Advanced Format in Legacy Infrastructures
More Transparent than Disruptive

Sponsored by IDEMA

Presented by Curtis E. Stevens

Agenda

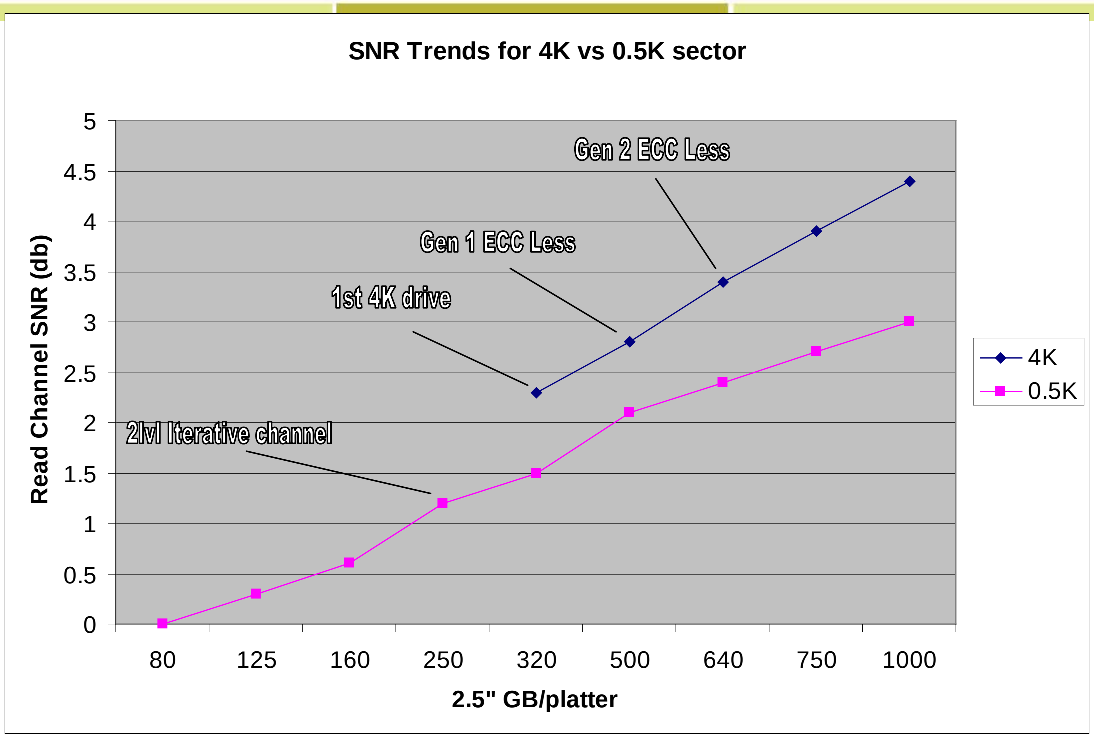
- AF History
- Enterprise AF
- Futures – SMR & LBA Indirection
- Hybrids & SSDs

AF History & Direction

- Advanced Format HDDs shipping for 18 months in AF 512e Formats – No more education required, industry adopting
- Presentation and educational efforts focused on Enterprise customers in last year
- AF 4Kn conversations are on track with most OEMs
  - Most customers are looking for well-defined transition paths between legacy 512b and 4kn to include 512e as the real solution
  - IDEMA member companies having to counter publications that say AF HDDs are less reliable than legacy with unneeded qualification validation
- AF has lead to meaningful conversations on SMR and more LBA indirection to the host
  - Enterprise customers excited about the possibilities

