Introduction to Open-Channel Solid State Drives and What's Next!

Matias Bjørling
Director, Solid-State System Software

September 25rd, 2018

Storage Developer Conference 2018, Santa Clara, CA
©2018 Western Digital Corporation or its affiliates. All rights reserved.

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.

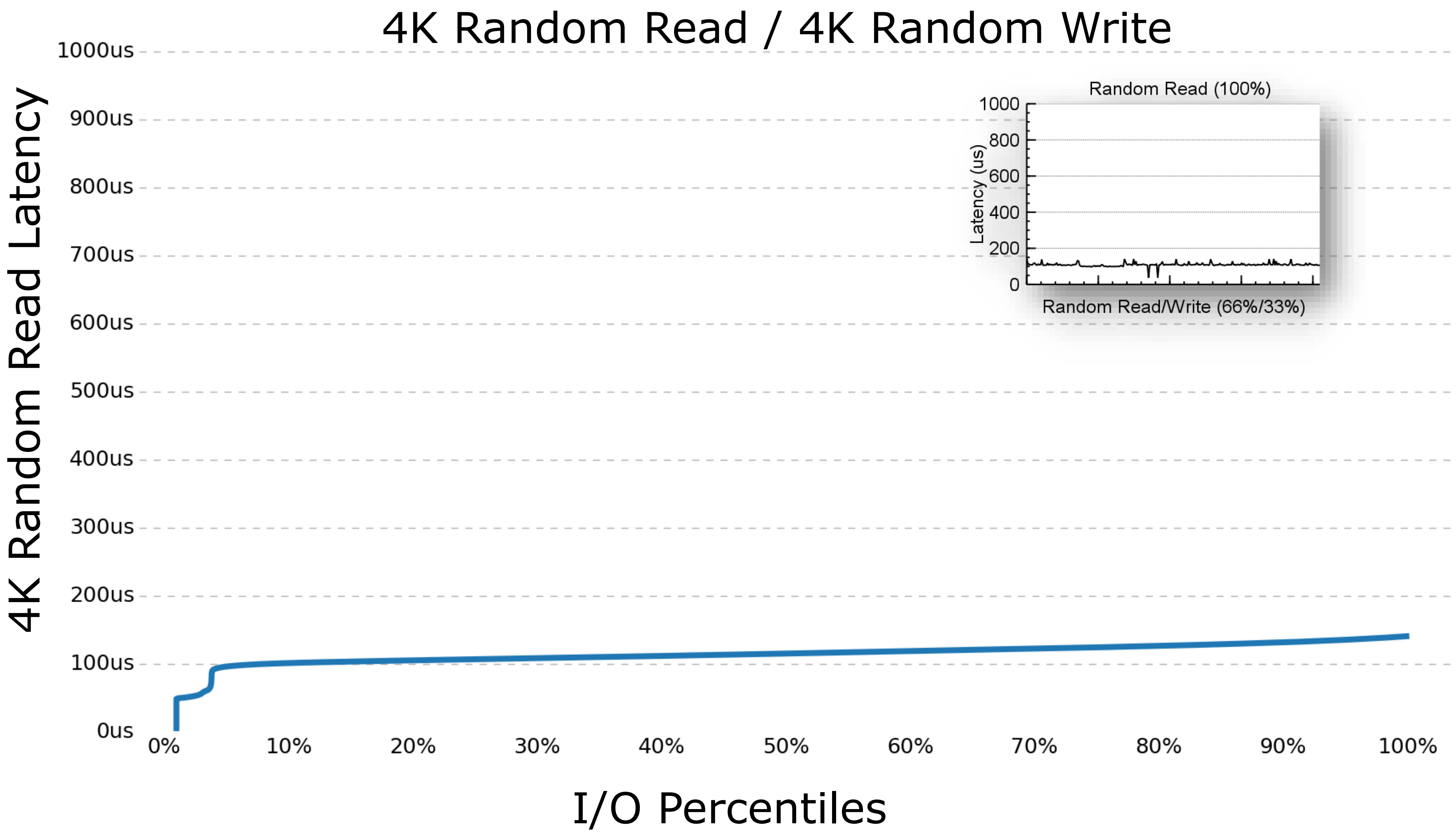
©2018 Western Digital Corporation or its affiliates. All rights reserved.

Agenda

1. Motivation
2. Interface
3. Eco-system
4. What’s Next? Standardization

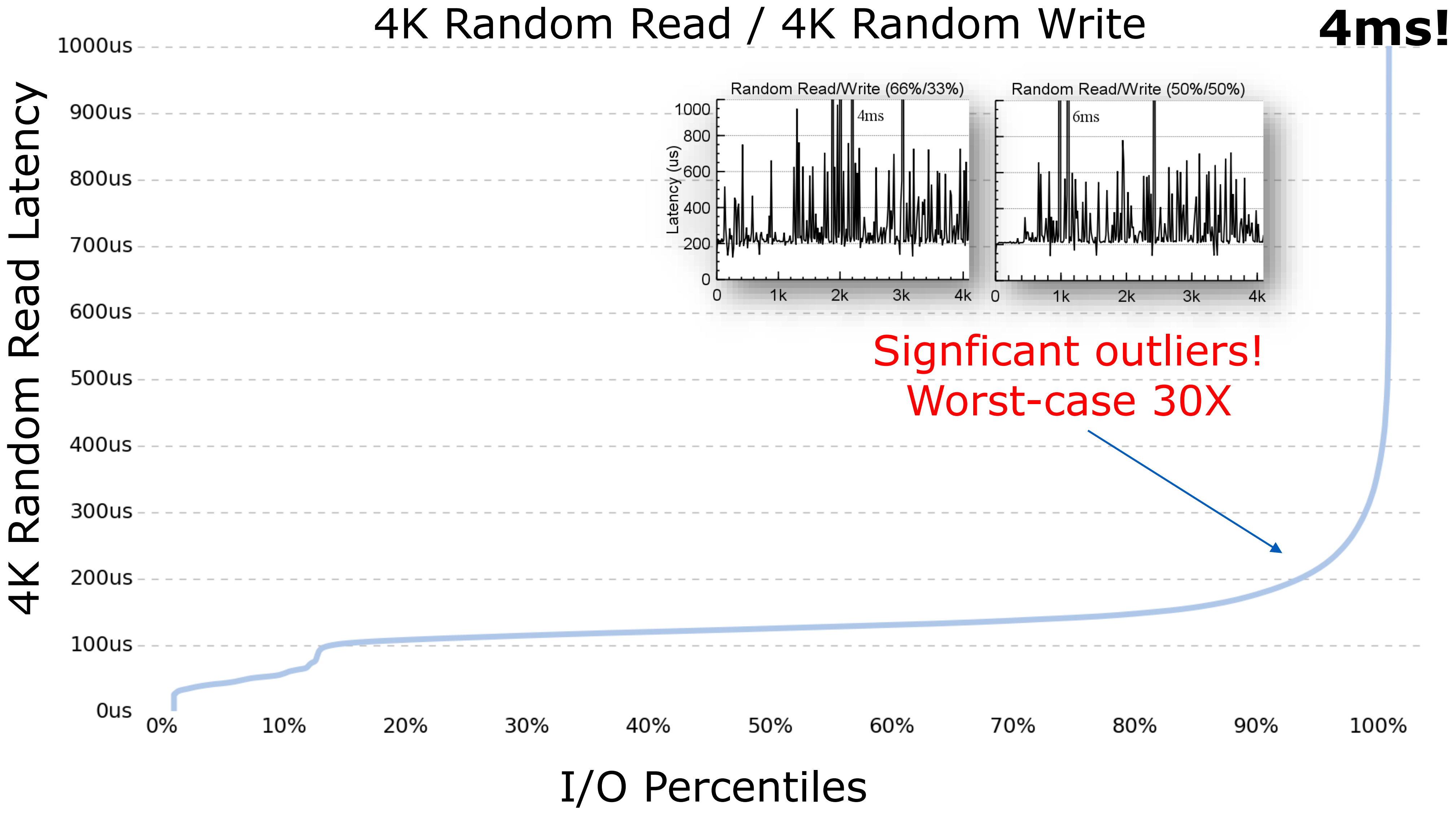
©2018 Western Digital Corporation or its affiliates. All rights reserved.

