SEAGATE
FLEX
Timothy Feldman
timothy.r.feldman@Seagate.com

Excerpt from *Disks for Data Centers*, Google Research

White paper for FAST 2016

Physical Design Options

…

3) Optimized SMR Implementation [capacity]

Shingled magnetic recording (SMR) allows a hard disk to achieve a higher areal density – typically 10-20% today, with potentially higher capacity gains in the future – by limiting the ability to perform random writes on the physical media. In particular writes must be done in order, like shingles, and the writes destroy the next track. As discussed, the value of a disk comes not just from its capacity, but also the number of useful I/O requests it can service. Because of the write restrictions imposed by SMR, when data is deleted, that deleted capacity can not be reused until the system copies the remaining live data in that SMR zone to another part of the disk, a form of garbage collection (GC).

Some data has lifetimes that are more predictable, which allows the storage system using a host managed SMR drive to reduce the GC overhead by grouping data with the same predicted lifetime into an SMR zone. Without this, the GC overhead in terms of both the IOPS and capacity becomes significant. Storing only "SMR friendly" data on the SMR disks could mitigate the problem, but this data is generally colder than the rest of the data, which means some of the IOPS on the SMR disks will be wasted. Worse, removing the SMR friendly data concentrates the IOPS requirements for the remaining hotter data on the conventional spindles ("conventional magnetic recording" or CMR), which increases the IOPS needed on the CMR disks. More broadly, SMR drives reveal a bias towards GB/$ improvements over IOPS/GB improvements.

One way of addressing this is to **mix CMR and SMR technologies in a single, hybrid hard drive**. This better mixes hot and cold data on the same disk and increases the chance of effectively using both the capacity and the IOPS. In particular, this allows new data to be stored in CMR, with SMR space used for 1) new data known to be long-lived or to have a predictable lifetime (when initially written), or 2) older data stored in CMR that has aged to the point where it is presumed to be long-lived and can be moved to SMR.

The simplest implementation of a hybrid drive is to use a mix of platters and heads, with some optimized for SMR, and the rest for CMR. This gives a permanent fixed ratio of CMR to SMR, which may become suboptimal over time. Alternatively, if a CMR optimized head is used for both SMR and CMR recording, the outer tracks could be used for CMR, while the inner tracks could be used for SMR. Note that the SMR capacity gain will be less in this case, but the CMR portion of the disk will see IOPS improvements due to the lower average seek distance.

Another way to reduce GC overhead is by relaxing the write restrictions imposed by defining a host-managed SMR abstraction. An example is "Caveat Scriptor" [11] , which allows the host to write randomly, knowing that it will destroy (specific) nearby data. The Caveat Scriptor proposal also supports other more interesting write patterns, including a circular buffer, and effectively allows the host to have dynamically resizable SMR zones. In the long term, an even more aggressive approach would be to allow dynamically configurable CMR and SMR tracks, which may be achievable with more advanced head-media interface (HMI) technologies.

research.google.com/pubs/pub44830.html (emphasis added)

Flex Redefines the Data Center Hard Drive

Seagate is leading the way in the development of a new dynamic recording method called Flex, to solve the evolving needs of hyperscale data centers. Flex will redefine the future of the data center hard drive. At the core, Flex vastly simplifies storage deployment and the storage supply chain, and enables data centers to more efficiently deploy and dynamically optimize installed storage to serve changing application requirements.

Flex enables a single hard drive to use multiple recording methods on the same drive media. For example, it will enable Dynamic Hybrid-SMR that blends SMR for higher capacity with CMR for higher performance, with optimized data locality, on a single drive. Then Flex lets the data center host configure how the data is stored on the drive's media, dynamically tuning the drive's data density to match the immediate needs of the entire data center system.

In a related development Google this week said it is sharing a proposal with the Open Compute Project (OCP) for Dynamic Hybrid Shingled Magnetic Recording (HSMR) HDD Product Requirements, with the aim of collaborating with hardware partners and customers to define a collection of Cloud hard drive features, behaviors and interfaces to adapt to exponentially growing storage needs in data centers.

"We are excited to see that Seagate has a technology solution for Dynamic Hybrid SMR on their 2019 roadmap," said Eric Brewer, VP of Infrastructure at Google and an author of last year's Disks for Datacenters white paper.