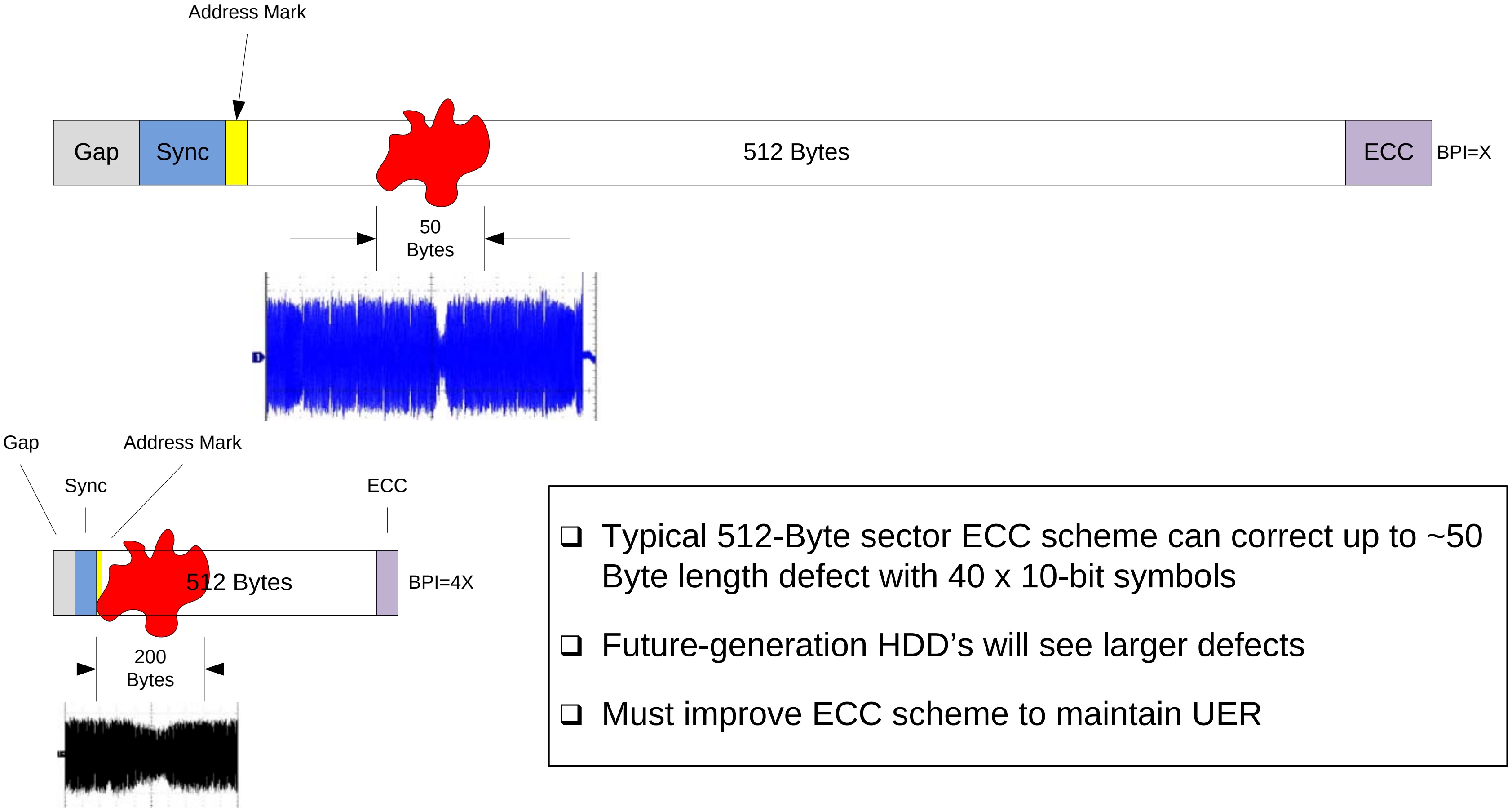
Enterprise AF

Why 4K media?



- 4K sector SNR advantages
  - Format efficiency gains due to removal of redundancy with a larger sector
    - PLO / gap reduction / overhead reduction
  - 4K SNR slope trend is much steeper than 0.5K slope
    - Limited coding tricks and signal processing gains available for 0.5K
    - Larger sector size enables longer codes and signal processing techniques

