Best Practices for Host Aware Flexible Data Placement (FDP) SW

Presenter: Dan Helmick, PhD

SAMSUNG

Outline

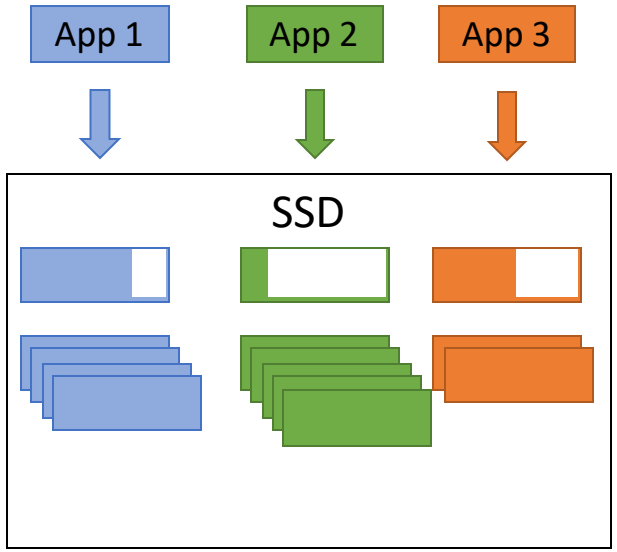
- Flexible Data Placement (FDP) – Overview
- Simplified SSD Composition
- FDP Workloads with WAF=1
- Intelligent Queries of Reclaim Unit Available Media Writes (RUAMW)
- Selecting Persistently Isolated vs Initially Isolated Reclaim Unit Handles (RUHs)
- Estimated Active Reclaim Unit Time Remaining (EARUTR)

SAMSUNG

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
- For further background on FDP, see TP4146a or Mike Allison's presentation during the NVMe Session

Logical View

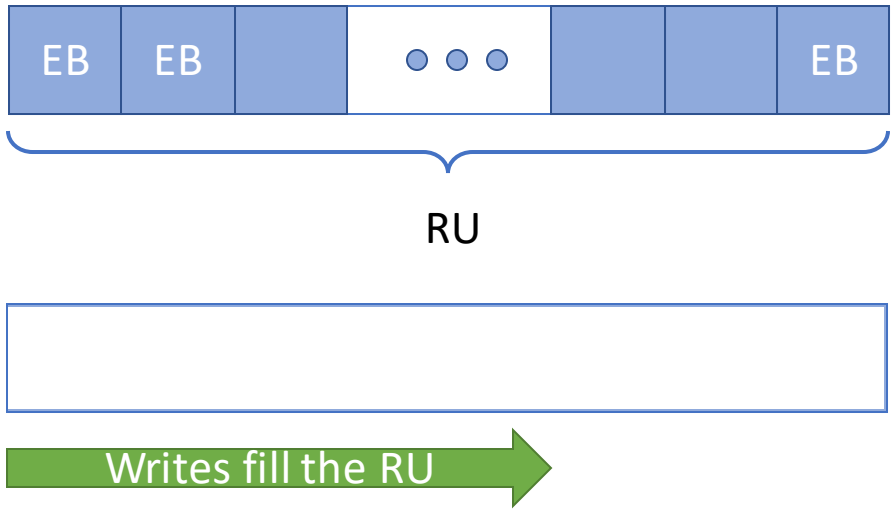


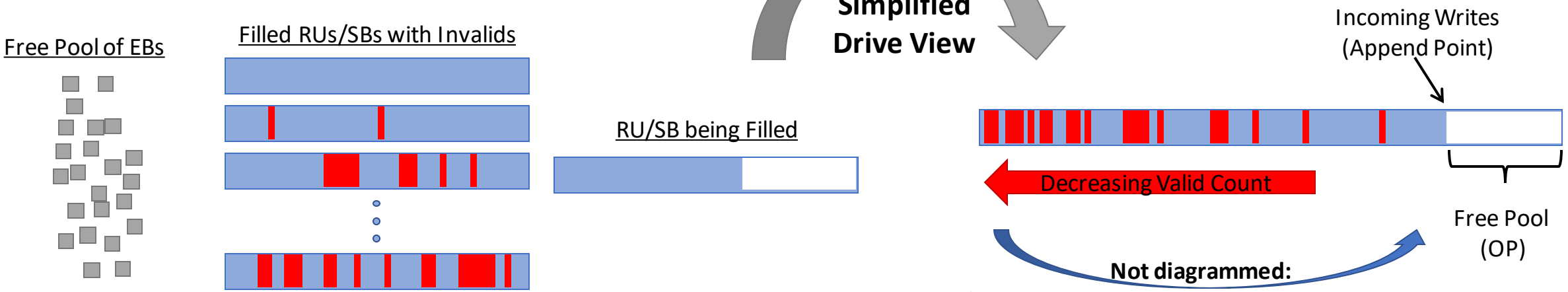
Streams	Flexible Data Placement (FDP)	Zoned Namespaces (ZNS)
Open Loop WAF==1	Polling for WAF==1	WAF==1 or Error
Backwards Compatible	Backwards Compatible	Not Backwards Compatible
Streams Granularity Size (SGS)	Reclaim Unit (RU) Size	Zone Capacity <= Zone Size
Placement and LBA disconnect	Placement and 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

