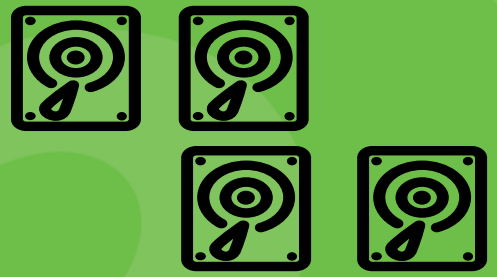
MAXIMIZING DATA'S POTENTIAL



Techniques for Shingled Disks

A View From Inside | Andrew Kowles, Firmware Group

SEAGATE

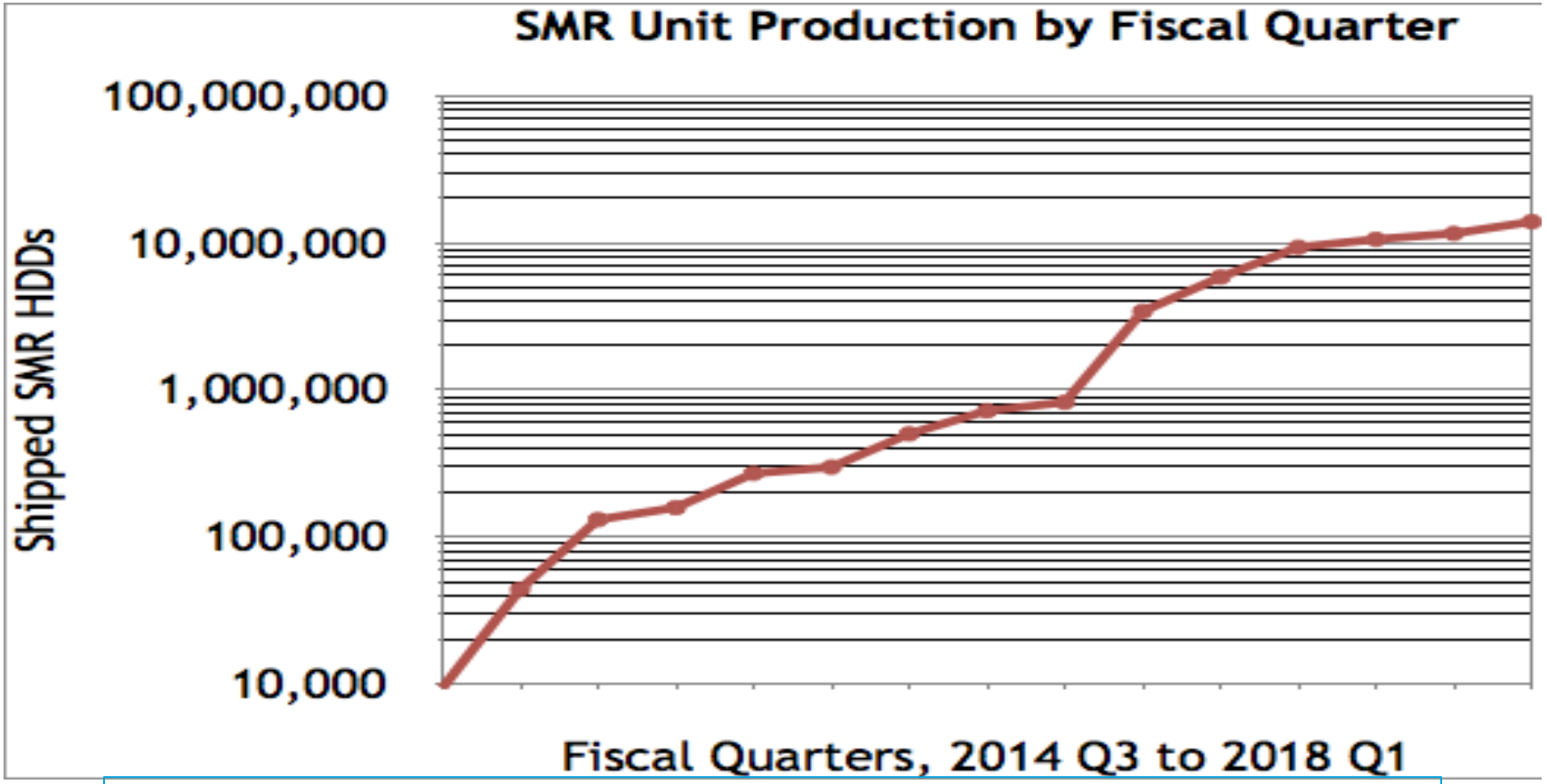
Outline

Why this presentation?