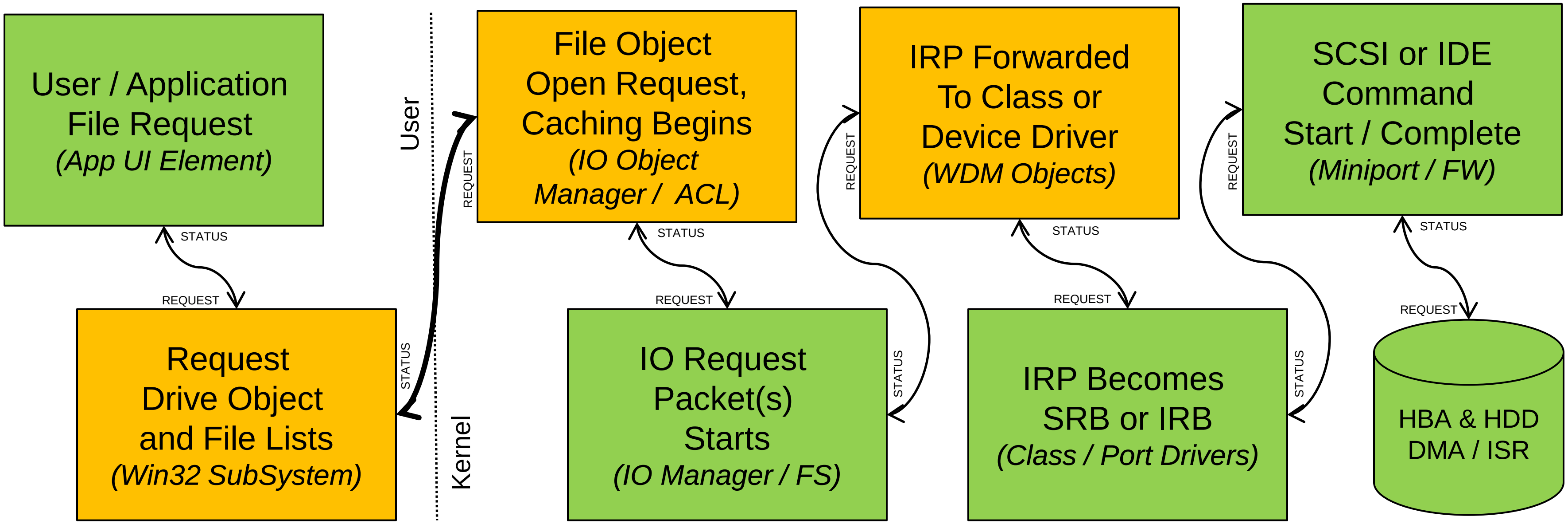
Another Motivation for 4K



- ❑ Typical 512-Byte sector ECC scheme can correct up to ~50 Byte length defect with 40 x 10-bit symbols
- ❑ Future-generation HDD's will see larger defects
- ❑ Must improve ECC scheme to maintain UER

Windows & LBA Attributes – An Example

- IOCTL_STORAGE_QUERY_PROPERTY -> STORAGE_ACCESS_ALIGNMENT_DESCRIPTOR
  - This API allows Applications or Filters to determine LBA size
  - Exposure is very low in aligned systems



User / Application File Request *(App UI Element)*

Request Drive Object and File Lists *(Win32 SubSystem)*

File Object Open Request, Caching Begins *(IO Object Manager / ACL)*

IO Request Packet(s) Starts *(IO Manager / FS)*

IRP Forwarded To Class or Device Driver *(WDM Objects)*

IRP Becomes SRB or IRB *(Class / Port Drivers)*

SCSI or IDE Command Start / Complete *(Miniport / FW)*

HBA & HDD DMA / ISR

AF is in production

- In systems distributed today by every major OEM
- It is used in both client and commercial systems
  - Also used in external storage
- Better behaved than originally thought
  - Unknowns were scary at first ☺
  - Concerns about data integrity have not been realized
- Expect an increase in AF adoption moving forward
  - Several years of AF ready OS releases now in place
  - AF aware applications are now in distribution
- The following non-IDEMA publications address data alignment
  - Alignment requirements have been a reality for decades

