STORAGE DEVELOPER CONFERENCE

SDC 23

BY Developers FOR Developers

Host Workloads Achieving WAF==1 in an FDP SSD

Presented by Dan Helmick, PhD

SAMSUNG

Agenda

- Background
  - FDP Overview
  - Visualizing Writes in an SSD
  - QD>1 impacts with FDP
- Some example WAF==1 workloads
  - Circular FIFO
  - Modified Circular Buffer
  - Log Structured File Systems
  - Probabilistic
  - Log Structured File Systems with Mismatched Host Extent and SSD RU

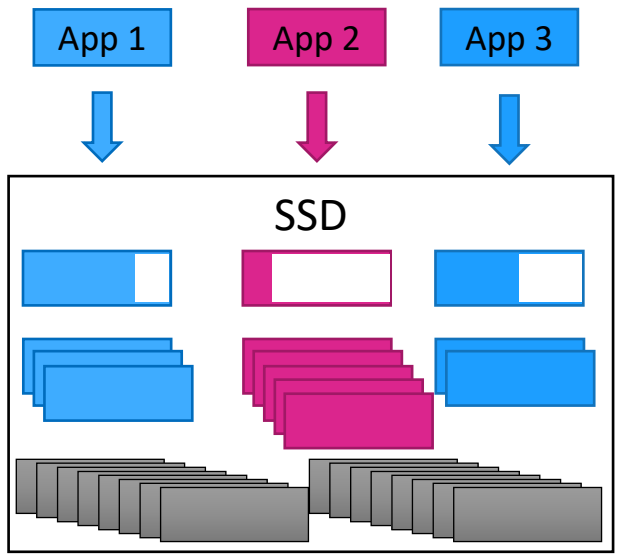
SAMSUNG

SDC 23

Flexible Data Placement (FDP) Overview

- Apps can direct write data to be co-located in an SSD
  - Possible for a VMM to set-up defaults for legacy VMs
- Filling and deallocating appropriately can achieve WAF==1

Logical View



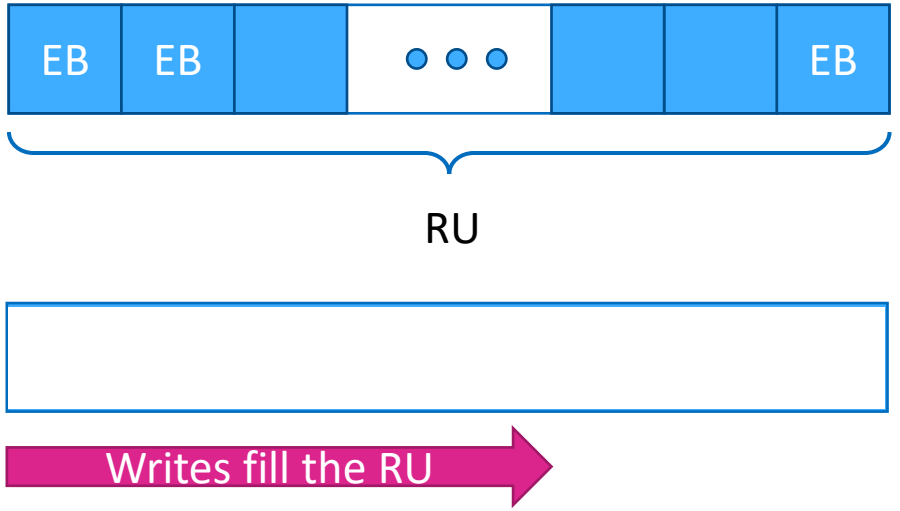
Streams	Flexible Data Placement (FDP)	Zoned Namespaces (ZNS)
Open Loop WAF==1	Polling for WAF==1	WAF==1 or Error
Backwards Compatible	Backwards Compatible	Not Backwards Compatible
Streams Granularity Size (SGS)	Reclaim Unit (RU) Size	Zone Capacity <= Zone Size
Placement and LBA disconnect	Placement and LBA disconnect	Placement and LBA relationship
QD>1 allowed	QD>1 allowed	QD>1 requires Zone Append
Full FTL mapping required	Full FTL mapping required	Potential for compacted FTL Mapping

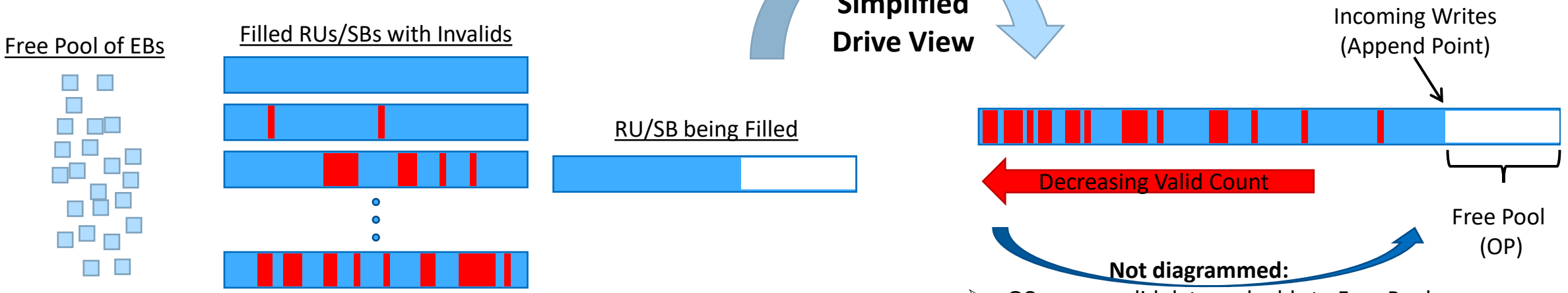
SAMSUNG

SDC 23

Simplified SSD Composition

- Reclaim Units (RUs) are composed of 1 or more Erase Blocks (EBs)
  - Ex: RU is equal to a SuperBlock (SB)
  - SB = 1 EB per Plane for every Die
- RU is filled in order even if the LBAs are out-of-order
- After filling an RU, a new set of empty EBs are selected to create a new RU
  - Rules may be applied in selecting EBs from the Free Pool
  - Ex: 1 EB per Plane for every Die to create a SB
- Diagramming a Conventional Drive = 1 RUH
  - Random traffic





