

OPEN

Compute Summit

January 28–29, 2014 San Jose

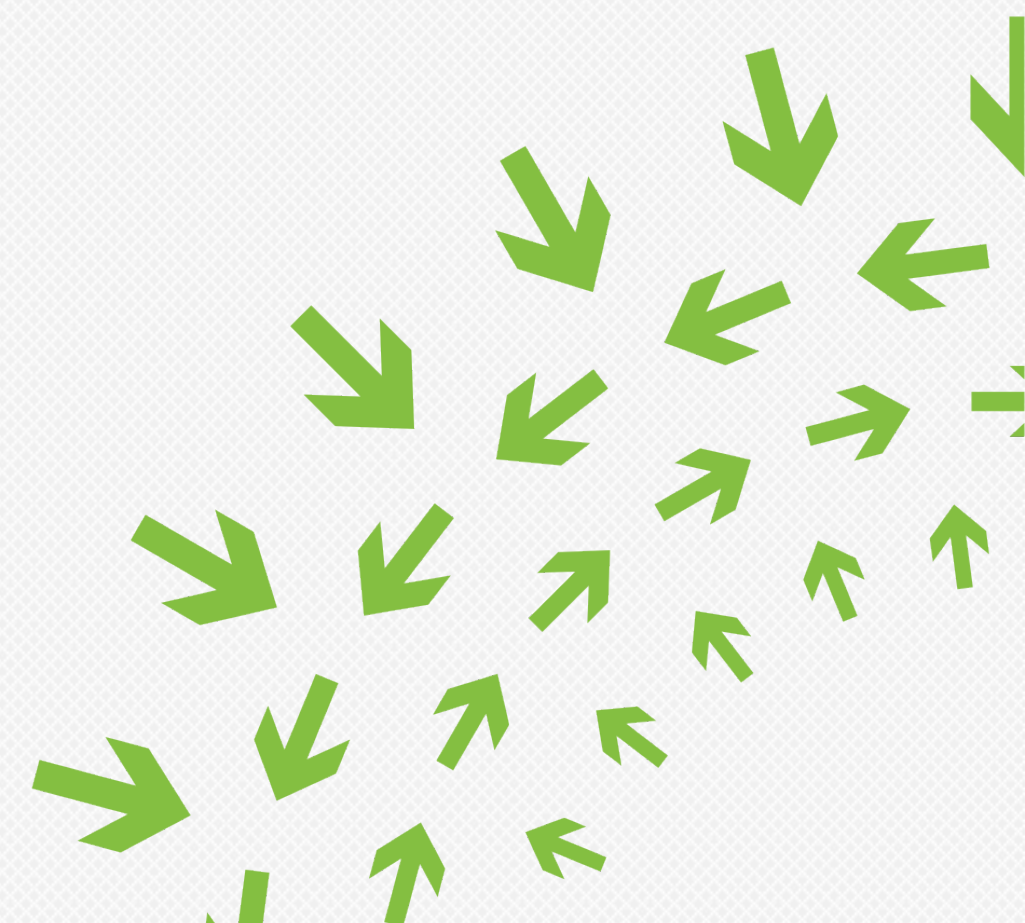




SMR Host Managed

Direction for Standardization

Albert Chen
Jim Malina
Western Digital Corporation

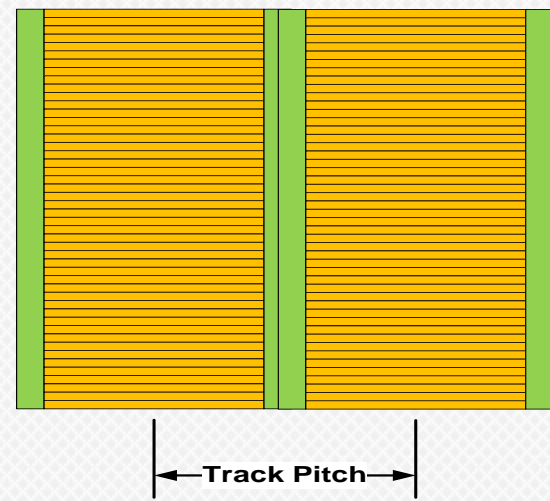


Agenda

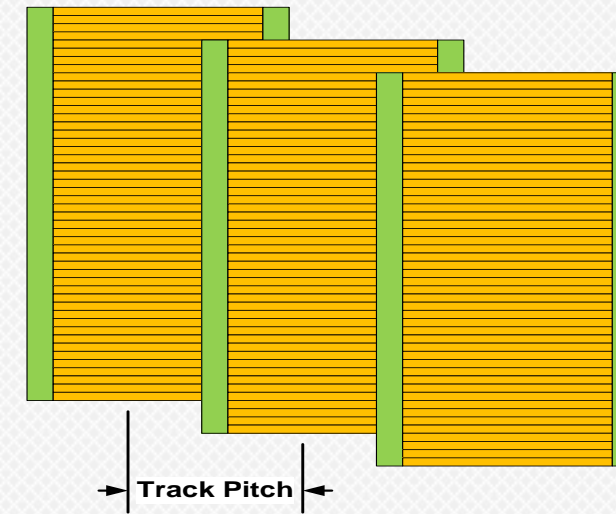
- SMR Technology
- SMR Alternatives
- WD Approach: Zone Block Commands
- Possible Implementations
 - Single Zone
 - Multiple Zone



SMR - What Changes:

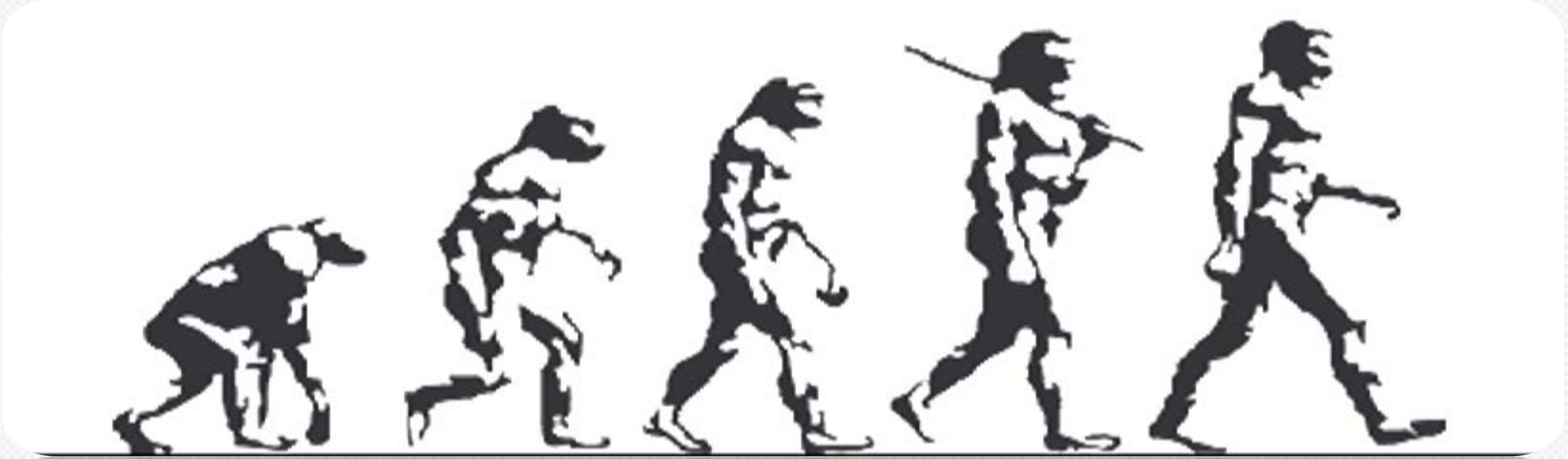


Random Write
Hot Read



Sequential Write
Hot Read





Firmware

Driver

Storage Stack

File system

Application



T10 SMR Types

SMR category	Description
Drive managed	No host changes. SMR device manages all requests. Performance is unpredictable in some workloads. Backward compatible
Host aware	Host uses new commands & information to optimize write behavior. If host sends sub-optimal requests the SMR device accepts the request but performance may become unpredictable . Backward compatible
ZBC – Zone Block Commands (previously restricted)	Host uses new commands & information to optimize write behavior. Performance is predictable . If host sends non-sequential write requests the SMR device rejects the request. Not backward compatible



SMR Alternatives

SMR Type	Drive Managed	Host Aware	Host Managed
Method	Indirection	Indirection & Sequential Write	Sequential write & Random write
Usage scenario	Client >100% duty cycle	Client & Data Center (serve two masters)	Client & Data Center SMR Friendly Storage Stack
Pros	Plug & play	Allows File Systems to mature	Reduced Complexity
Cons	Higher cost/complexity Unpredictable performance	Higher cost / complexity Unpredictable performance	SW support required Not backwards compatible