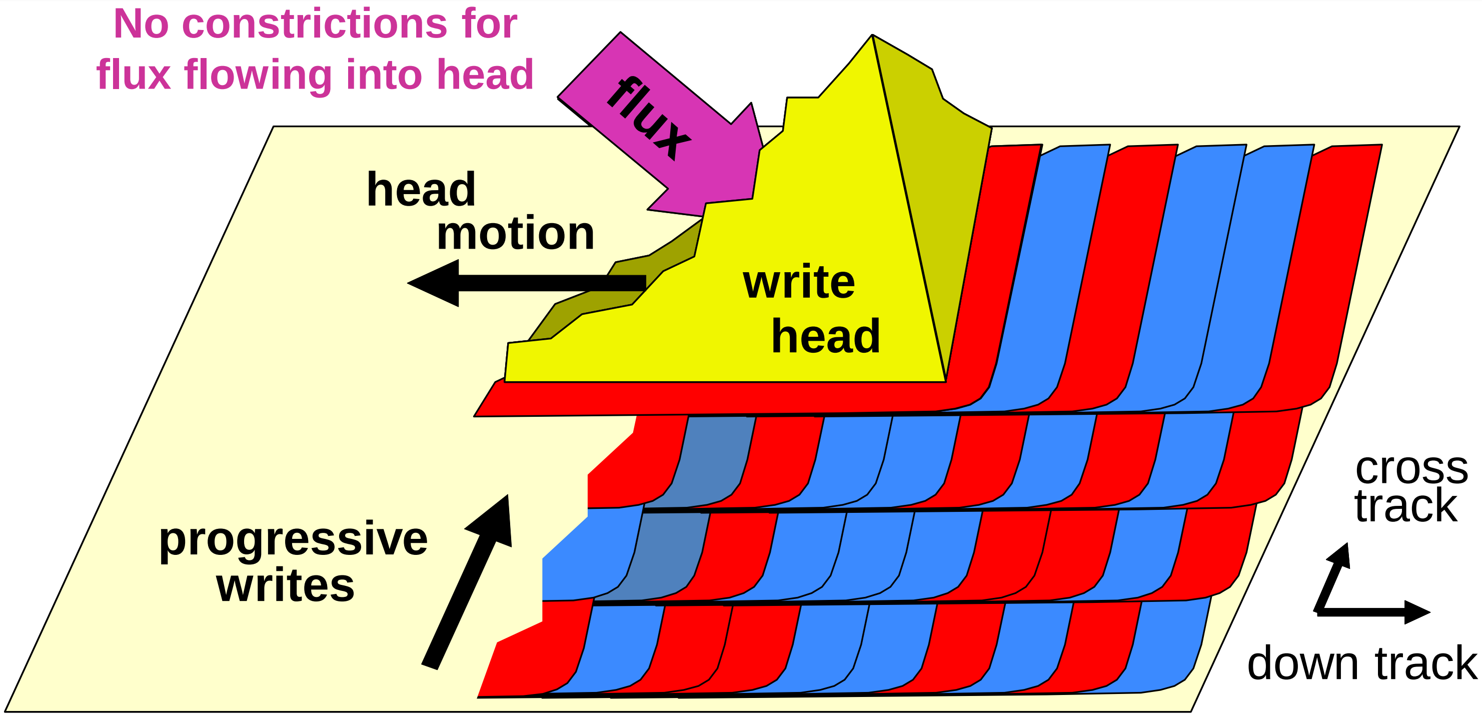
Awareness of Alignment

- SQL: http://msdn.microsoft.com/en-us/library/dd758814(v=sql.100).aspx
- VMWare: http://www.vmware.com/pdf/esx3_partition_align.pdf
- NTFS: http://ubuntuforums.org/archive/index.php/t-1565108.html
  - http://theether.net/kb/100104
- Microsoft: http://support.microsoft.com/kb/2385637
  - http://support.microsoft.com/kb/929491
- Other:
  - Wiki: http://en.wikipedia.org/wiki/XFS#Variable_block_sizes
  - Fat32: http://forum.easeus.com/viewtopic.php?p=24217&sid=18149153ac5fd04a17f576d0e30cc1ac
  - Oracle:
    - Unified Storage: http://blogs.oracle.com/dlutz/entry/partition_alignment_guidelines_for_unified
    - Solaris: http://osdude.wordpress.com/2010/09/10/aligning-solaris-x86-partitions-slices/
    - Solaris: http://www.oracle.com/technetwork/articles/systems-hardware-architecture/lun-alignment-163801.pdf
  - EMC Clariion: http://www.penguinpunk.net/blog/?p=499
  - Dell SQL: http://en.community.dell.com/support-forums/storage/f/1216/p/18697184/18820170.aspx
  - IBM Linux: http://www.ibm.com/developerworks/linux/library/l-4kb-sector-disks/

Futures – SMR & LBA Indirection

What is SMR?