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

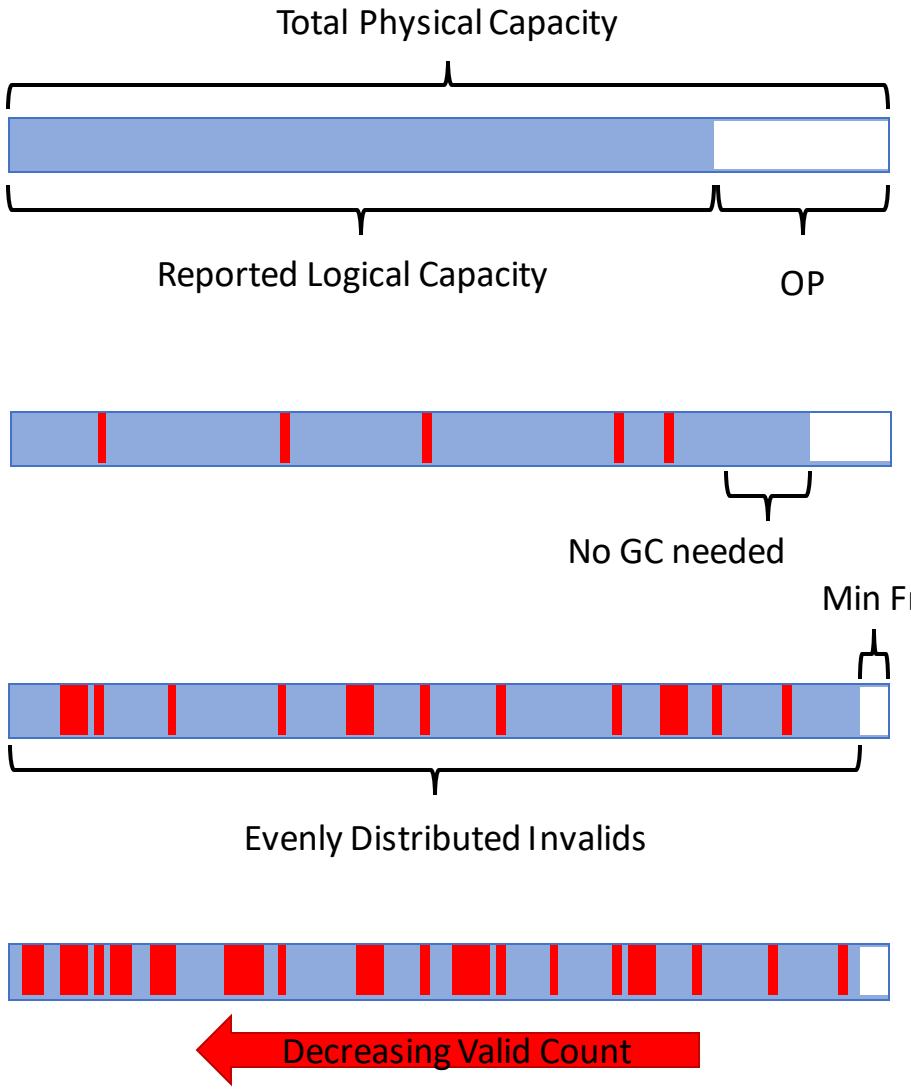
Visualized NAND and Performance

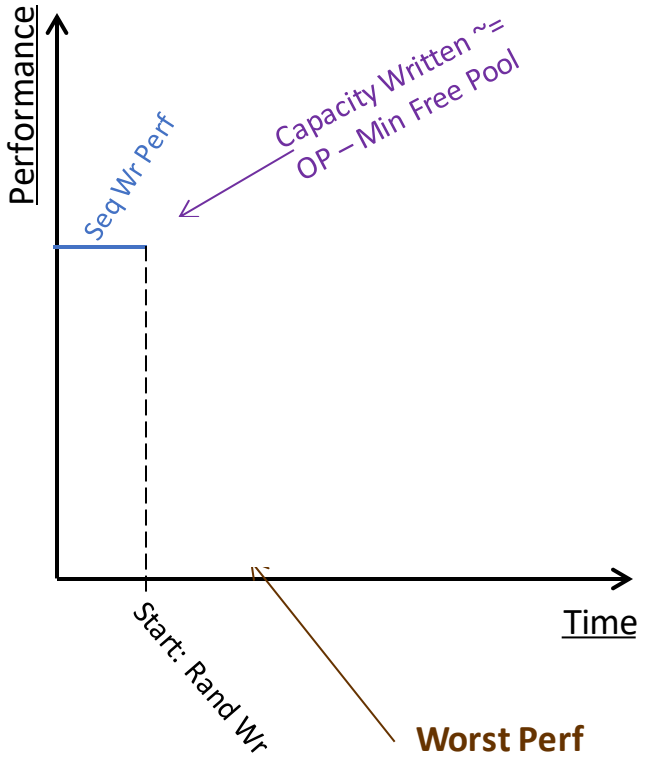
Transitioning Write Traffic: Sequential→Random

Preconditioning Visualization Example

