OpenZFS
DEV. SUMMIT

2014 / 10-11 NOV.
SAN FRANCISCO

Host Aware SMR

Timothy Feldman

Abstract

Host Aware SMR

An introduction to shingled magnetic recording for OpenZFS developers that are seeking an understanding of the technology, the standards that support it, and the opportunities to take advantage of the capabilities.

This presentation covers the various SMR device types: their models and theory of operations. A light exposure to the extensions being added to the T10 SCSI and T13 ATA standards is included.

Changes in Underlying Hardware

SMR

"*Shingled magnetic recording* (SMR) is one of the newest technologies contributing to the [increase of] density of the data placed on a disk drive.

"SMR devices can't be used with existing file systems without a major overhaul, but they're perfect for copy on write technologies used by NexentaStor as well as for key/value storage devices."

Underhahl and Novak. Software Defined Data Centers for Dummies, Nexenta Special Edition, Wiley.

Delivering Zettabytes

Largest capacity gains require SMR

Some applications run well on Drive Managed

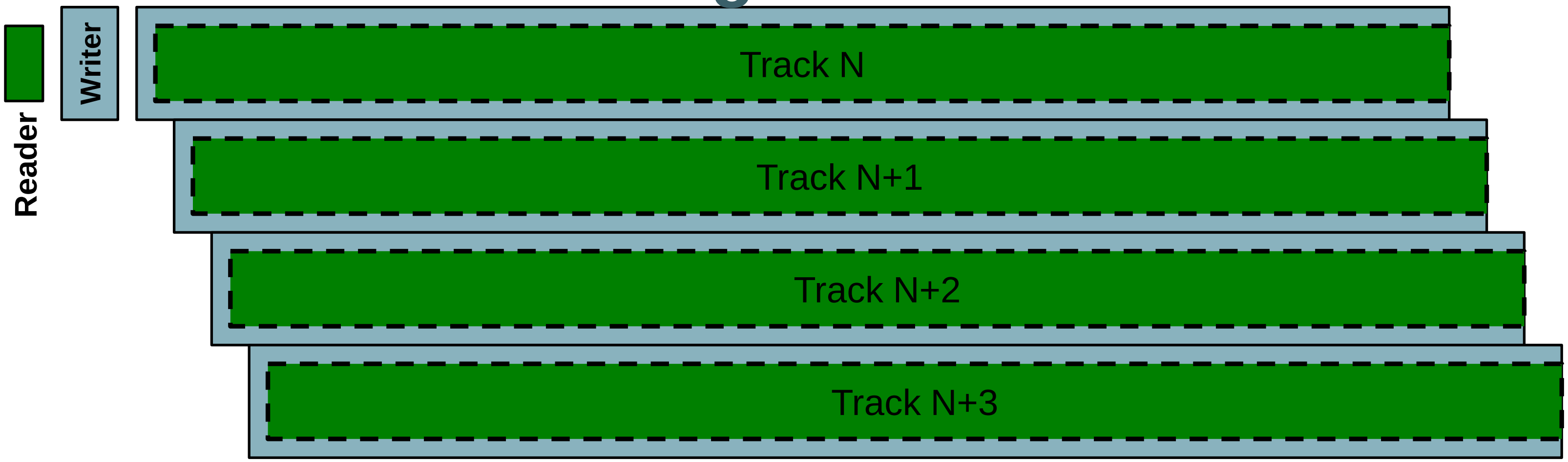
Other stacks will make use of Host Aware and upgraded filesystems