SMR write head geometry extends well beyond the track pitch



Wood, Williams, et al., IEEE TRANSACTIONS ON MAGNETICS, VOL. 45, NO. 2, FEBRUARY 2009

The larger SMR write head introduces the SMR Constraint

SMR Architecture

- Read Modified Write

  Read-modify-write operation first reads a portion of data from the disk, then modifies part of that portion with the host provided write data, and finally writes the whole portion back to disk

- Shingled Regions

  A group of tracks that is separated from neighboring shingled regions by a guard band. The purpose of the guard band is to prevent a write in a given region to interfere with data written on other regions. That isolation of interference guarantees that no read-modify-write will need to go beyond a region boundary

- Indirection

  An indirection system is a collection of data structures and algorithms that assigns physical locations to logical block addresses and retrieves physical locations of logical block addresses. In the case of SMR, the indirection system is to be designed to provide good read/write performance for a wide variety of natural workloads

- Hints

  Provide the device with information about the usage of data in the user area of the media. The purpose of hints is to allow the device to manage its data in a more intelligent way. Hints allow the device to increase reliability, performance and data integrity. Hints originate from the file-system

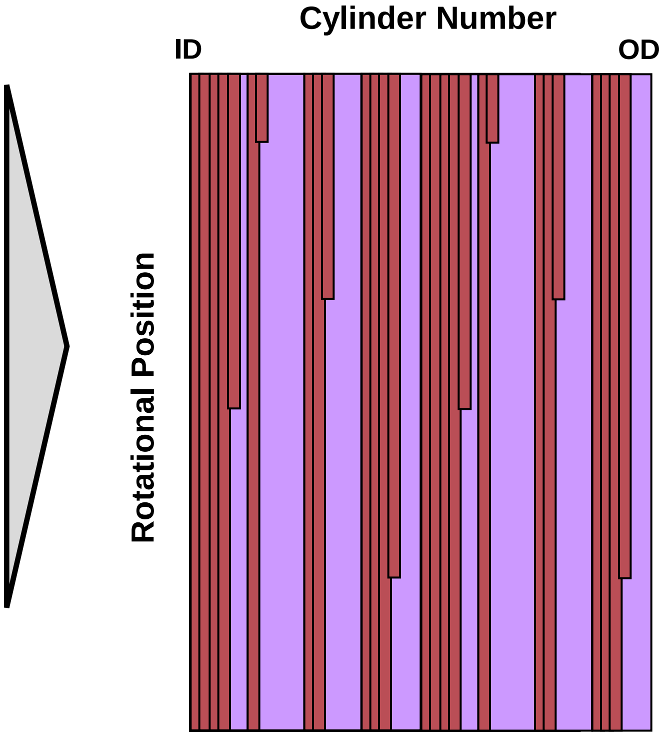
Shingled Regions

Implementing shingled regions is a good option to manage the random write performance impact of SMR

Simplistic View

(e.g. if all tracks are shingled)

- ❑ Large sequential writes should have reasonable performance
- ❑ Small random writes may have very poor performance