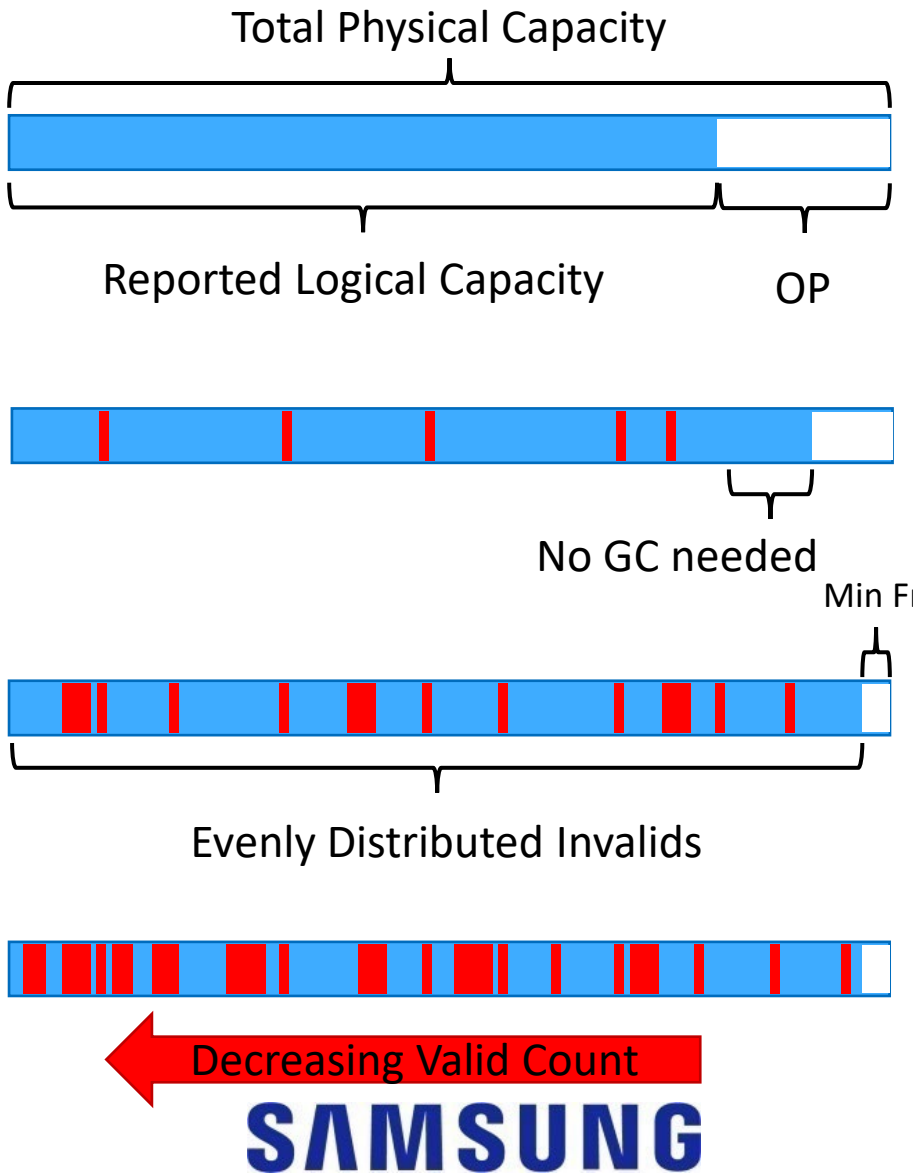
SAMSUNG

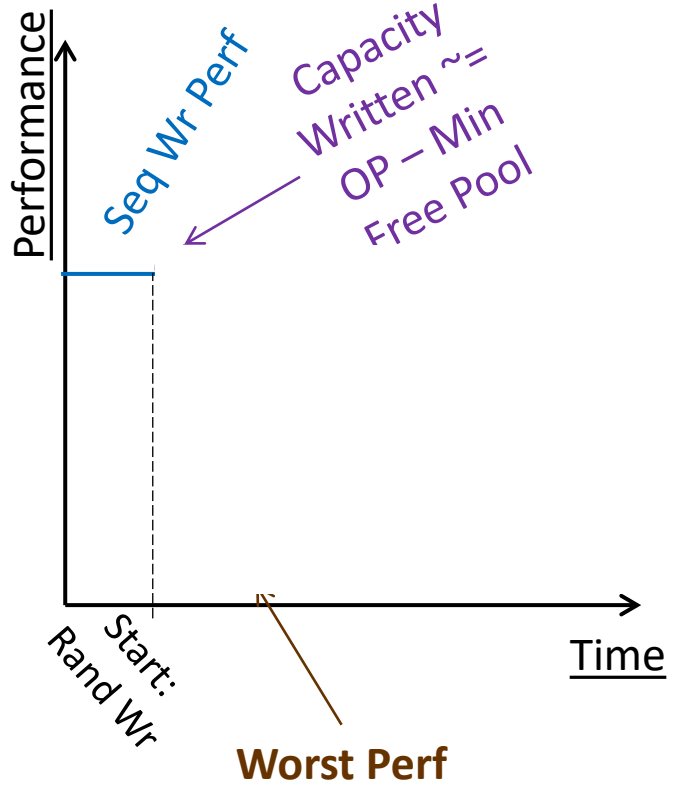
SDC 23

Visualized NAND and Performance Transitioning Write Traffic: Sequential→Random

Preconditioning Visualization Example

- Sequentially Written (Preconditioned):
- Random Writes start:
- Random Writes Reach Worst Case Performance:
- Random Write Steady State (SS):





- All SB/RU having roughly same invalid count
- High valid counts for any SB/RU that is selected

SAMSUNG

SDC 23

Extrapolating to FDP with Multiple RUHs

- Each RUH is a new append point in the NAND
- Characterization of Write behavior per RUH is required to understand SSD's WAF
  - WAF==1 on each RUH required for perfect drive WAF==1
- However, WAF improvements on each RUH benefit entire SSD
- Persistently Isolated vs Initially Isolated RUHs only matters for WAF>1
  - Not an emphasized discussion in this presentation