To Inform the Storage Ecosystem…

- **Now that all three HDD vendors have shipped SMR, it's time to open up to the rest of the system.**
  - **SMR is now broadly and deeply deployed**
- **Many of the problems for SMR HDDs are similar to SSD problems, which are being worked on openly by folks across the spectrum.**
  - **For example reducing device garbage collection via file systems.**
  - **For example addressing mapping table size for low cost devices via NVMe HMB**

Why? SMR Production (Seagate)



Total Seagate Shipments as of Feb, 2018: 85,217,016

SMR Review

A Brief Taxonomy

Drive Managed SMR

- Plug and Play aka Transparent SMR.
- STL is embedded inside the device.
- Not host side changes
- Suffers from observable performance challenges (In less than 10% of real client workloads)

Host Managed

- ZAC is T13
- ZBC is T10
- “Zoned Block Devices” is the generic term.
- ZAC-2 and ZBC-2 are the newest specs under review.
- ZAC/ZBC supports both Host Aware and Host Managed, but HM is preferred.

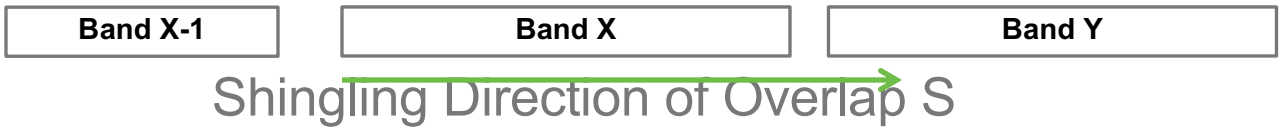
Flexible Magnetic Recording

- Extension to ZAC
- Allows both CMR and SMR recording rules on the same device, same SKU.
- Maximum configurability for Data Center Disks.
- Competing protocols which ned standardization: STX Flex (and Realms from WDC)

Zone Refresh for Reliability in HM Devices

High TPI Exacerbates Adjacent Track Interference…But Band Rewrites Mitigate…

- **Repeated writes of the same track tend to degrade the quality of adjacent tracks. This applies to both CMR and SMR drives**
- **HDD device firmware protects adjacent tracks using reliability systems for detection and refresh of degraded tracks.**
- **With SMR, TPI is increased which can reduce the number of writes it takes to degrade adjacent tracks.**
- **Zones in Zone Block Devices are laid out in physical bands, isolated from other physical bands.**
- **Zones are typically fixed to physical locations in HM-SMR devices.**
- **A 256MB zone refresh involves reading and writing hundred of MB of data…expensive!!**
- **Repeated reuse of the same logical zone can induce expensive zone refreshes.**
  - **The "Shingling Direction" causes asymmetry: Partial vs. Full refresh**
  - **Only need to refresh if there's valid data in those zones.**

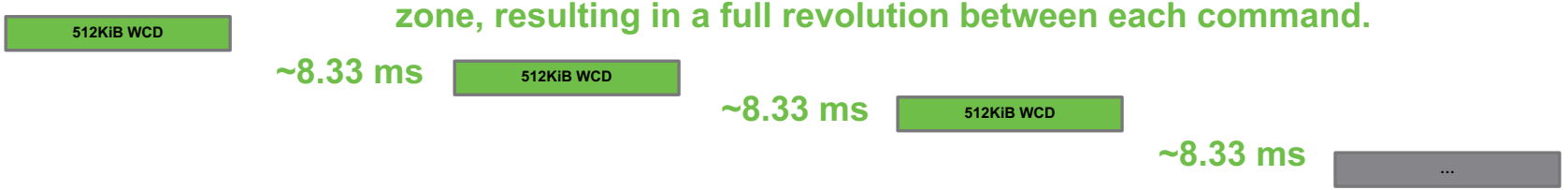


Linux and Zones and Command Queueing: A Problem

Out of Order Writes Will Abort with Host Managed SMR

Zone Block Linux Support Comes with a Catch. Queue Depth-per-Zone is limited to 1
This can reduce performance for sequential NCQ WCD workloads severely.
Here's how…Normally a sequential WCD QD=2 write can handle the following without lost revolutions.

