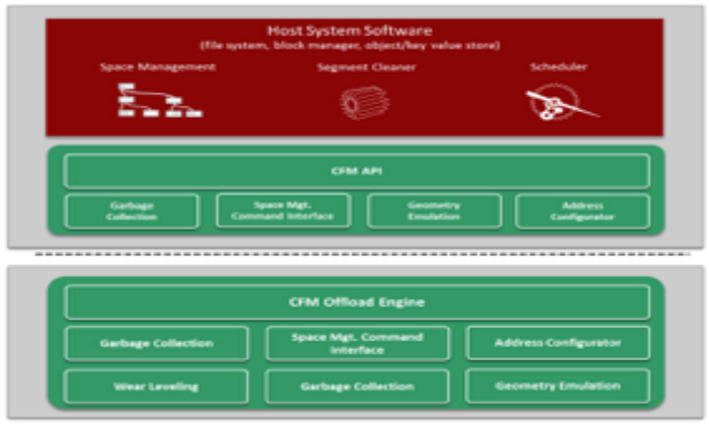
Integrating Cooperative Flash Management with SMR Technology for Optimized Tiering in Hybrid Systems

Alan Chen
Software Architect
Radian Memory Systems

Introduction

Who:
Radian Memory Systems offers Cooperative Flash Management as a replacement to conventional Flash Translation Layers



Audience:
This talk is oriented to storage stacks in data centers
Assumes a basic understanding of SMR and Flash

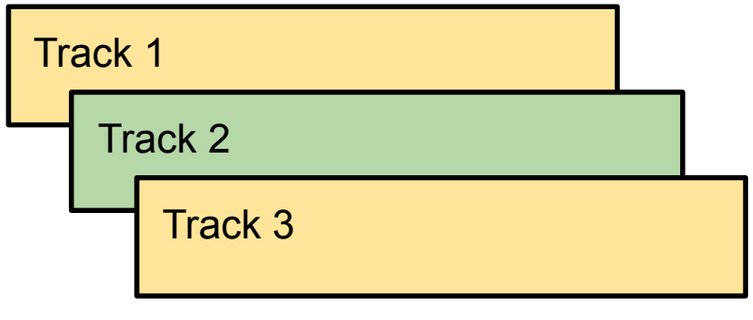
Why:
A number of filesystems doing SMR work, ext4, XFS, ZFS, proprietary stacks.
If you tackling SMR work or thinking about it, you might benefit in other ways by integrating Cooperative Flash Management (CFM)

Outline

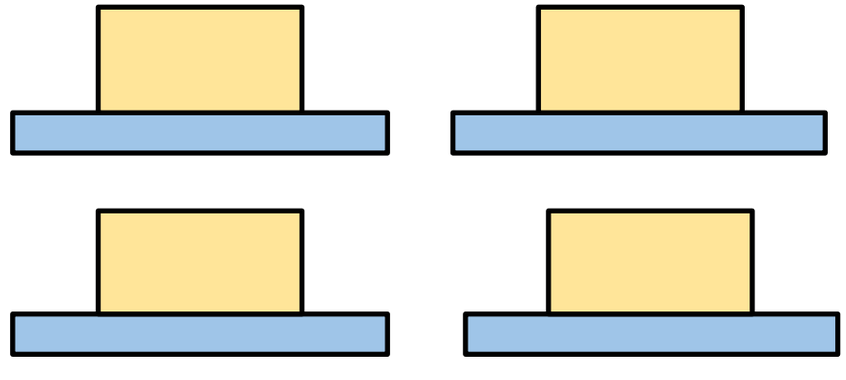
- Flash and SMR Commonalities
- What Is Cooperative Flash Management (CFM)
- xTL Translation Layer Benefits and Costs
- CFM and SMR APIs & Filesystems
- Hybrid System Design Considerations for SMR and CFM

Low Level SMR and Flash Commonalities

SMR Zones

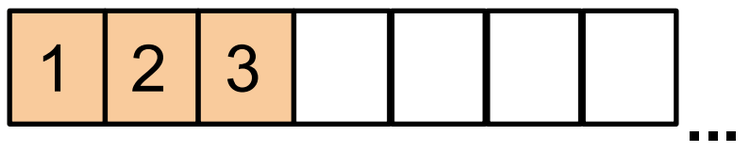


NAND Flash Blocks



Different physical structures, but logical operation of new medias are asymmetric

Random Read, Sequential Write, Erase/Reset before Overwrite



High Level CFM and SMR Commonalities

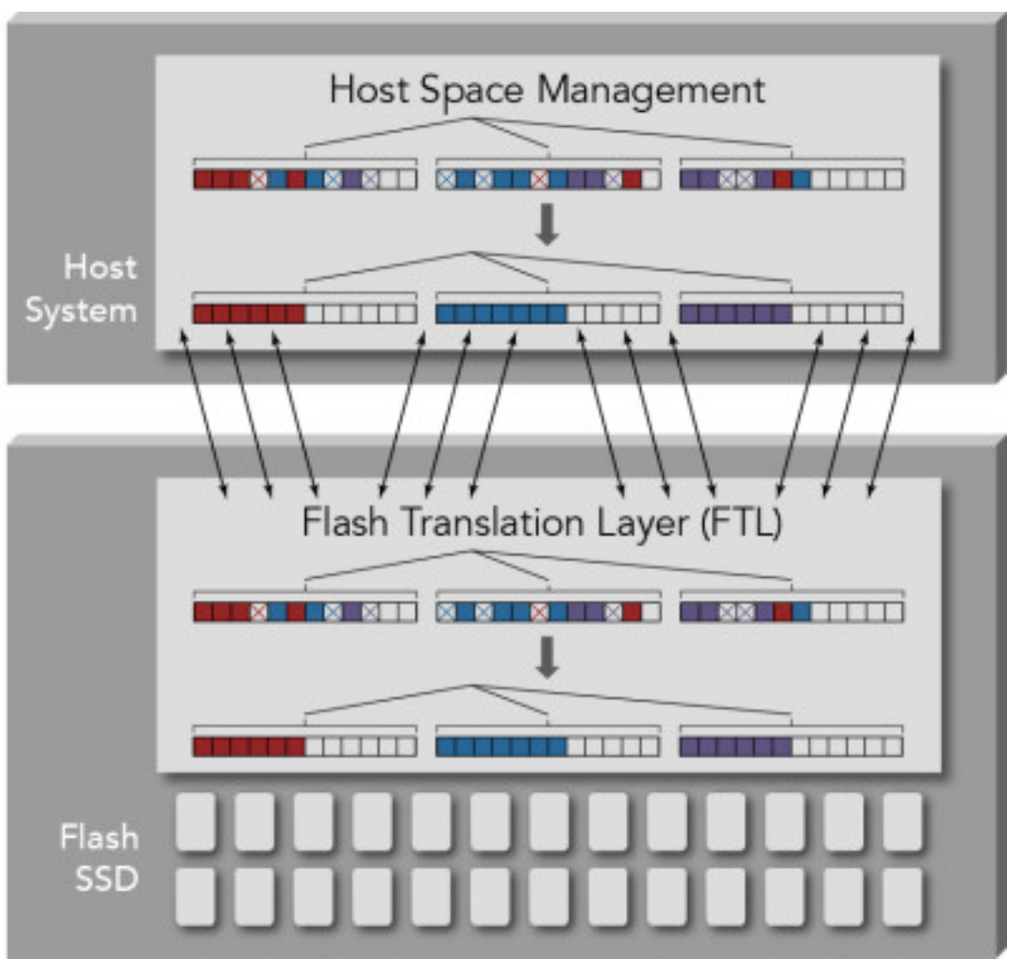
Benefit: Logical to Physical Translation Layer used to provide backwards compatibility to traditional block device interfaces