The development of Flex means that going forward, all media recording technologies can be tunable and optimizable, enabling our customers to balance the way storage is used to meet their own needs for performance, capacity and cost.

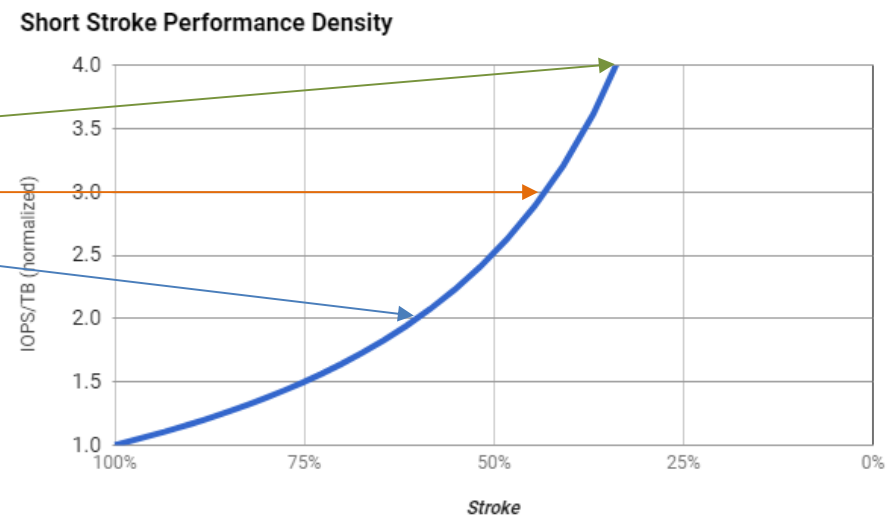
The move to Flex-enabled drives will also vastly simplify the supply chain for Seagate and for our partners. A customer can purchase a supply of drives based on its expected total capacity demand, and use the same drive for warm and cold storage. They can also use this same drive for any non-Flex-aware storage needs.

blog.seagate.com/intelligent/new-flex-dynamic-recording-method-redefines-data-center-hard-drive/

Problem Statement

- Hot data performance
  - Short stroking is powerful
    - 4x performance at 33% stroke
    - 3x performance at 44% stroke
    - 2x performance at 60% stroke

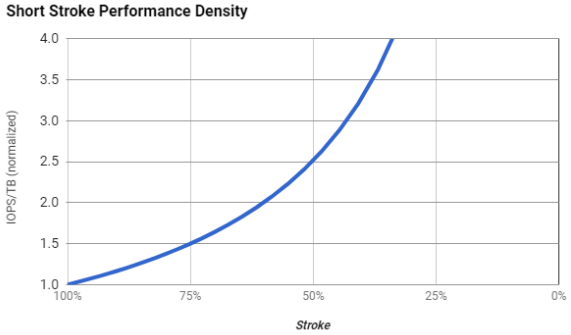
    but unused media represent wasted capital



Seagate Copyright © 2017

Problem Statement

- Hot data performance
  - Short stroking is powerful but unused media represent wasted capital



- Cold storage cost
  - SMR has advantages but under utilized actuators represent wasted capital

Seagate Copyright © 2017

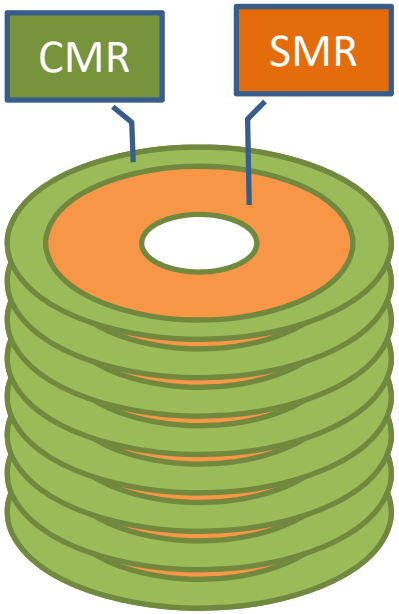
Problem Statement

- Hot data performance
  - Short stroking is powerful
- Cold storage cost
  - SMR has advantages
- Using two separate solutions is sub-optimal
  - Wasted capital
  - Deployment complexity: separate or heterogeneous clusters
  - Two drive types for forecasting, inventory, qualifications, …