0% Writes - Read Latency



©2018 Western Digital Corporation or its affiliates. All rights reserved.

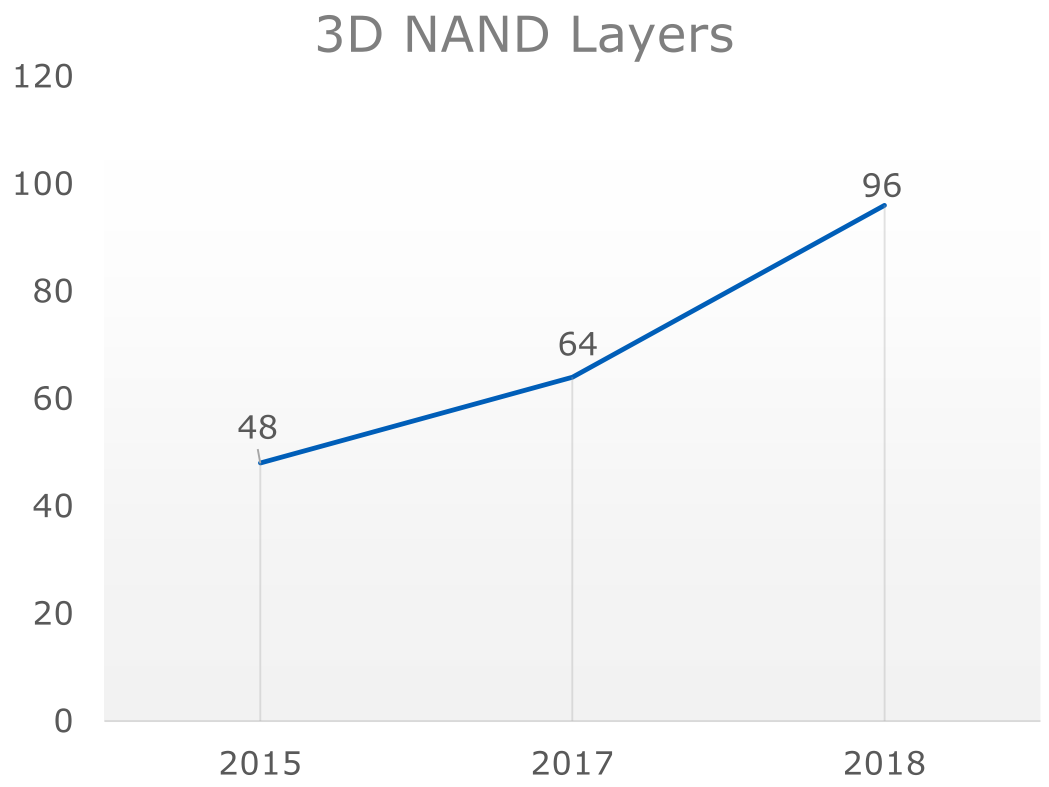
20% Writes - Read Latency

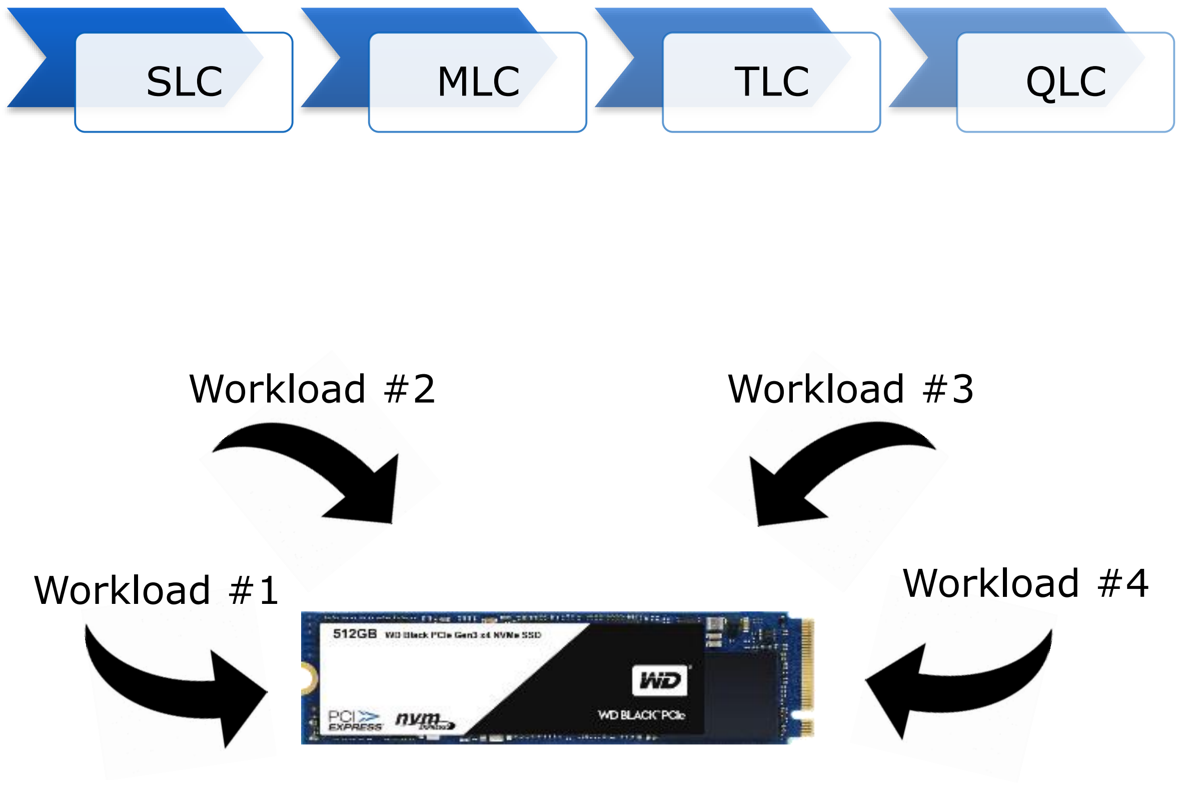