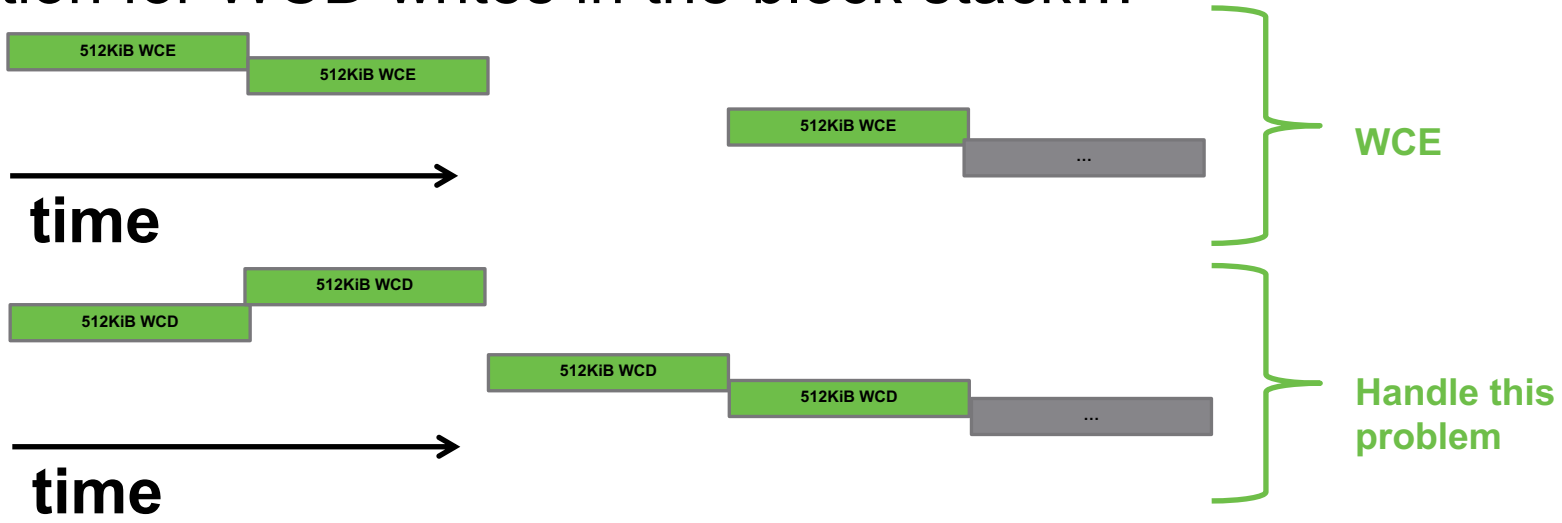
But Linux detects ZBDs and disallows more than a single tag per zone, resulting in a full revolution between each command.



Sequential Throughput Reduction of 40-70% is possible for HM-SMR HDDs without expensive compensation. The performance reduction is proportional to the transfer size of the commands.

Linux and Zones and Command Queueing: Solutions for Host Managed

Using WCE is a workaround. One solution is for a true solution for WCD writes in the block stack…



Call to Action!! Maybe we can talk about other solutions?
How to allow full Queuing in Zones??

Zone Management Commands

Guidance for the new API

- **Writes induce map updates.**
  - **And also Close Zone, Open Zone, Reset Write Pointer, and Finish Zone induce map updates.**
- **Map Updates need to be persisted.**
  - **Most devices can still handle uncontrolled power down with pending metadata updates, but it's stressful.**
- **Active Zones and Open Zones are Different!**
  - **It's subtle but important.**
  - **Close or finish zones which will become cold, to keep the number of open zones at or near the number of active zones receiving writes.**
- **Reset Write Pointer is like Trim.**
  - **Use it for as many zones as possible, as soon as possible.**
- **Report Zones Command Times are proportional to the number of zones being reported.**
  - **Minimizing the reporting of zones**

The Sustained Random Write Problem

Solved for Some DM-SMR Designs, but only for smaller writes.

