

Addressing Shingled Magnetic Recording drives with Linear Tape File System

Albert Chen and Jim Malina
SDC – Storage Developer's Conference 2013
SNIA

absolutely[™] 

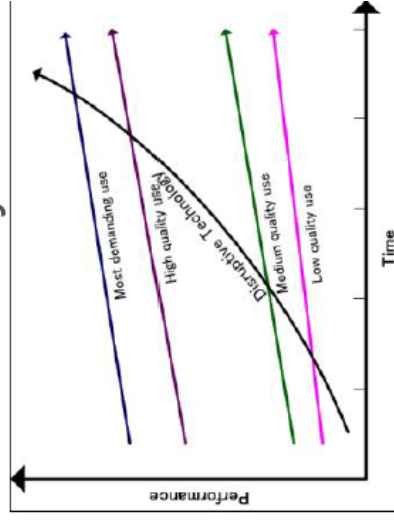
© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Host Managed SMR Agenda

- 1** Host Managed SMR Philosophy
- 2** Problem Statement - Opportunity
- 3** Approaches
- 4** Hardware: Archive / Cloud drive
- 5** Software: Filesystem
- 6** Development
- 7** Validation: Testing & Performance
- 8** Lesson's Learned / Feedback

Host Managed SMR Philosophy

- **Shingle Magnetic Recording** is a **disruptive technology** – Innovation that results in **variable performance** but adds **new value** proposition. Fringe customers may benefit and over time performance may increase.
- Identify markets ripe to benefit from SMR
 - Able to modify their storage stack
 - Write 100% sequentially (tape like)
 - Benefit from lower \$/GB



- "You've got to start with the customer experience and work backwards to the technology. You can't start with the technology and try to figure out where you're going to sell it" Steve Jobs

3 © 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Source available at <http://en.wikipedia.org/wiki/File:Disruptivetechnology.gif>



Problem Statement / Opportunity

- Background:
 - Archive / Cold storage is a new market.
 - Write once and forget. (read seldom, if ever).
 - Primary care about: TCO
 - Shingle Magnetic Recording is a new technology
 - Sequential write (only)
 - Path to lowest \$/GB if host can abide by one simple rule
 - Industry has discussed two approaches for SMR
 - Drive does data management – similar to FTL
 - Host does data management – via a SMR file system
- Problem Statement

- Solve the new market opportunity, Archive / cold storage with new recording technology that provides lowest \$/GB

Host Managed SMR Approaches

■ Drive managed SMR

- Similar to Flash Translation layers, data management is complicated.
 - Metadata, garbage collection, over provisioning, write amplification, electronics resources, variable performance, validation.

■ Host Managed SMR

- Similar to Flash File Systems, data management is complicated but there may be opportunity to leverage mature and popular file systems that write sequentially.
 - Drive will not be burdened by the resource cost of a drive managed system

Host Managed SMR

Hardware: Archive cloud drive

- Host Managed SMR, Single Zone, Host writes 100% sequentially – **Simplest SMR implementation**
- Intended for “Big Data”, Cold/Archive, Tape Replacement
- Use Case:
 - Write once, read seldom
 - Garbage Collect from Device to Device
- We built it
 - Proved AD gain, Build & test process
 - Proved Device Architecture, Reliability & Error handling

Software

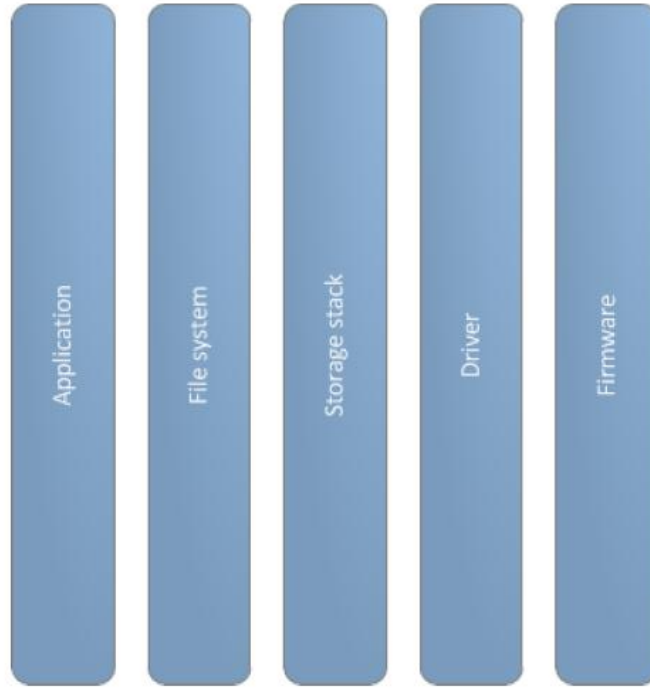
© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

absolutely™ 

Host I/O

- Sequential write only.
- No overwrite allowed.
- Garbage collection.
- Handle out of bounds access failure.

Abstraction layer



File system

- Industry usage/support.
- Open source/specification.
- User vs kernel space
 - Minimize risk
 - Ease of development/debug

LogFS

“LogFS is a Linux log-structured and scalable flash file system, intended for use on large devices of flash memory. It is written by Jörn Engel and in part sponsored by the CE Linux Forum.

LogFS is included in the mainline Linux kernel and was introduced in version 2.6.34, released on May 16, 2010.”

- Wikipedia (<http://en.wikipedia.org/wiki/LogFS>)

LTF5

The Linear Tape File System

David Pease*, Arnon Amir¹, Lucas Villa Real¹, Brian Biskoborn⁵, Michael Richmond⁶ and Atsushi Abel⁸

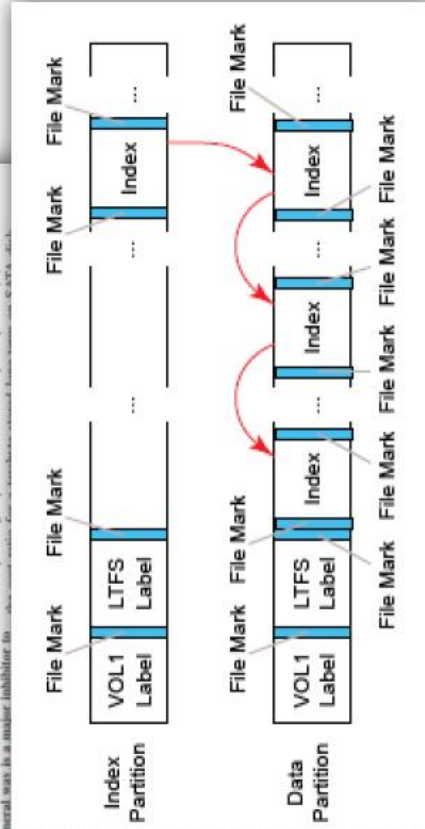
* IBM Almaden Research
pease@almaden.ibm.com
¹ arnon@almaden.ibm.com
² lucasvr@us.ibm.com
³ bbiskob@us.ibm.com
⁴ mar@almaden.ibm.com
⁵ IBM Yamato Lab
piste@jp.ibm.com

Abstract—While there are many financial and practical reasons to prefer tape storage over disk for various applications, the difficulty of using tape in a general way is a major inhibitor to its wider usage. We present a new generation of tape hardware and software that make tape using standard, familiar file system interfaces. The Linear Tape File System (LTFS) is a flexible, portable, and intuitive sharable media, such as a USB drive. A tape library can host dozens or even hundreds of drives operating in parallel. A case-study comparison [3] found that LTFS is a major inhibitor to its wider usage.