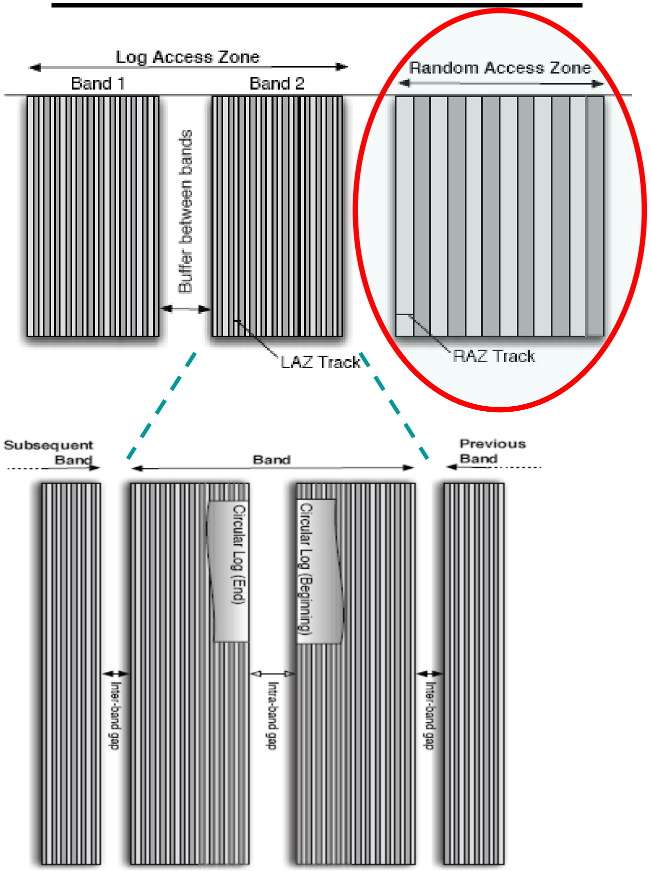


- ❑ HDD surface divided into regions of tracks
- ❑ Shingled writing within each region

Write Staging and Indirection

- Implementing a persistent 'Write Staging' solution will further improve the random write performance of SMR by delivering a persistent write caching effect
- Managing the 'Write Staging' likely involves background processes – data movement within the drive not in direct response to a host write
- A logical-to-physical indirection system will need to keep track of the data locations

One Possible Architecture



'Write Staging'

A. Amer et al. "Design Issues for a Shingled Write Disk System" MSST 2010

LBA Indirection

1. Host wants to rewrite **LBA-13** and **LBA-14**
2. Host seeks for next available unassigned block
3. Host assigns **LBA-13** and **LBA-14** to **PBA-32** and **PBA-33**

PBA-13 and **PBA-14**
Now unassigned

PBA-11 LBA-11	PBA-12 LBA-12	PBA-13 LBA-N/A	PBA-14 LBA-N/A
PBA-21 LBA-21	PBA-22 LBA-22	PBA-23 LBA-23	PBA-24 LBA-24
PBA-31 LBA-31	PBA-32 LBA-13	PBA-33 LBA-14	PBA-34 LBA-N/A
PBA-41 LBA-N/A	PBA-42 LBA-N/A	PBA-43 LBA-N/A	PBA-44 LBA-N/A