Agenda

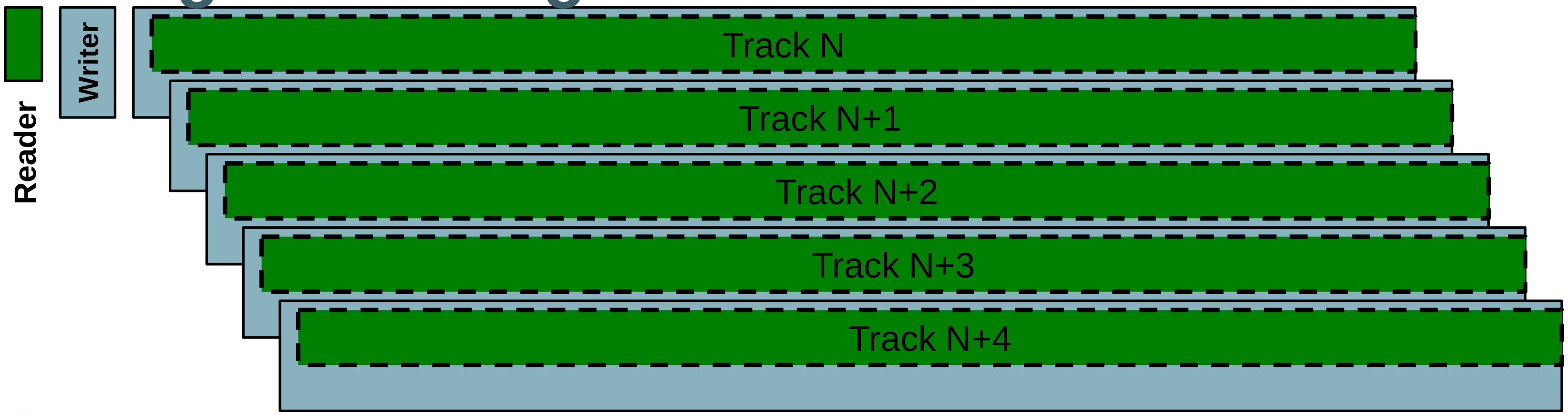
- Introduction to Shingled Magnetic Recording
- Autonomous "Drive Managed” SMR
- Host Aware
  - Problem statement
  - Goals
  - Methodology
  - Interface extensions
  - Usage model
  - Resources

Conventional versus SMR Writing

Conventional Writing



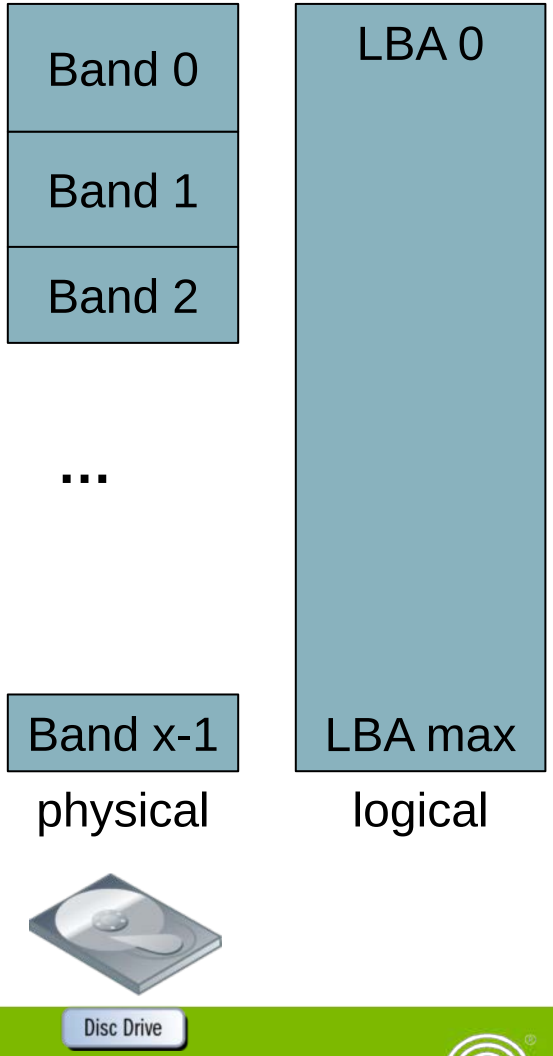
Shingled Writing



A disk as a set of bands

SMR Bands

- Physical construct
- Boundaries are not known outside the drive



Band 0

Band 1

Band 2

...