I. MOTIVATION
In today's digital world, more and more data is being stored on tape. The broadcast and media industries are changing to all-digital, file-based storage. In what is being replaced with file-based storage capacity. More important, the data is being stored on tape. The broadcast and media industries are changing to all-digital, file-based storage. In what is being replaced with file-based storage capacity. More important, the data is being stored on tape.

This transformation helps storage capacity. More important, the data is being stored on tape. The broadcast and media industries are changing to all-digital, file-based storage. In what is being replaced with file-based storage capacity. More important, the data is being stored on tape. The broadcast and media industries are changing to all-digital, file-based storage. In what is being replaced with file-based storage capacity. More important, the data is being stored on tape.

Within today's storage technology, data tape is still the

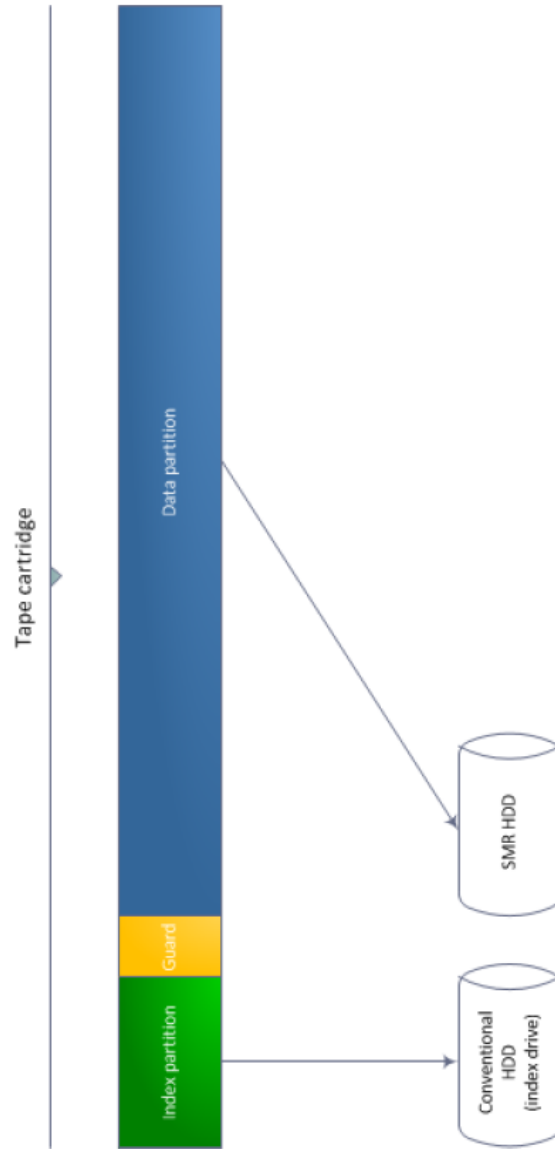


© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

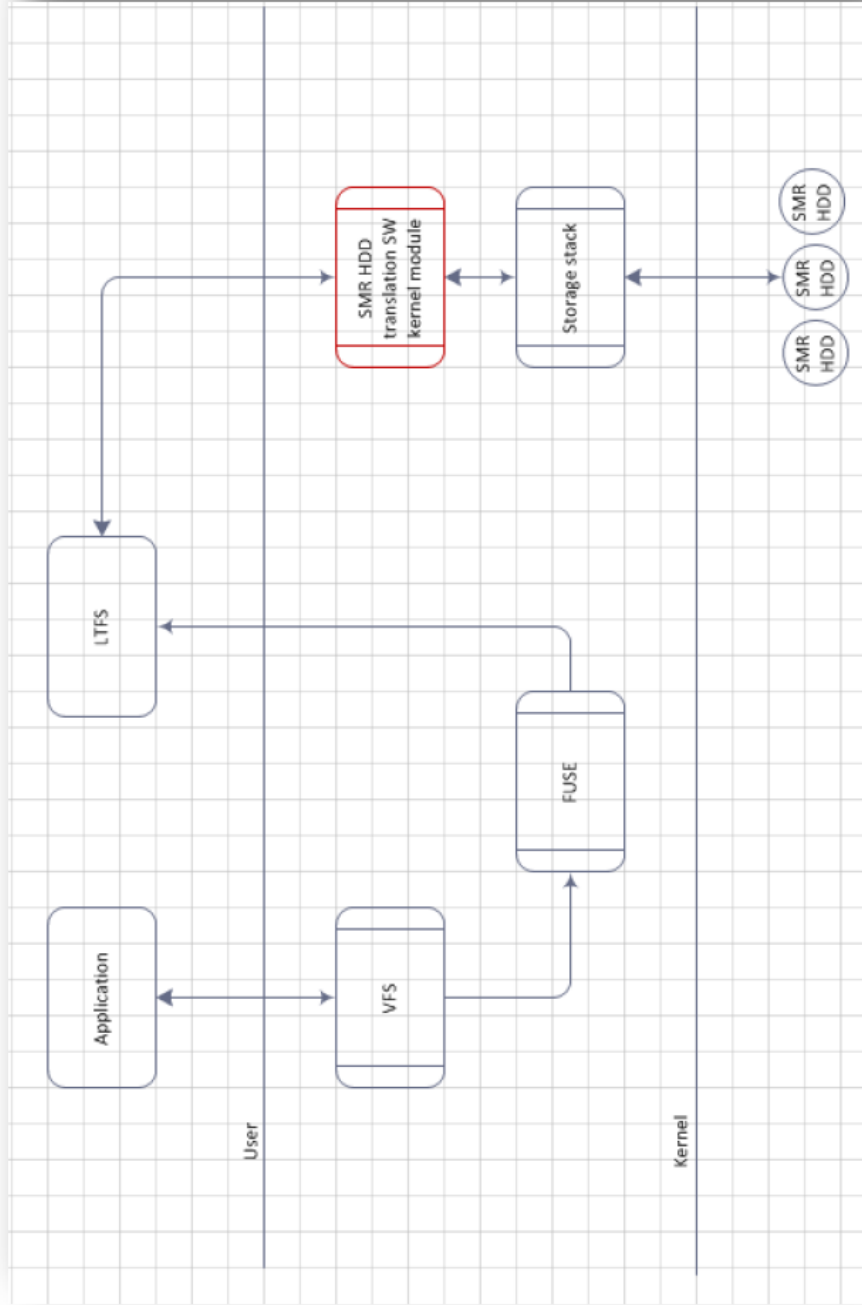
Images © 2010 IEEE <http://storageconference.org/2010/Papers/MSST/Pease.pdf>. © 2011 Hewlett-Packard, IBM and Quantum http://www.truistito.com/LTFS_Format_To%20Print.pdf.