WAF==1 on any RUH means GC path is not exercised

Figure Y: Initially Isolated Reclaim Unit Handles

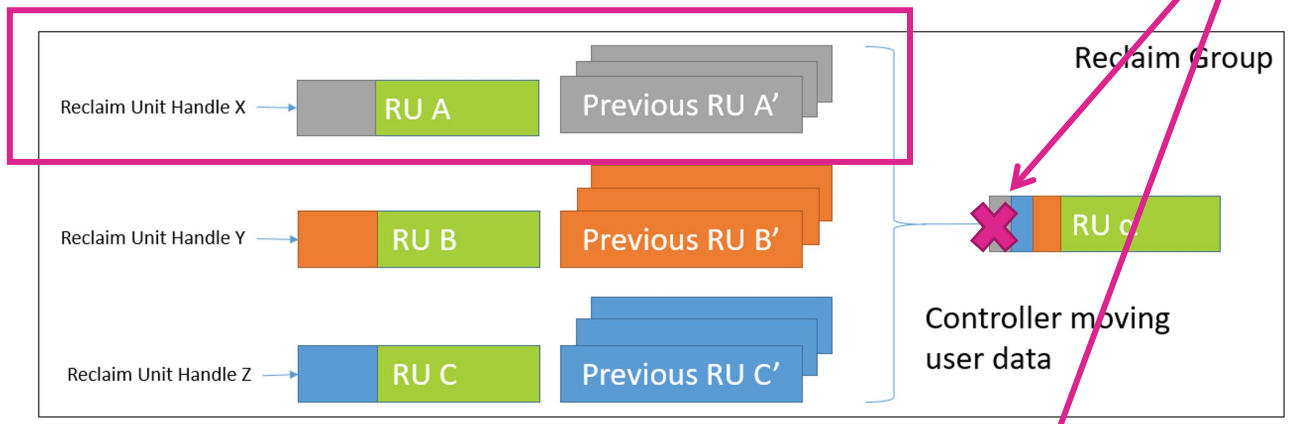
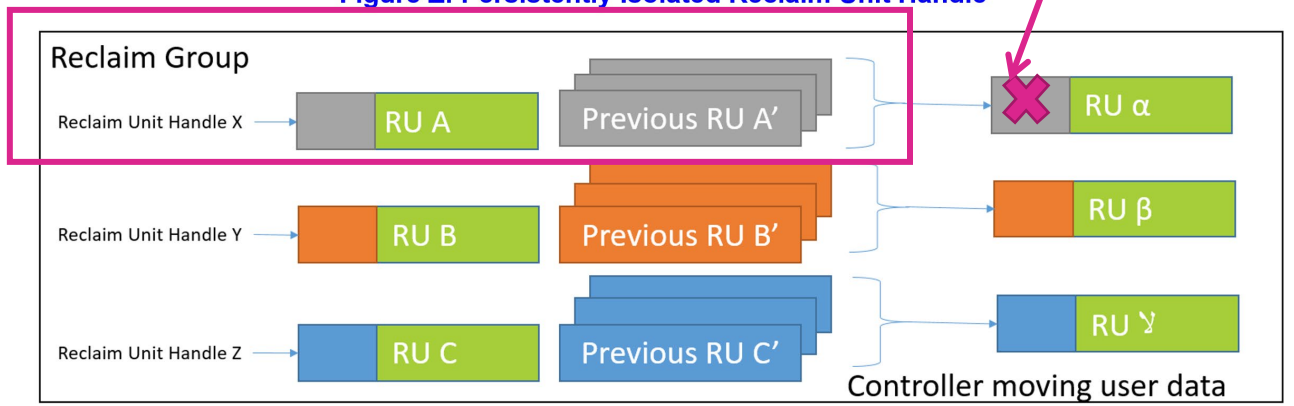