🟩 Unassigned

TRIM/UNMAP for SMR Drives

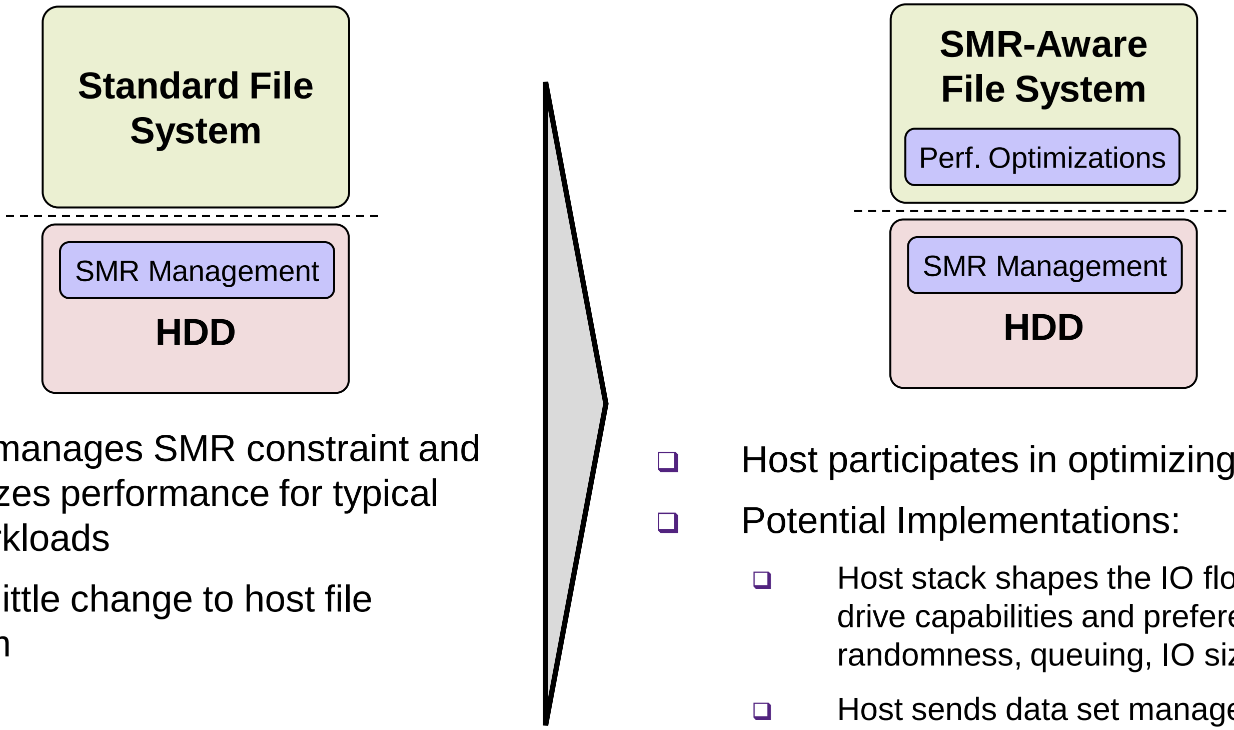
Similar to a SSD, a SMR drive's performance benefits from actively managing the instantiation of its logical blocks

Current Assumptions:

- SMR drives support the TRIM/UNMAP commands for basic dataset management
- A never written / freshly trimmed LB does not logically exist on the media
- Reads to a never written or freshly trimmed LBA return may return the original data, random data, or a sector filled with zeroes
- The drive reports its capabilities for how trimmed sectors are addressed
- The drive manages media health autonomously

Evolution of SMR

An SMR-aware host stack may be able to further optimize the system IO performance