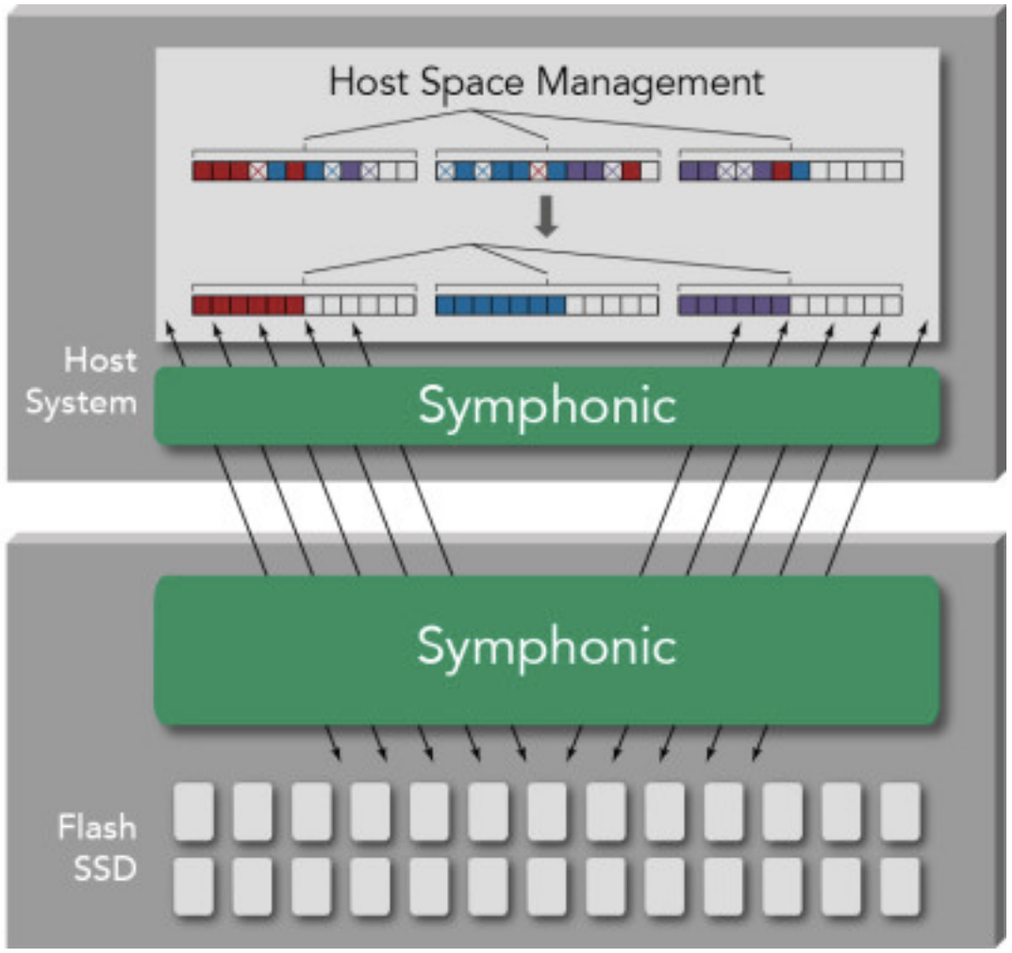
Provides an interface that allows Read, Write, and Random Overwrite

Priority	***Flash SSD***	***SMR Drive***
Backwards Compatibility	***Conventional SSD (FTL)***	***Drive Managed (D-MTL), Host Aware***
Performance & Cost	***Cooperative Flash Management (CFM)***	***Host Managed***

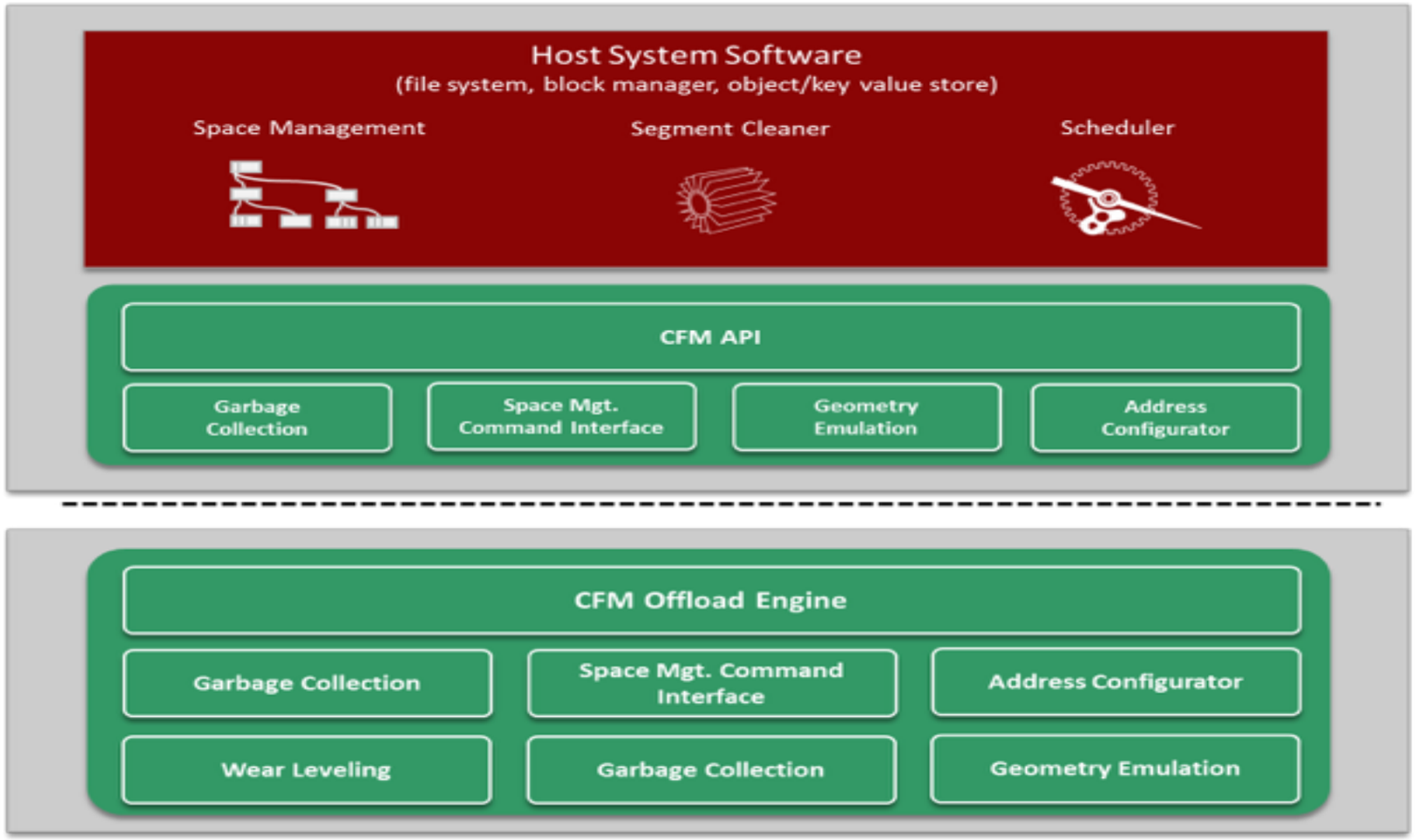
Both devices expend significant resources to provide the translation with suboptimal results due to data vs media lifetime disconnects.

Cooperative Flash Management Intro





Cooperative Flash Management Intro



- Abstraction of Low Level NAND
  - Bad Blocks
  - Wear Leveling
- Forward Compatibility
- Offload engine
- Address Configurator

What is Cooperative Flash Management?

Cooperative Flash Management (CFM) eliminates the FTL and exposes essential flash constraints, without raw flash headaches.

System operates in host-address space

System commands all drive operations - no asynchronous disconnects for garbage collection

Most similar to SMR host managed mode with similar goals

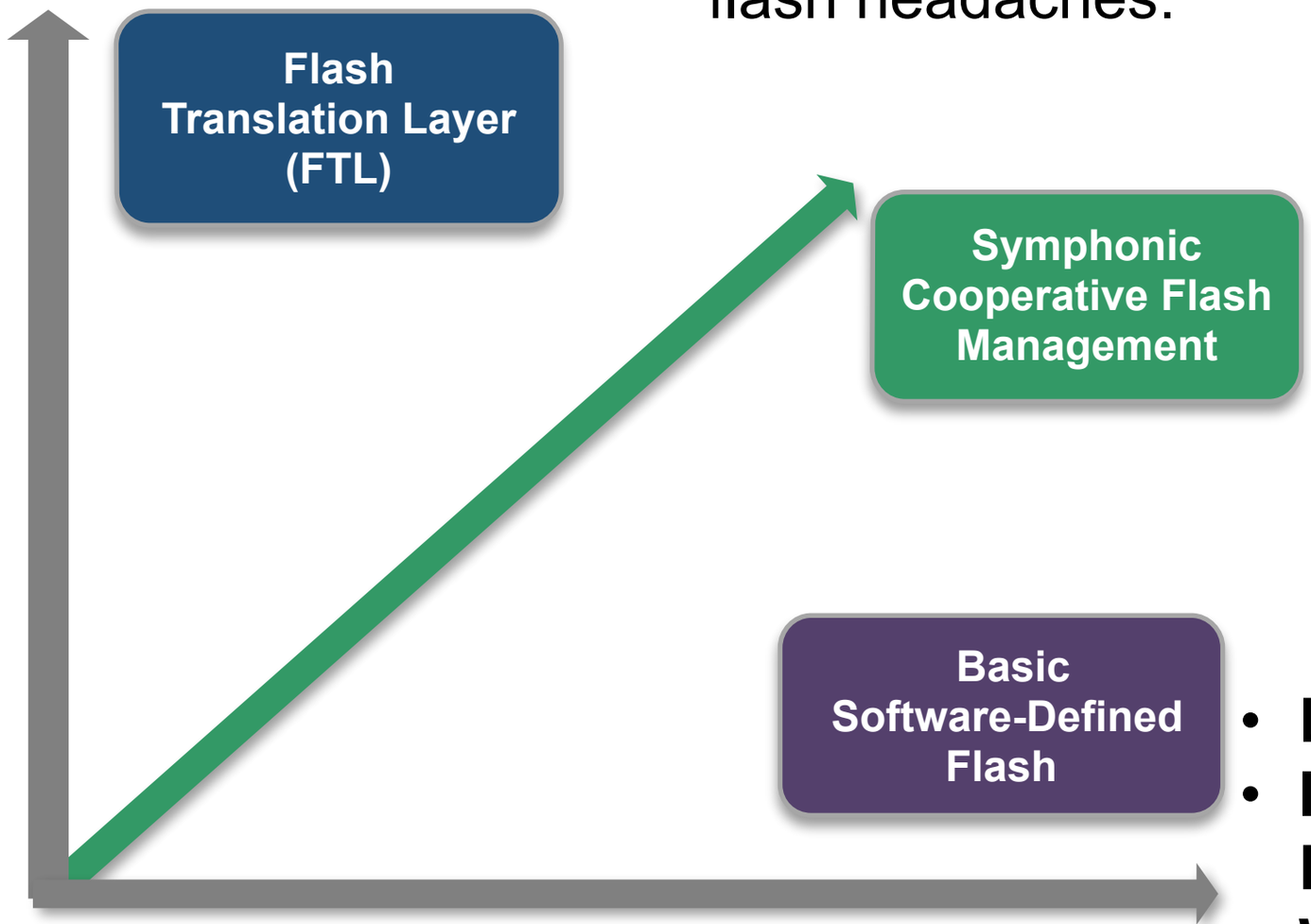
Radian manufactures a CFM device, the model RMS-250 Symphonic Drive.