Quest for a better drive... ZBC delivers

- More consistent performance across drive vendors
 - Drive performance comparable with conventional
 - Improved Data Integrity
 - No Write Amplification
 - Lower Complexity
 - Fewer Corner Cases
 - Fewer Drive Resources (uP MIPS, DRAM, Media Over Provisioning)
 - A solution with mature, well understood, production file systems already in the Linux kernel
 - Delivers a simplistic interface that enforces sequential requirements. Garbage collection is done in established host SW
 - Plus...
- Lower Power
 - Faster Turn Around
 - Access to application & system level semantics
 - Scale with Host HW
 - Simpler command set
 - Easier to Manage
 - Lower \$/GB



Host Managed – Zone Block Commands

- Standard being developed under INCITS T10
- New Device Type
- Same SBC read/write protocol
 - Uses subset of SBC commands plus new commands as necessary
 - Read any LBA at any time
- Two types of zones – random write and sequential write only (all zones may be read random or sequentially)
 - Estimated total budget for random writeable zones: < 0.1% total capacity
 - First and last zones (zone with LBA 0 and LBA Max) may be random writeable
 - All other zones are sequential write only
 - Random write zones are 512 byte LBA accessible
 - Sequential write zones are 4K byte aligned
- Zone size
 - Zone size(s) fixed at the factory. Current thinking is 256MB.
- Writes to LBAs in sequential write zones are sequential write only
 - Non-sequential write commands within a zone return an error
 - Sequential write zones have a sequential write pointer
 - Report sequential write pointers for sequential write zones
- All read/write LBA access performance is comparable to conventional HDD

SMR Zoned Devices

Potential Implementations



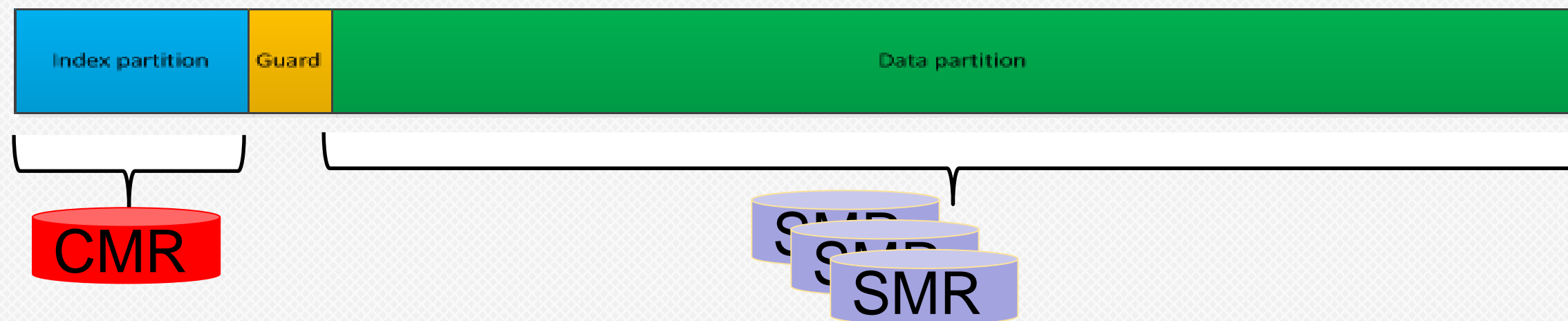
Host Managed: Single Zone Device

- Best possible SMR benefit (Capacity gain)
- Drive must be written 100% sequentially
- Host may “reset” the pointer to LBA 0



Host Managed: Single Zone Device Example

- File System – LTFS (Linear Tape File System)
 - Developed by IBM
 - Open Sourced – single device version
 - Requires 2 partitions
 - “Small” Index partition (re-writeable)
 - “Large” Data Partition (sequential write only)

