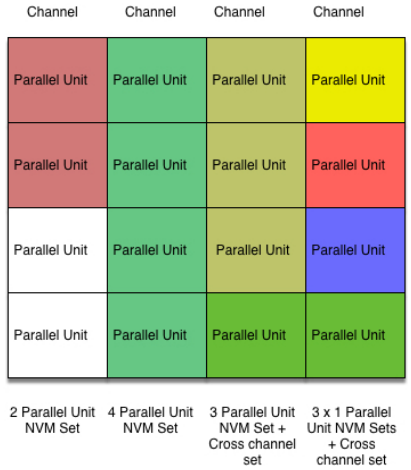
Zoned Namespaces (e.g. Chunks)



Today's talk

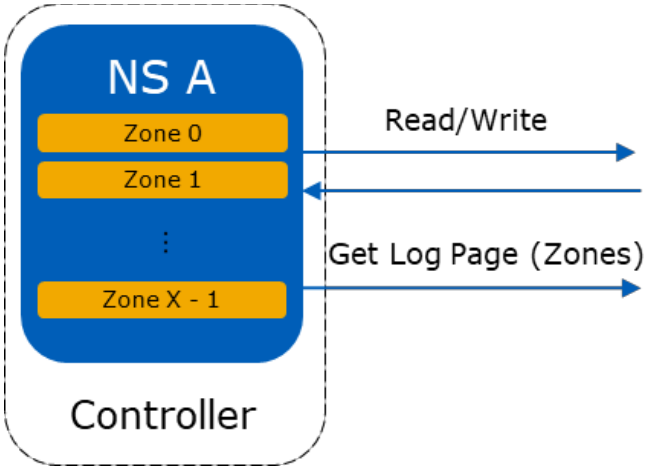
Endurance Group Mgmt (e.g. Parallel Units)

- Explicit Provisioning of Device Topology
- Build on IO Determinism



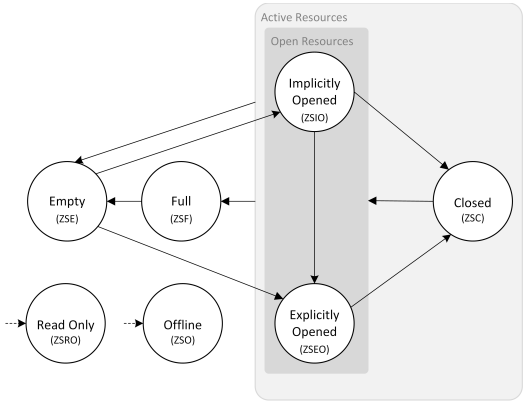
Zoned Namespaces (ZNS)

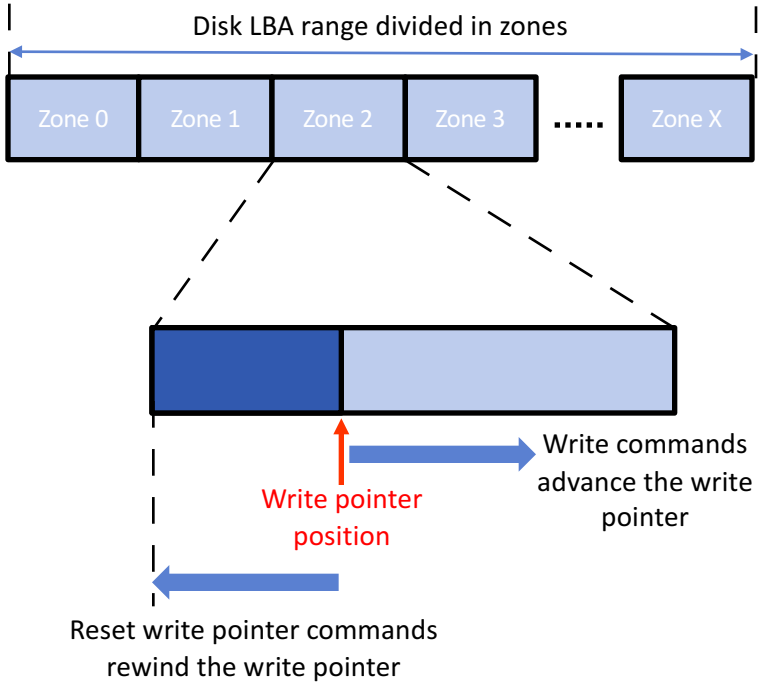
- **Technical Proposal in the NVMe working group**
- **Standardizes zone interface as an approach to:**
  - Reduce device-side write amplification
  - Reduce over-provisioning
    - *"Note that excessive over-provisioning is similar to early replacement -- in both cases you buy more devices."*
      - Mark Callaghan, Facebook
  - Reduce DRAM in SSDs
    - Highest cost after NAND itself
  - Improve latency outliers and throughput
    - Less device-side data movement
    - Tail at scale
  - Enable software eco-system.
    - Everyone benefits from improvements!



