A Decade of Data Placement

Matias Bjørling, Distinguished Engineer, Western Digital

the Future of Memory and Storage

Western Digital

Exponential Data Growth

- The cloud-native era has intensified demand for high-performance and low-cost storage solutions
- To reduce the $/GB cost, data centers are actively seeking to utilize as much as their available storage capacity,
- As they increase their storage utilization, SSD-based storage increases its internal write amplification, leading to:
  - Excess write activity, primarily due to SSD garbage collection
  - Reduced endurance as more writes wear out SSD's media faster
  - Higher infrastructure cost through increased power consumption

"To achieve these levels of device-level write amplification (1.1x & 1.4x), flash is typically overprovisioned by 50% (…) but reducing flash overprovisioning while maintaining the current level of performance is an open challenge at Facebook."

Source: CacheLib Caching Engine: Design and Experiences at Scale. OSDI 2020

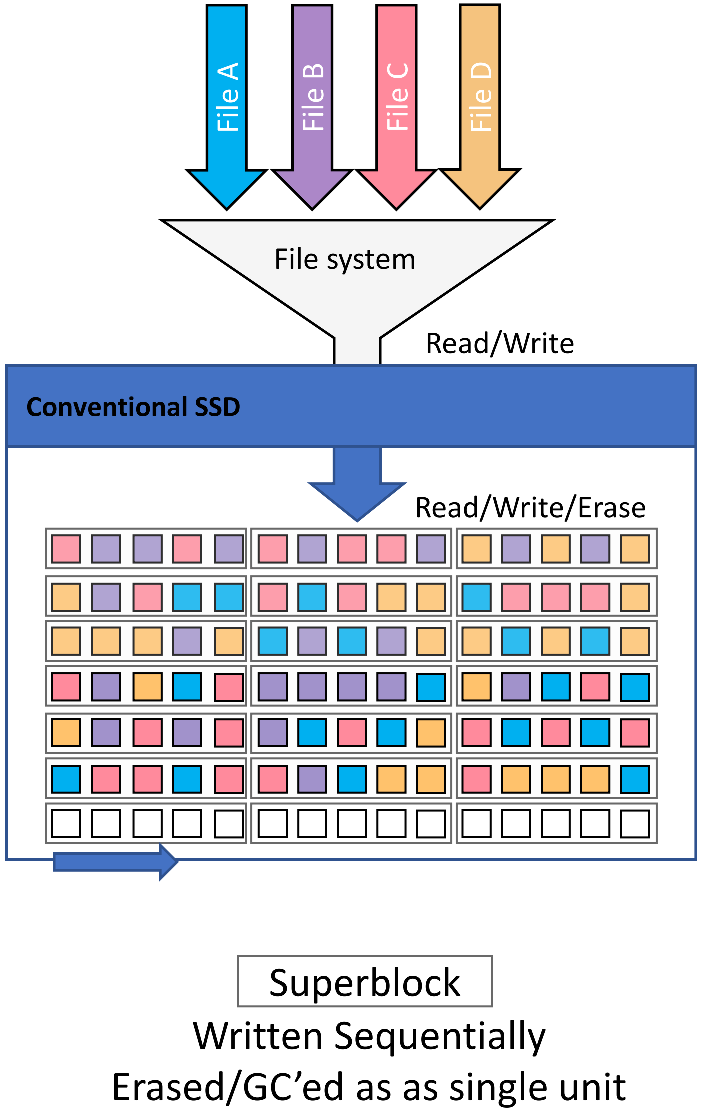
Achieve high performance through **extreme over-provisioning (e.g., 50%)**, but at the expense of **twice the media cost**.

FMS

Western Digital.

Write Amplification?

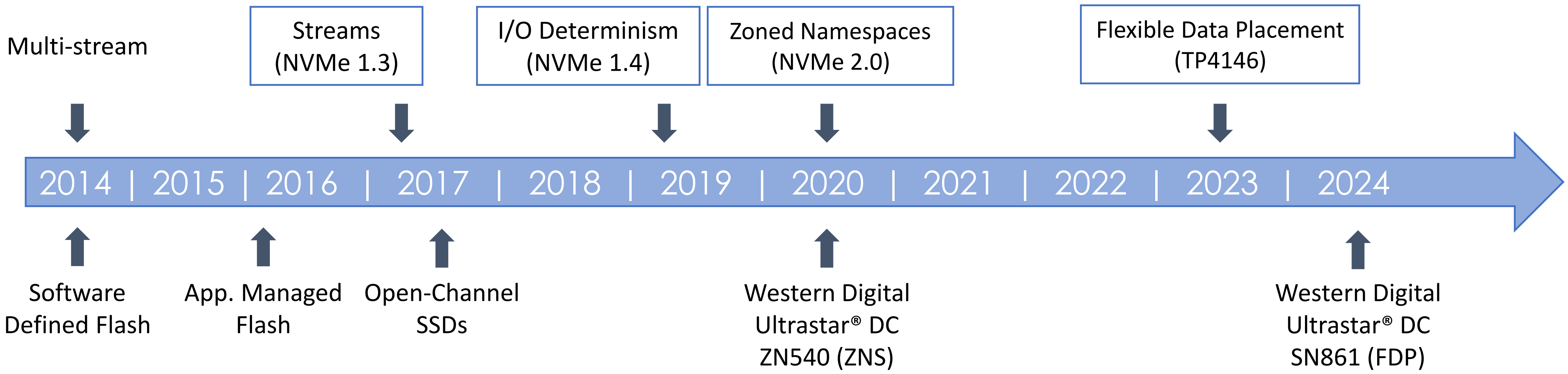
- Write amplification results from a mismatch between the host interface and the failure to align the SSD's media interface (NAND flash)
- **Conventional** ways to reduce write amp.
  - Trim/Unmap/DSM (Dealloc.)
  - Host and device over-provisioning
- **Data Placement**
  - Active research topic
    - Multi-stream (2014), Software-Defined Flash (2014), Open-Channel SSDs (2014, 2017), Application Managed Flash (2016), many more
  - Standardization
    - Streams, I/O Determinism, Zoned Namespaces, Flexible Data Placement



