Storage with **INTENSE** Network Growth (**SWING**) 

*Tom Coughlin, Coughlin Associates*  
**JANUARY 4 & 5, 2015**  
*RIVIERA HOTEL CONVENTION CENTER, LAS VEGAS*
Outline

• SV 2015 Agenda
• SWING
• HDDs
• Flash Memory and Solid State Storage
• Magnetic Tape
• Cloud Storage
• FaceBook
• Conclusions and Future Events
Intense Network Growth

• Video production, 4K distribution, OTT VOD are chewing up storage and bandwidth—Exabyte workflows within 10 years!
• The “Internet of Things” will expand our control of the world but requires big data analytics, bandwidth
  – security and privacy (where is the data kept?)
• Data center cloud storage is driving open standard and specialized storage growth and mobile applications
Day 1 - Sunday January 4, 2015
7:30 to 8:00 AM  Continental Breakfast
8:00 to 8:15 AM  Introduction: Tom Coughlin, Coughlin Associates
8:15 to 9:45 AM  A1: Creative Storage: Looking for Storage for High Resolution Content Capture and Production
9:45 to 10:00 AM  Morning Break
10:00 to 10:30 AM  Keynote 1 Jeff Qin, Facebook
10:30 AM to 12:00 PM  B1: Finding and Keeping It Safe: Protecting, Finding, Storing and Recovering Personal and Commercial Content
12:00 to 1:30 PM  Lunch and Exhibits
1:30 to 2:00 PM  Keynote 2 Michael Lee, Industrial Internet Consortium
2:00 to 2:45 PM  C1: Analyst Perspectives: What Will Cloud Storage Do to Local Storage?
2:45 to 3:15 PM  Afternoon Break
3:15 to 4:00 PM  D1: Speed is the Need: High Performance Data Center Fabrics to Speed Networking
4:00 to 5:00 PM  E1: Evolve or Die!: Storage Developments Drive New Storage System Options
5:00 to 6:00PM  F1: New Hollywood: Technology Allows Us to Do More with Less
6:00 to 8:00 PM  Reception

Day 2 - Monday January 5, 2015
7:30 to 8:00 AM  Continental Breakfast
8:00 to 8:45AM A2:  Storage Visions 2015 Rising Stars, Young Engineers Panel
8:45 to 9:15 AM  Keynote 3 Zack Deiri, Samsung Semiconductor Inc.
9:15 to 9:30 AM  Morning Break
9:30 to 10:30 AM  B2: Piping Up—Wires and Spectrum—Transporting Big Data into the Future
10:30 AM to 12:00 PM  C2: Saving the Best to Last: Long Term Content Protection and Archiving
12:00 AM to 1:00 PM  Lunch and Exhibits
1:00 to 2:30 PM  D2: The Sky’s the Limit: Opportunities and Challenges for Consumer and Enterprise Cloud Storage
2:30 to 3:00 PM  Keynote 4 Shawn DuBravac, Consumer Electronics Association
3:00 to 4:30 PM  E2: New Applications, New Technologies—What’s Ahead for Content Storage?
4:30 to 5:00 PM  Afternoon Break
5:00 to 6:30 PM  F2: Storage for Consumers: At Home and in the Car
6:30 PM  Conference Ends
HDDs
SAS versus Kinetic Open Storage

SAS
- Standard form factor
- 2 SAS ports
- SCSI command set
  - data = read (LBA, count)
  - write (LBA, count, data)
  - LBA :: [0, max]
  - data :: count * 512 bytes
  - CRC on cmd and PI on block

Kinetic Open Storage
- Standard form factor
- 2 Ethernet ports (same connector)
- Kinetic key/value API
  - value = get (key)
  - put (key, value)
  - delete (key)
  - key :: 1 byte to 4 KiB
  - value :: 0 bytes to 1 MiB
  - HMAC on cmd and SHA on value

© Coughlin Associates, 2015
## SMR Types (HGST)

<table>
<thead>
<tr>
<th>SMR category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive managed (Autonomous)</td>
<td>No host changes. SMR device manages all requests. <strong>Performance is unpredictable in some workloads.</strong> Backward compatible</td>
</tr>
<tr>
<td>Host aware</td>
<td>Host uses new commands &amp; information to optimize write behavior. If host sends sub-optimal requests the SMR device accepts the request but performance may become unpredictable. Backward compatible</td>
</tr>
<tr>
<td>Host Managed</td>
<td>Host uses new commands &amp; information to optimize write behavior. <strong>Performance is predictable.</strong> If host sends sub-optimal requests the SMR device rejects the request. <strong>Not backward compatible</strong></td>
</tr>
</tbody>
</table>