Flash Memory Summit

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



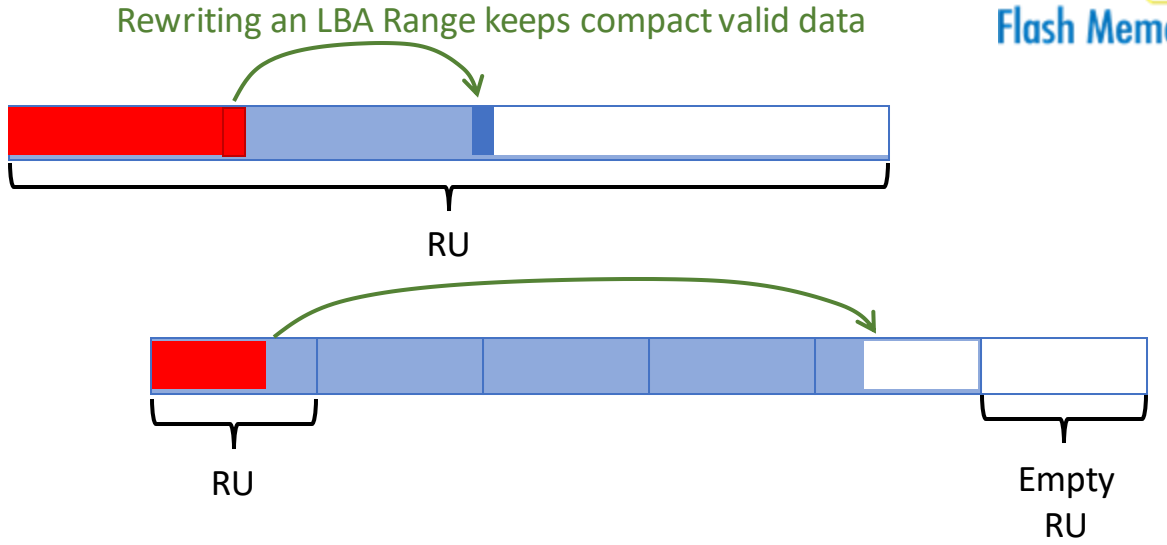


Worst Perf

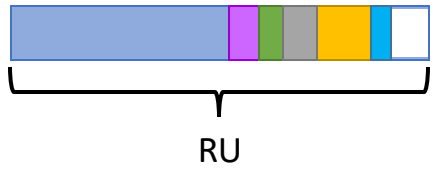
- All SB/RU having roughly same invalid count
- High valid counts for any SB/RU that is selected