This class of workloads causes a significant challenge to DM-SMR.

Dynamically mapped designs with sophisticated multilayered writeback caching can "solve" the problem, with some exceptions, as we will see.

Host Managed SMR makes random updates illegal, but we don't need to go fully to HM-SMR to make improvements.

FS Serialization log-structuring is good for all storage devices, SSDs, CMR, and SMR HDDs!

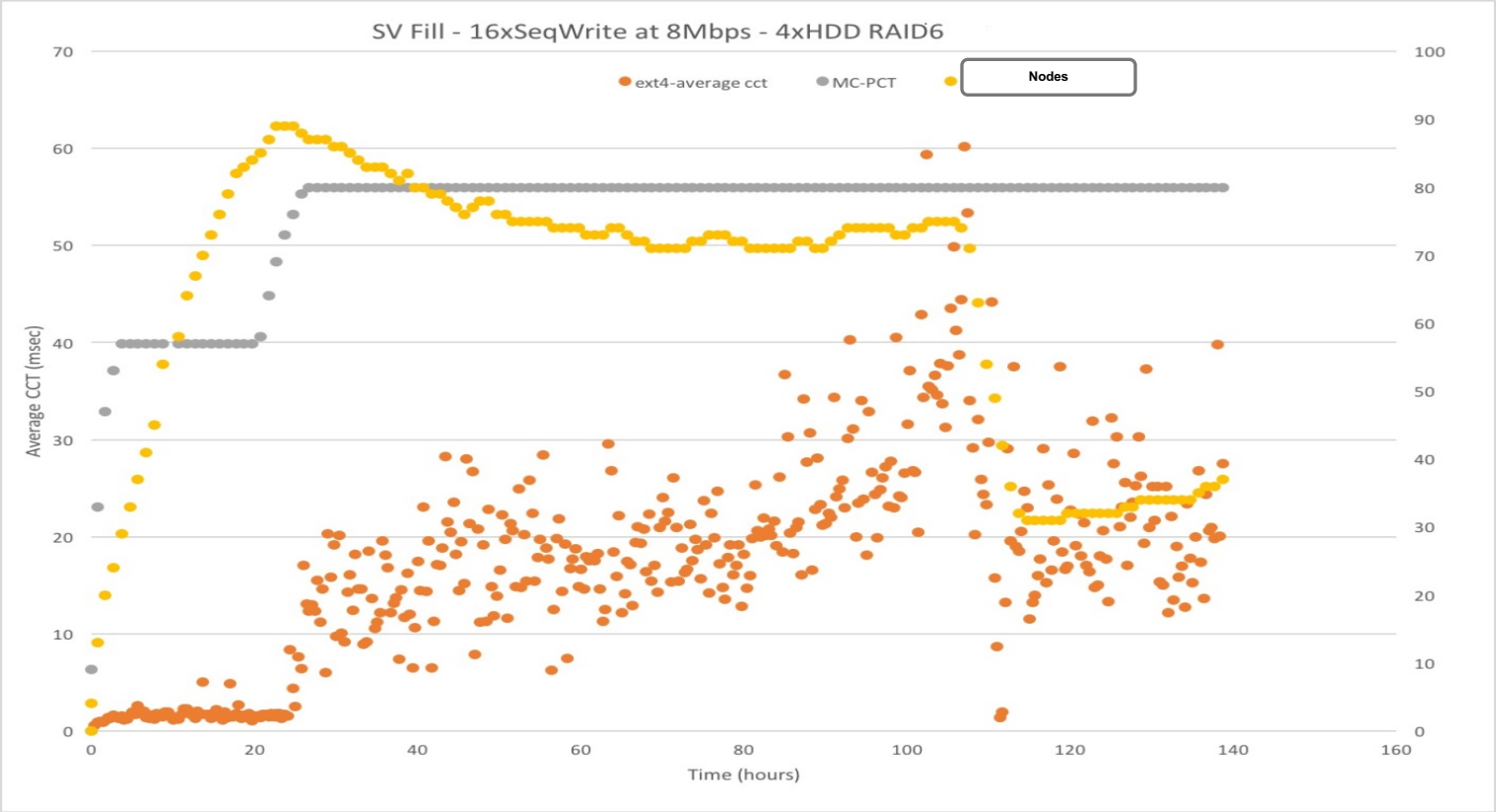
FS metadata is a major culprit in real world workoads.