©2018 Western Digital Corporation or its affiliates. All rights reserved.

NAND Chip Density Continues to Grow

While Cost/GB decreases

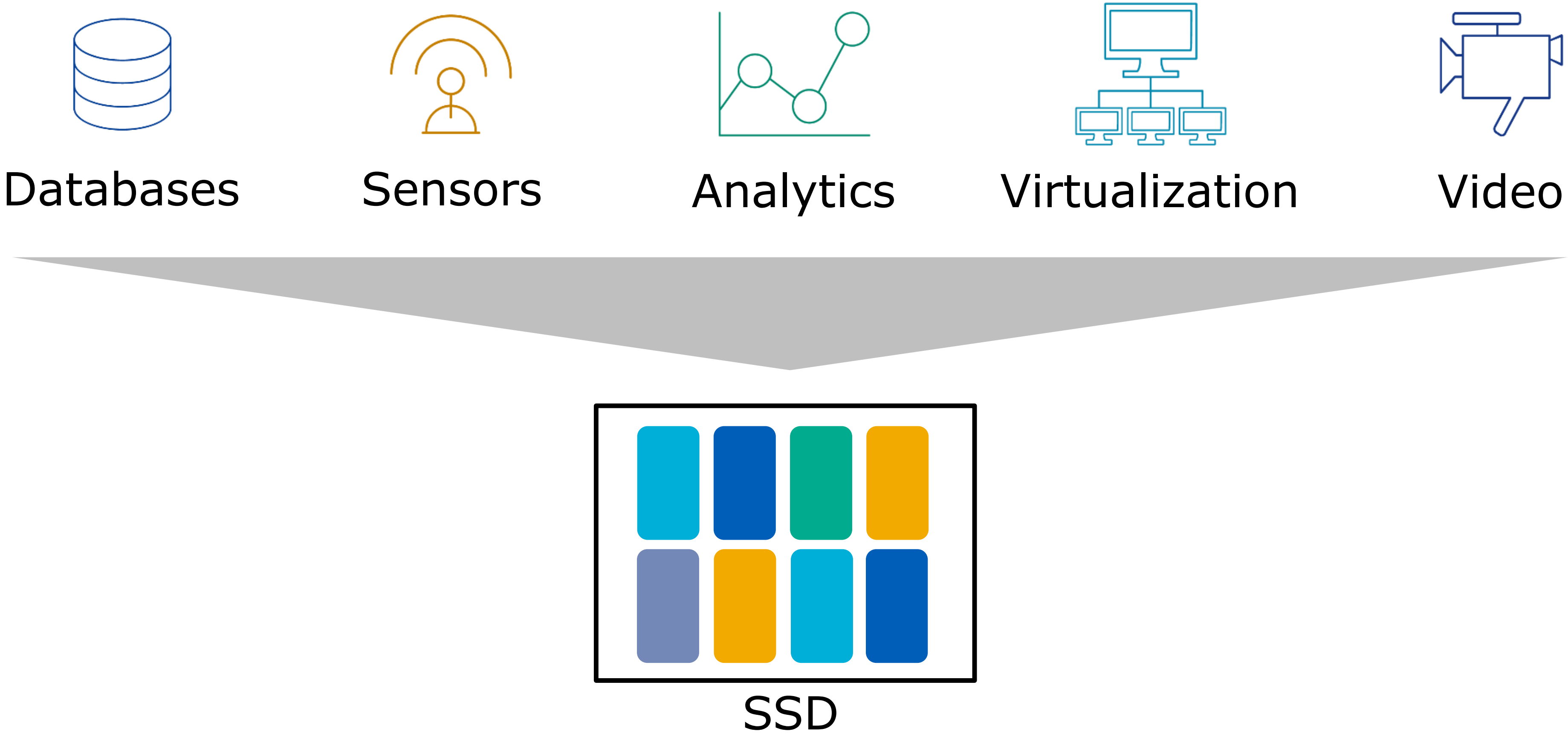




©2018 Western Digital Corporation or its affiliates. All rights reserved.