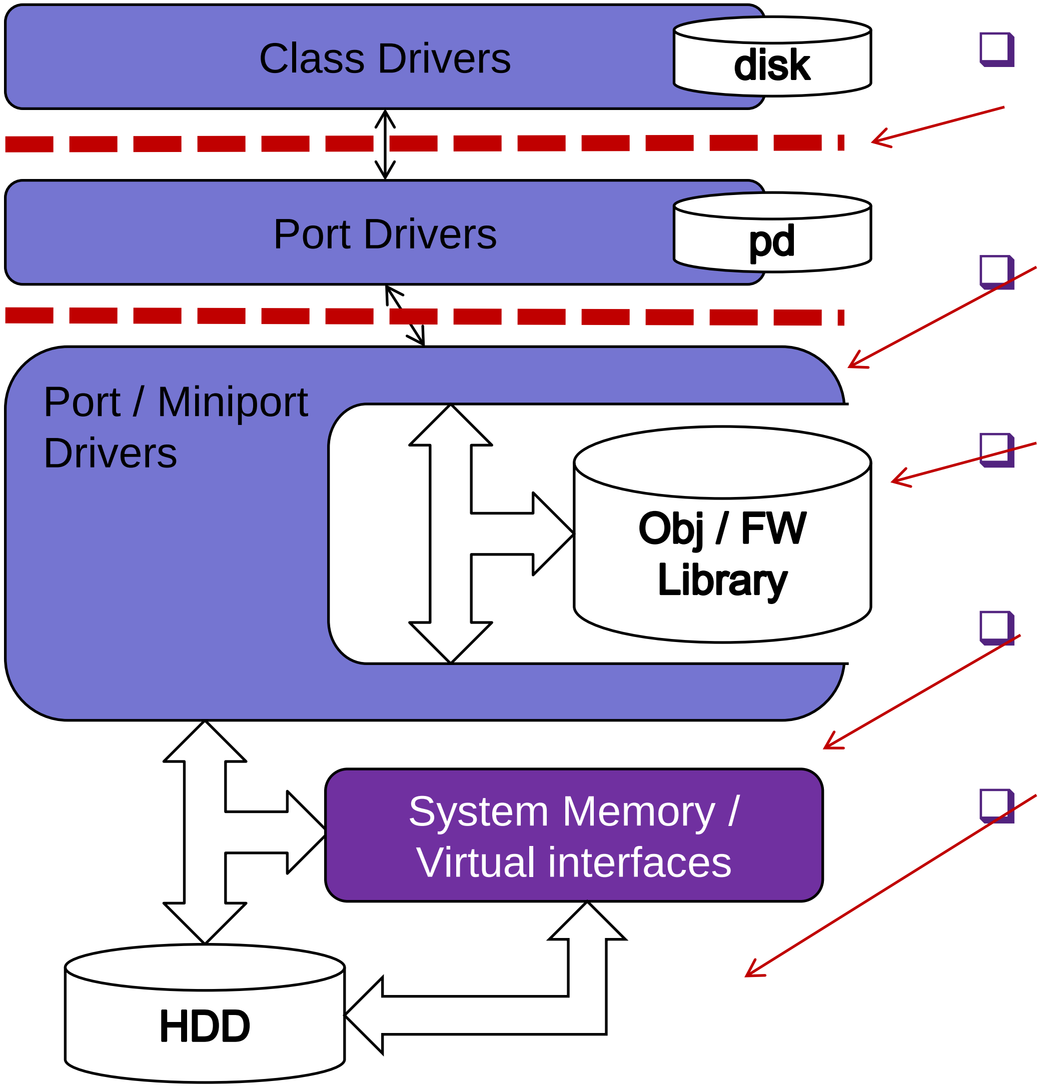


- Drive manages SMR constraint and optimizes performance for typical IO workloads
- No or little change to host file system

- Host participates in optimizing drive performance
- Potential Implementations:
  - Host stack shapes the IO flow to match advertised drive capabilities and preferences, e.g. for randomness, queuing, IO size, alignment, idle time
  - Host sends data set management ‘hints’ into drive

IDEMA will continue to work on industry alignment for SMR

OS Hinting for LBA Indirection



- **Filter Interfaces**
  - Defined by 3rd Party or OS Vendor
- **Port / Miniport Driver**
  - Are definitions needed?
- **Object Interface(s)**
  - Defined by IDEMA or Standards
- **Virtual Interfaces**
  - Defined by 3rd Party
- **HDD Interfaces**
  - Defined by IDEMA or Standards

Today in SMR

- Proposal coming into September T10 meeting
  - Defines “runtime” hinting, in protocol
  - Defines “change your mind” (AKA preconfiguration)
  - Proposes initial hints
- Initial hints
  - Boot Material – Material available quickly at power-on
  - Access latency – Differentiates material that is needed quickly vs material that may take a little longer
  - Access frequency – Differentiates material that is only intended to be used once, or many times

Hybrids & SSDs

Ecosystem and Storage

- Everything new and exciting trending towards
  - Super thin, ultra-portable
  - Instant on, phone or TV like performance
  - Little to no power use at idle
- Most recent Hard Disk Drive trends still pushing capacity
  - Driven by existing market demands, explains AF
  - Desktop and Laptop markets still growing
  - Little to no tolerance for legacy to stop trend
  - Power management control expanding with TPM
- AF HDD and SSD features converging in compute space
  - New markets are important to both
  - Hybrid Hard Drives bridging the gap for capacity
    - Hybrid and SSD both use same NAND

In Conclusion

- AF Works ☺
- Visit the IDEMA AF booth
- Developer kits are available
  - Visit IDEMA.org
  - Visit the booth
  - Talk to your local HDD supplier
- AF is setting the stage for future development in the eco-system
- AF is transparent ☺