**T10/T13 ZBC/ZAC**

**ZBC = Zoned Block Commands**

**ZAC = Zoned ATA Commands**

© Coughlin Associates, 2015
Zoned Block Device Drive Types

• Drive Managed
  – Drive autonomously hides all SMR issues
  – Backward compatible

• 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
Issues and Challenges in the Data Recovery Lab (Drive Savers)

- Helium HDDs required new lab processes
  - Helium reflow process was developed for reinjection after rebuild
  - 7 platter design required new tools for headstack removal with much less space to maneuver
  - Testing expensive with HDD price premiums
  - All their 3.5” will be He moving forward

- SMR HDDs will require new methods
  - Just publicly shipping SMR labeled 8TB Seagate
  - Added complexity with translation layer and garbage collection
  - Preliminary testing and recoverability in lab is good

- SMR + He HDDs on the horizon
  - HGST 10TB coming soon

- Firmware
  - Reverse engineering and manipulations of failed modules
  - Will require assistance from OEMs
Issues and Challenges in the Data Recovery Lab 2

• Time to image increases
  • Media damage or others issues can slow this process significantly
  • 6TB He HGST SATA takes ~9 hours to image without errors
  • May be more cases for data extraction from source device without image

• Time to rebuild RAIDs increases
  • Secondary failure during rebuild can be complicated
  • RAIDs are typically recovered in degraded state or virtualized when possible

• Amount of data to recover increases
  • File counts in same cases at 10+ million per volume
  • More VMs to recover
  • More time needed to backup
Flash Memory & Solid State Storage
Implications for SSDs

• High Speed SSDs for Hot Data or Metadata
  – PCIe Gen 3 / NVMe Host Interfaces
  – Example: Samsung SM1715

• Low Cost Read-Intensive SSDs
  – Enterprise TLC (enabled by V-NAND)
  – Read-Intensive SSDs (<= 1 DWPD)
  – Example: Samsung PM853T

© Coughlin Associates, 2015
Small and Flexible Storage (Marvel)

- M.2 is a small form factor card
  - 22 mm wide & from 30 mm to 110 mm in length
  - Supports a variety of functions/interfaces, including WiFi, USB, PCIe, SATA

- A SATA M.2 will typically be an SSD
  - Will see use in enterprise as cache and boot drive
Memory Channel Storage Technology (Diablo)

- Data remains local to the processor, application, and system memory
- Distance and contention issues are eliminated
Direct Connection Between Driver and Device

Legacy SSD Driver → I/O Interface → Very Slow

NVMe Driver → Memory Space → Faster. But Still Has Extra Data Hop

MCS DRIVER

Fastest. Direct Connection. No Extra Hops.
‘3D V-NAND created the fastest ATA 600 SSD’

‘Vertical NAND opens up a whole new world when we look at SSD endurance, density, battery life for portables, and last but not least, SSD performance.’

‘3D V-NAND ensures better performance and durability’

© Coughlin Associates, 2015
Comparison of Memory Technologies

<table>
<thead>
<tr>
<th>Normalized Log Scale</th>
<th>$10^0$</th>
<th>$10^1$</th>
<th>$10^2$</th>
<th>$10^3$</th>
<th>$10^4$</th>
<th>$10^5$</th>
<th>$10^6$</th>
<th>$10^7$</th>
<th>$10^8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>SRAM</td>
<td>DRAM</td>
<td>NAND</td>
<td>DISK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MB = 1unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power/GB</td>
<td>NAND</td>
<td>DRAM</td>
<td>SRAM</td>
<td>DISK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10mW = 1unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost/GB</td>
<td>DISK</td>
<td>NAND</td>
<td>DRAM</td>
<td>SRAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10c = 1unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latency</td>
<td>SRAM</td>
<td>DRAM</td>
<td>NAND</td>
<td>DISK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ns = 1unit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next Gen Memory must fill the DRAM & Flash latency gap.
Magnetic Tape
Exploring Object Storage

What is Object Storage

• An emerging alternative to file based systems; ideal for storing large volumes of unstructured data.
• Decuples data from its physical medium or location
• Employs the inclusion of Meta Data, and Universal ID
• Its flat & infinite namespace make possible for large scale storage
• Provides a foundation for data longevity techniques

How do you talk to Object Storage

• RESTful API
• Client server model
• Gateways/Appliances

© Coughlin Associates, 2015
How Does Deep Storage Work?