What is Cooperative Flash Management?

Radian MEMORY SYSTEMS

- **Integration**
- **Backward / Forward Compatibility**
- **Reliability**
- **Scalability**

Cooperative Flash Management (CFM) eliminates the FTL and exposes essential flash constraints, without raw flash headaches.



- **Latency / QoS**
- **IOPS / Bandwidth**
- **Wear Out**
- **Cost**

FTL vs CFM vs Raw NAND Flash

Radian MEMORY SYSTEMS

	FTL SSD	*Raw SW Flash*	*Radian CFM*	*SMR HM*
Address Space	***Fully virtual***	***Physical***	***Host Physical Space***	***Host***
Garbage Collection	***Transparent Asynchronous***	***Host based***	***Host control, device offload***	***Host***
Wear Leveling	***Transparent Asynchronous***	***Host based***	***Host control, device offload***	***-***
Overprovisioning	***10-30%***	***3%***	***3%***	***Spare Sectors Only***
Multi Drive Scalability	***Excellent***	***Poor host resource***	***Excellent***	***Excellent***
Upgradability	***Excellent***	***Problematic***	***Excellent***	***Excellent***

What is the Cost of the Translation Layer?

Flash SSD	*SMR Drive*
Conventional SSD (FTL)	***Drive Managed (D-MTL) Host Aware***
Cooperative Flash Management	***Host Managed***

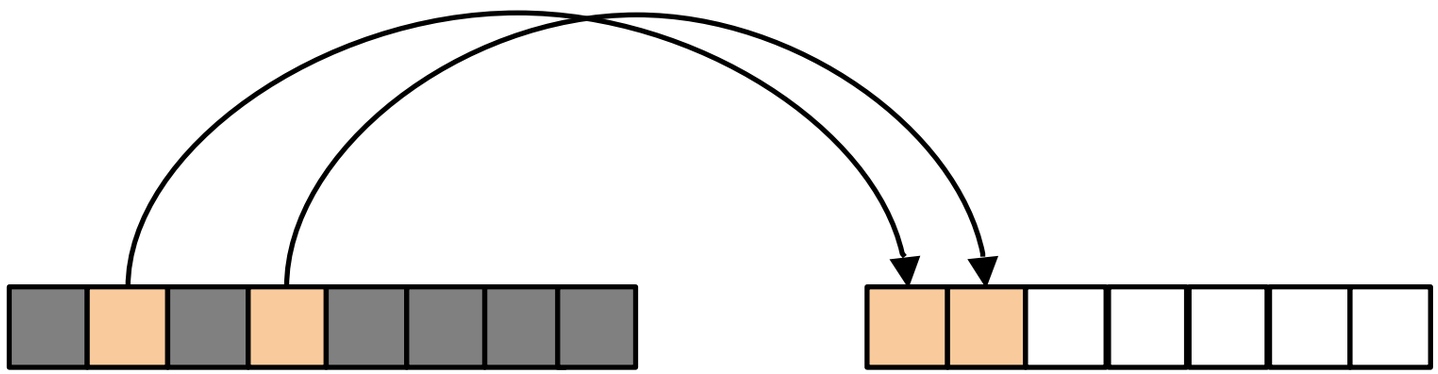
Translation layers provide nice backwards compatibility.

Why leave the comfortable world of the xTL?

Garbage Collection & Write Amplification

Write Amplification is undesirable, but is fundamental to Flash & SMR Media:

Random Read, Sequential Write, Erase/Reset before Overwrite



Stale Valid Unwritten

Valid data - data written by the filesystem which is still referenced

Stale data - data previously written by the filesystem which is no longer needed (but is still present)

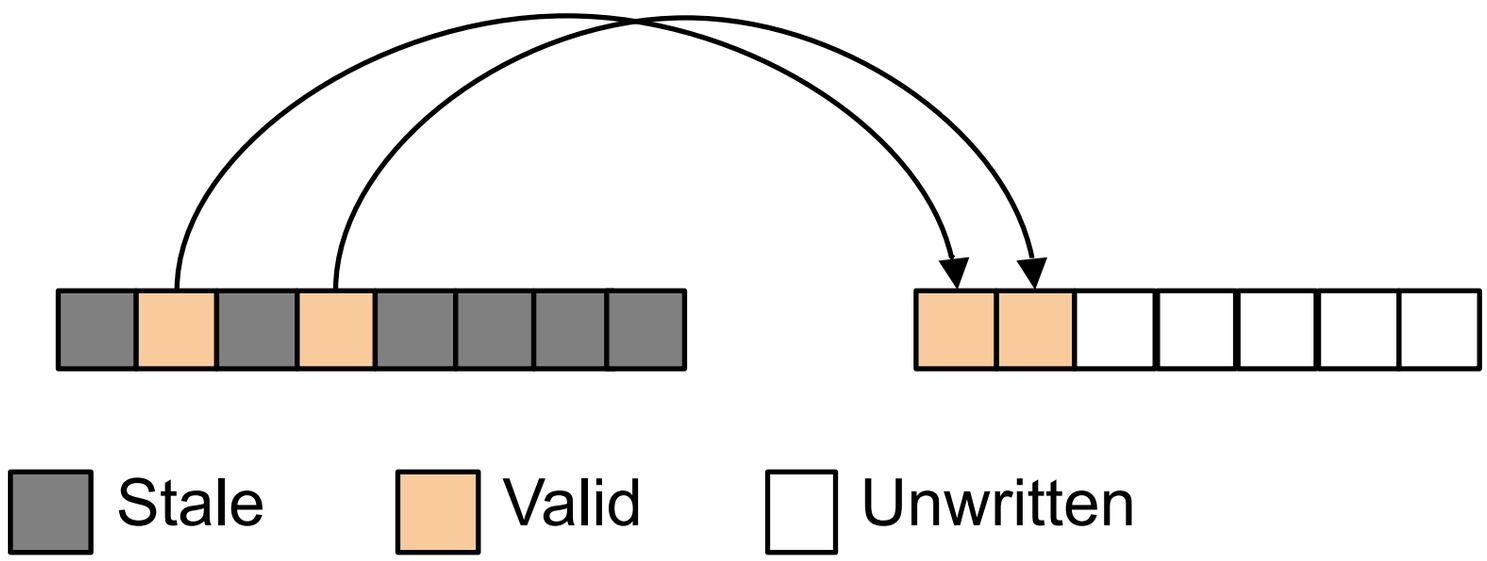
Unwritten - unwritten area of the EU or Zone

GC/WA: Rough Performance for Raw Media

Mixed IO Case	*NAND Flash (NVMe)*	*SMR (SAS)**
"Segment" Size	*2-8 MiB (Block Erase Unit)*	*100-300 MiB (Zone)*
Parallelism	*per Flash die (128 - 256 ish per drive)*	*per Spindle (1 per drive), per Head (~12 per drive)*
Write Rate	*5-30 MiB/sec/die (1 die, unbuffered)*	*140-300 MiB/sec*
Read Rate	*1+ GiB/sec/die (1 die)*	*140-300 MiB/sec*
Write 'Access' Time	*2 msec (Tprog)*	*3 avg [0.3 - 8] msec (seek)*
Read 'Access' Time	*100 [50 - 300] usec (Tread)*	*3 avg [0.3 - 8] msec (seek)*