Band x-1

physical

LBA 0

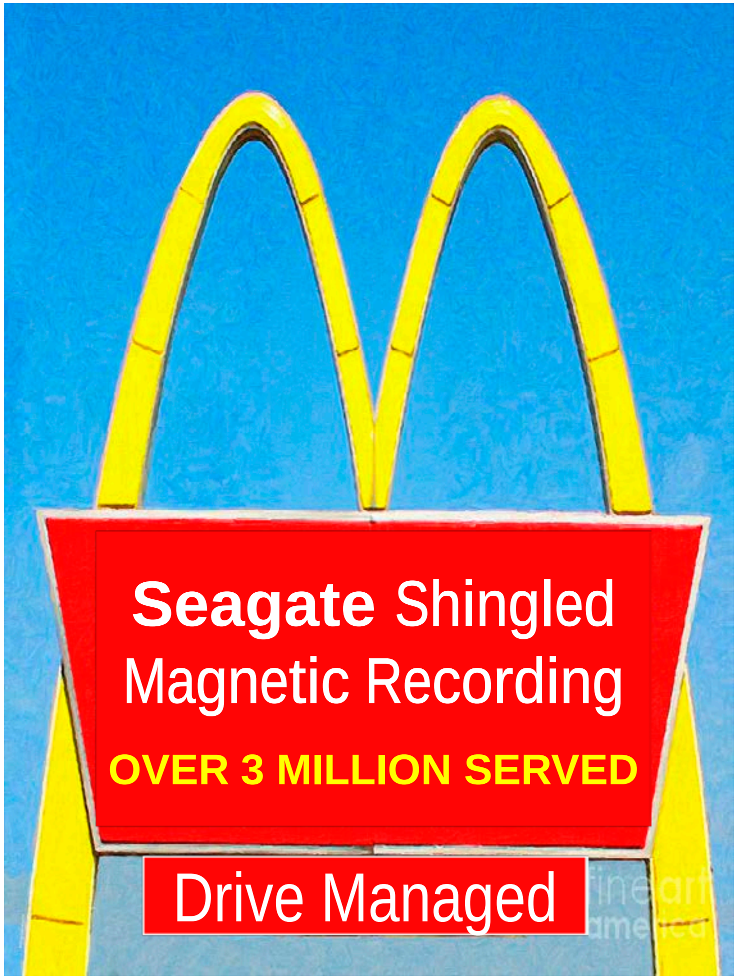
LBA max

logical

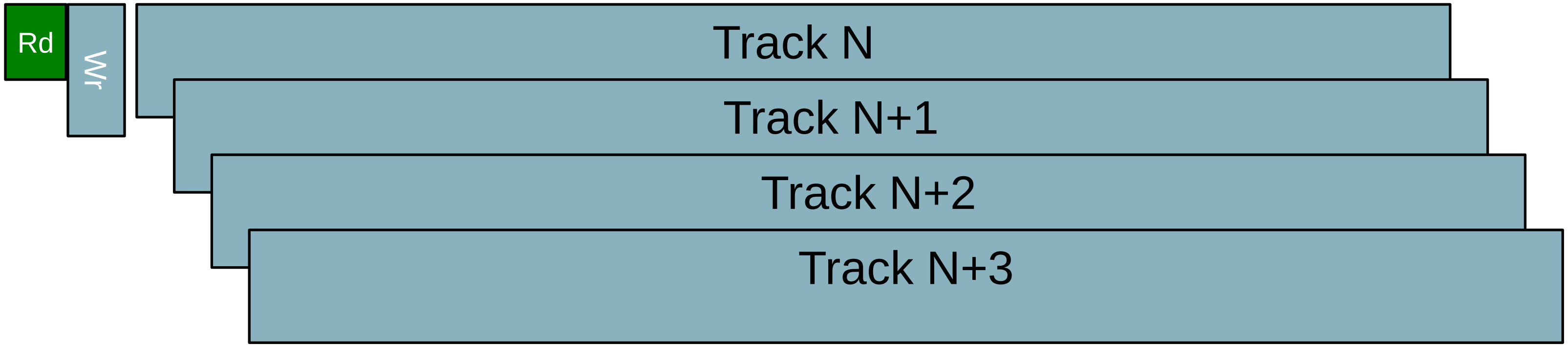
Drive Managed SMR

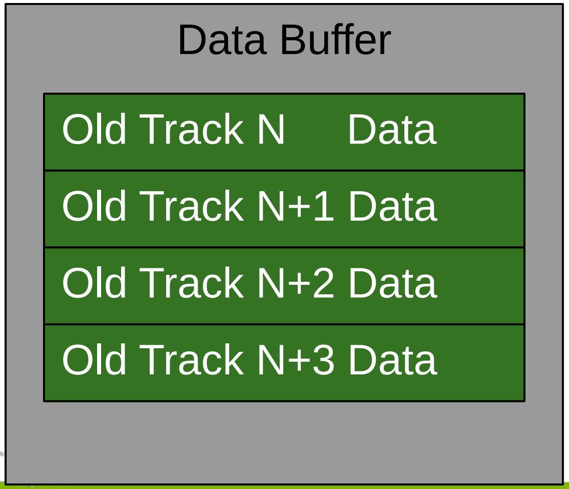
The first SMR drive type

- Drive autonomously hides all SMR issues
- Backward compatible



Updating a band with new data





Old Track N Data

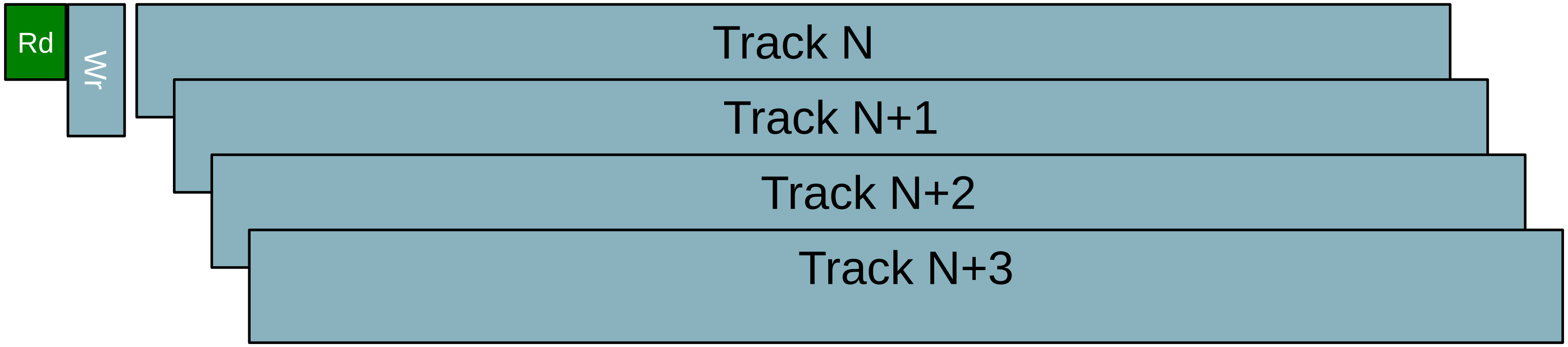
Old Track N+1 Data