Figure Z: Persistently Isolated Reclaim Unit Handle

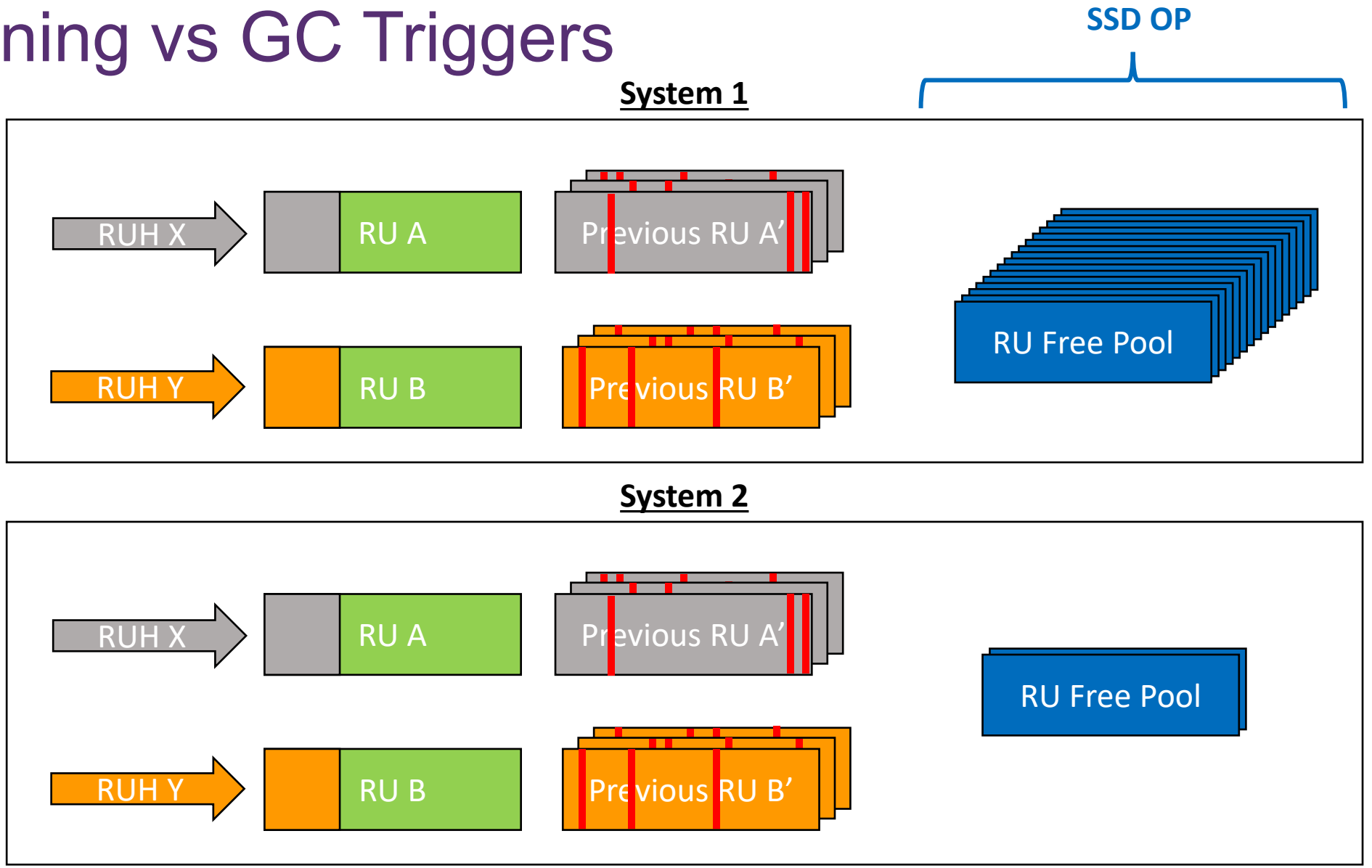


SAMSUNG

SDC 23

Over Provisioning vs GC Triggers

- Similarities
  - Valid Count in Previously Filled RUs
  - Incoming RUH Traffic
- Differences
  - System 2 is low on OP
  - Small amounts of non-optimal traffic will result in very high WAF
- SSD OP provides protection
  - **Buffers** against **race conditions** of traffic ordering (Writes vs Deallocates)
  - **Protects** against **minor imperfections** in Host optimized traffic



SAMSUNG

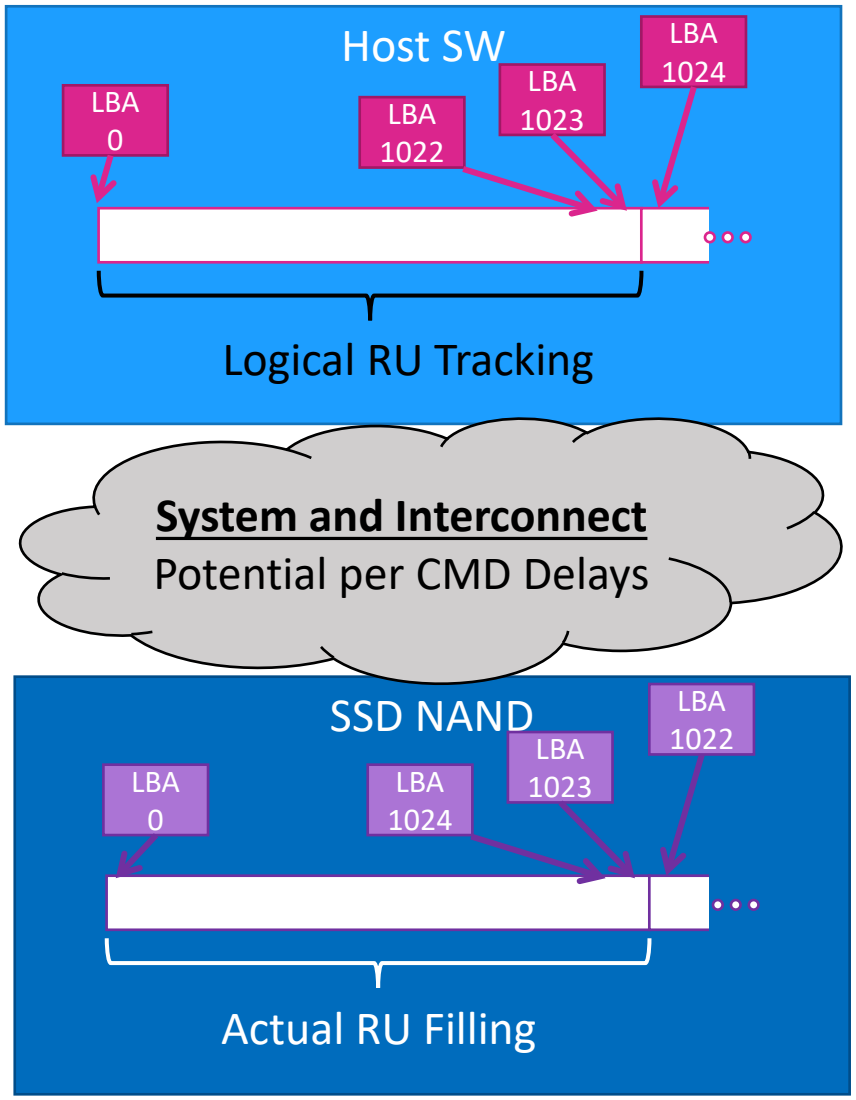
SDC 23

Background: QD>1 System Race Conditions

- High QD has a chance of out-of-order processing
- This can create a disconnect of
  - Expected RU as tracked by Host SW
  - Actual RU as placed on SSD NAND
- Through the length of the RU, this doesn't matter.
  - But at RU boundaries can potentially create orphan LBAs

- Example:
  1. For each LBA in range [0, 9999] (Write LBA)
  2. Deallocate Logical RU of range [0 1023]
     - Problem: LBA 1024 was placed in an older RU because it arrived earlier than LBA 1022 and 1023

- Mitigations
  - Run Host SW as QD == 1
  - Wait for all completions of a Logical RU to return before starting a new RU
  - Accept Risk
    - Increased Host OP and/or SSD OP to protect the system against errant GC



SAMSUNG

SDC 23

Some example WAF==1 workloads

- Circular FIFO
- Probabilistic
- Modified Circular Buffer
- Log Structured File Systems

SAMSUNG

SDC 23

Circular FIFO

- Looping over any LBA Range
  - LBA Range is constant
  - Deallocate or direct overwrite of LBA acceptable
- Any length in relation to RU
- New empty RUs appended as needed