** SMR performance is estimated for a host-managed mode looking a SAS drives and SMR Archive drive performance*

xTL Latency Single GC Latency



Rough cost in latency to garbage collect a single NAND EU vs a Zone at 25% validity

Flash: (~ 1 MiB / 4MiB)

single die 25 page sized read+write transfers => **53 ms**

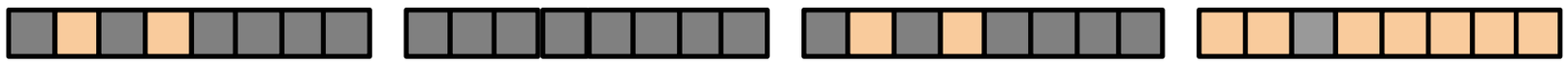
SMR: (~ 30 MiB / 125 MiB)

read + write => **200ms +** (3 ms avg seek time x # seeks)

Other GC Factors

Flash

- High parallelism - the EU is on one of many flash dies, smaller granularity
- Higher likelihood that the valid percentage of EU is smaller
- Disadvantage of write endurance limits

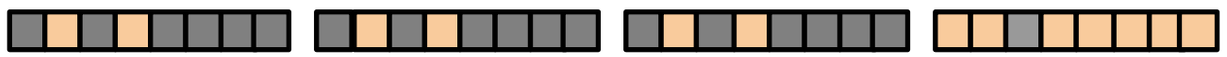


SMR

- Lower parallelism
- Good transfer speed once started, disadvantage of larger granularity
- Likelihood of relatively larger valid data percentage in Zone
- Seek cost introduces variance due to relative zone layout

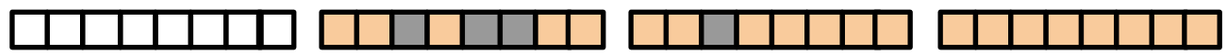
xTL Over Provisioning

Managing accounting of many Garbage Collection operations requires extra space (over provisioning) or extra operations.