For example, a Surveillance system with multiple camera streams can work well with an SMR disk…except for those pesky EXT inode updates (See EXT4-Lazy paper for possible solution).

Streaming workoads are synergistic with SMR disks in terms of both cost and compatibiliy, with just a few tweaks

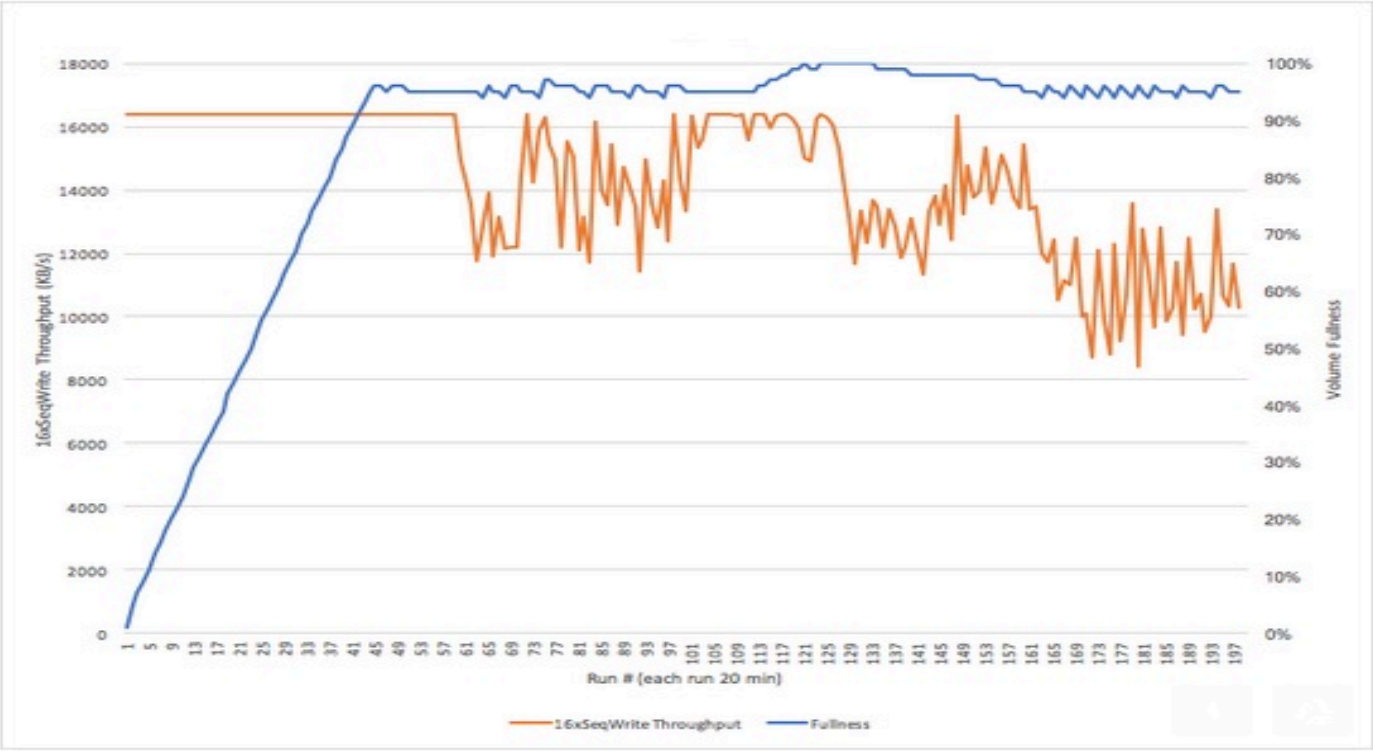
For example, a RAID5/6 SV system with 128K stripe size. The RAID parity updates induce excessive SMR work.