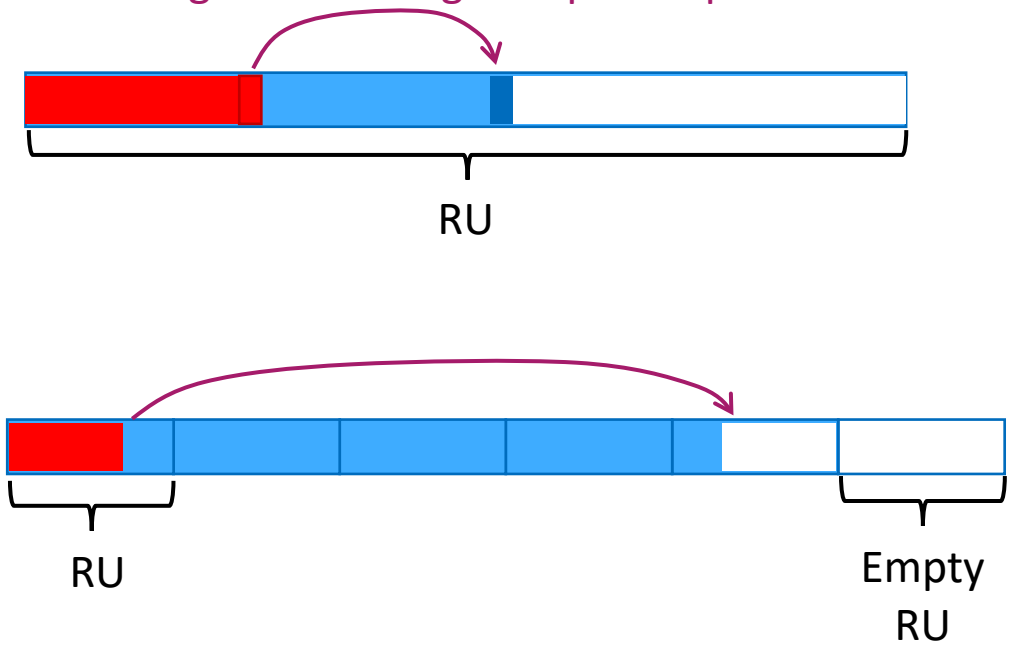
- Implementation concerns:
  - If QD>1, race conditions can alter RU association
    - Particularly at RU boundaries
  - Some drive architectures are exposed to different delays:
    - Deallocate then Write of LBAs
    - Direct overwrite of LBAs

- Recommendations:
  - Allow both SSD and Host OP
  - **SSD OP:** Some SSD OP reduces the probability the emptying RU will need to be used for the newest RU
  - **Host OP:** Deallocations far ahead of the LBA's overwrite enable the most consistent cross-vendor behaviors

Example: Circling over LBAs 1-4

1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4

Rewriting an LBA Range keeps compact valid data

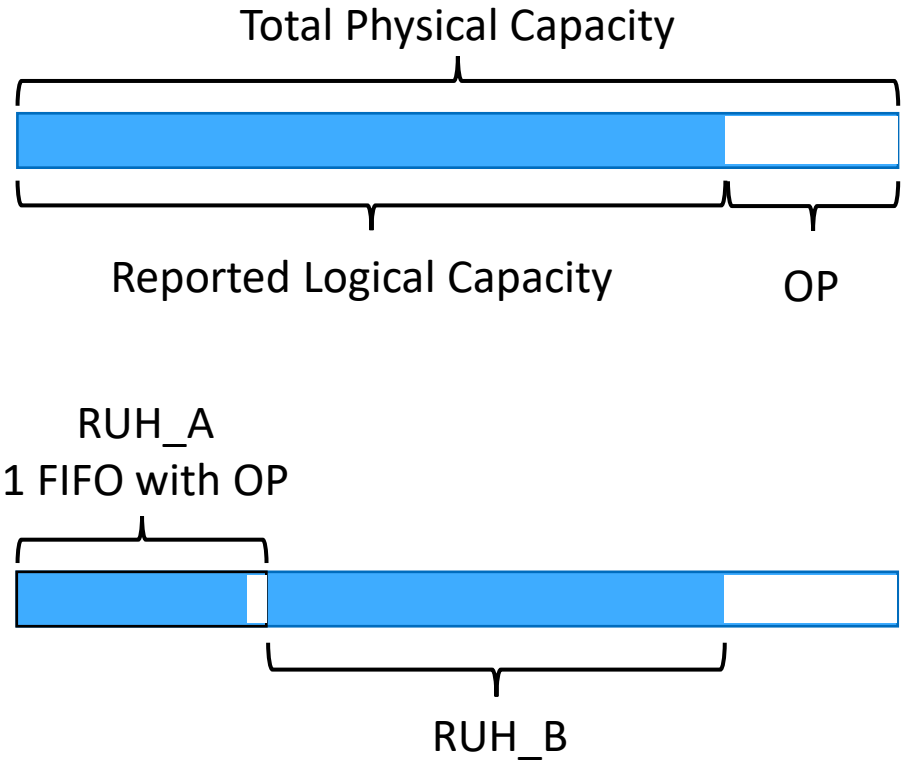


SAMSUNG

SDC 23

Visualizing Multiple Circular FIFOs

- Sequentially Written (Preconditioned):
  - This is a single Circular FIFO
- 2 Circular FIFOs written to 2 RUHs
  - Each FIFO is written compactly on the NAND
  - Each FIFO consumes minor OP
    - Shown visually on RUH_A
  - Majority of OP remains available for drive wide benefits. Examples:
    - WAF reduction
    - Endurance extension
    - NAND handling
- Every compactly written RUH preserves the SSD OP