FMS

Western Digital

A Decade of Data Placement



Data Placement Benefits

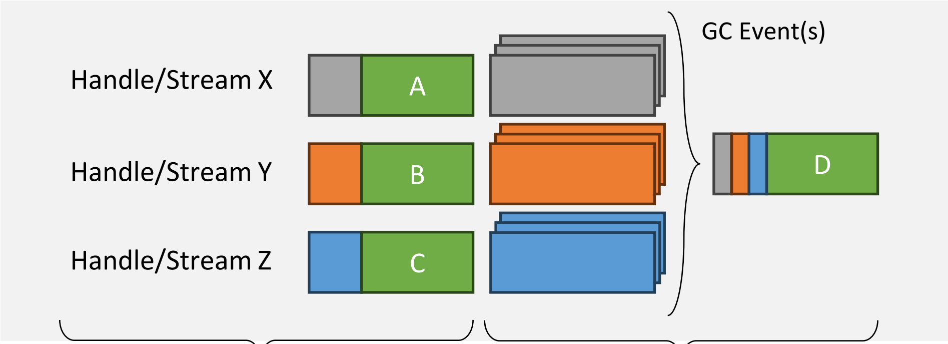
Enhanced performance: Lower write amplification translates to faster write speeds and better Quality of Service (QoS) performance

Reduced overprovisioning: Data placement allows for greater utilization of an SSD’s raw capacity

Increased endurance: Less wear and tear on the SSD’s lifespan

Standardized NVMe® Interfaces

Streams (2017) & FDP (2023)

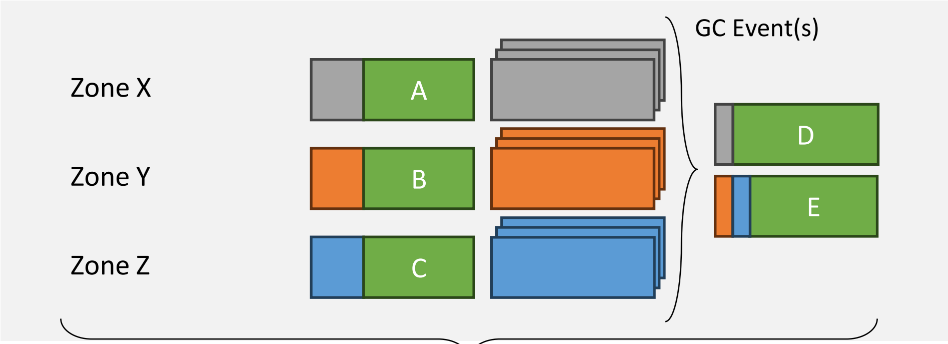


Host owns initial data placement

SSD owns later data placement (e.g., due to GC events)

Potentially less host involvement
WAF >= 1, OP Required (e.g., 7%)

Zoned Namespaces (2020)



Host "owns" all data placement
Redirects user data how it fits current and future data placement

More host involvement
WAF = 1, No OP (0%)