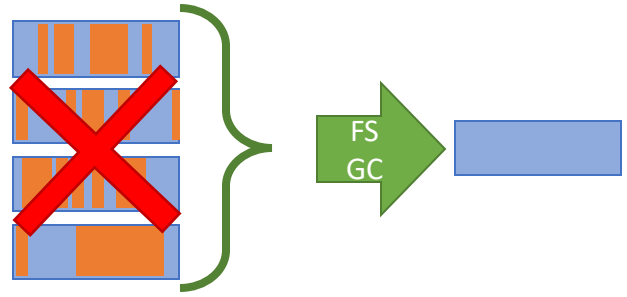
SAMSUNG

FDP Workloads with WAF=1



- Circular FIFO
  - Looping over any LBA Range
  - LBA Range is constant
  - Any length in relation to RU
  - New empty RUs appended as needed
- Log Structured File Systems (FS)
  - Objects Appended to fill an RU
  - Host GC aligned with Drive GC activity
  - Full RU deallocates aligned with FS
    - FDP SSDs accept misalignment, but may not receive WAF=1
    - *See next slide discussing deallocate assurances*

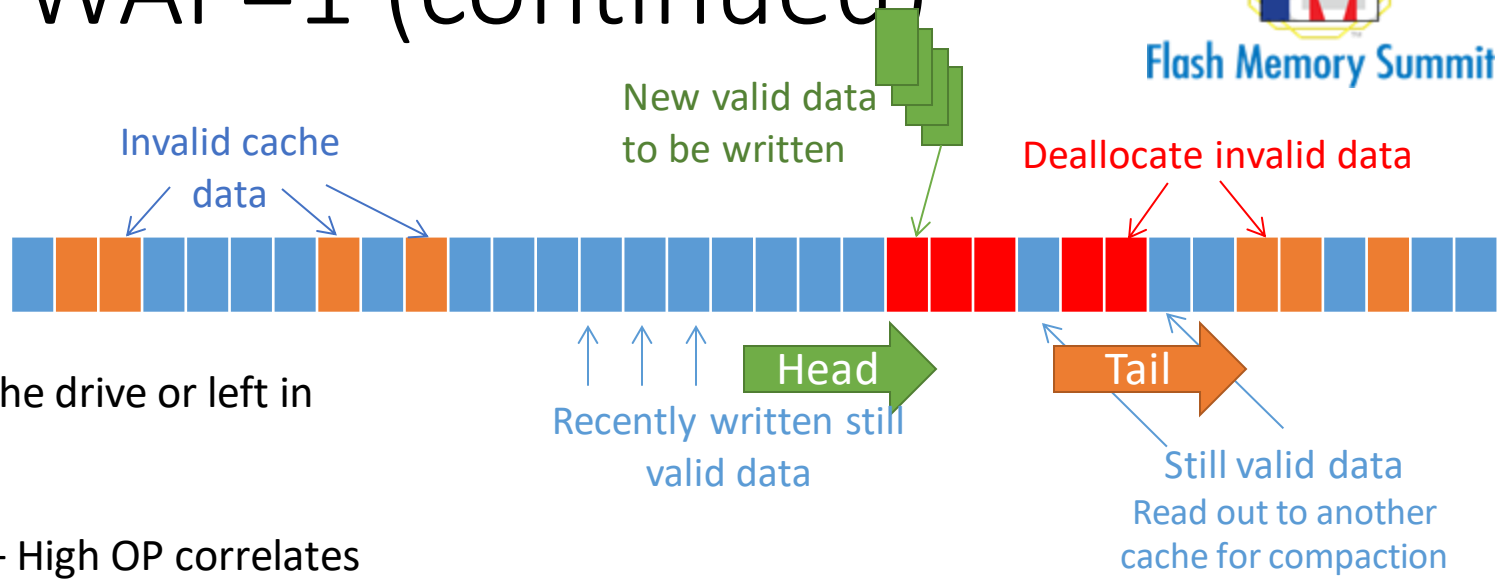


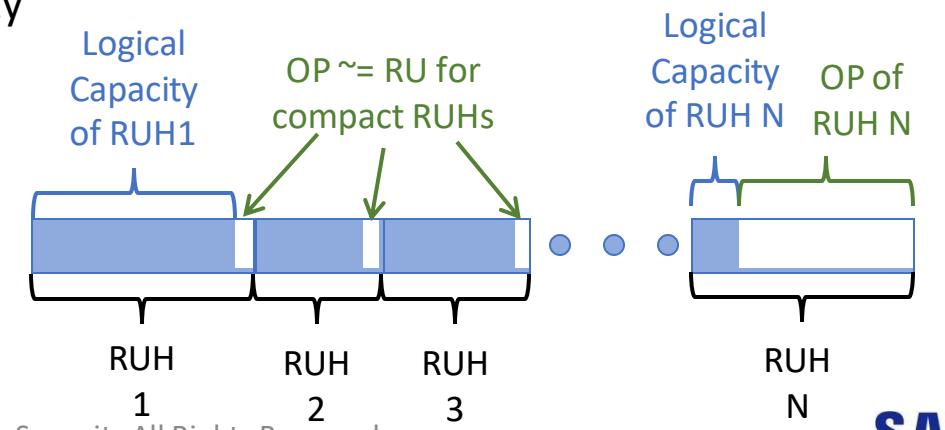


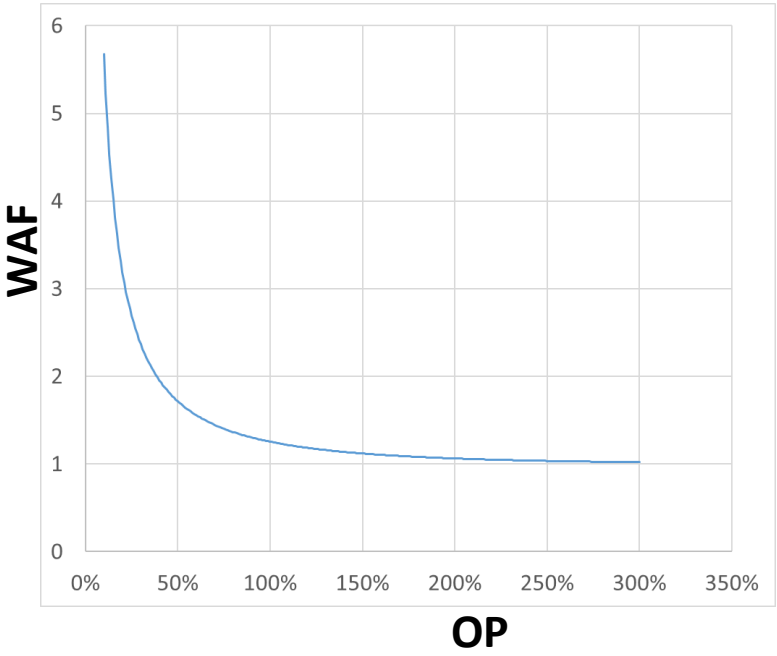
SAMSUNG

FDP Workloads with WAF=1 (continued)

- 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
  - Invalid cache entries can be deallocated to the drive or left in place
- Probabilistic
  - **Low WAF** is 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