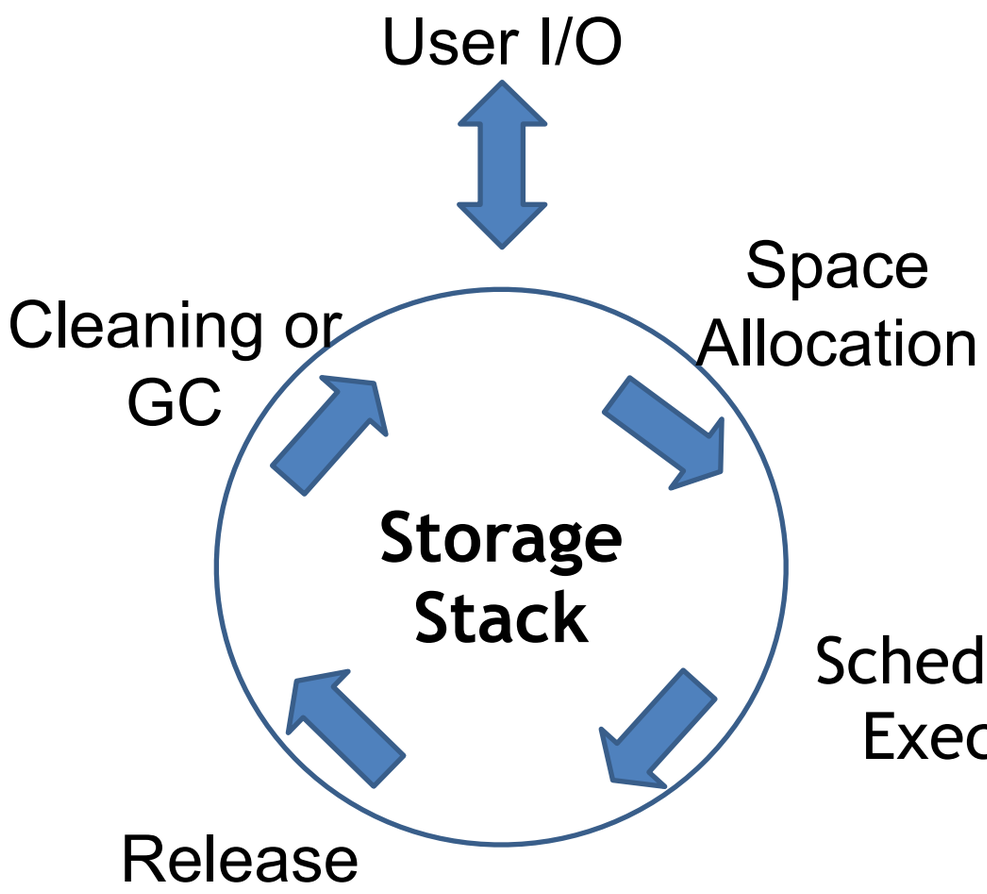
Stuck - nowhere to go

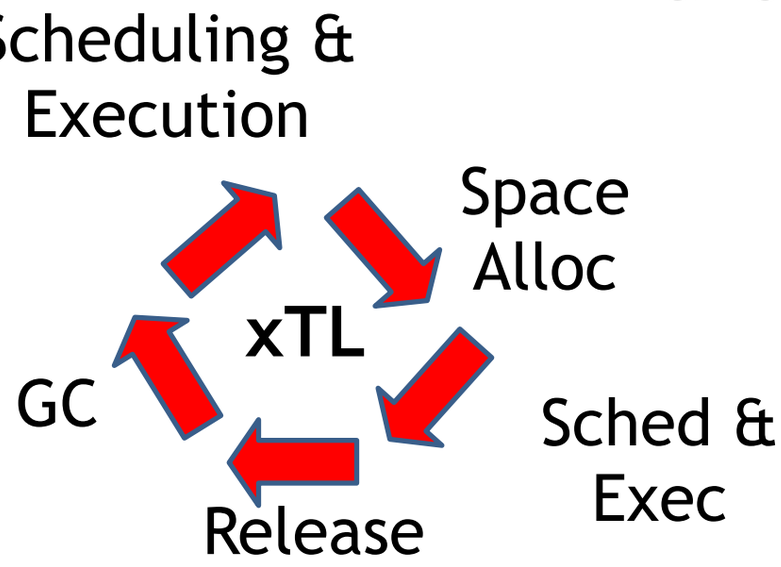
Must maintain at least one open zone



Performance/Write Amp Concerns? Must maintain more open zones

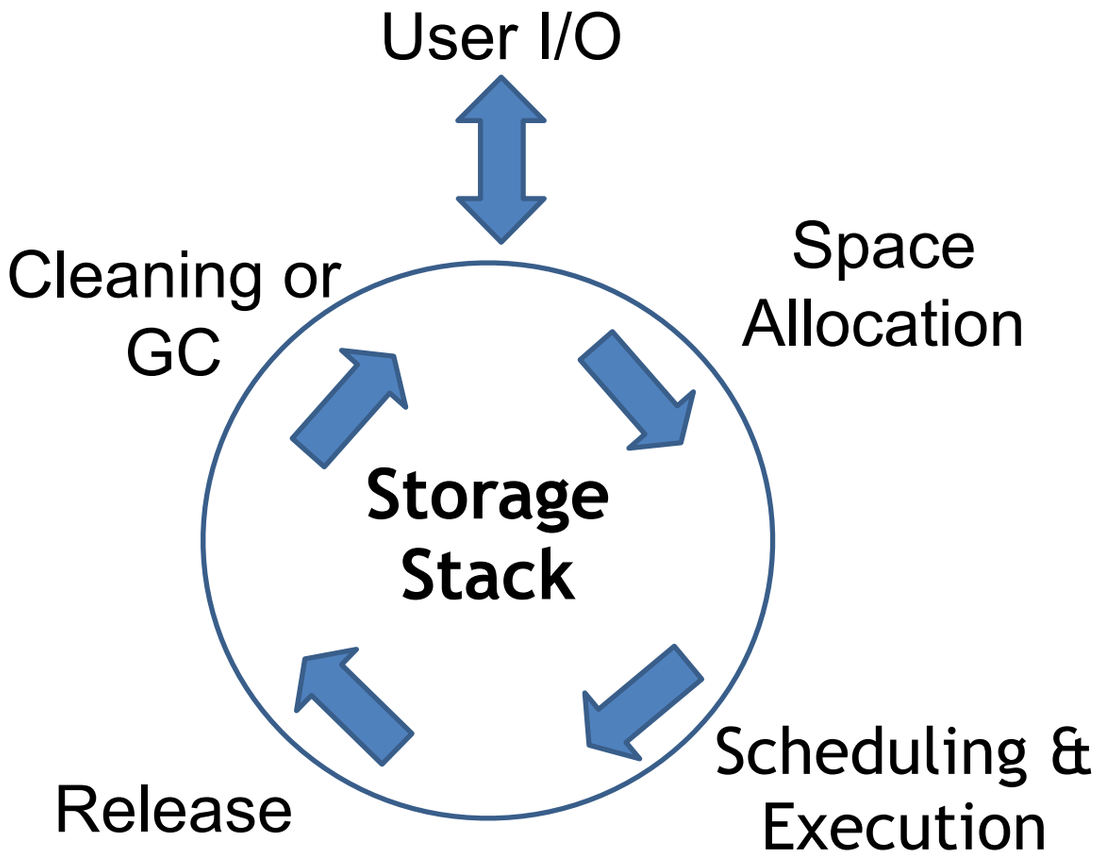
Translation Layer System Level





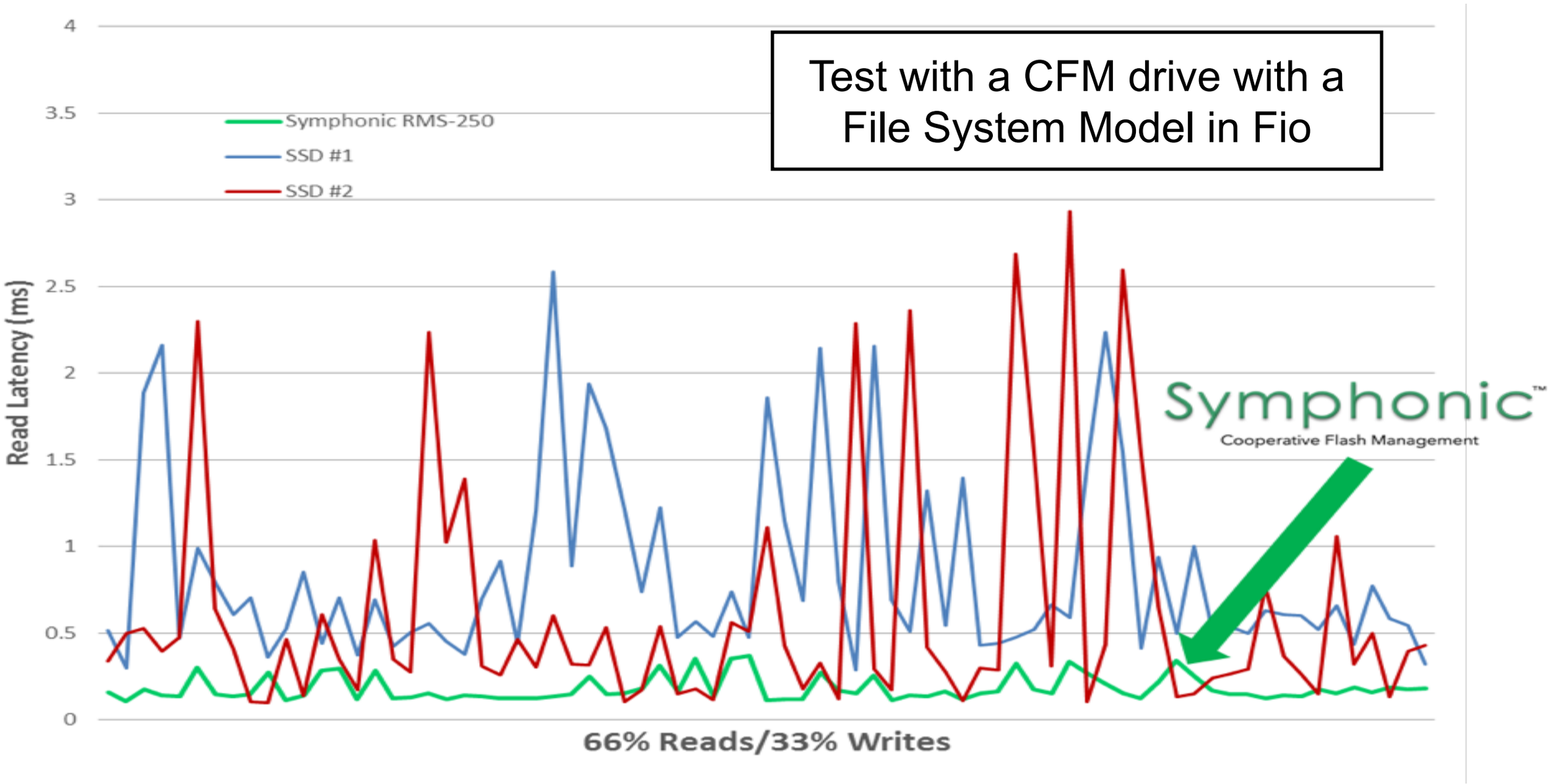
- Alignment of design choices and block sizes are needed to optimize the performance of a storage stack
- Performance and latency is lost with every disconnect
- FTLs ownership of the internal lifecycle of flash causes a large amount of disconnect

Translation Layer System Level

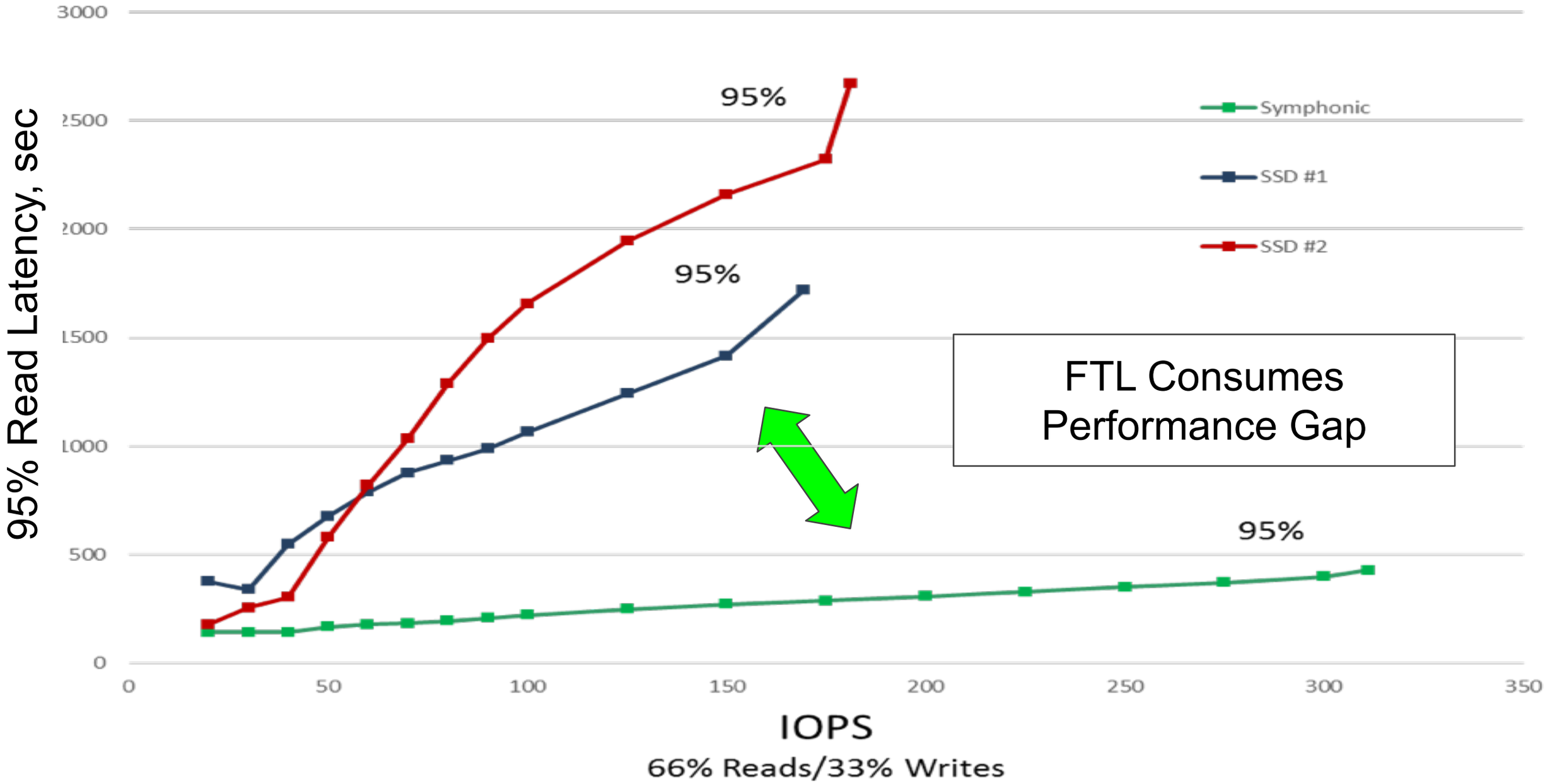


Alignment of design choices along the lifecycle delivers increased efficiency and reduced latency