absolutely™
wD

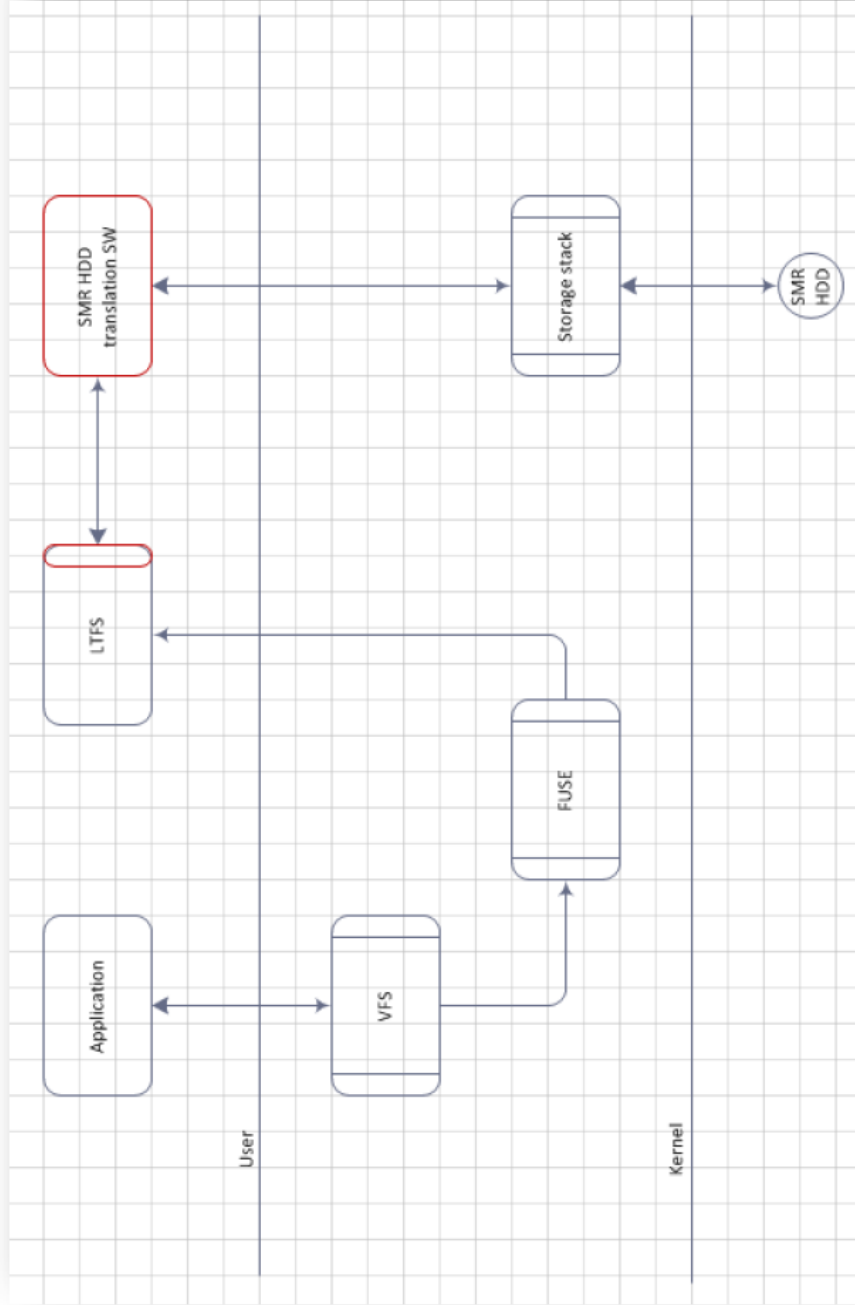
Partition



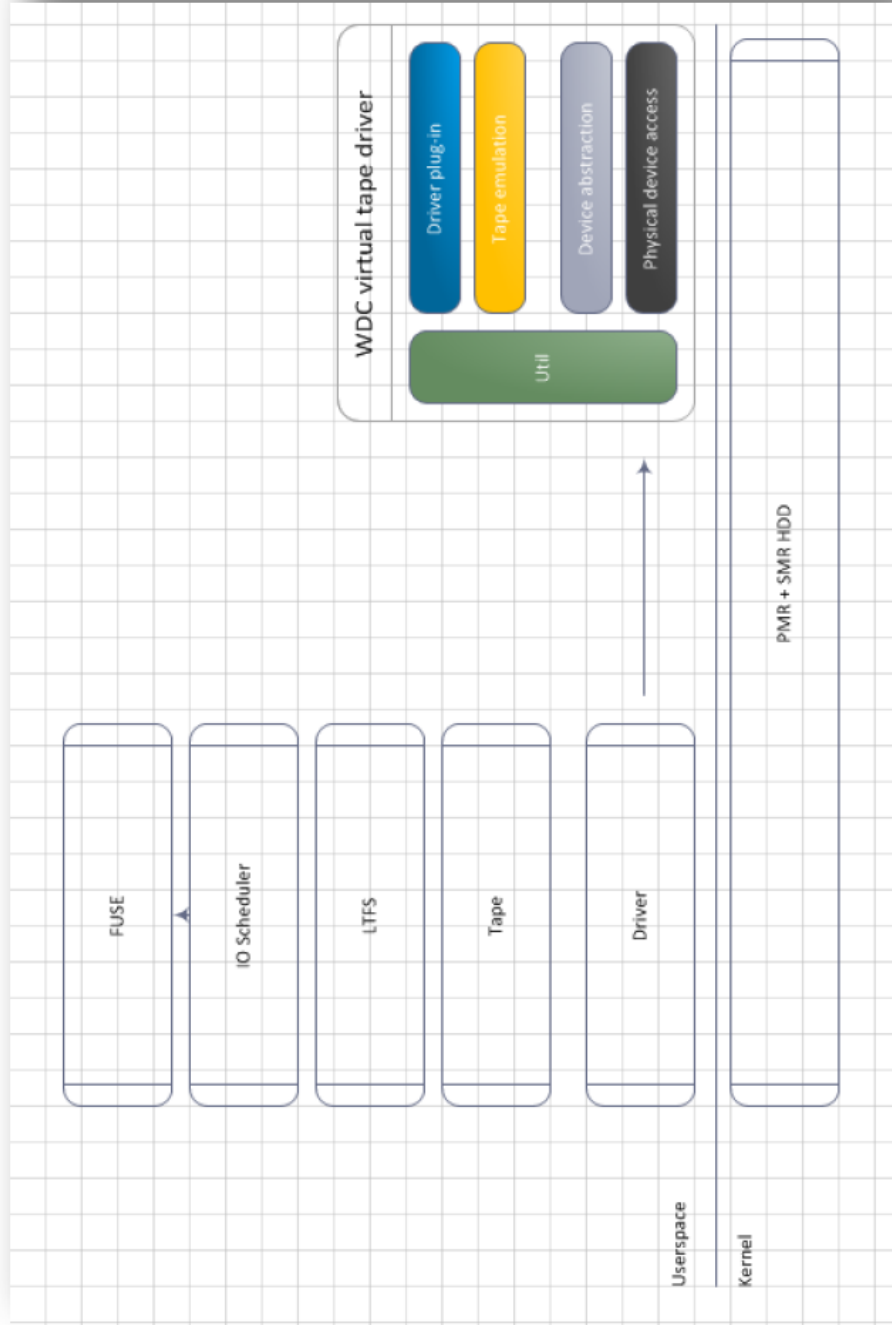
Kernel module



User space



Code layout



Development

LTFS 1.2.5 with Ubuntu 10.04.4 LTS

Lucid Lynx

© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED



Coding

- No changes to LTFS source
- Backend tape driver
- ~5000 LOC in 25 files
- 2 month POC
- Work in progress