Seagate Copyright © 2017

The Flex Idea

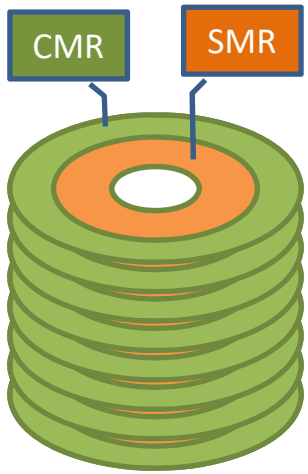
- One drive type that supports both CMR and SMR
- Start as 100% CMR
  - backward compatible to legacy software
- Field configurable to part CMR, part SMR
  - unlimited reconfigurations



Seagate Copyright © 2017

Customer Value Proposition

- **Maximize performance** for hot data
- **Minimize cost** for cold storage
  - Accommodate the growth of cold and write-seldom
- **Simplify supply chain**
  - Same drive for hot and cold
- **Deployment flexibility**
  - Homogenous infrastructure
  - Any ratio of CMR to SMR
  - Dynamic configuration for future needs



Seagate Copyright © 2017

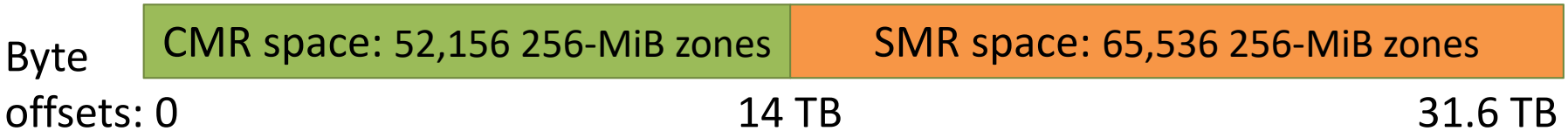
Flex API Goals and Requirements

1. Provide backward compatibility
   - Legacy software can use a drive in a 100% CMR state
   - This state is expected to be the shipping state
   - This state can be regained for repurposing for legacy software
2. Leverage existing protocols
   - Extend the zoned block device model
3. Support ATA and SCSI
   - Use separate, non-overlapping LBA ranges
   - Support both queued and non-queued forms of new commands, and 16-byte CDBs
   - Co-develop for straight wire SCSI to ATA Translation (SAT)
4. Allow diverse configurations
   - Conversions are fine-grained
   - Multiple seams are allowed
   - Different applications can use different configuration policies
5. The results of a conversion are deterministic and discoverable before the conversion
6. Conversions are fast
   - Media that comes online is written by the application before it has to serve read requests
   - The number of write frontiers is bounded in both CMR and SMR space
7. A conversion does not invalidate data stored on media not involved in that conversion
8. Conversions have prerequisites to help ensure copying of valid data and erasure of secure data
9. The API is extensible to future recording methodologies beyond CMR and SMR
10. Since dynamic recording is not mutually exclusive with security, allow security feature sets to be enabled

Seagate Copyright © 2017

Flex LBA Layout

- Example for CMR=14 TB, SMR=17.6 TB



LBAs:

```
CMRmin=0
CMRmax=6 5DDF FFFF
SMRmin=6 5DE0 0000
SMRmax=E 5DDF FFFF
```

Notes:

- CMR space must start at LBA 0
- SMR space must start after CMRmax
- There may or may not be unused LBAs between CMRmax and SMRmin

Seagate Copyright © 2017

Flex Zones

- Each zone is a 256-MiB extent of LBAs
  - Both the CMR and SMR spaces are made up of abutting zones
  - Each zone has its own write pointer indicating its write frontier
- Zones are the granularity of conversion
  - Each zone is online or offline
  - Online zones are provisioned with media; offline zones are not
  - A conversion command specifies a single extent of zones to take offline or to bring online
  - Application allocation unit is any integer number of zones