Ubiquitous Workloads

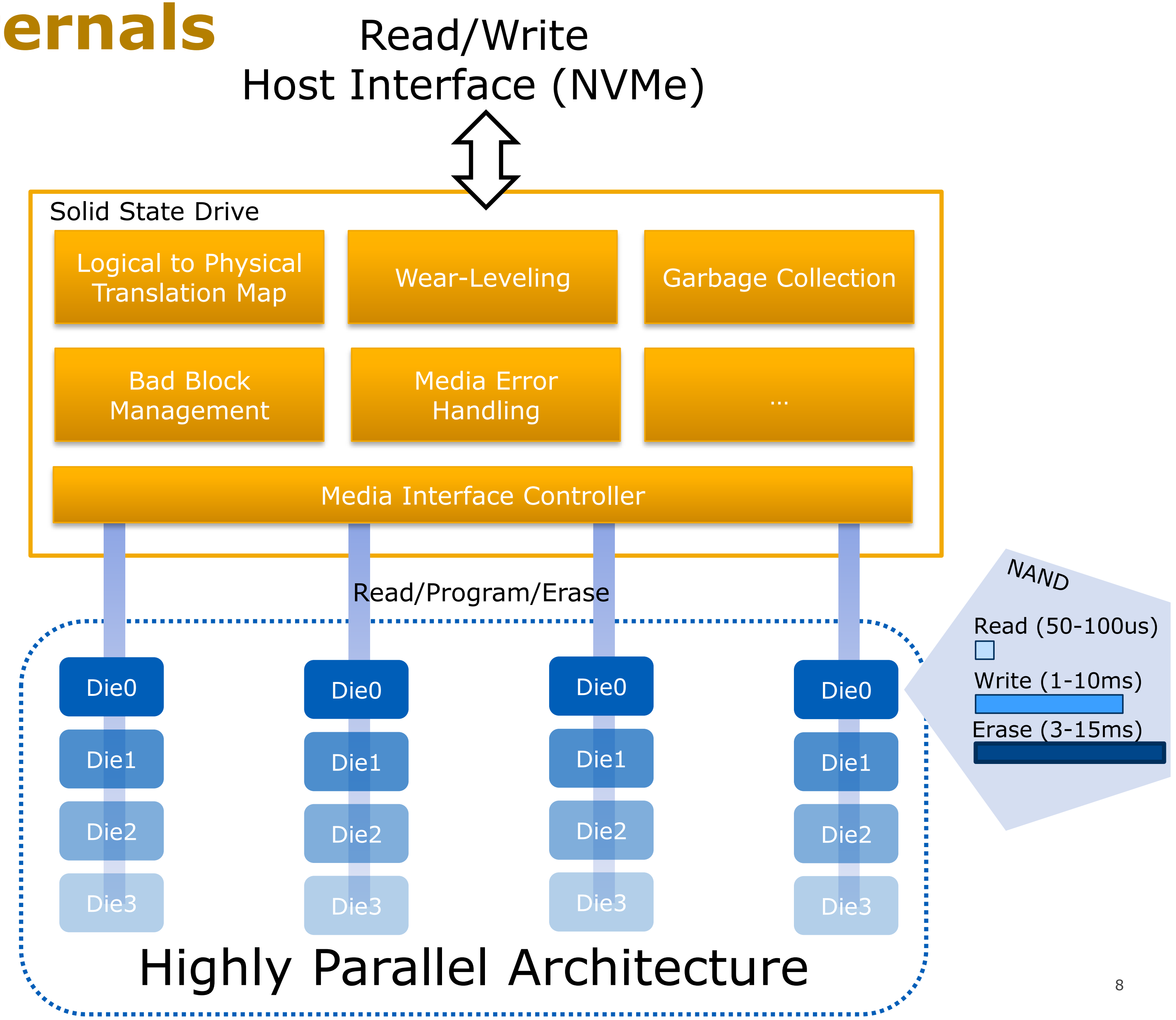
Efficiency of the Cloud requires many different workloads of a single SSD



©2018 Western Digital Corporation or its affiliates. All rights reserved.

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

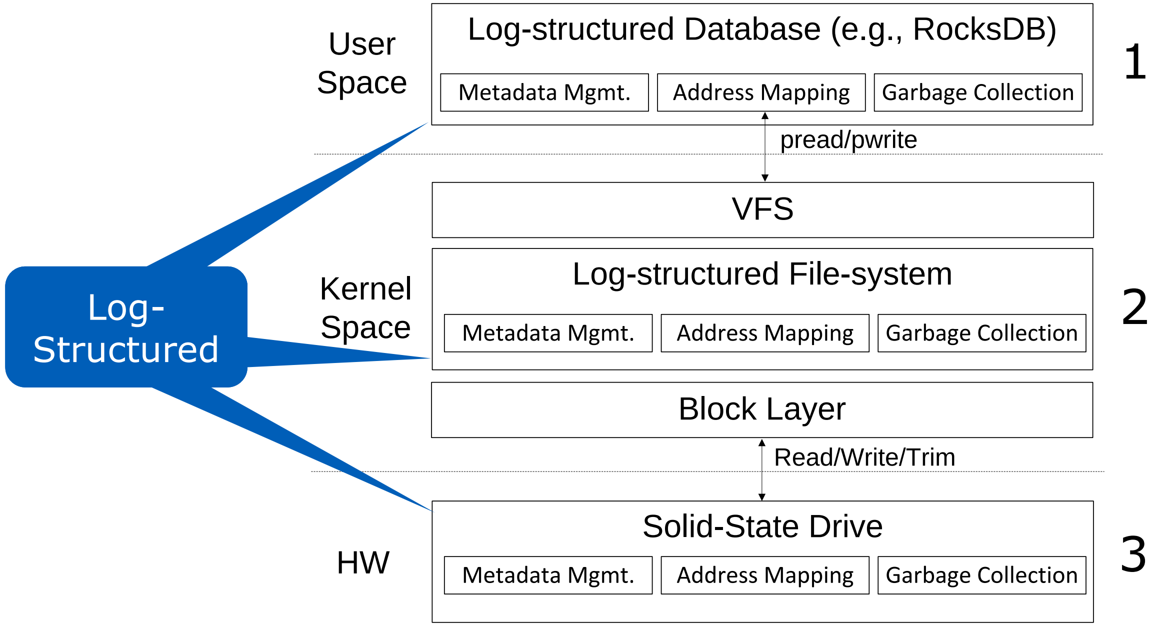


©2018 Western Digital Corporation or its affiliates. All rights reserved.

Single-User Workloads

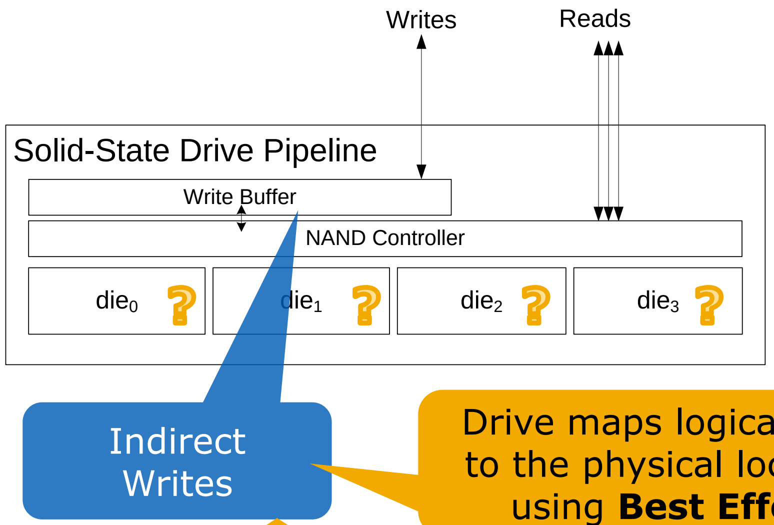
Indirection *and* ***Indirect Writes*** *causes outliers*