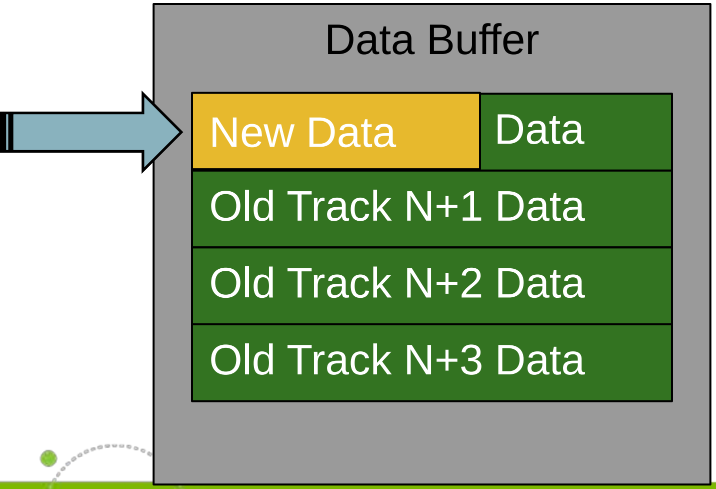
Old Track N+2 Data

Old Track N+3 Data

1. Read old data

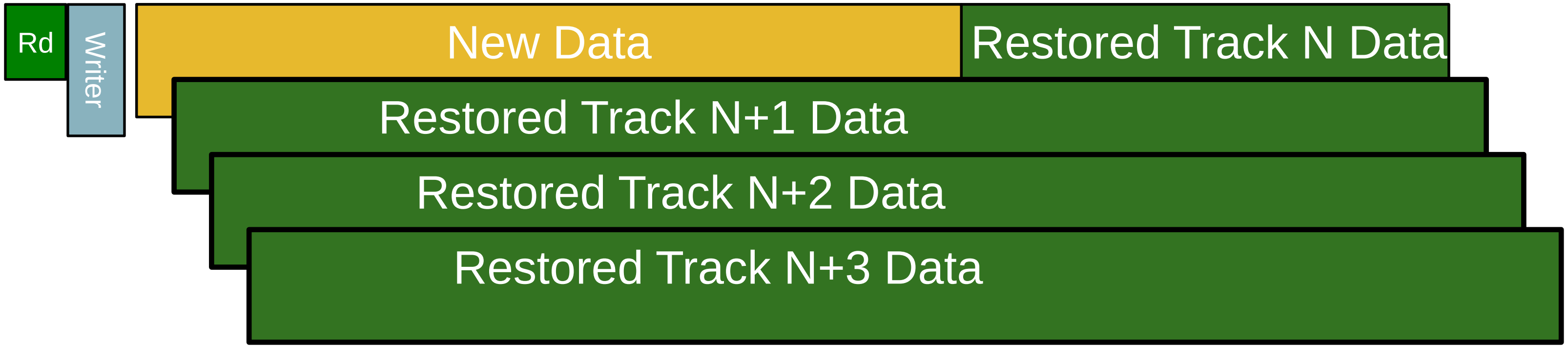
Updating a band with new data





1. Read old data
2. Merge with new data

Updating a band with new data



Rd

Writer

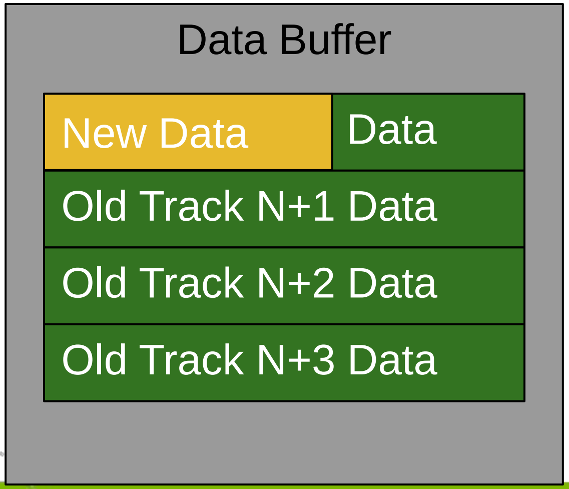
New Data

Restored Track N Data

Restored Track N+1 Data

Restored Track N+2 Data

Restored Track N+3 Data