Symphonic CFM Performance

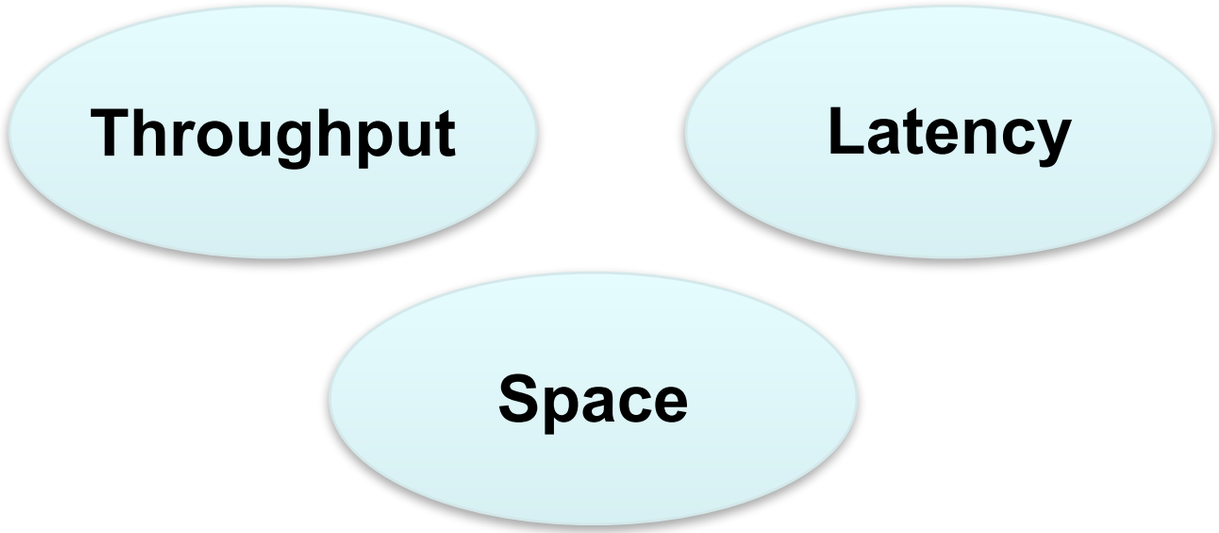


FTL Cost in Performance



Raw Media Performance and xTL

Flash Translation Layers (FTL) and Drive-Managed Translation Layers (DMTL) expend raw media performance to provide for a backward-compatible block overwrite paradigm.



How do we get closer to the raw media performance?

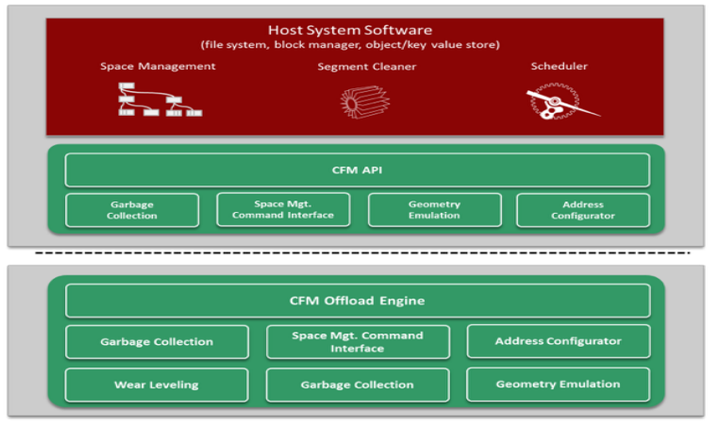
For Flash, Cooperative Flash Management (Symphonic)
For SMR, host managed mode

SMR Host Mode & CFM Commonalities For File Systems / Storage Stacks

SMR ZBC vs CFM API

Zoned Block Commands (ZBC)
- SCSI T10
- ATA T13 (ZAC)
- Applies to SMR Media
- Possible Driver and OS level modifications

Cooperative Flash Management
- Host Layer API
- NVMe
- NVMe Vendor Commands
- Applies Flash Media & NVRAM
- No Driver or OS specific modifications



CFM & SMR API Commonalities

Key Crossover Concepts in the APIs

CFM	*SMR ZBC*
Segments	***Zones***
Non-Overwriting Access	***Sequential, Non-Overwriting***
Segment / EU Erase	***Write Pointer Reset***

A CFM Segment is a configurable grouping of Flash resources equivalent of an SMR Zone

File System Awareness

Log structured filesystems look ideal for CFM and SMR operation

But even Log structured / CoW are rarely 100% “non-overwriting”

- metadata locations needs to be reconsidered
  - may need to layout to specific areas
  - may need to reduce range of metadata overwrites
  - advanced storage stacks may separate or distribute metadata

- overwrite is tempting even in non-metadata areas
  - File systems/storage stack must be rewritten to remove legacy overwrite optimizations

Other Commonalities: File System Lifetimes

Radian MEMORY SYSTEMS

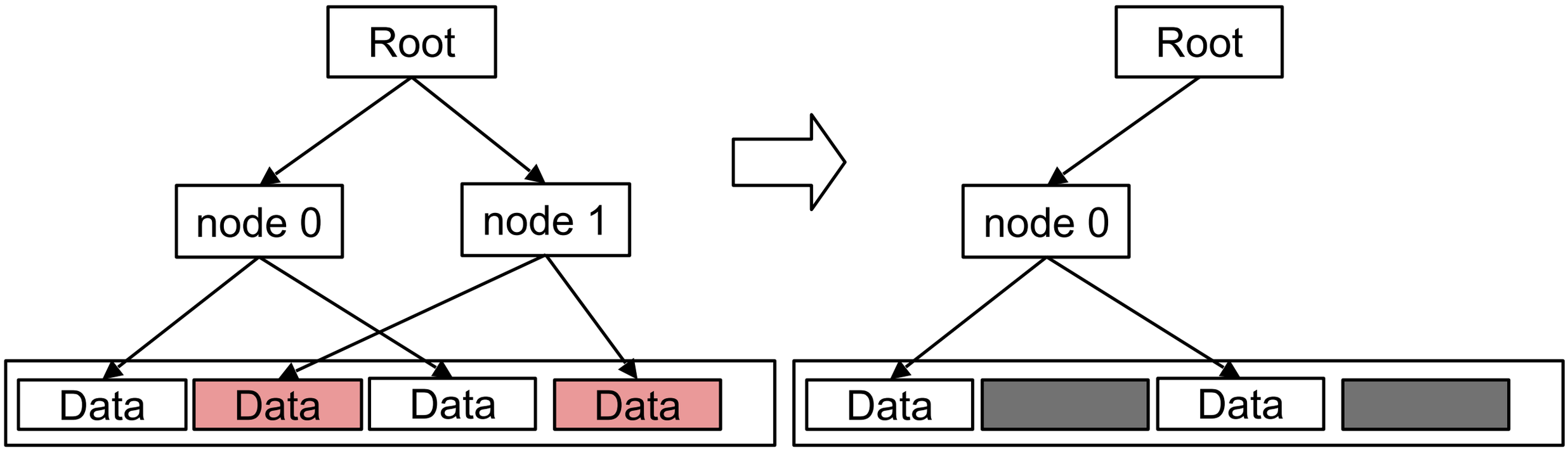
Alignment of FS (meta)data blocks vs media ‘segments’ are different

- Lifetime of data (FS standpoint) vs
- Lifetime of media segments (drive standpoint)

Leads to lifetime management work - moving of data that is still valid, but the zone/segment needs to be reset to reclaim free space.

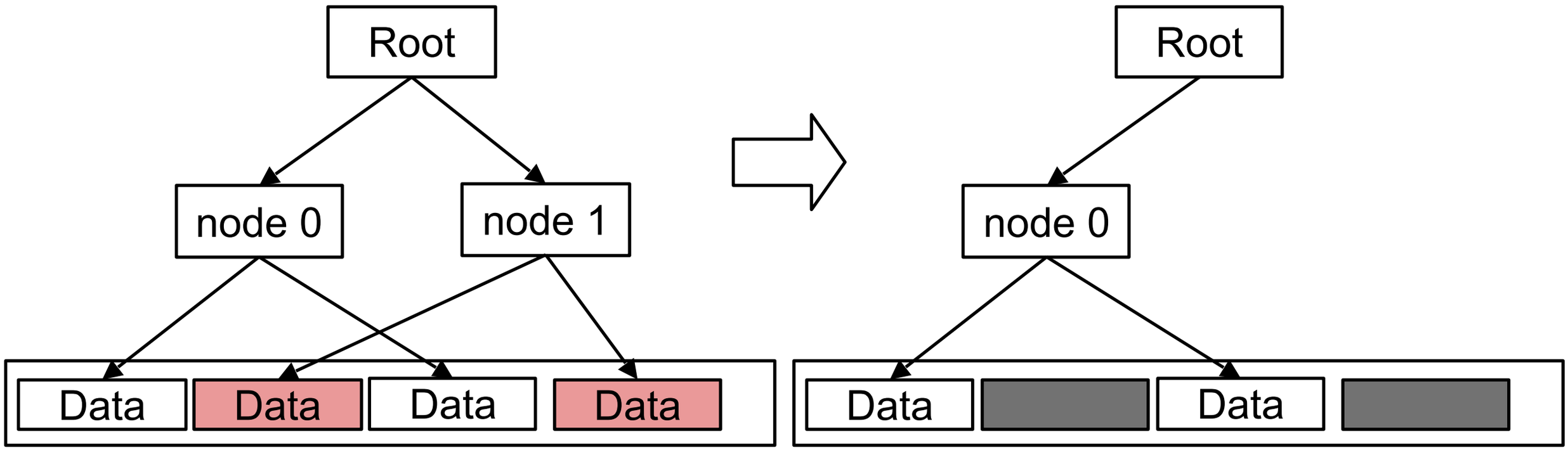
Number and % valid for GC transfers minimized vs FS and Drive independently managing lifetime.