Validation: Testing & Development

© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

absolutely™ 

Validation at each level

- Directory

- File

- Device

POC setup



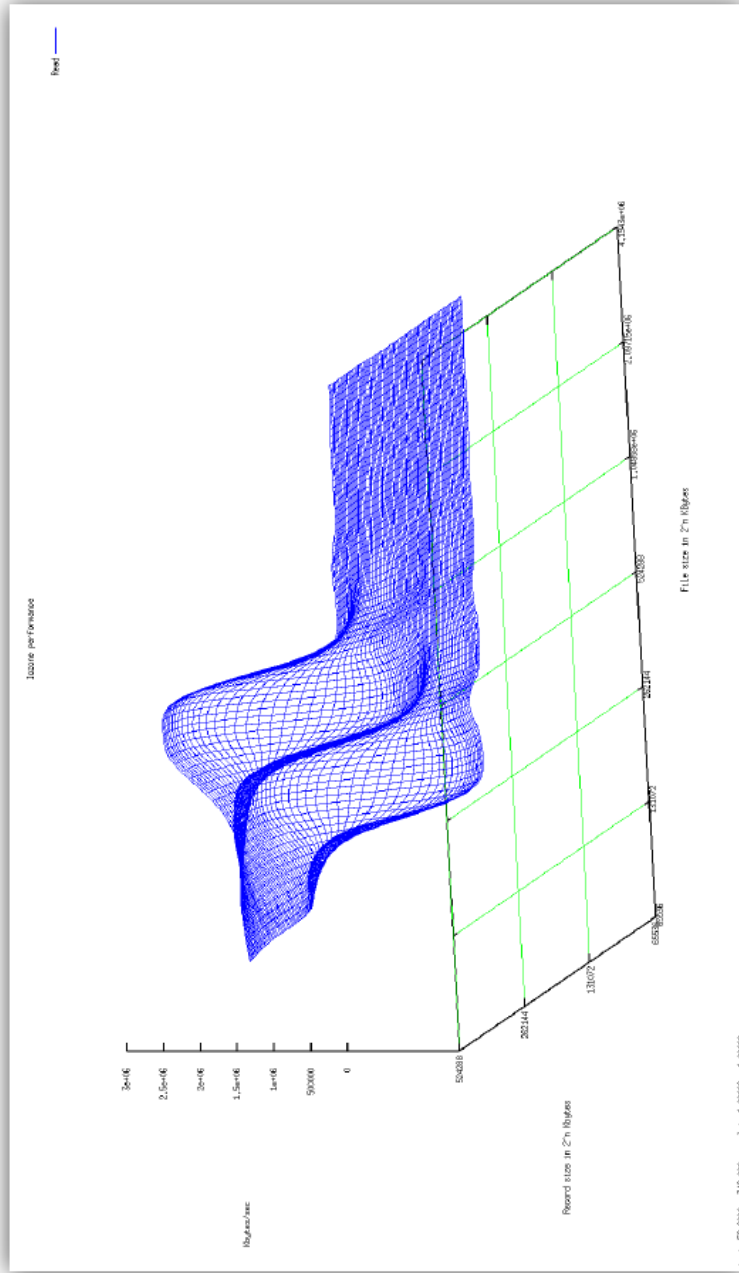
absolutely² 

© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Parameters

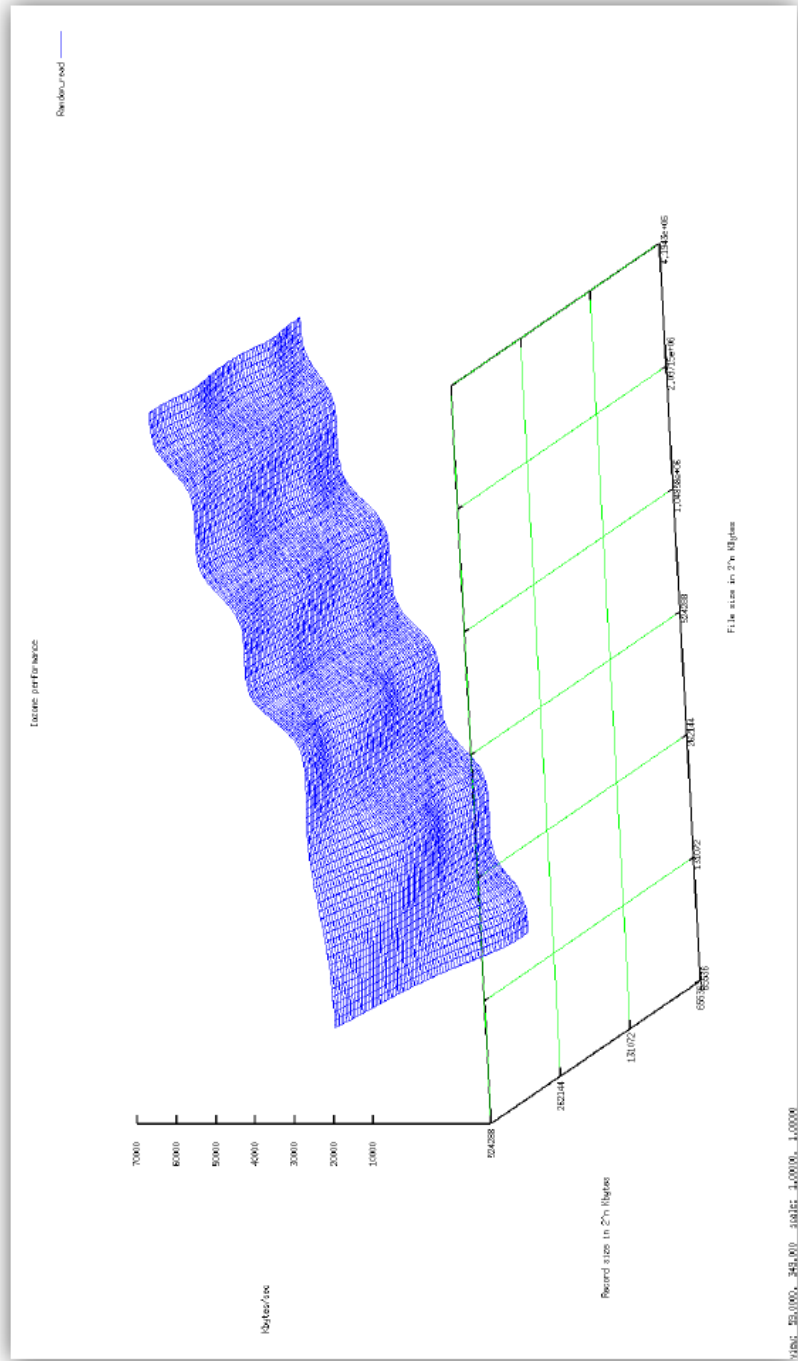
- File size: 500MB to 4GB
- Record size: 100MB to 500MB
- O_SYNC
- Purge processor cache before each file operation
- Unmount LTFS between each test
- Intel Corporation 5 Series/3400 Series Chipset 6 port SATA AHCI Controller (rev 06)