Data Buffer

New Data

Data

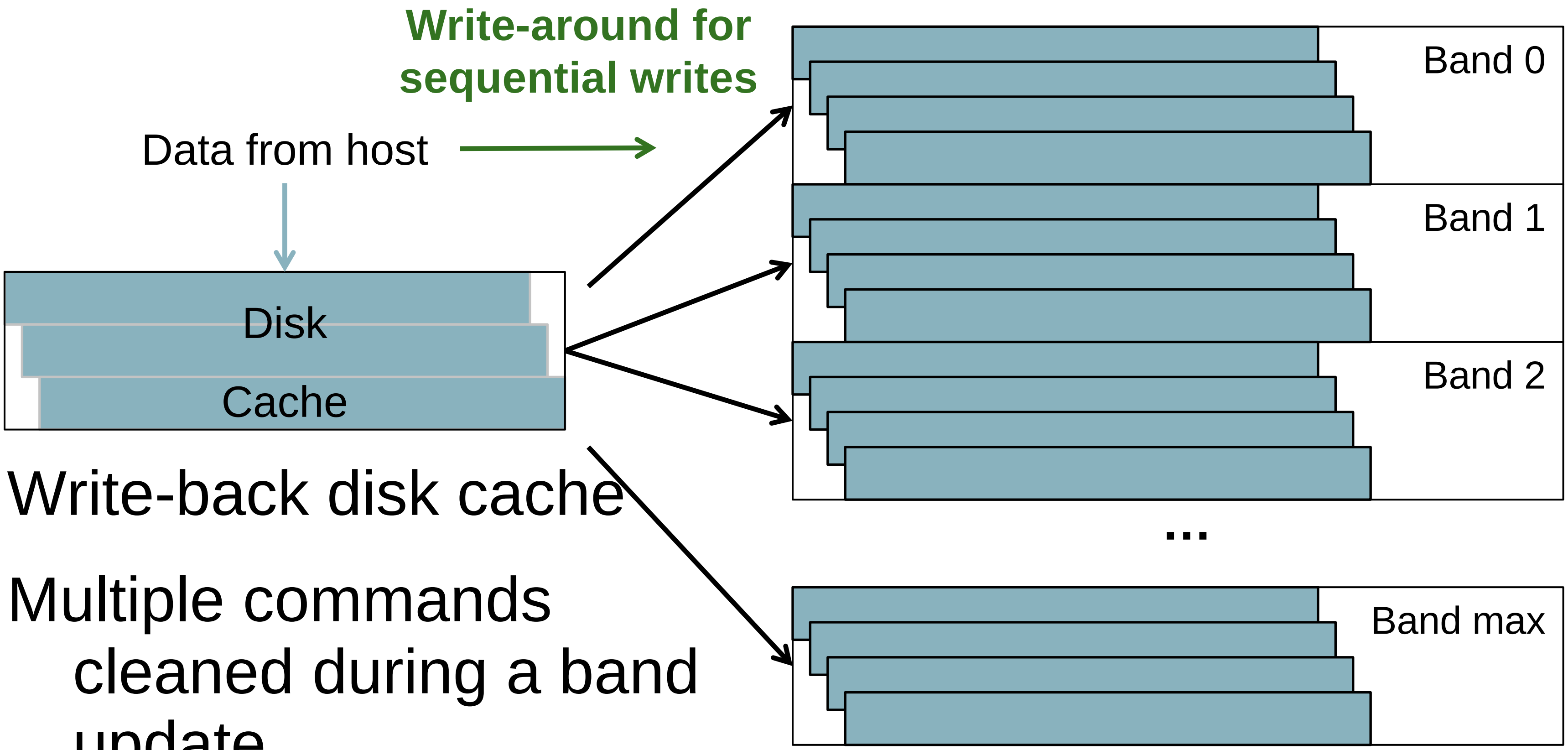
Old Track N+1 Data

Old Track N+2 Data

Old Track N+3 Data

1. Read old data
2. Merge with new data
3. Write new data, refreshing old data

Improving Write Performance



Write-back disk cache

Multiple commands cleaned during a band update

Drive Managed SMR

Advantages

- No host changes required
- Large write-back disk cache
  - Fast bursty random writes
  - Efficient cache cleaning for high spatially density
- Write-around for sequential writes
  - Conventional performance at media data rate
- Extremely effective in Personal Compute

Problem Statement

Advantages

- No host changes required
- Large write-back disk cache
- Write-around for sequential writes
- Extremely effective in Personal Compute