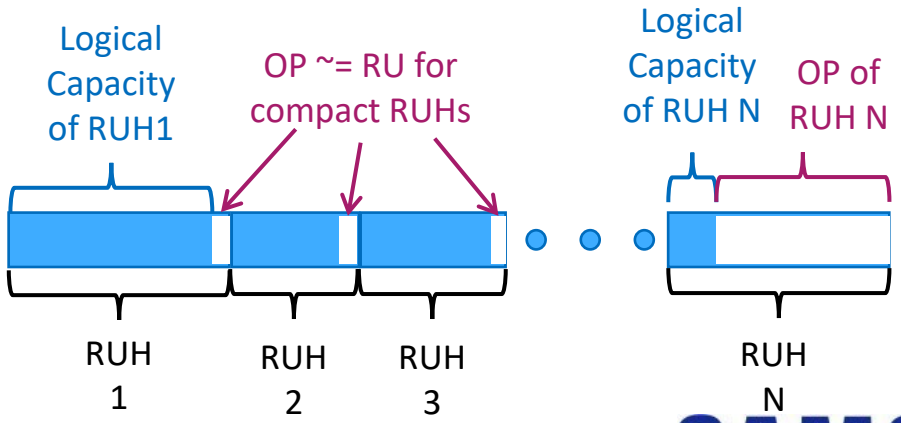


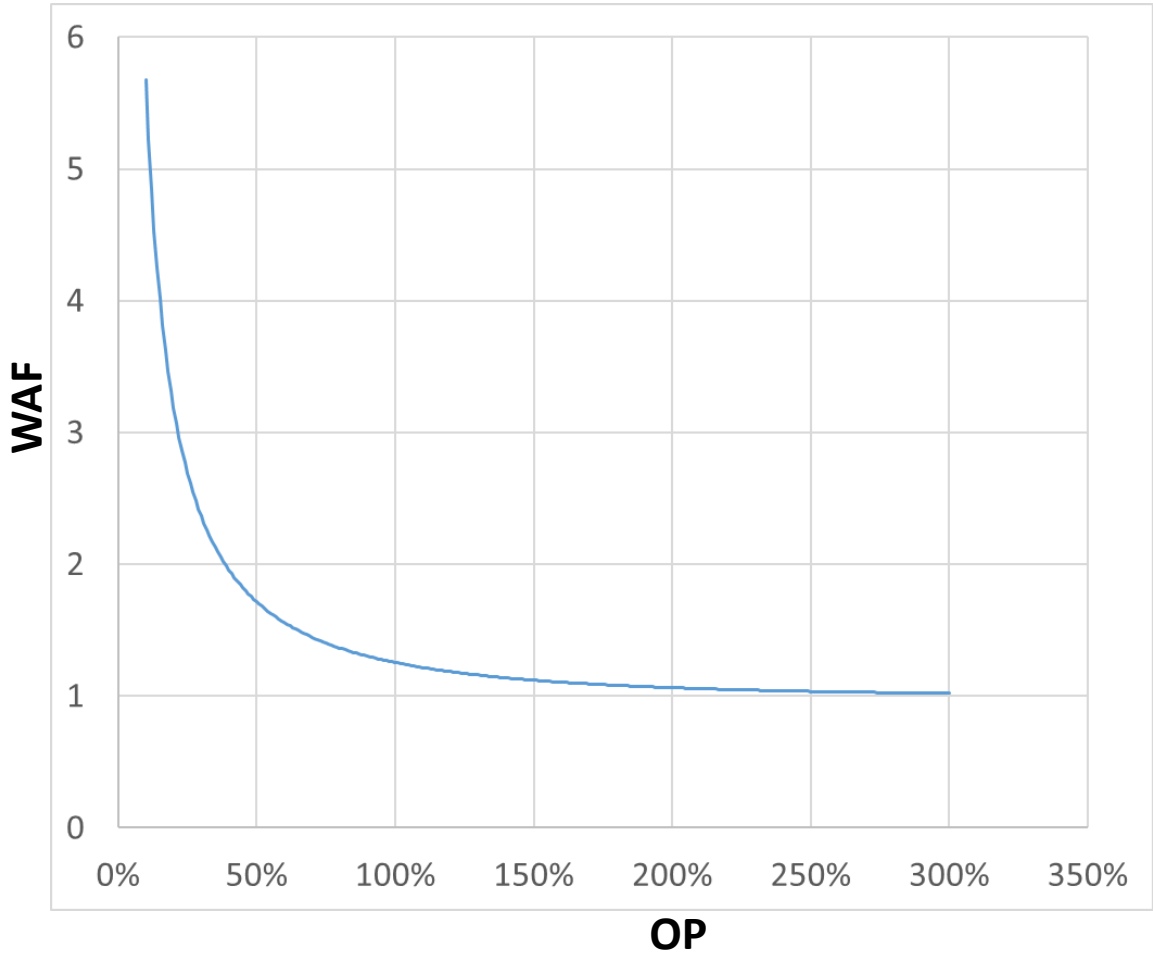
SAMSUNG

SDC 23

Probabilistic

- **Low WAF** can be achieved through probabilities – High OP correlates to low WAF
- Several well behaved RUHs allow poorly behaved RUHs to consume more OP
  - Overall system improvements!
- RUH N illustrates a small logical capacity using a large physical capacity