Performance – Sequential Read

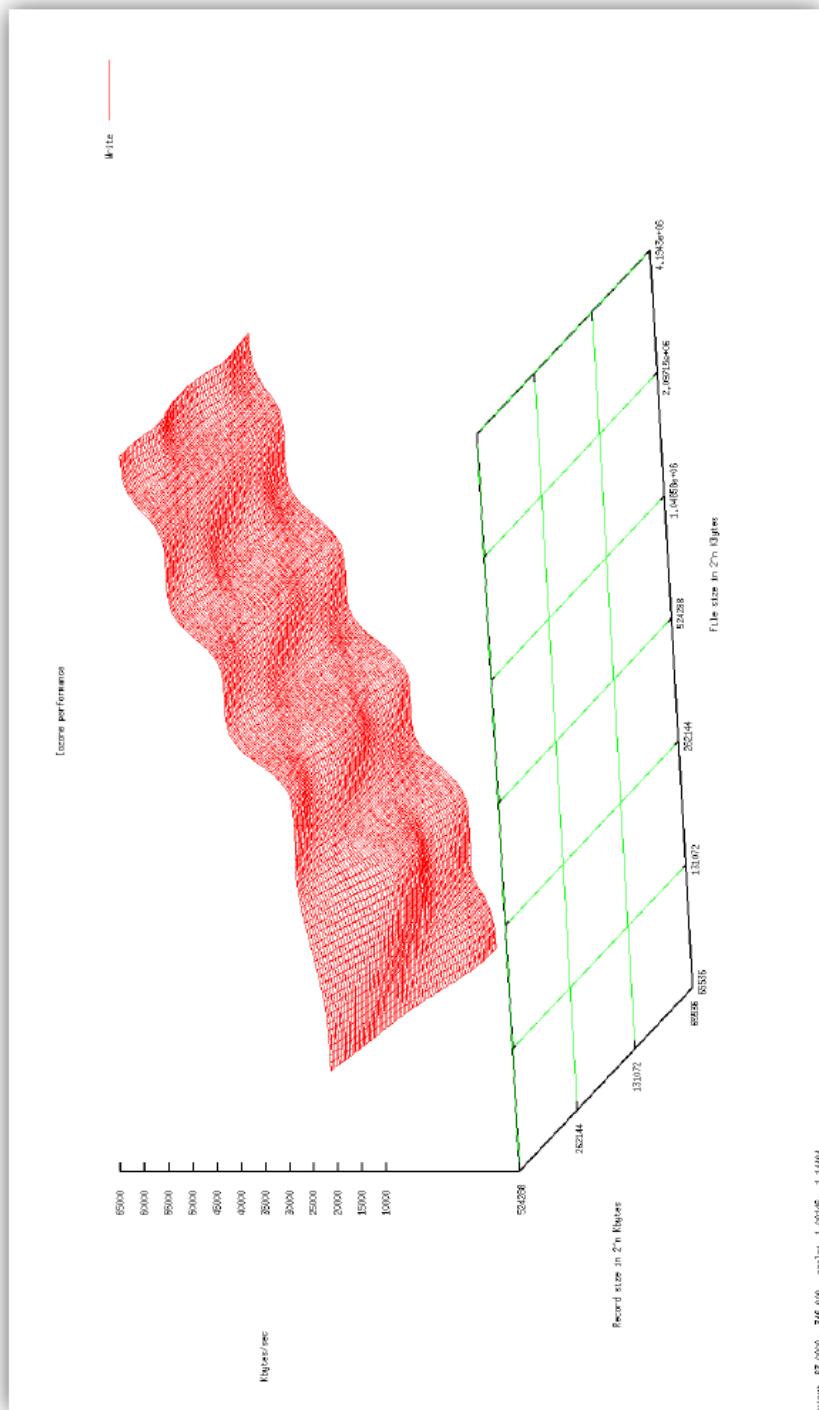


© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Performance – Sequential Write