Applications w/DS3 Client MAM/PAM/EDIT Workflow MGR (Create/Capture/ Edit)

DS3 API

Objects Created

Inside DS3 Server:
- Full abstraction of storage
- Managing all aspects
- SSDs for object cataloging
- HDDs for Cache/buffer

© Coughlin Associates, 2015
Archive eXchange Format (AXF)

- AXF is a universal standard for the wrapping (encapsulation), storage, transport and preservation of any type of file assets

- AXF is like an advanced ZIP which encapsulates any number of files of any type, metadata and a universal file system – an “object store”

- AXF is IT-centric and applies to all types and generations of storage technologies

- The first open standard targeting the storage, preservation and transport of file assets

- AXF fully defines (and constrains) implementations for increased interoperability
Digital equivalent to a “time capsule”
Provides for Original Content Backup
  – Which begins immediately
  – On the Set / In the Field / In the Lab
Yields Immediate & Local Preservation
  – On Original Mediums & Alternative Media
Content Cataloging w/integrated metadata
  – Long Term Asset Management, Repurposing, versioning
Needs to Maintain “Life Cycle” Continuity

© Coughlin Associates, 2015
Cloud Storage
Extended Online (Quantum) Defined...

• Seamlessly extend storage and workflow operations with unique Object Storage
  - Extends online storage to non real-time workflow operations
  - Eliminates performance impact on real-time operations – edit, etc.
  - Provides access to transcode, delivery, archive, ingest
  - Allows content and assets to be seamlessly and quickly assessed
  - Provides higher resiliency than RAID with lower latency than tape
  - Extends online for global access and geo-spread (replication)
  - Nearly unlimited scalability self-healing and self-protecting
  - No migration required – maintains content for long term access
  - Can be used in combination with on–premise and cloud based workflows
A Cloud Experience, Everywhere (Zadara)

At Service Providers

Storage

Replication

In Colocation

Storage

On Premises

Replication

Replication

Replication

Storage

Storage

Storage

© Coughlin Associates, 2015
Collaborate, Distribute and Federate (DDN)

Overview

• Single, federated global namespace
• Multi-site - ingest, production, distribution, collaboration, & archive
• Access through production or directly from object storage
• Availability and protection controlled by policy
• DR for Free

© Coughlin Associates, 2015
World Cup “Near Live” solution

- Multi-screen capture by EVS
- Scale Out Transcoding by Elemental
- High-Speed Transfer by Aspera
- FASP

10 Mbps x24x2, 200ms / 2%
480 Mbps total ingest

© Coughlin Associates, 2015
FaceBook
Storage in Facebook Infrastructure

Front-End Cluster
- Web
- Newsfeed
- Ads
- Cache

Flash Backend Cluster
- User DB
- Timeline DB
- Ads DB

Service Cluster
- Search
- Instagram
- Messaging

HDD Storage Cluster
- Photo/Video
- HDFS
- Cold Storage

© Coughlin Associates, 2015
Longer Useful Life

Today servers are kept in production for 3 years.

With disaggregated rack:
Compute – 3 years for power efficiency
RAM sled – 5 years or more
Disk sled – 4 to 5 years depending on usage
Flash sled – 6 years depending on write volume
Optical Archive Racks for Cold Storage

– Optical has longer life expectancy: 10-15 years
– 1PB rack density now, 2~4PB in a few years
– Making Cold Storage agonistic to media types
Conclusions

• The 2015 Storage Visions theme is “Storage With Intense Network Growth” and we had a variety of sessions exploring growth drivers and opportunities
• All storage technologies developing—flash big enabler
• Clouds are becoming an important part of mainstream storage
• Companies like Facebook are leading the way
• Major changes in memory, storage, compression and applications will enable new market opportunities…
Mark Your Calendar Now!

June 30, 2015

At the Doubletree, Culver City, CA

Creative Storage™

2015 Conference

Call for Papers and Sponsors

Now Open

CreativeStorage.org

An Entertainment Storage Alliance™ Event
Mark Your Calendar Now!

January 4 & 5, 2016

Storage Visions® 2016

StorageVisions.com

AN ENTERTAINMENT STORAGE ALLIANCE™ EVENT
Thanks