Space Management Considerations



File contained in node 1 is deleted, but invalid data blocks cannot be rewritten until the entire segment is erased/reset. However, data blocks connected to node 0 still need to be kept.

Space Management Considerations

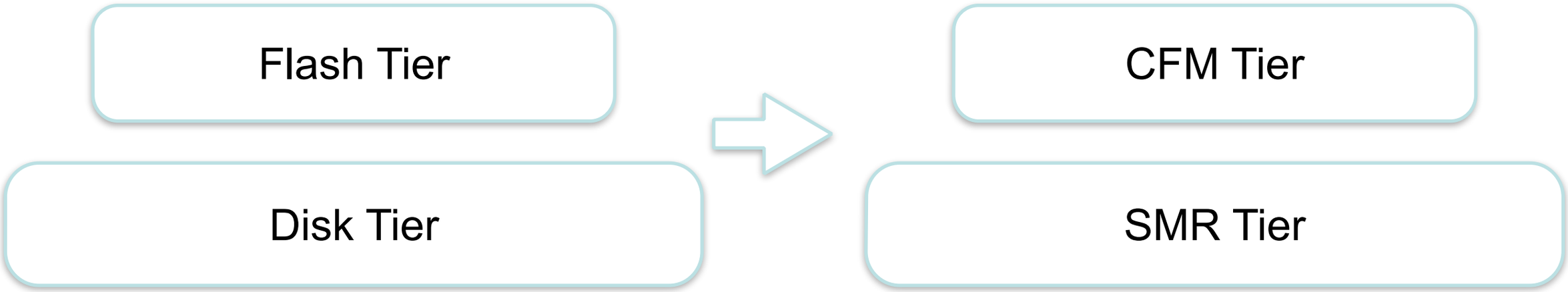


SMR: ~100-200 MiB sized zones
CFM: ASL Segments are configurable ~4 MiB and up

Radian MEMORY SYSTEMS

Hybrid System Design Considerations

A Notional Hybrid System

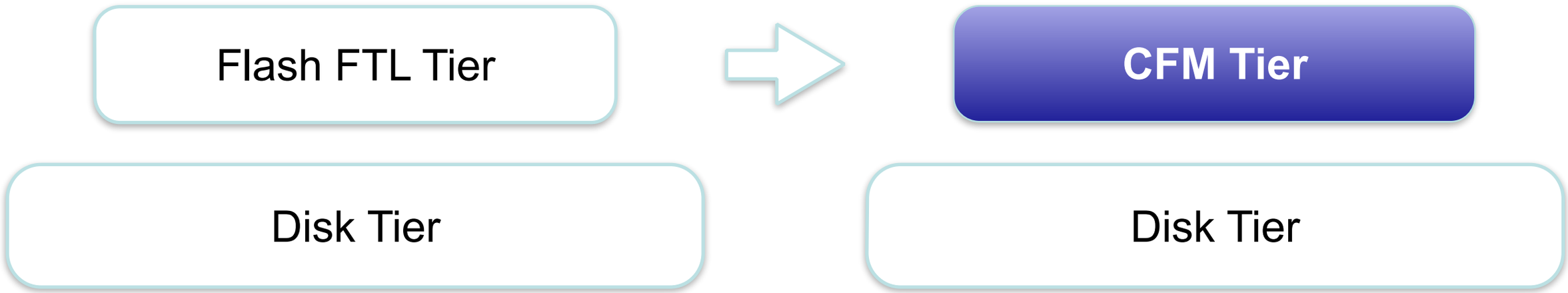


CFM allows the flash tier of the system to increase performance with finer control of latency and data lifetime.

HDD SMR allows the disk tier of the system to increase the storage depth at exceptionally low cost

This allows expansion of the system envelope in two directions and overall performance to increase.

CFM Tier Considerations



FTLs perform data relocation (garbage collection) transparently to the host system

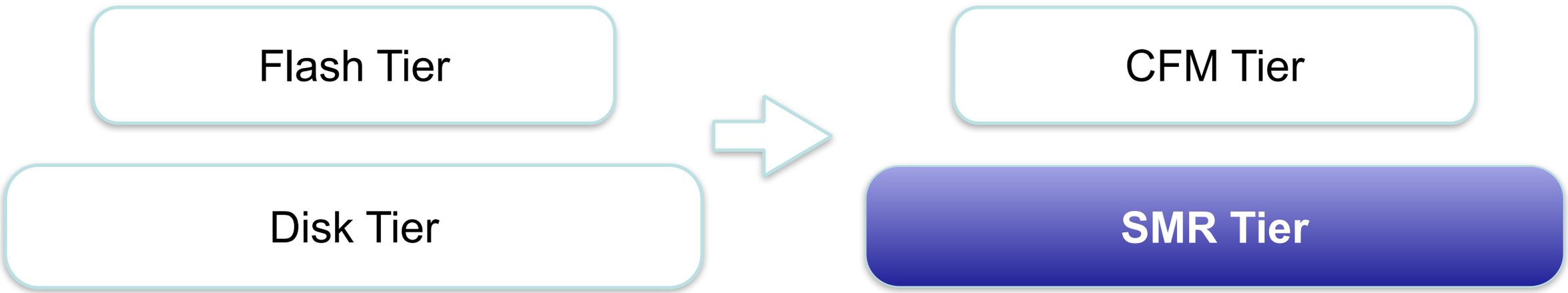
CFM provides hosts visibility into the Flash data relocation process

Initial efficiency from making storage stack CFM aware

Further efficiency in system with new design options for hosts to more aggressively and actively insert hot/cold intelligence into requisite garbage collection processes

Amortizes costs of relocation more efficiently – price is already being paid

Hybrid System Design Considerations



Advantages are even more compelling with SMR

Storage stack handling both CFM and SMR is aware of media alignment sizes

Like Flash, optimal management of SMR requires host intelligence for relocating valid data

Hybrid System Design Considerations

Unoptimized Host

Asymmetric Optimized Host

Flash Tier

CFM Tier

Disk Tier

SMR Tier

CFM + Host managed SMR creates more efficiencies, amortizes fixed costs, and opens yet larger space for system designers

Unoptimized Hosts value backwards compatibility over cost and performance

Combined asymmetric aware storage stack

Integrated Stack: Degrees of Freedom

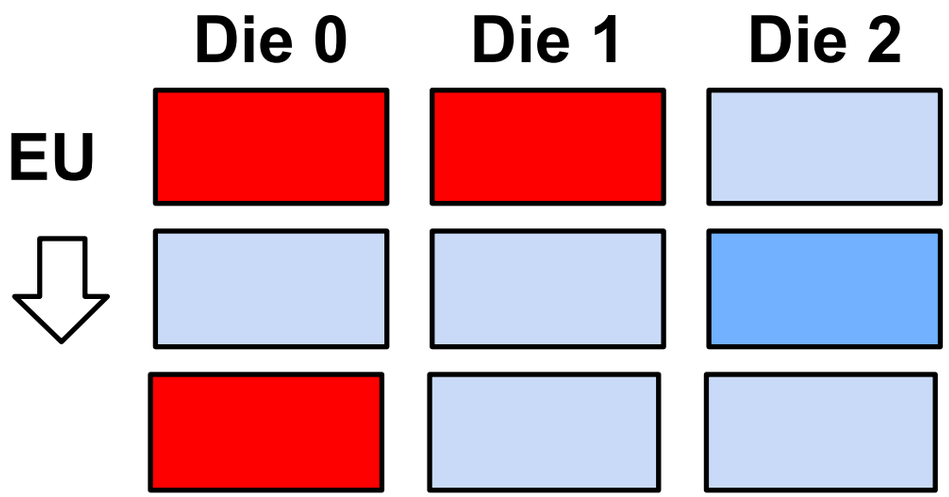
Solving these high level issues for one media, paves the way to operate with the other, SMR Host Managed integration allows advanced Flash integration with CFM.