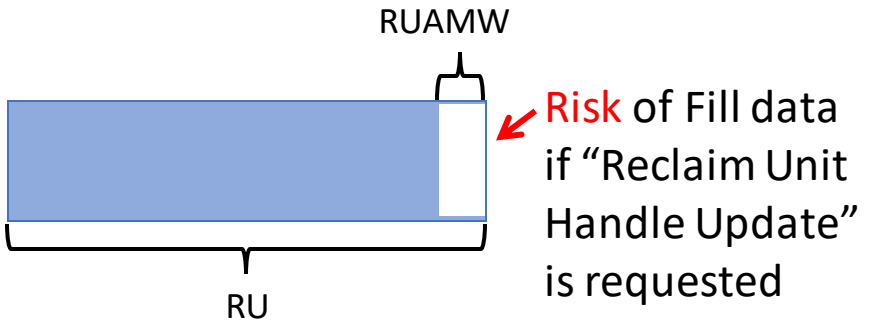


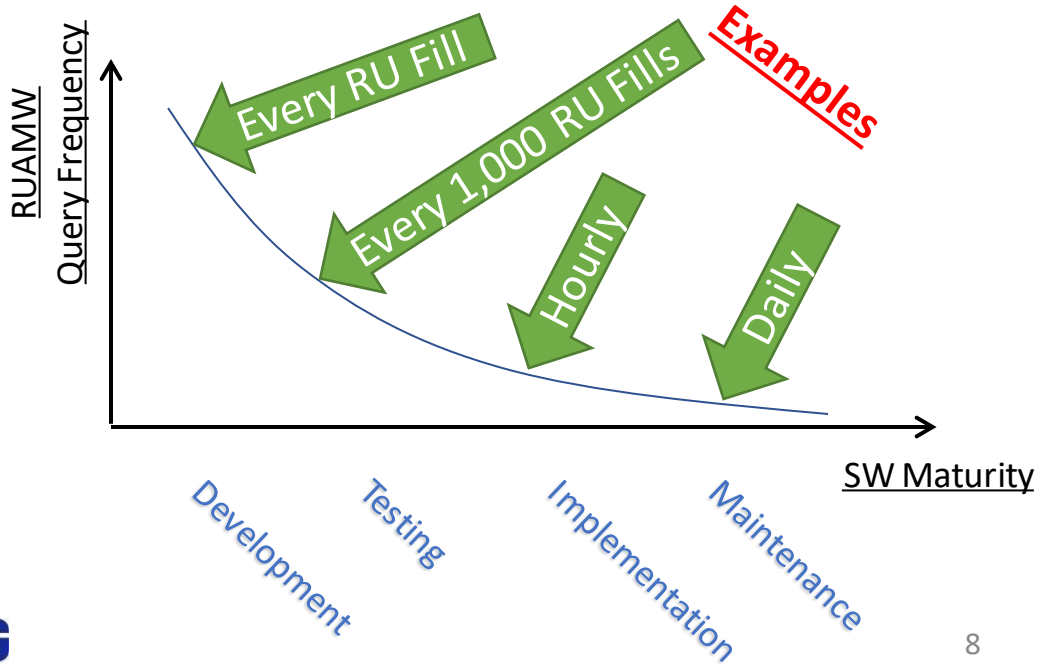


SAMSUNG

Intelligent Queries of RUAMW

- Reclaim Unit Available Media Writes (RUAMW)
  - Returned in the RUH Status Descriptor
  - Count of LBAs until RU is full
  - Check:
    Host_RU_Estimated_remaining_count == RUAMW
- Recommendations
  - Query RUAMW when RU is **almost** Full
    - If misalignment between Host and Drive expectations, reaction may be to reinitialize RU counts to zero.
    - "Reclaim Unit Handle Update" forces smallest amount of Fill data and WAF impact
  - Frequency of RUAMW queries should relate to the SW Maturity
  - Customer requirement: All RUs shall be equal to Reclaim Unit Nominal Size (RUNS) for the life of the drive.





SAMSUNG

Selecting Persistently Isolated vs Initially Isolated RUHs

Initially Isolated	Persistently Isolated
WAF=1 for tuned traffic	WAF=1 for tuned traffic
Most WAF reductions available due to OP	Most robust WAF to poorly tuned traffic
Minimal OP loss for each additional RUH	Larger OP losses per RUH (Incoming RU, GC RU, Tracking, …)
Nominal FDP Development and Validation	Increased Development and Validation Time
Minimal Cost Increases per RUH (Capacitors, buffers, …)	(Same) * (tracking overheads)
Probability of Die Collisions increases per RUH	(Same) * WAF_per_RUH

- Considerations at Acquisition
  - Persistently Isolated can inflate cost, power, and development time
  - Inflated (Persistently or Initially) RUH counts can inflate die collisions for reduced performance or increase power consumed
- Prioritized Recommendations in SW Deployment
  1. Use all the Persistently Isolated RUHs for the highest value applications
  2. Use all Initially Isolated RUHs available
  3. Tiered sharing of RUHs among applications based on 1) capacity and 2) write frequency
  4. Migrate the highest value applications to FDP friendly workloads and Initially Isolated RUHs – See prior slides