Challenges

- Disk cache is a limited resource
  - Full cache has slow random write performance
  - Larger cache has areal density cost
- Cleaning is complex
  - Large command latency tails
- Write-around is limited
  - Sequential detection is non-trivial
    - Multiple streams versus random
    - Long inter-command time versus end of stream
    - Multiple tracks are needed
      - Write one track after the next track is buffered or queued

SMR Drive Types

The rest of the story

Drive Managed

- Drive autonomously hides all SMR issues
- Backward compatible

Permissive

Host Aware

- Superset of Drive Managed and Host Managed
- Backward compatible
- Extensions to ATA and SCSI command sets

Host Managed

- Extensions to ATA and SCSI command sets
- Error conditions for some reads and writes
- Not backward compatible
- New device type

Restrictive

Host Aware SMR Solution

Goals

- Performance parity with conventional disks
  - Constrained, intended use cases
  - Trivial sequential detection
- Minimal interface changes
  - A few new commands a parameters
  - No changes to Read and Write commands
- General purpose
- Enable more markets
  - Grow beyond Personal Compute

Host Aware SMR Solution

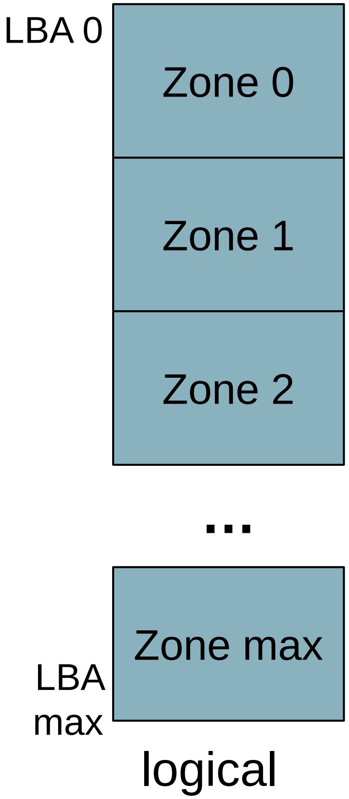
Achieving the goals

Goals

- Parity with conventional disks
- Minimal interface changes
- Enable more markets

Methodology

- Zones
  - Logical address ranges exposed to host
- Write Pointers
  - Location of sequential writing
- Host controls zone life cycle
  - Tell drive what sectors are not in use, “unwritten”
- Expose key device capabilities
  - Number of active sequential streams at full performance
  - Amount of random write space at full performance



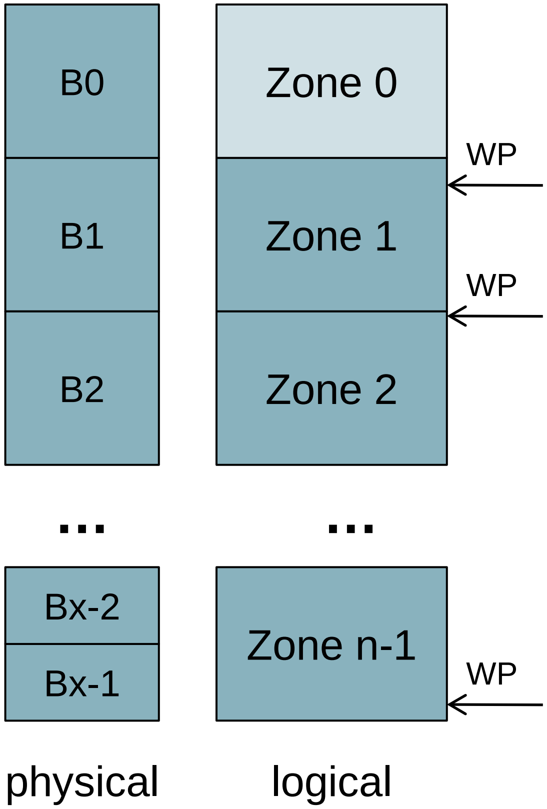
Zones

SMR Bands

- Physical construct
- Boundaries are not known outside the drive

Zones

- Logical space is divided into zones
  1. Conventional zones
  2. Write pointer zones
     - Each zone has its own Write Pointer
     - Each zone has its own state