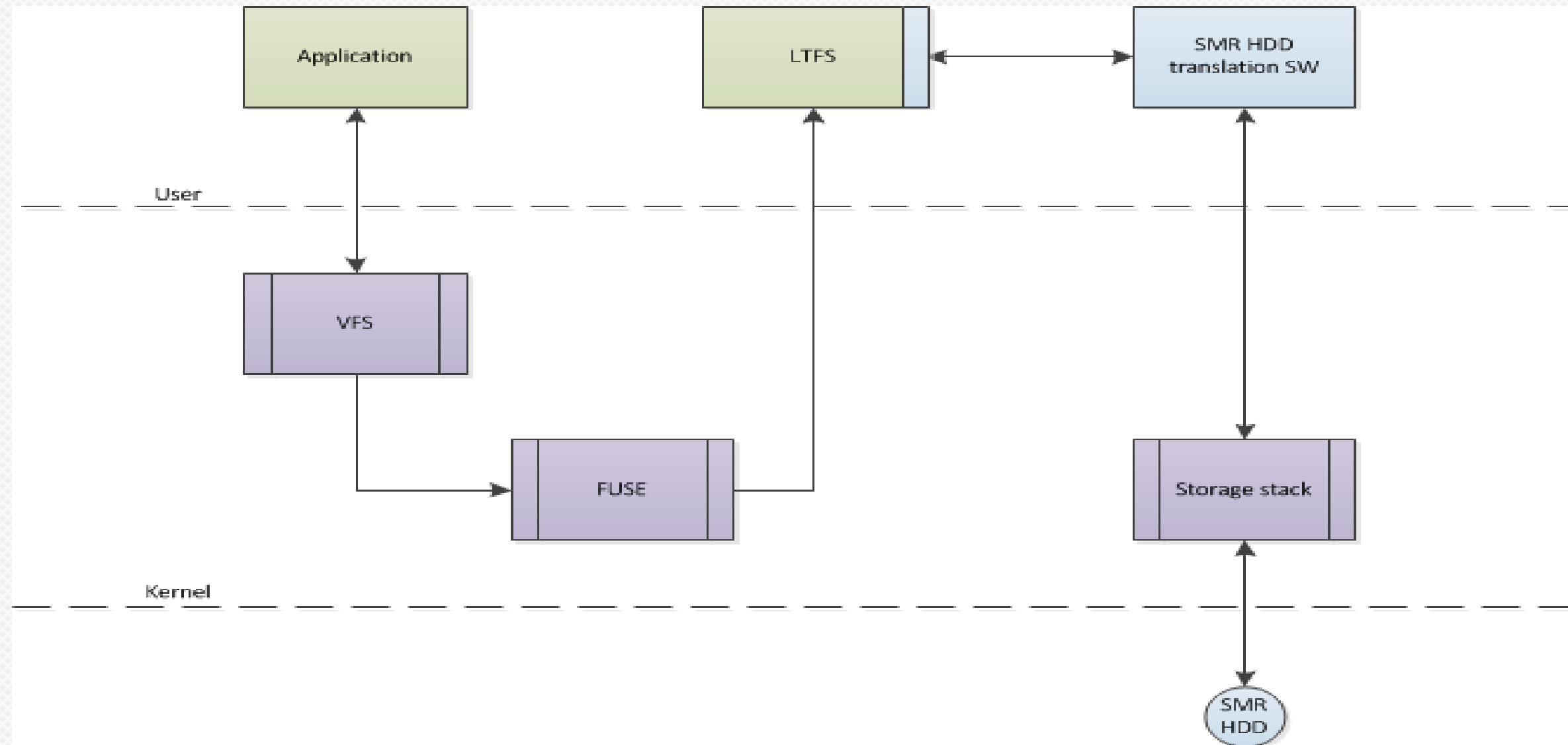


- Optimized for object store




Single Zone Device Implementation – Cloud Archive Application

Userspace (e.g. Hadoop) application to handle SMR semantics



Just for Archive?

But...a single zone may be used beyond archive!




**Gecko: A Contention-Oblivious
Design for Cloud Storage**

HotStorage Talk on June 13, 2012

Ji-Yong Shin
Cornell University

In collaboration with Mahesh Balakrishnan (MSR SVC), Tudor Marian (Google),
Lakshmi Ganesh (UT Austin), and Hakim Weatherspoon (Cornell)

 Cornell University
Department of Computer Science

<https://www.usenix.org/conference/fast13/contention-oblivious-disk-arrays-cloud-storage>

<https://www.usenix.org/system/files/conference/hotstorage12/hotstorage12-final57.pdf>



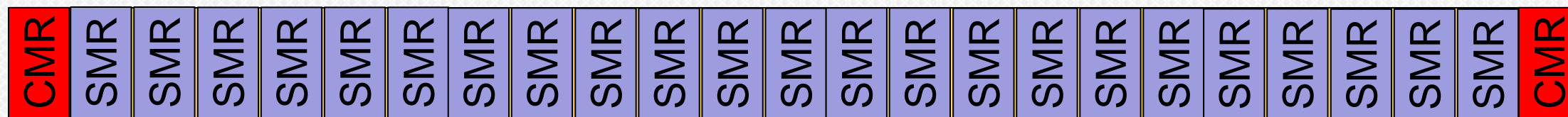
Host Managed: Multiple Zone Device Example

- Drive Partitioned into zones
 - Two zone types – random writeable and sequential writable
 - Sequential write zones have individual sequential write pointers
 - Estimated ratio of random : sequential capacity (1:10000)
- Drive reports sequential write pointers
- Host “resets” the pointer to first LBA in zone
- Most Flexible Use Case Support for ZBC