Host: Log-on-Log



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

Device: Indirect Writes



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

Host is oblivious to data placement due to the indirection

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Open-Channel SSDs

I/O Isolation

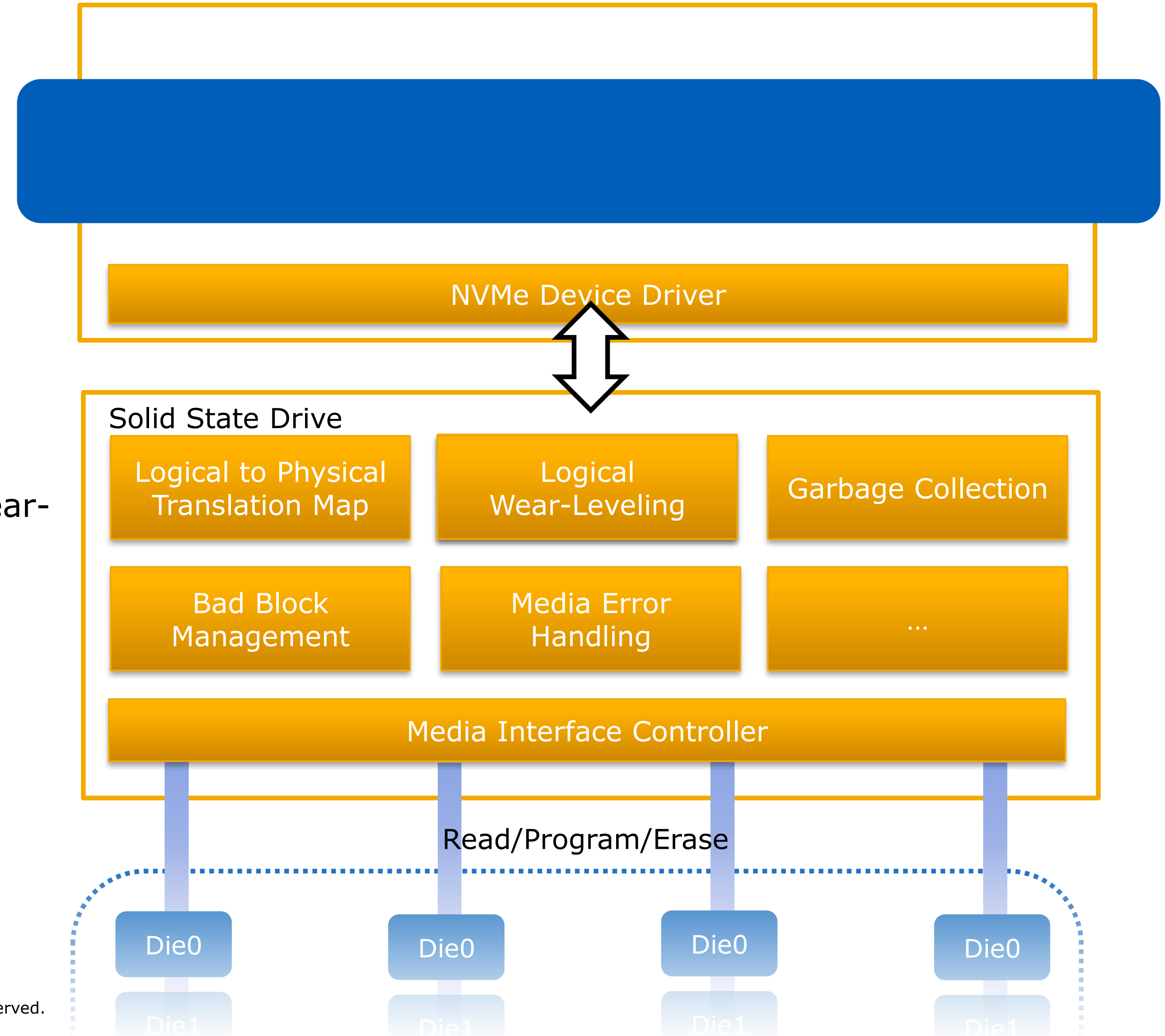
Predictable Latency

Data Placement &
I/O Scheduling

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Solid State Drive Internals

- Host Responsibility
  - Logical to Physical Translation Map
  - Garbage Collection
  - Logical Wear-leveling
    - Hint to host to place hot/cold data
- Integration
  - Block device
    - Host-side FTL that does L2P, GC, Logical Wear-leveling
    - Similar overhead to traditional SSDs
  - Applications
    - Databases and File-systems



©2018 Western Digital Corporation or its affiliates. All rights reserved.

Concepts in an Open-Channel SSD

Interface Blocks

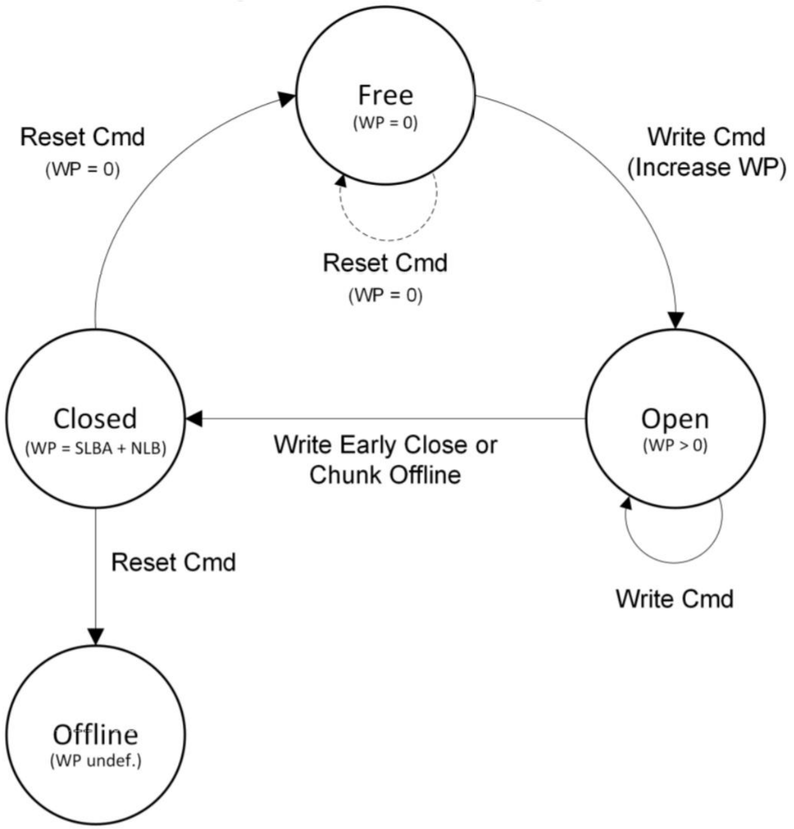
- Chunks
  - Sequential write only LBA ranges
  - Align writes to internal block sizes
- Hierarchical addressing
  - A sparse addressing scheme projected onto the NVMe™ LBA address space
- Host-assisted Media Refresh
  - Improve I/O predictability
- Host-assisted Wear-leveling
  - Improve wear-leveling

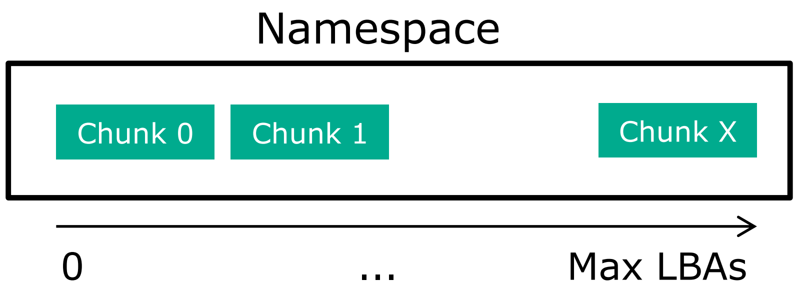
©2018 Western Digital Corporation or its affiliates. All rights reserved.