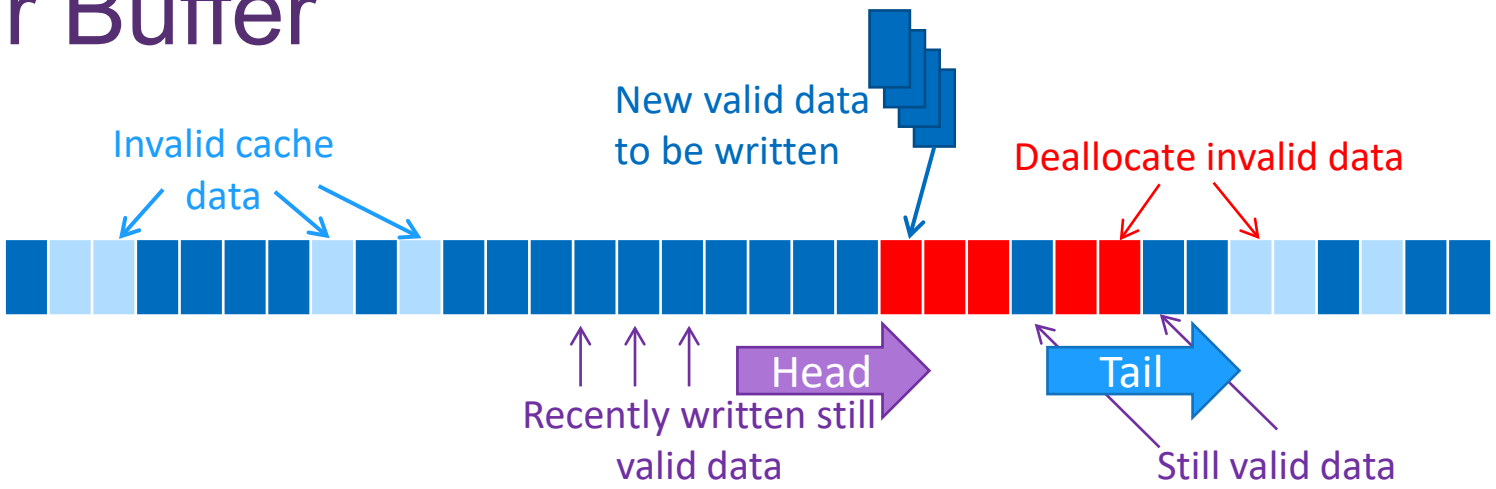




SAMSUNG

SDC 23

Modified Circular Buffer



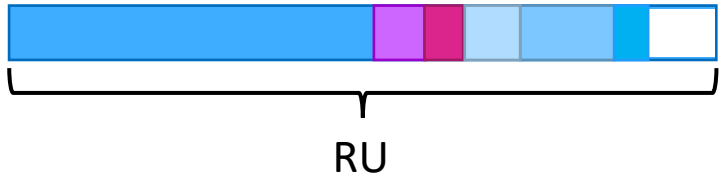
- WAF==1 through Deallocate assurances
- Common example is Cache management
  - Head: Appends incoming cache entries
  - Tail: Reads out still valid cache entries → Transitioning them to invalid
- Options: Invalid cache entries can be deallocated to the drive or left in place

SAMSUNG

SDC 23

Log Structured File Systems (FS)

- Objects Appended to fill an RU
  - Emphasizes writing sequentially to storage
  - Helps both HDDs and SSDs
- Reads are Random
  - SSD strength
- Variations are found everywhere by different names
  - Blobs
  - Zones
  - Slices
  - Extents
  - Data Volumes
  - Data Nodes
  - SSTables



Host Extent used in this presentation

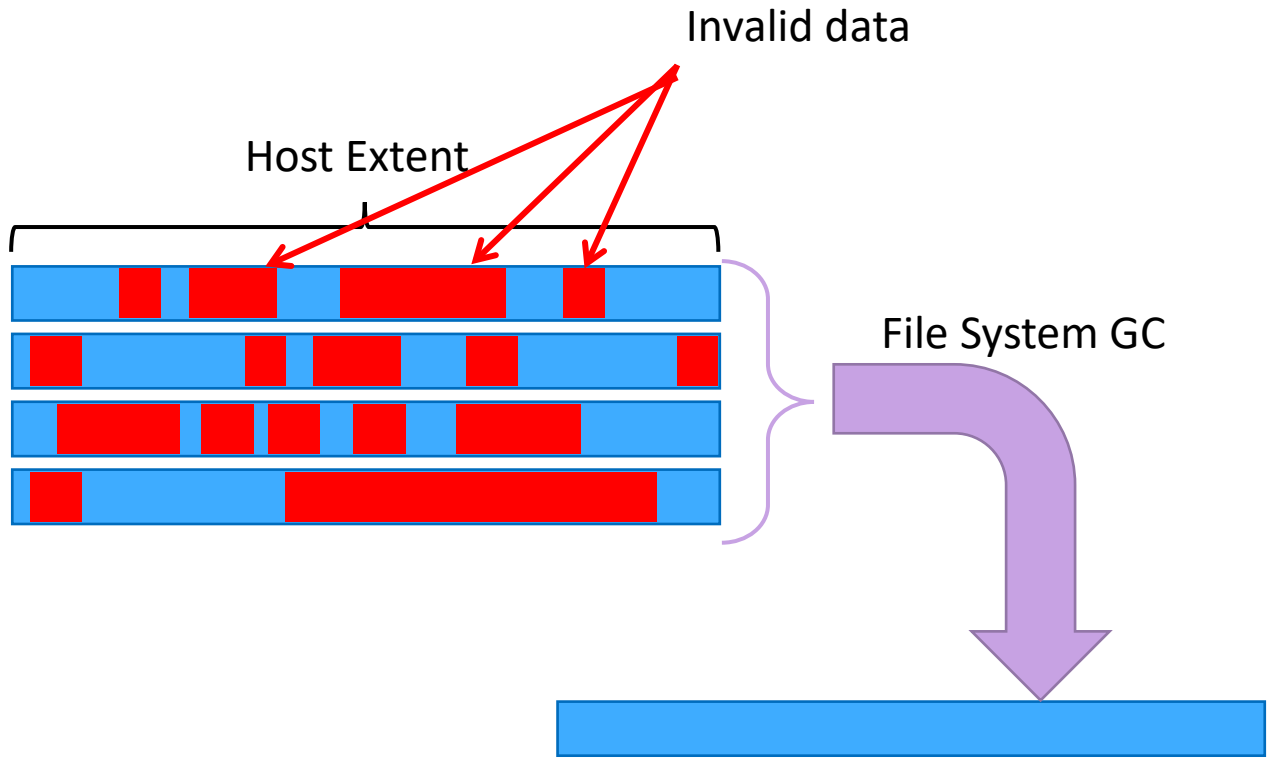
- Higher level protections may be applied at the system level
  - RAID or Erasure Codes
  - CRC

SAMSUNG

SDC 23

Log Structured File Systems Interacting with an FDP SSD

- When Host Extent == RU
  - Host GC aligned with Drive GC activity
  - Deallocates are a critical part of achieving WAF==1
- Full RU deallocates aligned with FS
  - Invalid objects **may** be communicated to SSD
- Implementation
  - Object-to-RU endings can be misaligned if QD>1
  - Object deallocates are not required to be communicated to SSD
- Recommendations
  - Allow both SSD and Host OP
  - **SSD OP:** enables robust operation without object deallocates communicated to SSD
  - **Host OP and SSD OP:** can both compensate for race-conditions on Object-to-RU placement