But, some differences should be accounted for in a stack handling both CFM & SMR-HM:

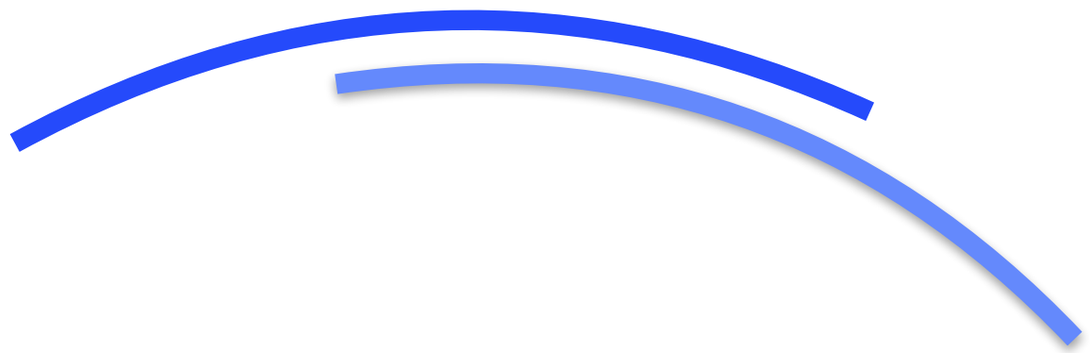
- Parallelism differences might drive different scheduling strategies
- Different segment sizes and rewritable areas sizes might drive different layout decisions
- Different performance might also drive layout decisions

Hybrid System Media Layout

Flash Layout



SMR Layout

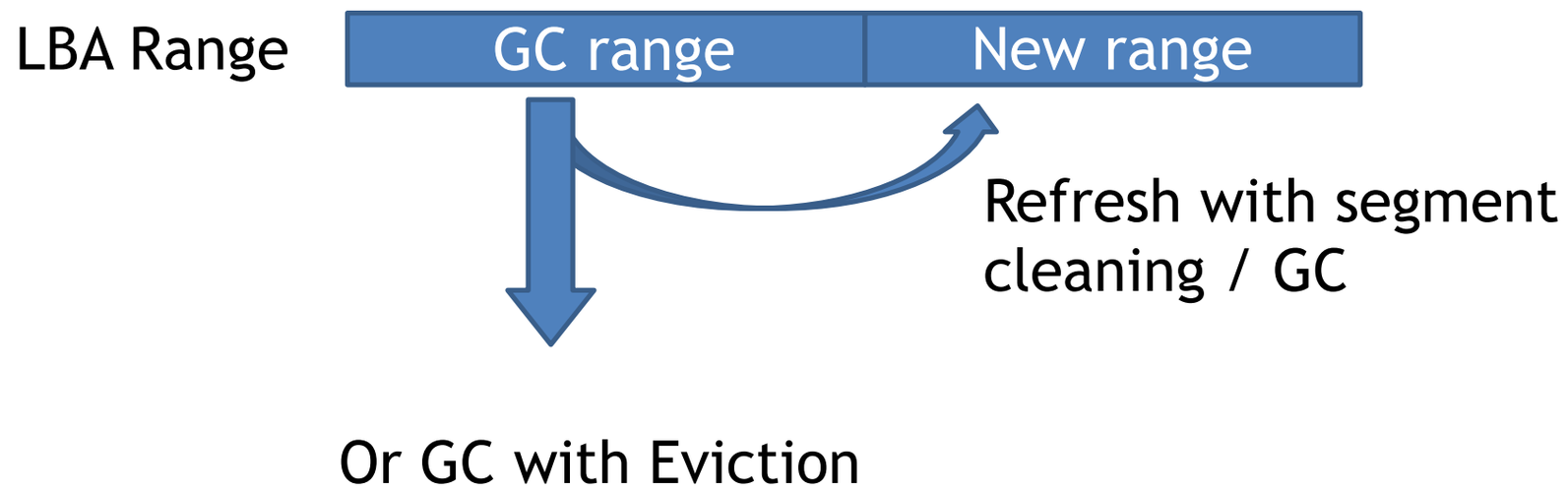


Open layout allows arrangement by relative probability of access (rel. hot / cold)
Or to optimize write areas or minimize seek distances
Or other custom layout strategies

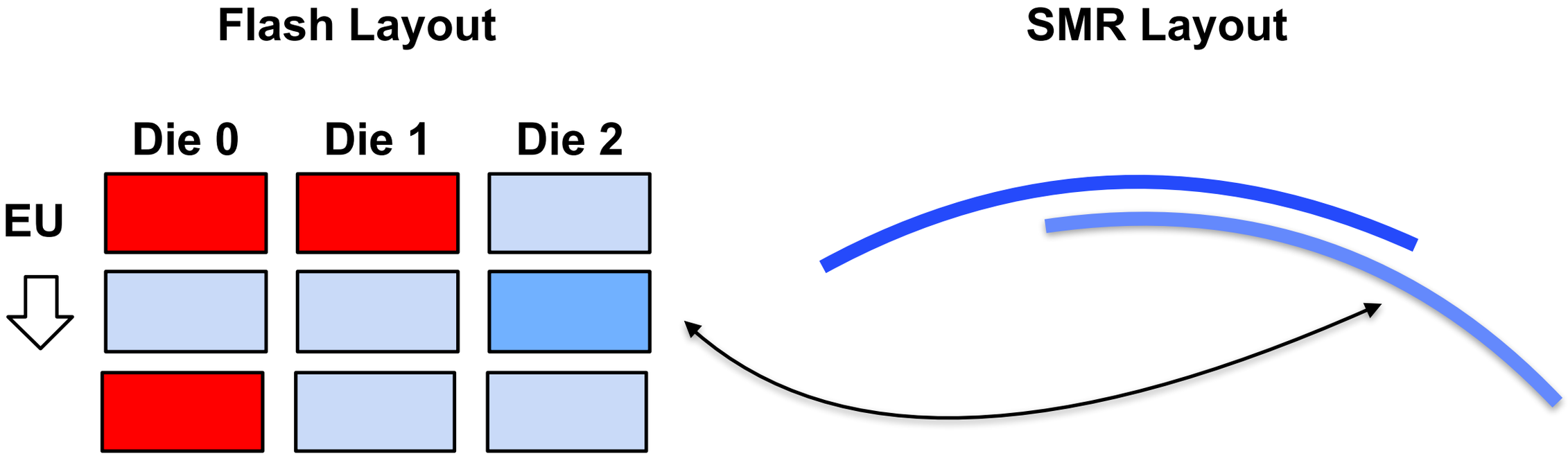
Opportunity for Tier Transitions

Host may also coordinate system garbage collection / segment cleaning with incoming segment fill, data tier eviction, or other system processes

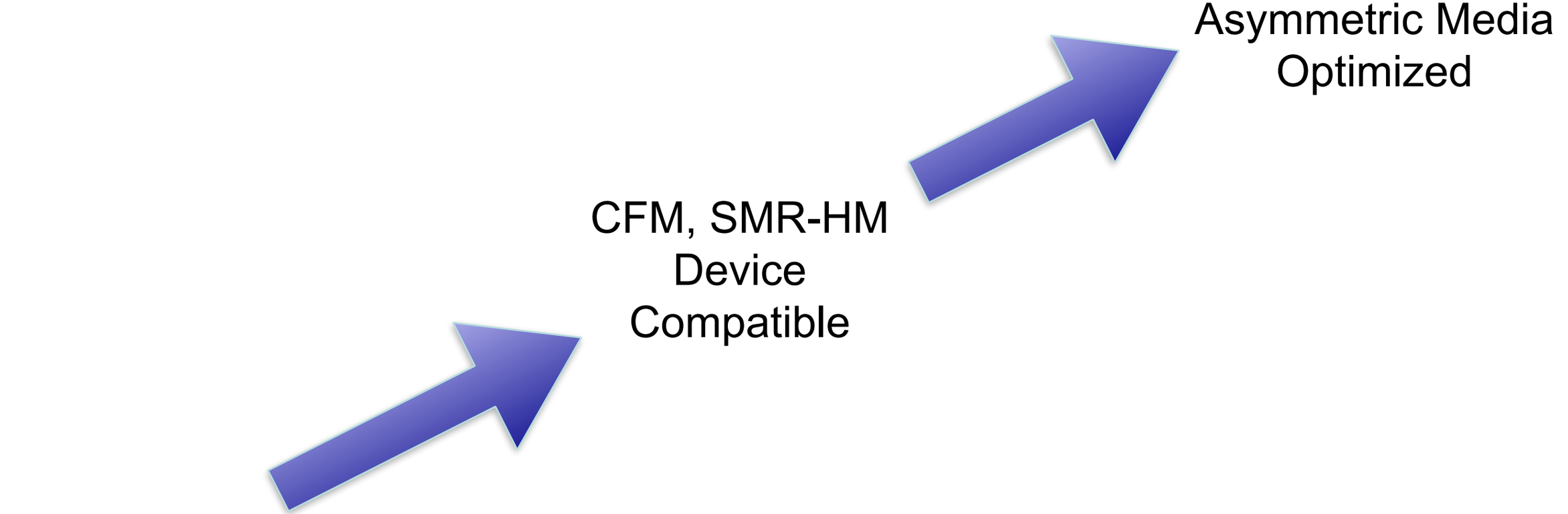
Convert some required GC operations to system operations to improve efficiency



Hybrid System Tier Movements



Hybrid Considerations



Thanks!

Questions?

Contact Info
Alan Chen
achen@radianmemory.com
http://www.radianmemory.com