Seagate Copyright © 2017

Flex Zones In SMR Space

- Existing write pointer zone types
  - Sequential write preferred (Host Aware)

  or

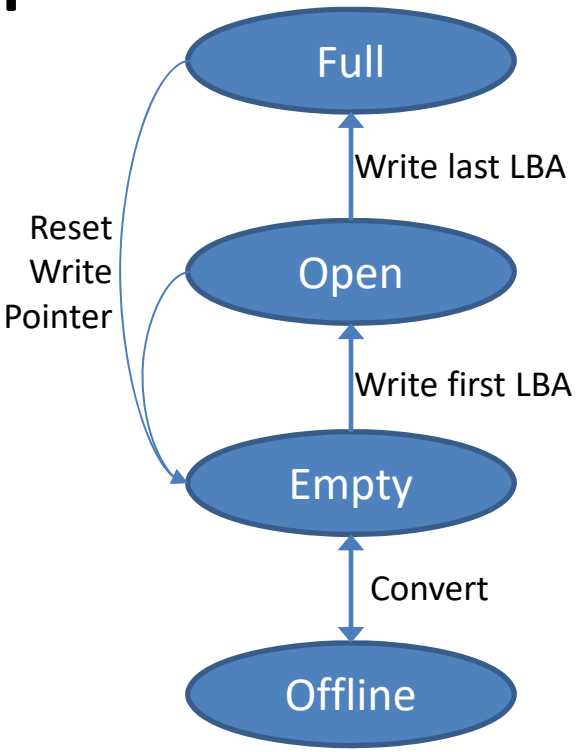
  - Sequential write required (Host Managed)

  A drive may support only one of the two, or both

Seagate Copyright © 2017

Flex Zones In CMR Space

- A new zone type for CMR space
  - 256 MiB
  - Commands: Reset Write Pointer
    - No Close, Finish or Open command support
  - States: Offline, Empty, Implicitly Opened, Full
    - State transitions:
      - Conversions cause transitions between Offline and Empty
      - Writes at write pointer transitions from Empty to I. Open, I. Open to Full
      - Reset Write Pointer causes transitions to Empty
    - No limit to number of zones in a given state
      - All zones can be Implicitly Open
    - No Closed or Explicitly Opened states
  - Write pointer: one per zone
  - Read and write rules
    - Writes must not start above the write pointer
      - Only one transition between initialized and uninitialized
    - Reads above the write pointer may be restricted or unrestricted
      - Restricted versus Unrestricted is reported as a device parameter