Figure Y: Initially Isolated Reclaim Unit Handles

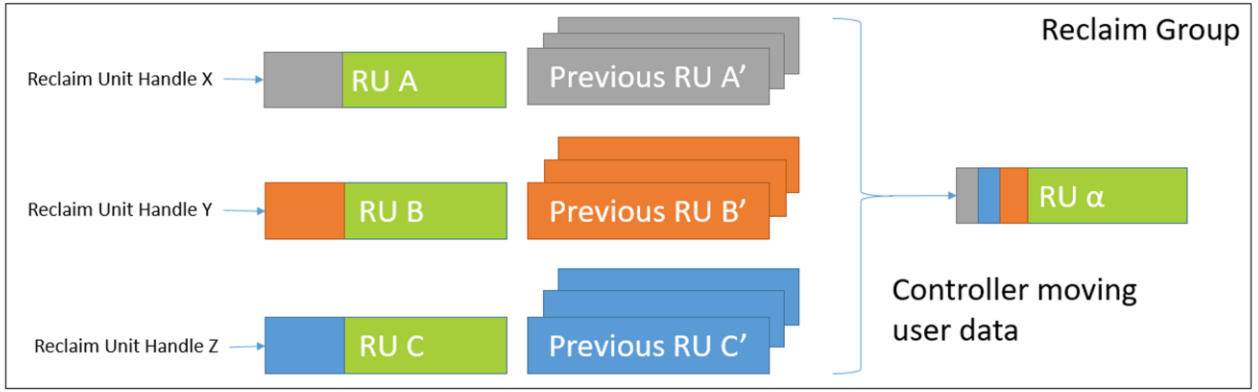
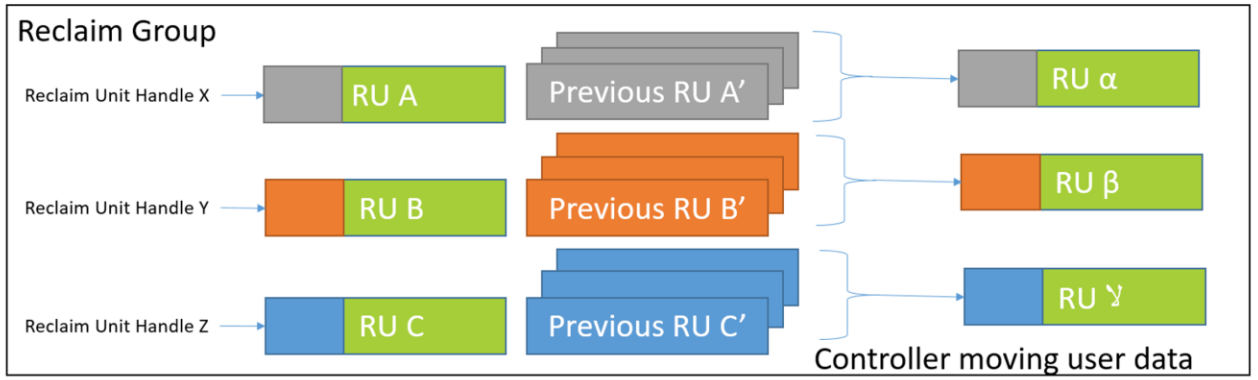


Figure Z: Persistently Isolated Reclaim Unit Handle



SAMSUNG

Estimated Active Reclaim Unit Time Remaining (EARUTR)

7:4	**Estimated Active Reclaim Unit Time Remaining (EARUTR):** This field indicates an estimate of the time in seconds that the Reclaim Unit currently referenced by the Reclaim Unit Handle is allowed to remain referenced by that Reclaim Unit Handle (refer to the Flexible Data Placement section in the NVM Express Base Specification) before the controller may modify the Reclaim Unit Handle to reference a different Reclaim Unit. This value is the remaining time at the time the I/O Management Receive command is processed by the controller. If this field is cleared to 0h, then no time is reported.

- Related:
  - Estimated Reclaim Unit Time Limit (ERUTL)
- NAND Impactors
  - EB to EB variations
  - Beginning of Life vs Ending of Life variations
  - Measurements can be noisy
  - Environmental impactors (Temperature)
  - Neighboring traffic (Disturb)
- Some optional SSD Mitigations
  - 1 drive setting for the life of the drive may be conservative
  - Improved estimates through RU open time
  - Sense if RU transition is required prior to triggering activity
- Recommended SW Practices
  - Do not assume EARUTR or ERUTL is accurate
  - Do not assume EARUTR will always decrease
  - Do not assume EARUTR will transition to a new RUH upon reaching zero
- Use ERUTL and EARUTR as informative
  - ERUTL set data management policies
  - EARUTR for general scheduling updates after a power fail or other scenario
  - Avoid managing ERUTL or EARUTR as if they are errors
- Drives will already minimize these moves as much as possible
  - The cost of a miss is 1 PE cycle for 1 RU. Low.

SAMSUNG

Conclusions

- FDP enables several choices for improved WAF
  - Both Write constrained and Deallocate constrained can achieve WAF=1
  - Imperfections in traffic may still yield low WAF due to drive OP
- RUAMW query frequencies at varying SW Development stages were discussed
- Helped inform decisions on Persistently Isolated vs Initially Isolated RUHs at acquisition and deployment
- Estimated Reclaim Unit Times
  - ERUTL can inform data center policy decisions
  - EARUTR can prioritize traffic. Ex: Recovery after a power fail

SAMSUNG