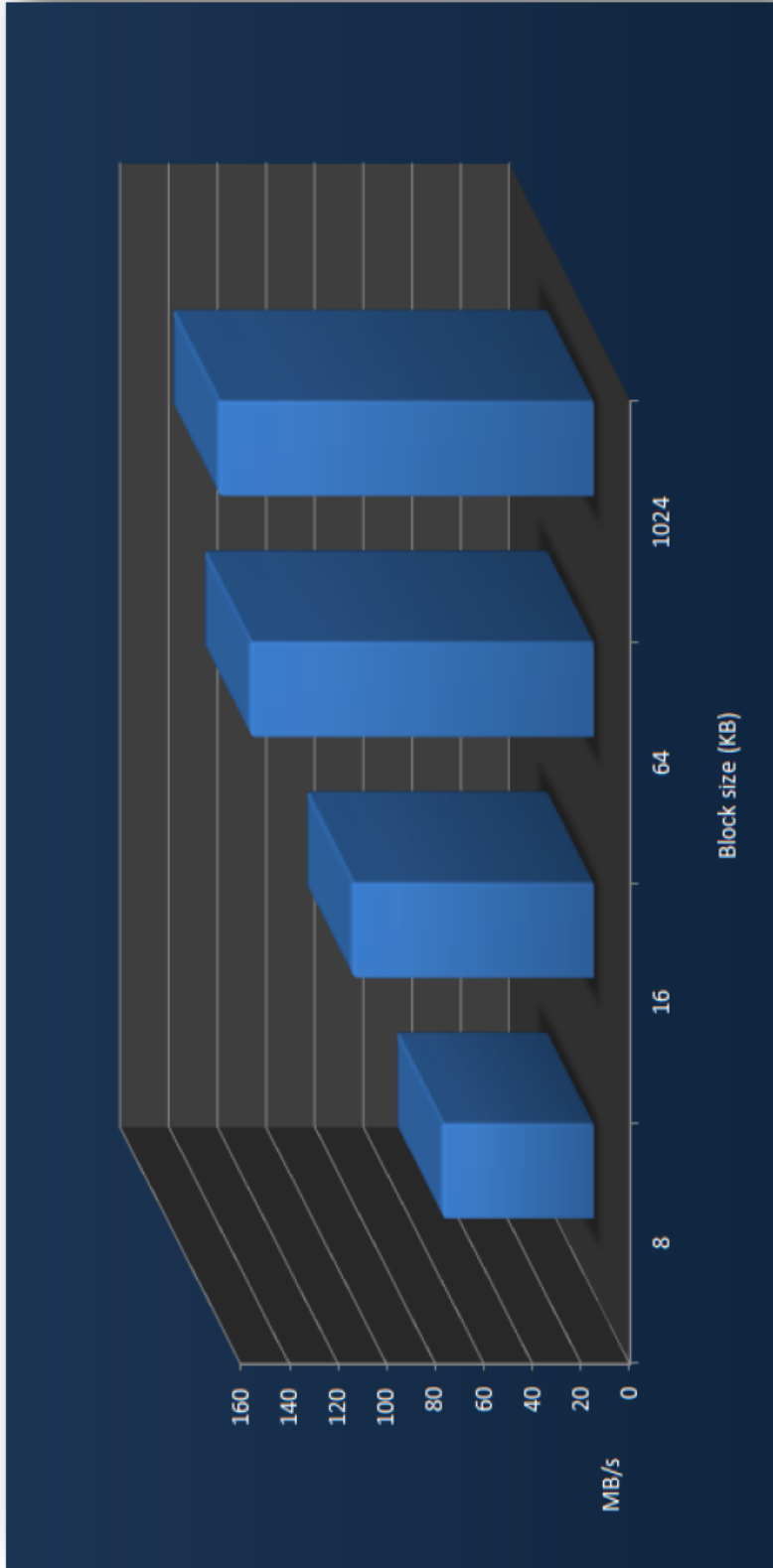
Performance – Sequential Write



© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

absolutely™

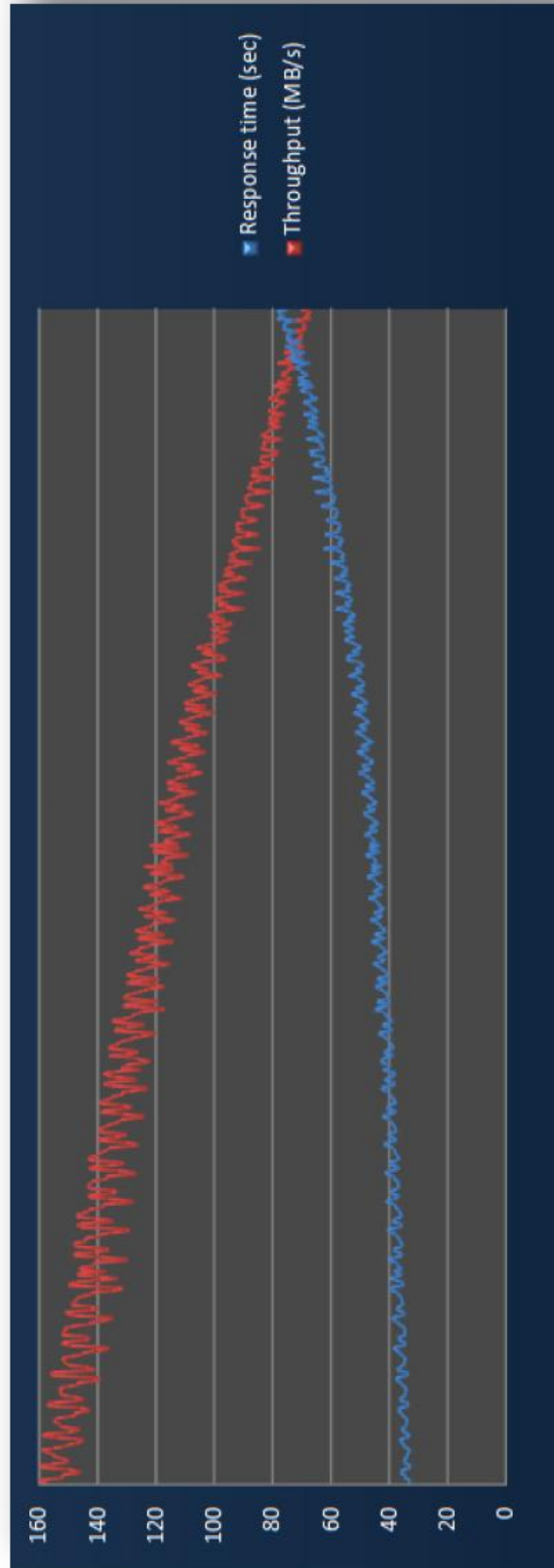
Throughput $f(\text{block size})$



© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED



Throughput $f(\text{LBA})$



© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Write sequential: ~116MB/s

© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

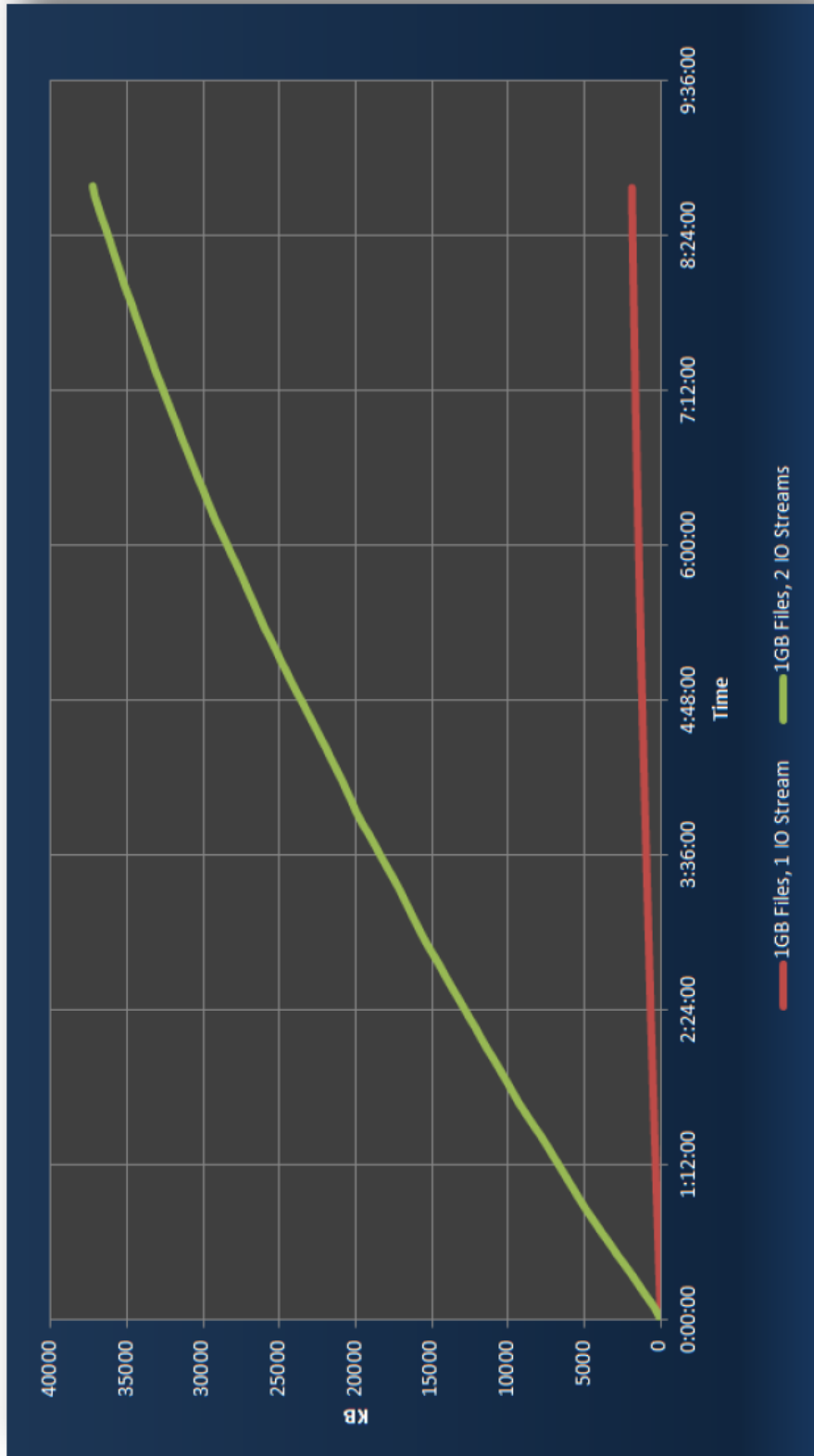


Metadata



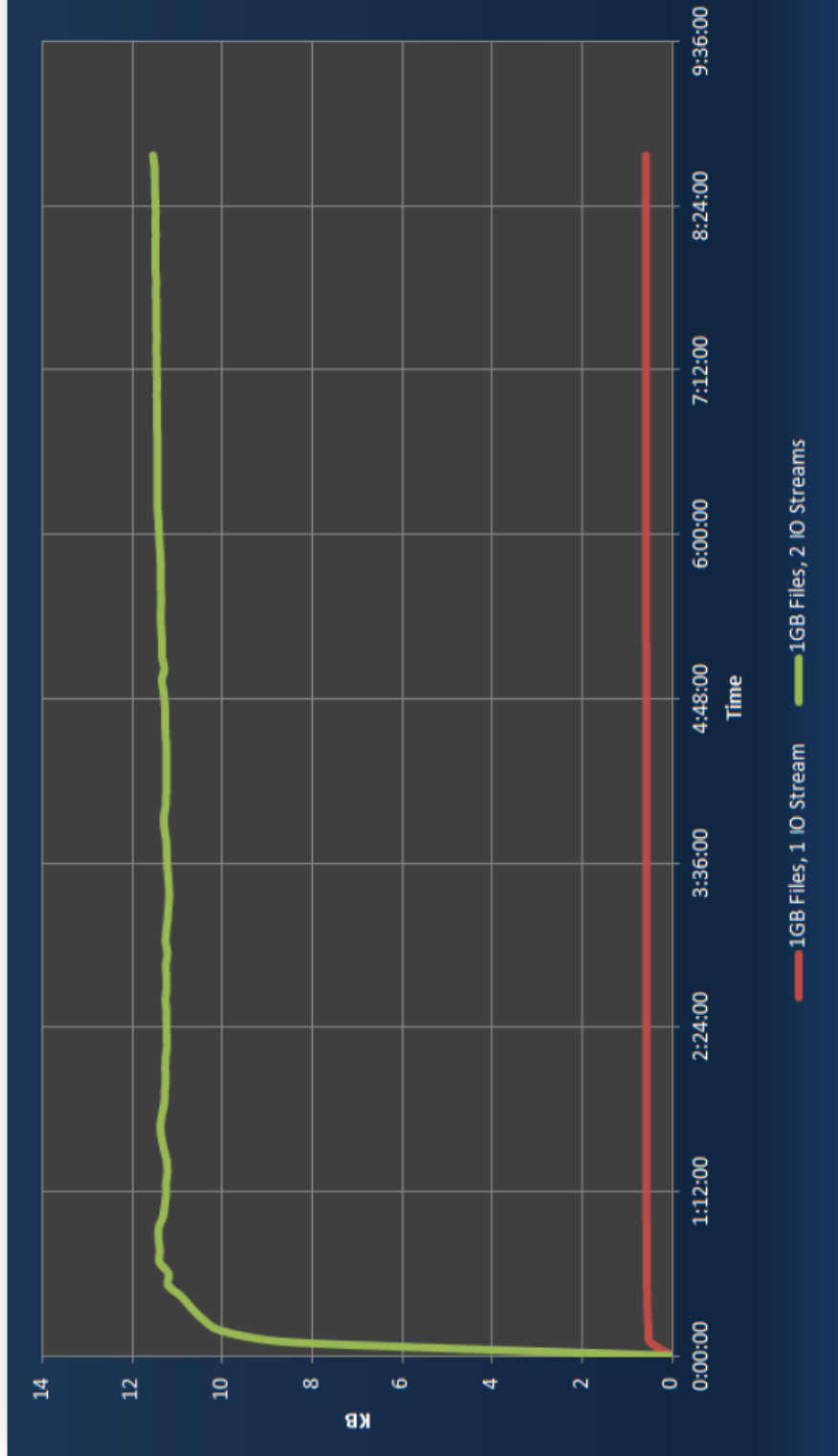
© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Index size growth $f(t)$