Seagate Copyright © 2017

Flex Configuration Examples

100% CMR

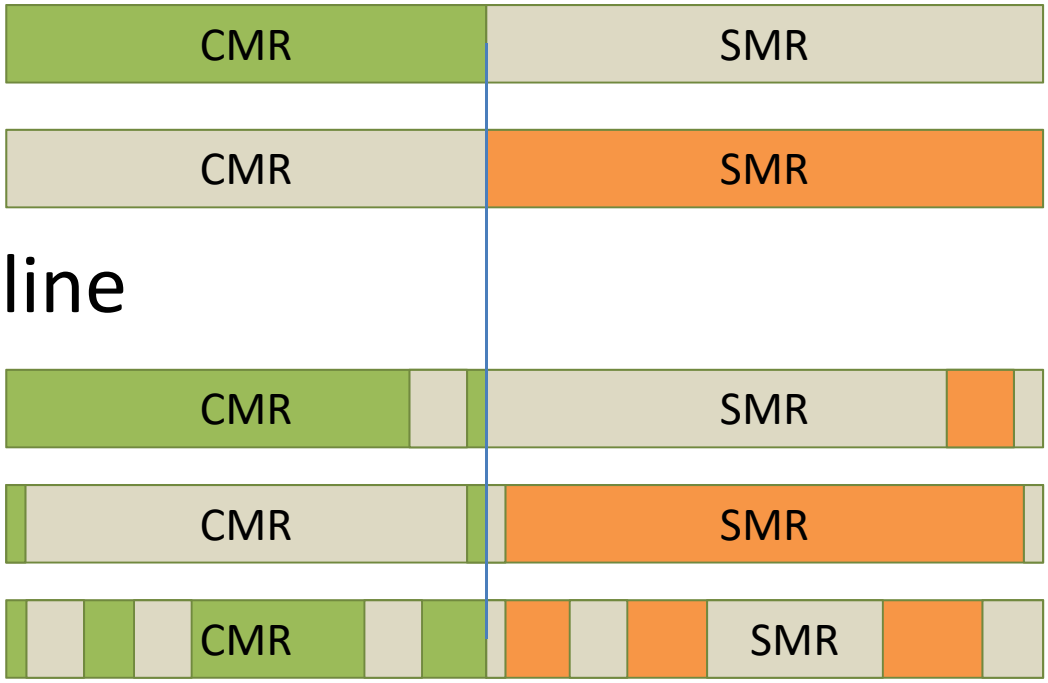
100% SMR

1 SMR extent online

small

large

>1 SMR extents

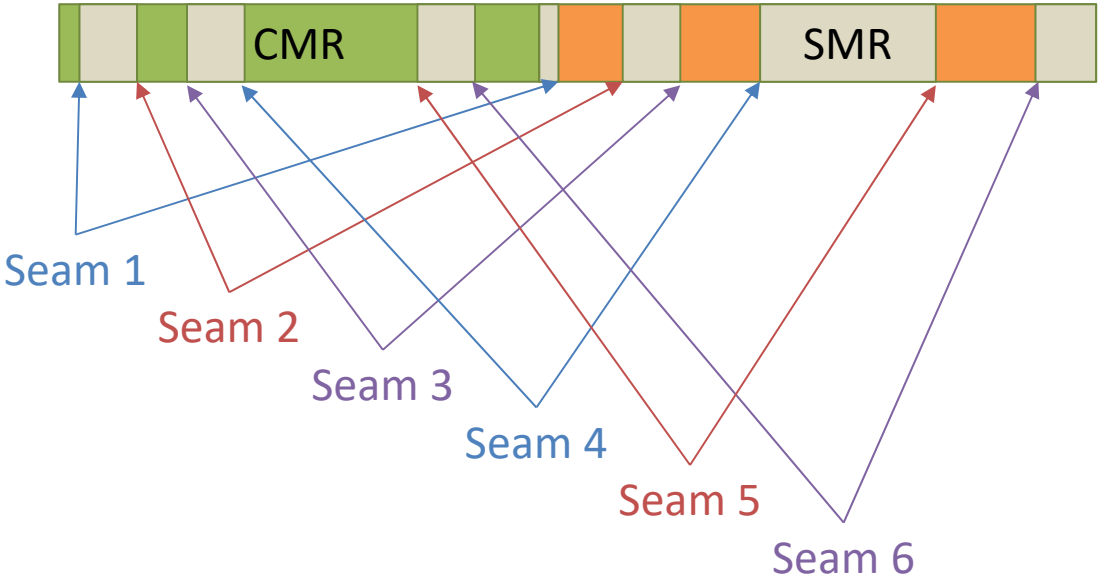


Seagate Copyright © 2017

Flex Multiple Seam Example

3 SMR extents

6 seams



Seagate Copyright © 2017

Flex Conversion Sequence

Convert from CMR to SMR

- Query to find extent of zones associated with the extent the host might convert *[1]
- Copy valid data out of the CMR extent*
- Erase the CMR extent*[2]
- Issue Reset Write Pointer to the CMR extent [1]
- Invoke conversion operation [1]
  - CMR extent changes to Offline
  - SMR extent comes online in Empty state
- Writes to the online SMR zones must comply with write rules of the SMR zone type
  - Writes to sequential write required zones must start at the write pointer

Convert from SMR to CMR

- Query to find extent of zones associated with the extent the host might convert *[1]
- Copy valid data out of the SMR extent*
- Erase the SMR extent*[2]
- Issue Reset Write Pointer to the SMR extent [1]
- Invoke conversion operation [1]
  - SMR extent changes to Offline state
  - CMR extent comes online in Empty state
- Writes to the online CMR zones must comply with write rules of the CMR zone type [3]
  - Writes to CMR zones must start at or below the write pointer

[1] single command

[2] overwrite, crypto erase, or other method

[3] new write pointer zone type

* optional

Seagate Copyright © 2017

Flex Conversion Notes

- All zones are convertible
- Query returns the same information as Convert – the new state of extents that will/did change
- Query result may show the same size for CMR and SMR
  - For instance, at zero skew the density may be the same
- Reset Write Pointer is used as a handshake that the copy and erase steps, if needed, have occurred
- The full commands sequence is not required to be issued
  - Issuing Query does not guarantee a Convert will be issued
  - The application can return to online uses at any point before issuing Convert
- Convert is atomic
  - Either both extents have changed states to Offline and Empty, or there is no change in state
  - This applies to unexpected power loss
- CMR space coming online as Empty allows for quick command completion
  - Host has the prerogative of initializing with a pattern or with application data

Seagate Copyright © 2017