Zoned Namespaces similar to ZBC/ZAC for SMR HDDs

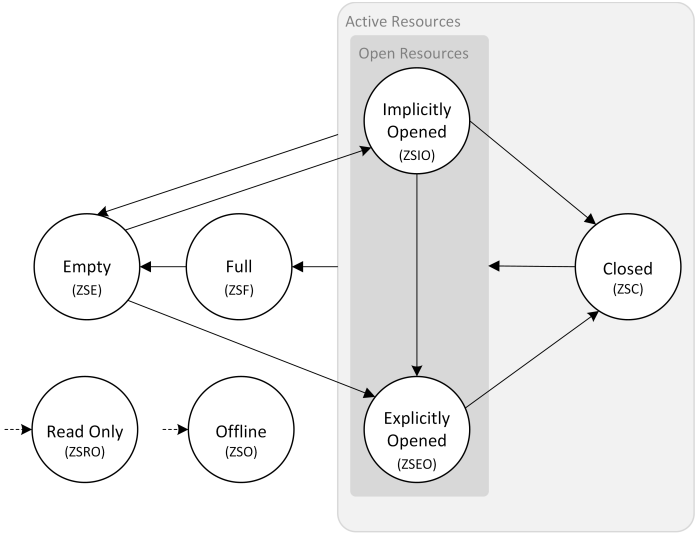
- Storage capacity is divided into zones
- Each zone is written sequentially
- Interface optimized for SSDs
  - Align with media characteristics
    - Zone size aligned to media (E.g., NAND block sizes)
    - Zone capacity aligned to physical media sizes
  - Reduce NAND media erase cycles (Write amp.)

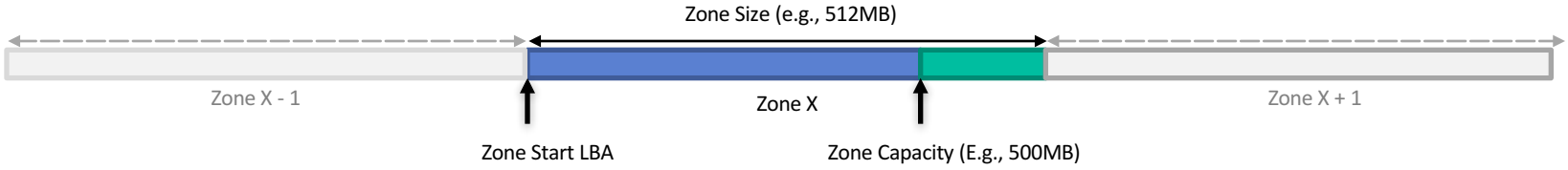




Zone Information

- Zone State
  - Empty, Implicitly Opened, Explicitly Opened, Closed, Full, Read Only, Offline
  - Empty -> Open -> Full -> Empty -> ….
- Zone Reset
  - Full -> Empty
- Zone Size & Zone Capacity
  - Zone Size is fixed
  - Zone Capacity is the writeable area within a zone



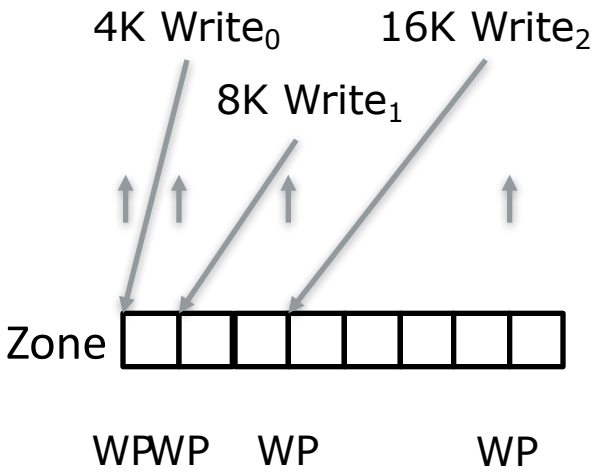


Zone Append

- Low scalability on multiple writers to a zone
  - Write Queue Depth per Zone = 1
  - IOPS: *80K vs 880K on Qemu* and *300K vs 1400K on metal*
- ZAC/ZBC requires strict write ordering
- Limits write performance and increases host overhead
- Big challenge with software eco-system, HBAs, etc.
- Introducing Zone Append
- Append data to a zone without defining offset
  - Drive returns where data was written in the zone