Write In-Place File Systems and Content Streaming



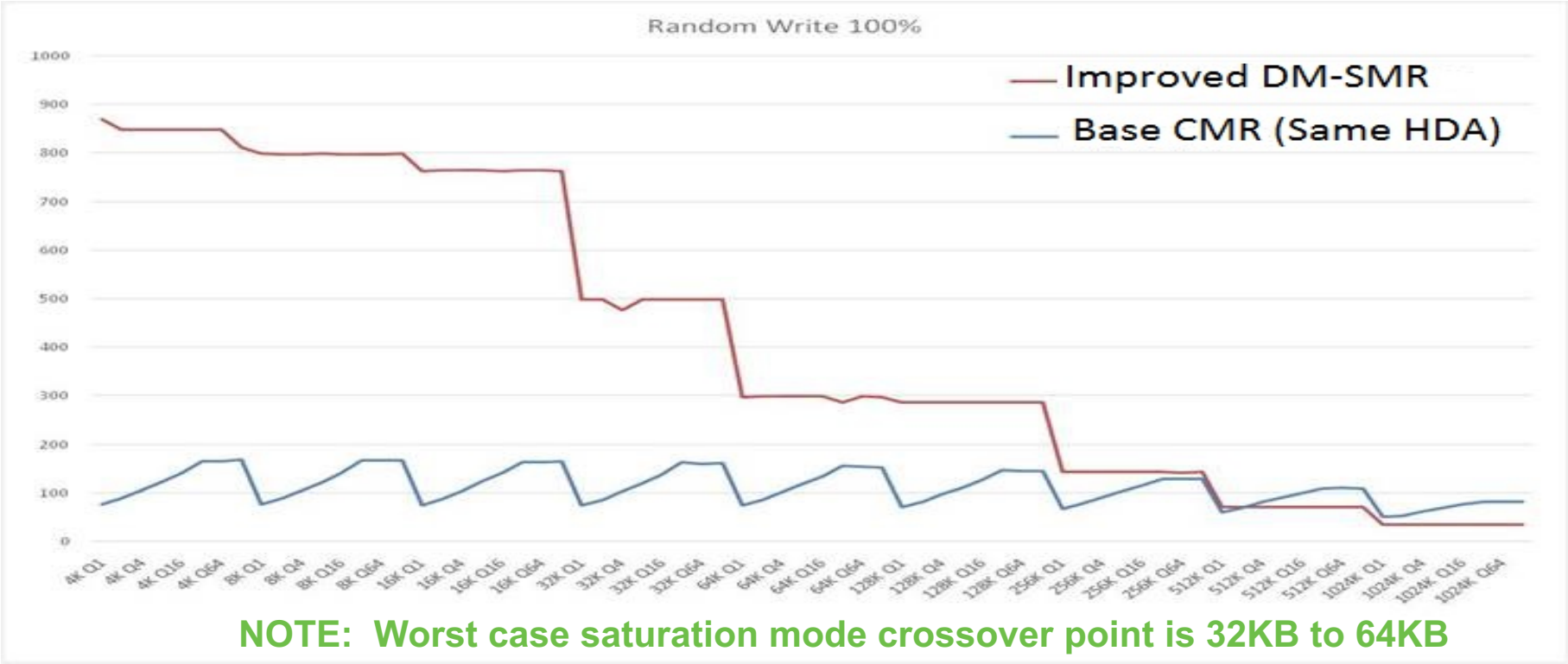
Capacity Utilization vs. Performance

This is One Example of State Dependent Performance for DM-SMR



The Sustained Random Write Problem 2

Solved with Some Designs, for Some Operating Points



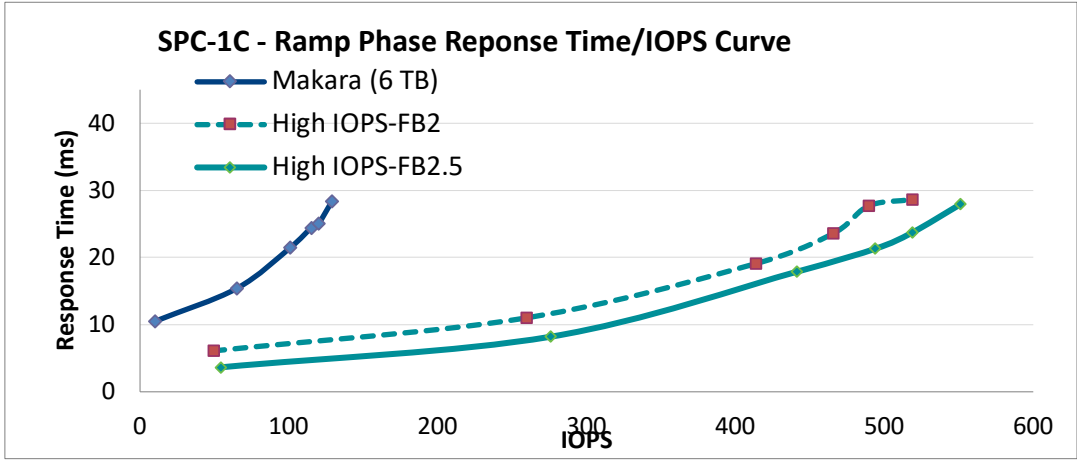
NOTE: Worst case saturation mode crossover point is 32KB to 64KB

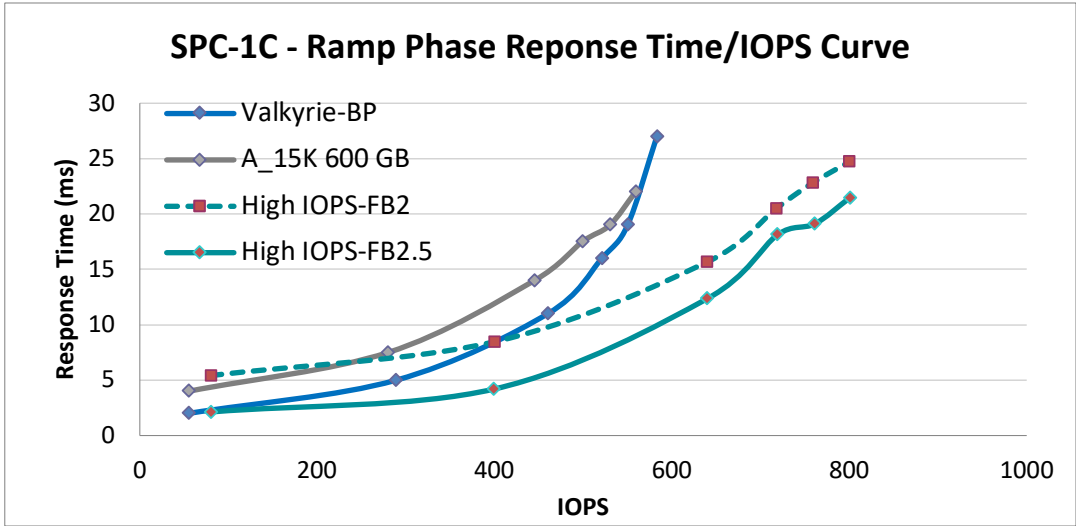
SPC-1C Single Drive Performance Summary

Unsaturated Improved DM-SMR Wins…

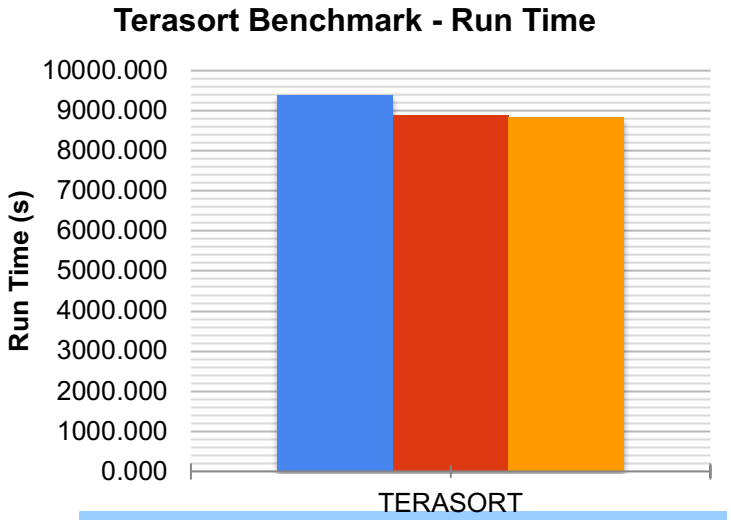
- SPC-1C measures the performance under sustained real world complex IOs simulating OLTP (on-line transaction processing) workloads.
- For MuleX drive since we can complete the test (~ 2 hrs) without saturating the media cache (MC) we get a 4x performance improvement
- For MuleY results the media cache is saturated but we still get a 40% performance improvement since write transfer lengths in the range where high IOPS outperforms base drive
- Hi IOPS/Improved DM-SMR ***can still be defeated by large block counts due to the OP requirement. 5% is used here..***