Host Managed: Multiple Zone Device Example

- File System – NilFS (Newly Implemented Logging FS)
 - Developed by NTT Laboratories
 - Open Sourced – single device version
 - Zone aware, performs garbage collection, power safe, Kernel 2.6.32 +
 - Writes sequentially within sequential zones
 - Requires
 - Zones of equal, power of 2 size
 - Two rewritable zones – for super block



- Garbage Collection
 - Performed from zone to zone (within device)

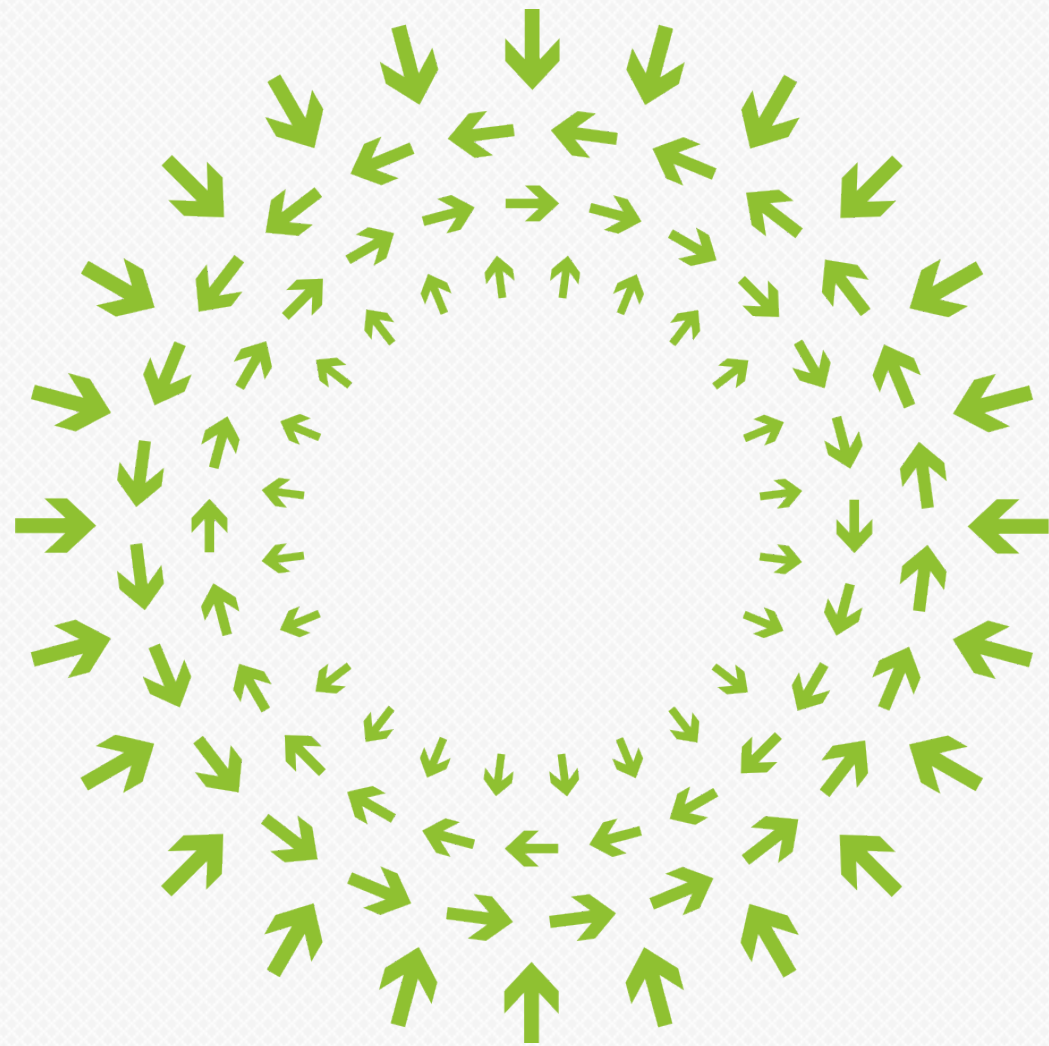


Next Steps

- WD has several configurations of SMR drives available that are compliant with the ZBC specification – request to participate in our early testing
- WD is collaborating with File System and Application Developers to broaden deployment alternatives – contact WD to participate in these collaborations.
- Review INCITS T10 ZBC Project Specification Draft

(Contact James Borden (james.borden@wdc.com) for more information)





OPEN

Compute Summit

January 28–29, 2014 San Jose