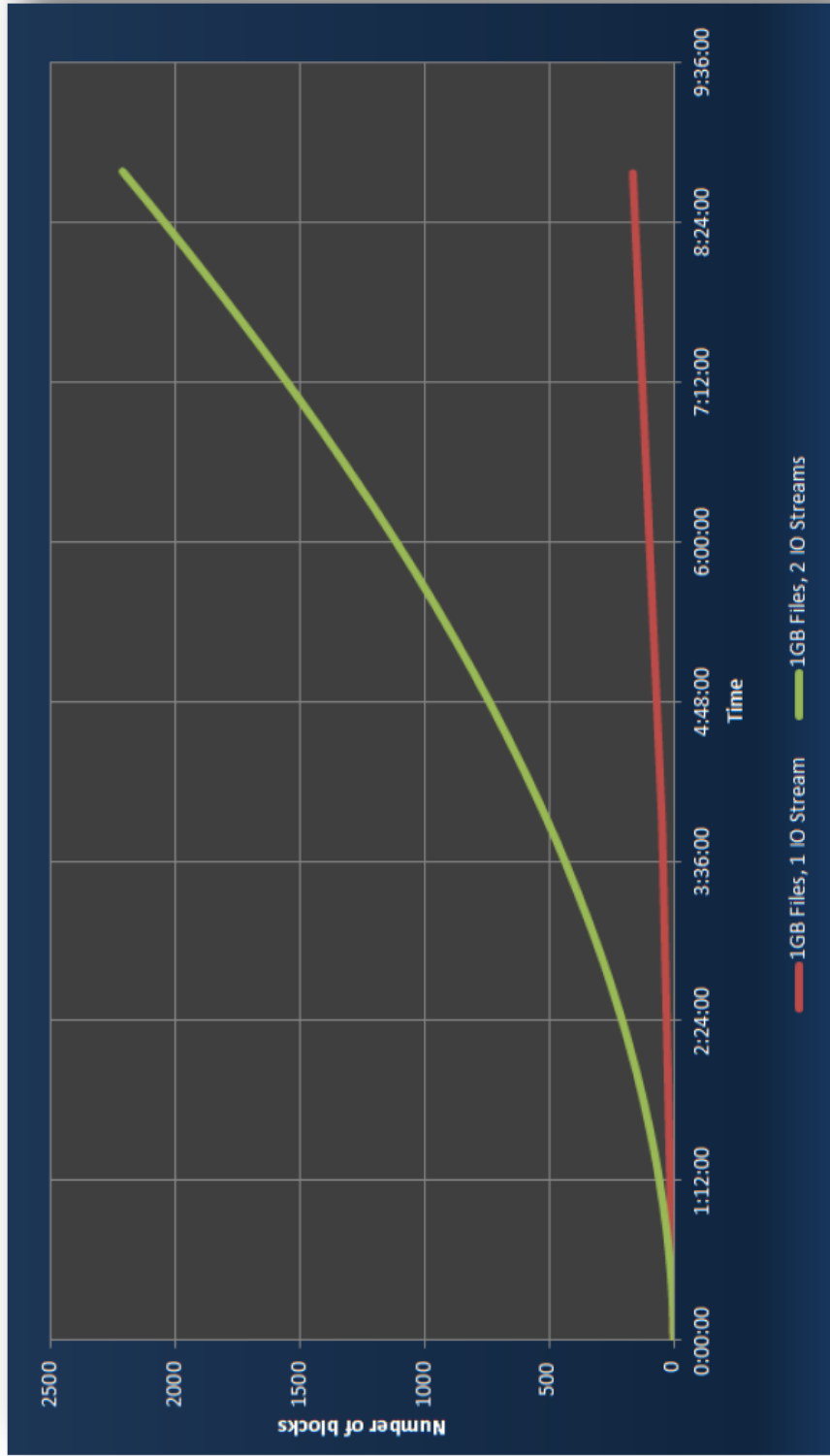
© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Index size/file $f(t)$



© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

Index block allocation (1MB) $f(t)$

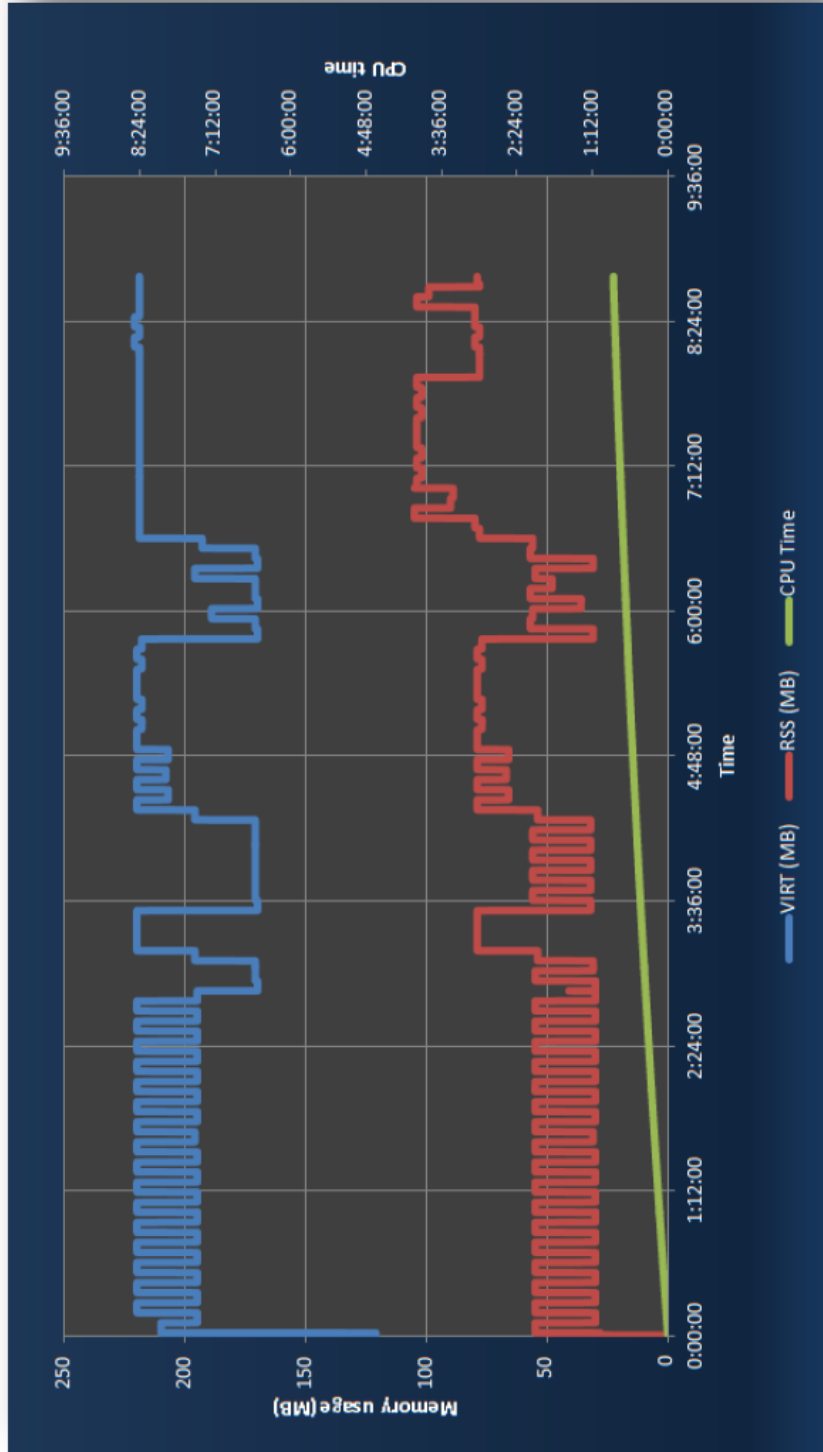


© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

System information

- Dell PowerEdge T110
- 2GB 1333 MHz DRAM
- Intel Xeon E3-1220 CPU
- LSI SAS9207-8e HBA

CPU and memory usage $f(t)$



© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED

What have we learned?

© 2013 WESTERN DIGITAL TECHNOLOGIES, INC. ALL RIGHTS RESERVED



1. LTFS interoperability
2. Performance
3. Testing

Host managed SMR Feedback

- Customers “dig” it - easy to use
- Writes are 100% sequential – fastest way to store data on HDD
- Read access at full HDD speeds (random and sequential)
- Smallest overprovisioned solution – **lowest \$/GB**
- Simple design – lower cost – higher reliability