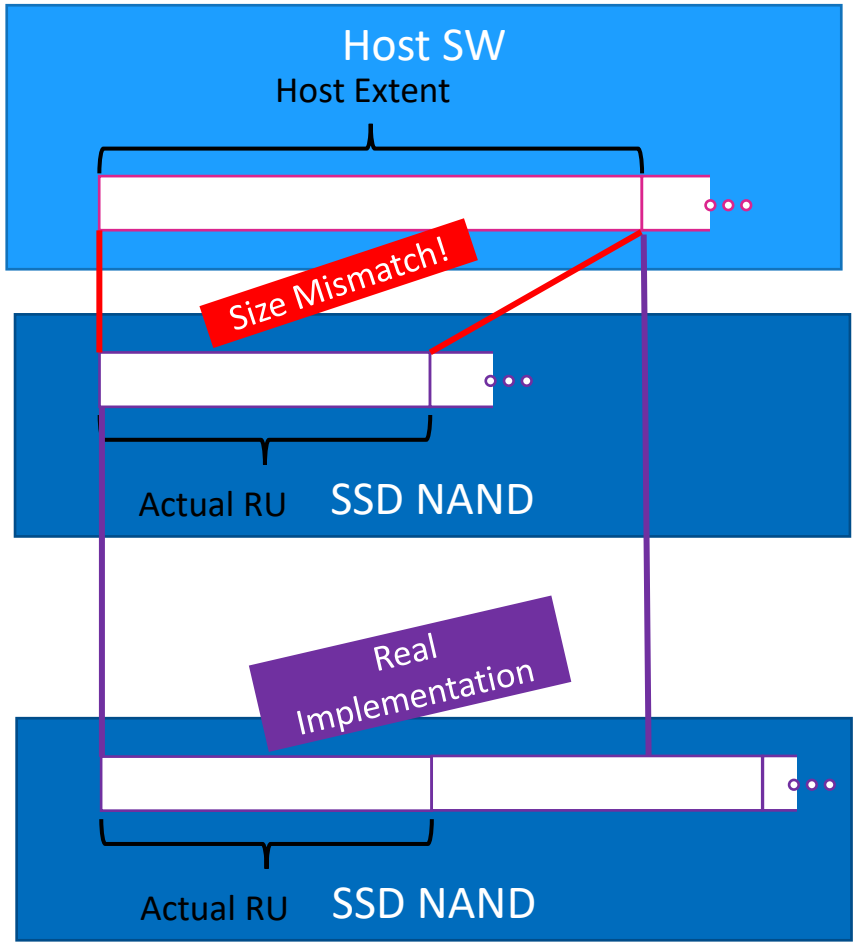


SAMSUNG

SDC 23

Size Mismatch: Host Extent vs SSD RU

- Log Structured File System built with Host Extents rather than RU matching
  - Host Extent may not match SSD RU size
- Reasons Host Extent may not match SSD RU
  - Vendor-to-Vendor mismatch
  - Generation over Generation SSD RU changes
  - SW developed separate from SSDs
- Some Critical Findings
  - WAF==1 singularities
    - Host Extent = N * (SSD RU), where N = 1, 2, …
    - Deallocating a Host Extent frees up several SSD RUs
  - Large Host Extents improve WAF
  - System OP is always a helpful tool to leverage



SAMSUNG

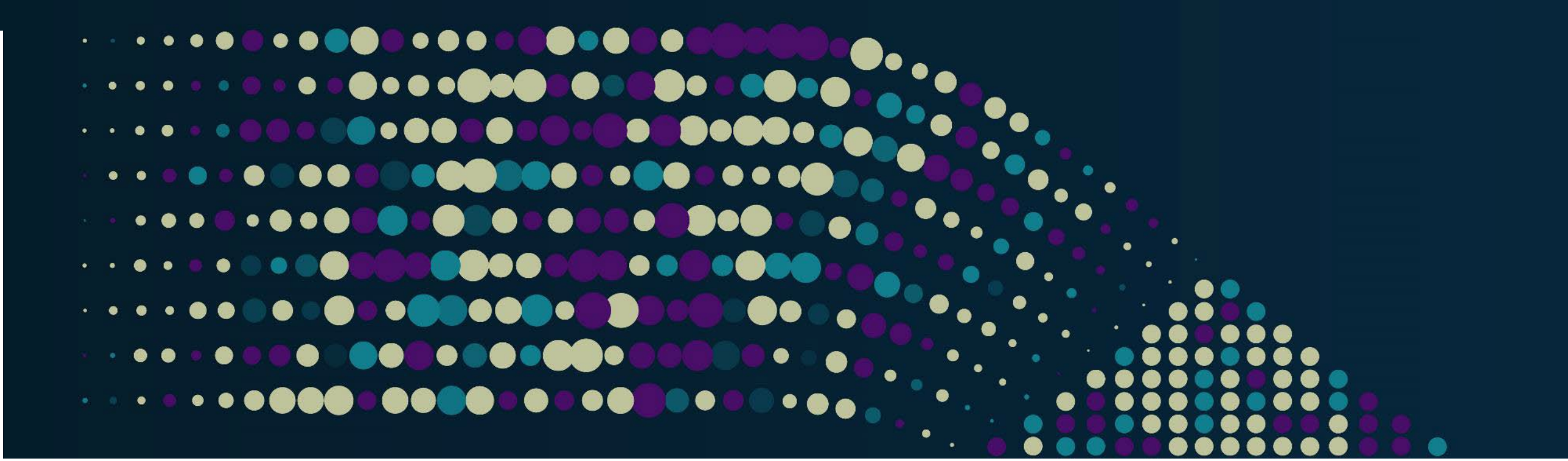
SDC 23

Conclusions

- Various WAF==1 workloads are possible
  - Circular FIFO
  - Probabilistic
  - Modified Circular Buffer
  - Log Structured File Systems with Host Extents a multiple of SSD RU
- Write, Overwrite, and/or Deallocate assurances are all reasonable methods of reaching WAF==1
- Enable System OP (Host OP and/or SSD OP)
  - Compensates for QD>1 out of ordering
- Deallocate far before LBA re-use to cover delay differences in SSD implementations

SAMSUNG

SDC 23



Please take a moment to rate this session.

Your feedback is important to us.

SAMSUNG

SDC 23