Chunks #1

Enable orders of magnitude reduction of device-side DRAM

- A chunk is a range of LBAs where writes must be sequential.
  - Reduces DRAM for L2P table by orders of magnitude
  - Hot/Cold data separation
- Rewrite requires a 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.
- Same device model as the ZAC/ZBC standards.
- Similar device model to be standardized in NVMe (I'll come back to this)

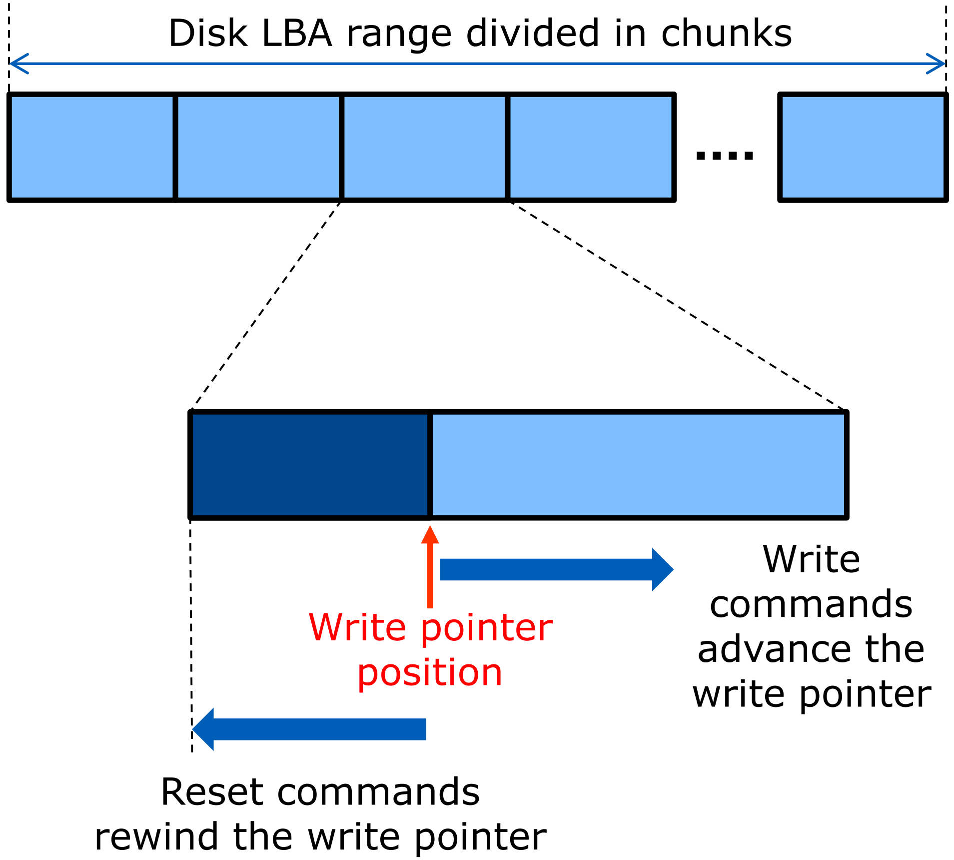




©2018 Western Digital Corporation or its affiliates. All rights reserved.

Chunks #2

- Drive capacity divided into chunks
- Chunk types
  - Conventional
    - Random or Sequential
  - Sequential Write Required
    - Chunk must be written sequential only
    - Must be reset entirely before being rewritten



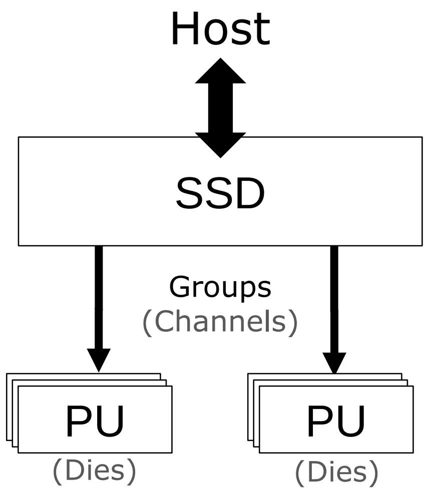
©2018 Western Digital Corporation or its affiliates. All rights reserved.

Hierarchical Addressing

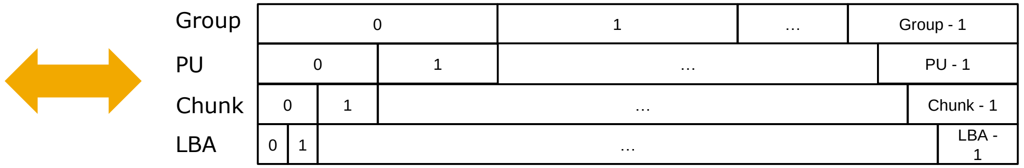
Channels and Dies are mapped to Logical Groups and Parallel Units

- Expose device parallelism through Groups/Parallel Units
  - One or a group of dies are exposed as parallel units to the host
  - Parallel units are a logical representation