Write Pointer Zones

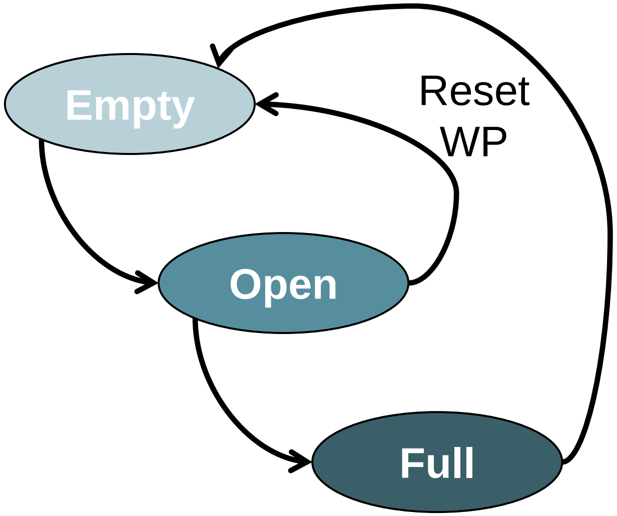
Writes at the write pointer have conventional performance

- Write pointer automatically advances

Writes not at the write pointer handled like Drive Managed

- Write pointer may or may not advance

Issue Reset Write Pointer before re-writing



Empty

- Write pointer is at start of zone

Open

- Write pointer is mid-zone

Full

- No write pointer value

New Commands

REPORT ZONES

- Reports configuration and current state of zones
  - Type, Size, Start LBA, State, Write Pointer

  Size = 256 MiB
- SAME flag in returned header specifies that all zones are the same size and type

  SAME = true
- Report can be restricted by type or state
- No method to change the configuration in the field

RESET WRITE POINTER

- Resets the write pointer of a zone to the start
- All LBAs in that zone become unwritten

Host Aware Device Capability Parameters

In New ATA Log or SCSI Vital Product Data Page

Open zones

- Optimal Number Of Open Sequential Write Preferred Zones

  The largest number of zones that should be open for best performance

  = 128 zones

Random write zones

- Optimal Number Of Non-Sequentially Written Sequential Write Preferred Zones

  The largest number of zones that should be randomly written for best performance

  Arbitrary set of zones, no configuration needed

  = 16 zones

Host Aware Signature

How do you tell that a drive is a Host Aware device?

SCSI (SAS)

- HAW_ZBC = true
  - New bit in Block Device Characteristics VPD page

ATA (SATA)

- Host Aware Feature Set = true

Host Aware SMR Solution

Intended Usage Model

1. Use REPORT ZONES and parameters to determine configuration
2. Assign random write zones as needed
   - Limit to the device’s Random Zones capabilities
   - Don’t care about Write Pointer values
   - Don’t issue Reset Write Pointer
3. Use the rest of the zones for sequential writing
   - Write Pointer is implicitly known
4. Control the number of Open zones
   - Limit to the device’s Open zones capability
5. Garbage collect to evacuate zones for re-use
   1. Copy non-stale data to an open zone
   2. Issue Reset Write Pointer
   3. Move zone to free pool

Resources

feldman@seagate.com	Sample drives
t10.org	ZBC – SCSI Zoned Block Commands letter ballot starting soon
t13.org	ZAC – ATA Zoned ATA Commands
github.com/hgst/libzbc	Linux user space libraries ZBC emulation
git.kernel.org/pub/scm/linux/kernel/git/hare	Linux SCSI layer components in review
Linux Vault Conference	March 11-12, 2015 Boston

The Future of Cheap & Deep

Capacity gains require SMR

Some applications run well on Drive Managed

Other stacks will make use of Host Aware and upgraded filesystems

- Leveraging the intended usage model