FMS

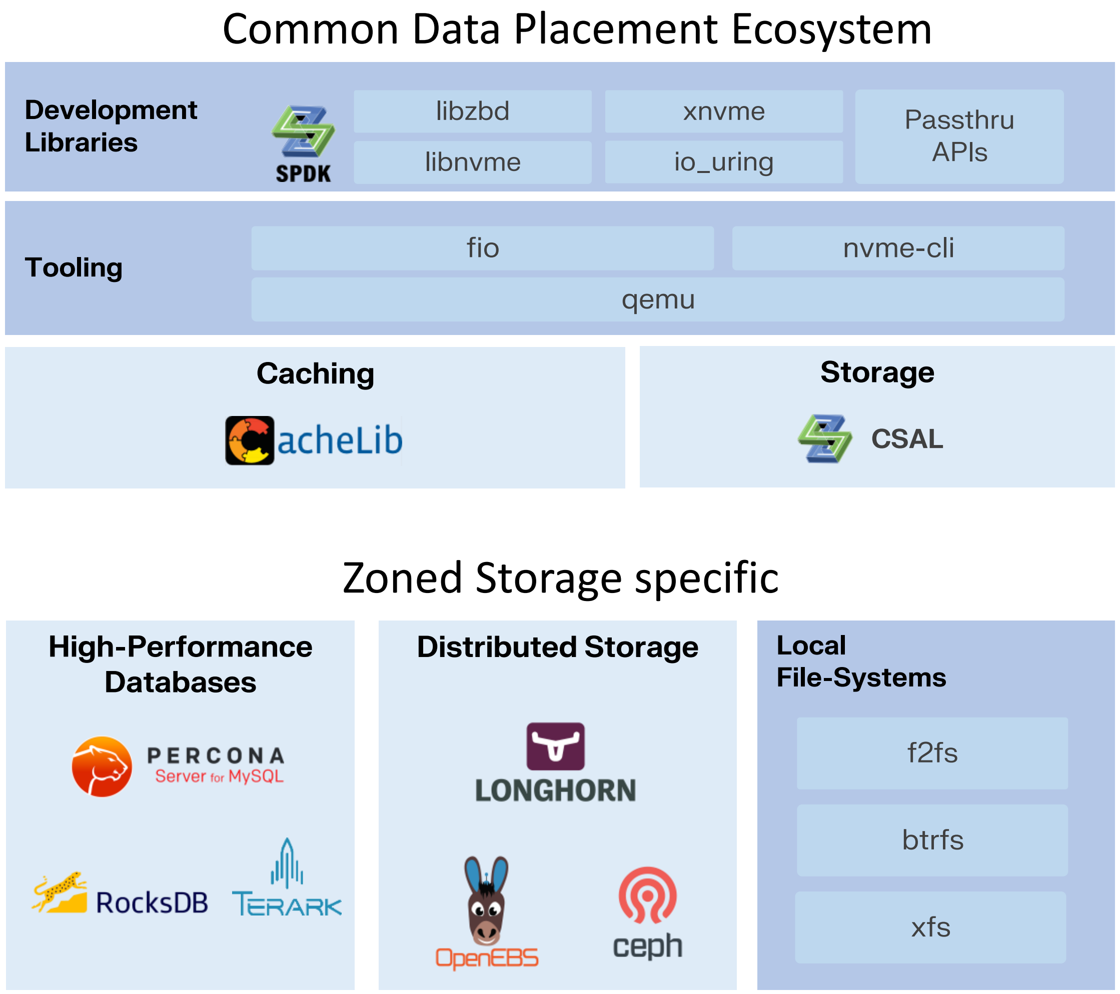
Western Digital.

Data Placement Interfaces

	Streams	FDP (Stream)	FDP (Full host Integration)	Zoned Namespaces
WAF Expectation	WAF >= 1			WAF = 1
Encapsulation	Stream/Reclaim Unit Handle ID		Reclaim Unit Handle	Zones (Set of LBAs)
Unit Writable Capacity	Unbounded		Approximate	Fixed
Finish Unit	N/A		Yes (Update Handle/Zone Finish)	
Reset Unit	DSM (Dealloc)		Multiple DSM (Dealloc) to invalidate data within an expected reclaim unit (if data is written non-seq)	Zone Reset
Placement Tracking	N/A		Each write LBAs tracked to allow accurate deallocs. An implementation may write sequentially to reduce tracking overhead.	Data placement is tracked through zones.
Unit State Communication	N/A		Asynchronous (Host probes state continuously from device)	Synchronous (Host and device always in sync on unit's state)
How to write	Write Cmd + Stream Id	Write Cmd + Reclaim Unit Id	Write Cmd + Reclaim Unit Id Continuously monitor through log pages: - Change in Reclaim Unit Avail. Media Writes (e.g., every 100 writes) - FDP Events (RU not fully written to cap., media reallocated)	Write Cmd
Example of open-source use-cases	RocksDB (support removed from Linux kernel in 2022)	CacheLib, xfs (In the works)	TBD	Applications: RocksDB, CacheLib, MySQL, Ceph File-Systems: f2fs, btrfs, xfs

Data Placement Ecosystem

- The software ecosystem for data placement continues to move forward
- Flexible Data Placement
  - Support added to core tools (qemu, fio, SPDK, …)
  - Utilized through passthru kernel APIs
  - RocksDB & XFS write hint passthru in progress
- Zoned Storage (SMR, ZNS, Zoned UFS)
  - Broad enablement due earlier standardization and multiple storage device types
  - Utilized through native kernel APIs
  - Native XFS support in progress



Session Talks

- William Cheng, Silicon Motion
  - FDP Benefits in QLC Applications: A Case Study
- Mariusz Barczak, Solidigm
  - Cloud Storage Acceleration Layer (CSAL): Leveraging Gen5 FDP NVMe Technologies
- Rory Bolt, KIOXIA
  - FDP: What Every Storage Architect Should Know!
- Jonmichael Hands, FADU
  - FDP Performance in VMs with Multiple NVMe Namespaces: Case Studies
- Panel Discussion