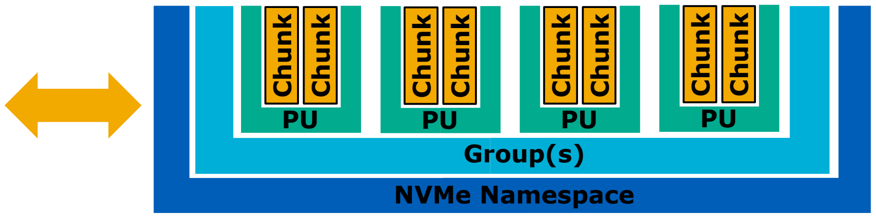
Physical



LBA Address Space



Logical

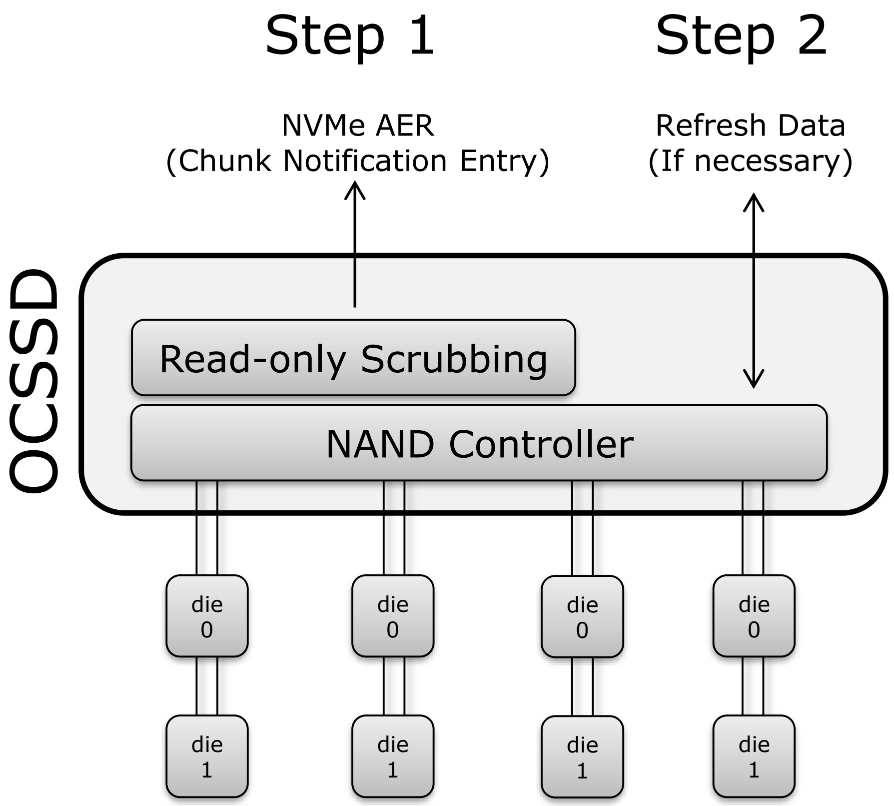


©2018 Western Digital Corporation or its affiliates. All rights reserved.

Host-assisted Media Refresh

Enable host to assist SSD data refresh

- SSDs refreshes its data periodically to maintain reliability. It does this through a data scrubbing process
  - Internal read and writes make the drive I/O latencies unpredictable.
  - Writes dominates I/O outliers
- 2-step Data Refresh
  - Device to only perform the data scrubbing read part - Data movement is managed by host
  - Increases predictability of the drive. Host manages refresh strategy
    - Should it refresh? Is there a copy elsewhere?

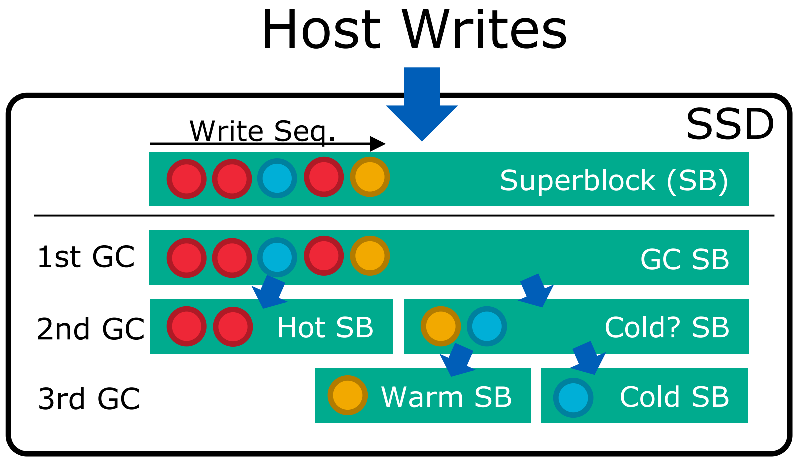


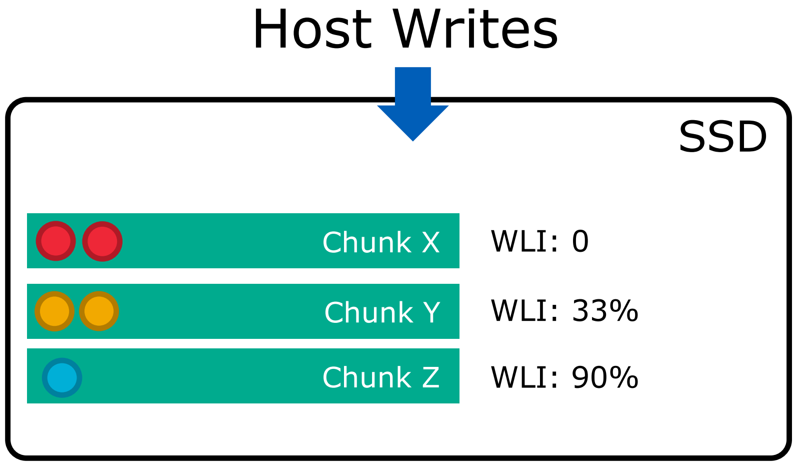
©2018 Western Digital Corporation or its affiliates. All rights reserved.

Host-assisted Wear-Leveling

Enable host to separate Hot/Cold data to Chunks depending on wear

- SSDs typically does not know the temperature of newly written data
  - Placing hot and cold data together increases write amplication
  - Write amplication is often 4-5X for SSDs with no optimizations
- Chunk characteristics
  - Limited reset cycles (as NAND blocks has limited erase cycles)
  - Place cold data on chunks that are nearer end-of-life and use younger chunks for hot data
- Approach
  - Introduce per-chunk relative wear-level indicator (WLI)
  - Host knows workload and places data w.r.t. to WLI
  - Reduces garbage collection→ **Increases lifetime, I/O latency, and performance**