- **SPC-1C performance significantly better than a conventional drive – workload dominated by small block random writes where Hi IOPS outperforms conventional drive**

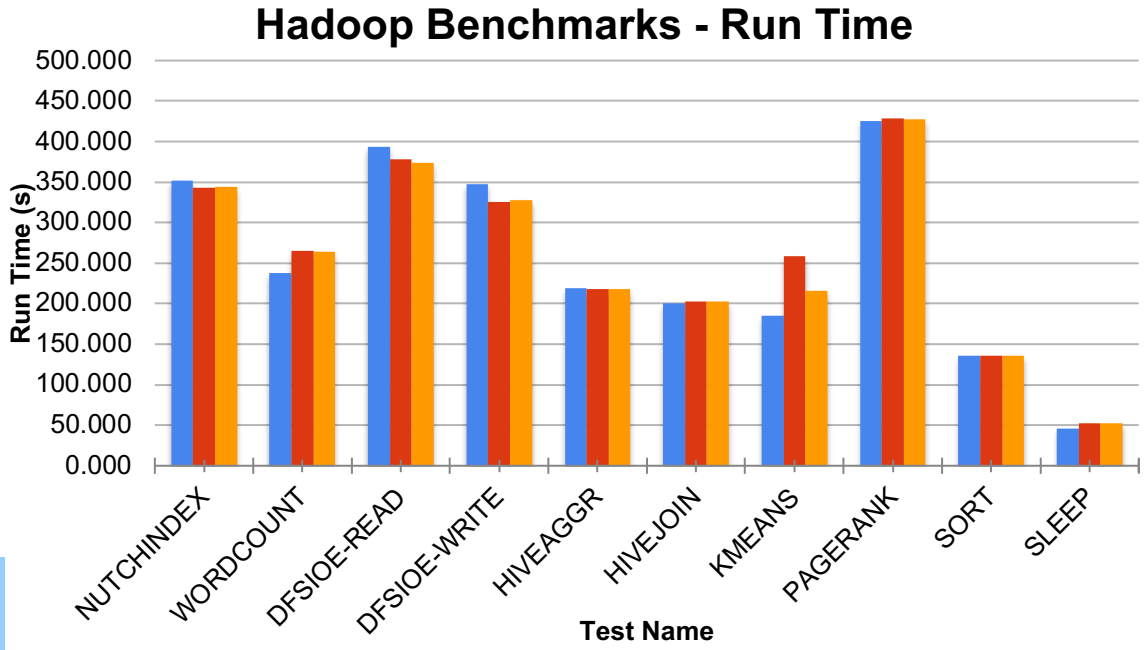




Hadoop Results for Improved DM-SMR



-Plots show time to complete test Lower better



- Improved SMR drive is at parity (within 5%) with the base CMR drive for these set of benchmarks
- There are still issues, but this slide shows internal data (Blue=SMR, Red=CMR, Yellow=CMR-SSHD).

HM-SMR

- Even Zone Reuse (with understanding of the zone layout and isolation ATI asymmetry)
- Solve the Linux Queue Depth per Zone problem
- Think About Zone Management Commands and Zone Metadata deeply to improve the device performance and reliability
- Engage with Seagate early and often. We'd like to answer your questions and help the community wrangle with ZBDs

DM-SMR

- Avoid Large Block Random Writes
- Lengthen Short Sequential Streams
- Avoid Intermingling Metadata and Content. Differentiate the commands with FUA.
- Use WCE instead of WCD for content.
- Use FUA instead of Flush Cache.
- Use Trim to minimize cache and capacity utilization.

Thank You

Come see me with your questions, please.

And use this email hotline for your questions:

andrew.kowles@seagate.com

SEAGATE