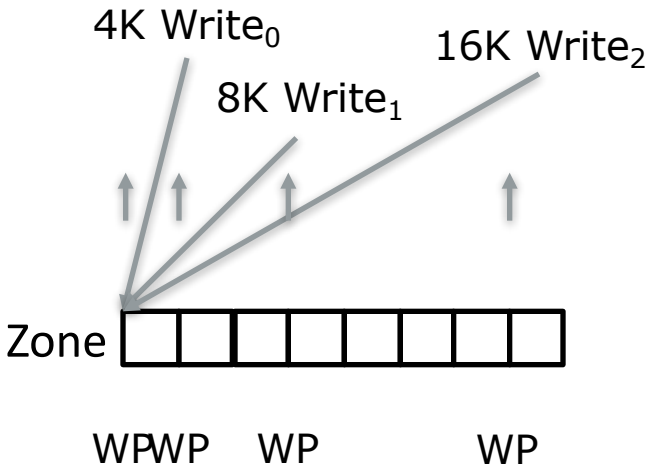
Zone Write Example

3x Writes (4K, 8K, 16K) – Queue Depth = 1



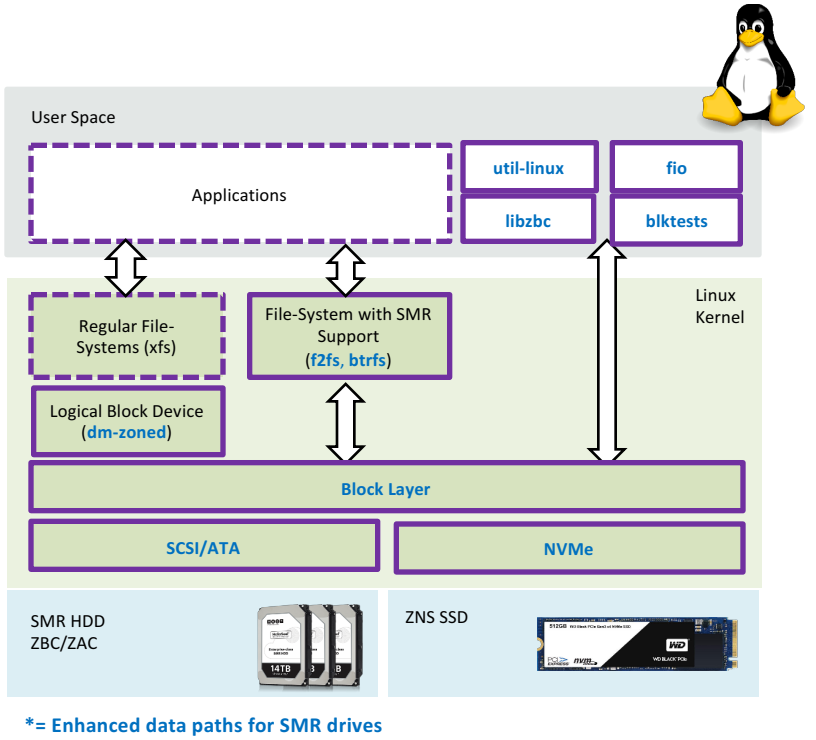
Zone Append Example

3x Writes (4K, 8K, 16K) – Queue Depth = 3



ZNS: Synergy w/ ZAC/ZBC software ecosystem

- Existing ZAC/ZBC-aware file systems & device mappers “just work”
  - Few changes to support to ZNS
- Reuse existing work already applied with for ZAC/ZBC hard-drives (SMR)
- Integrate directly with file-systems
  - No host-side FTL
  - No 1GB DRAM per 1TB Media requirement
  - Better utilization of SSD
- Code is already in production at hyper-scalers and available in the Linux eco-system.



*= Enhanced data paths for SMR drives

Mature Software Stack

ZNS to use existing storage stack

- User-space libraries
  - Libzbd
  - Nvme-cli
  - Blktests
  - Util-linux (blkzone)
  - fio
  - libzns
- Kernel-space libraries
  - NVMe support for Zones
  - XFS, Btrfs, F2FS, dm-zoned, etc…
- Qemu with ZNS support

ZBC State Model

- States (Empty, Implicitly Opened, Explicitly Opened, Full, Offline)
- Each zone
  - Size
  - Capacity
  - Write Pointer
  - And other fields
- Zone Management command
  - ZM Open, ZM Close, ZM Reset, ZM Finish