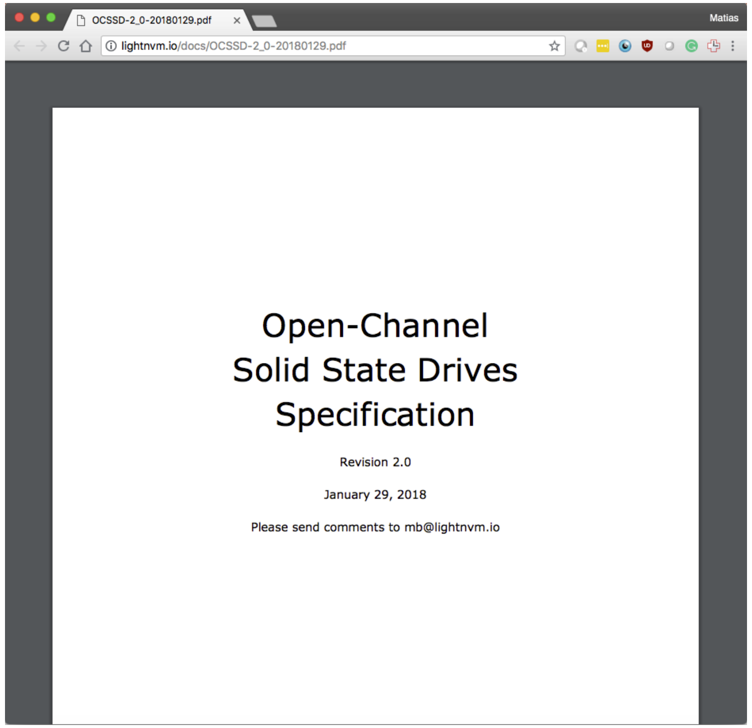


©2018 Western Digital Corporation or its affiliates. All rights reserved.

Interface Summary

The concepts together provide

- **I/O Isolation** through the use of Groups & Parallel Units
- **Fine-grained data refresh** managed by the host
- **Reduce write amplification** by enabling host to place hot/cold data efficiently
- **DRAM & Over-provisioning** reduction through append-only Chunks
- Direct-to-media to **avoid expensive internal data movement**



Specification available at http://lightnvm.io

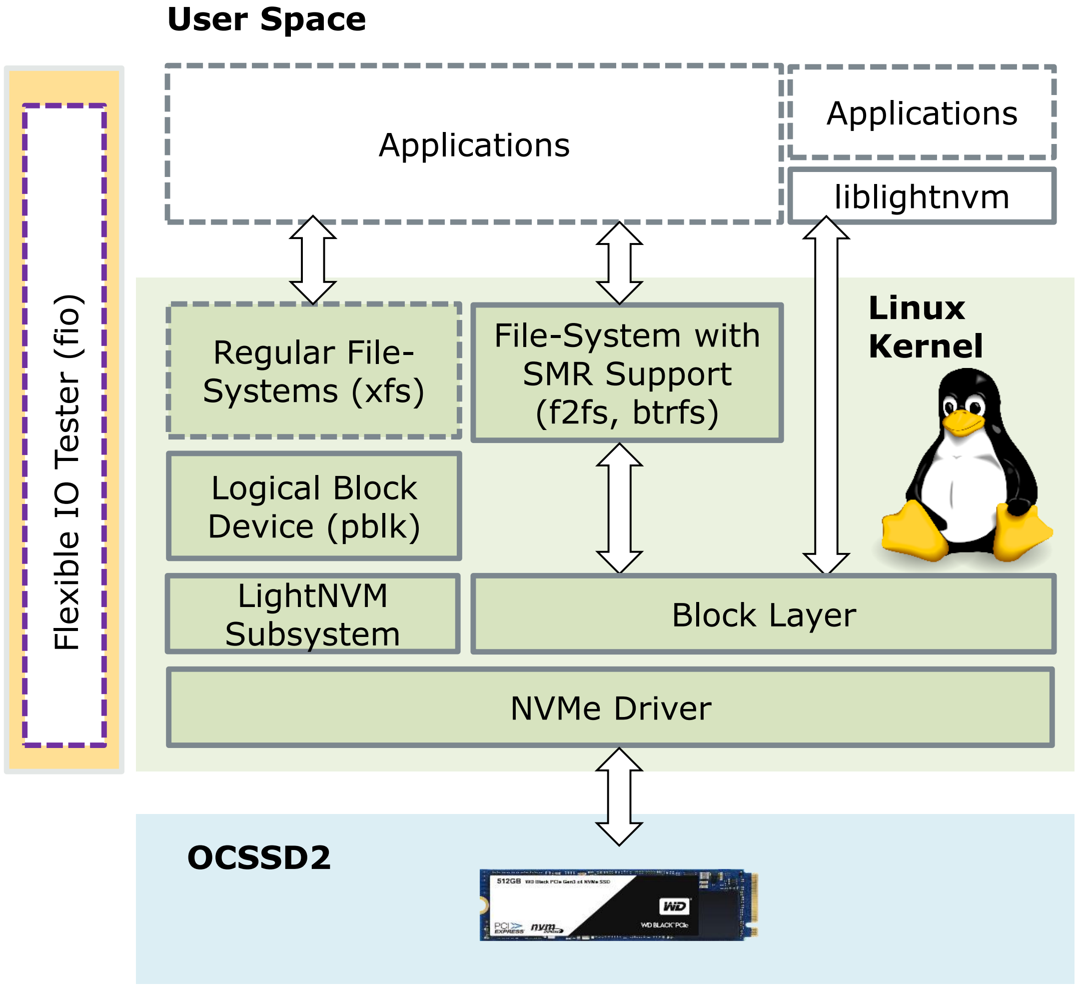
Western Digital

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Eco-system

Large eco-system through Zoned Block Devices and OCSSD

- Linux Kernel®
  - NVMe Device Driver
    - Detection of OCSSDs
    - Support for 1.2 and 2.0 specification
    - Register with LightNVM subsystem
    - Register as a Zoned Block Devices (patches available)
  - LightNVM Subsystem
    - Core functionality
    - Target management
  - Target interface
    - Enumerate, get geometry, I/O interface, etc.
    - pblk host-side FTL – Map OCSSD to Block Device
- User-space
  - libzbc, fio (ZBD support), liblightnvm
  - SPDK



©2018 Western Digital Corporation or its affiliates. All rights reserved.

Open-Source Software Contributions

- 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).
- SPDK Support for OCSSD (June 2018)
- Fio with Zone support (August 2018)
- Upcoming
  - OCSSD as a Zoned Block Device (Patches available)
  - RAIL – XOR support for lower latency
  - 2.0a revision

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Tools and Libraries

LightNVM: The Linux Open-Channel SSD Subsystem

https://www.usenix.org/conference/fast17/technical-sessions/presentation/bjorling

LightNVM

http://lightnvm.io

LightNVM Linux kernel Subsystem

https://github.com/OpenChannelSSD/linux

liblightnvm

https://github.com/OpenChannelSSD/liblightnvm

QEMU NVMe with Open-Channel SSD Support

https://github.com/OpenChannelSSD/qemu-nvme

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Western Digital®

©2018 Western Digital Corporation or its affiliates. All rights reserved.

Western Digital and the Western Digital logo are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries. The NVMe word mark is a trademark of NVM Express, Inc. All other marks are the property of their respective owners.