Rethinking RAID

Dwain Sims
dsims@bayleafnc.org

Secure Computing with Apache Struts

Dwain Sims
dsims@bayleafnc.org

Who is this guy?

MS Computer Science, West Virginia University

16 Years in Silicon Valley

- Lockheed
- Sun Microsystems

12 Years in Linux High Availability

5 Years in Flash Storage

- Fusion-io
- SanDisk
- Western Digital

Inspiration

Storage is going through a Revolution

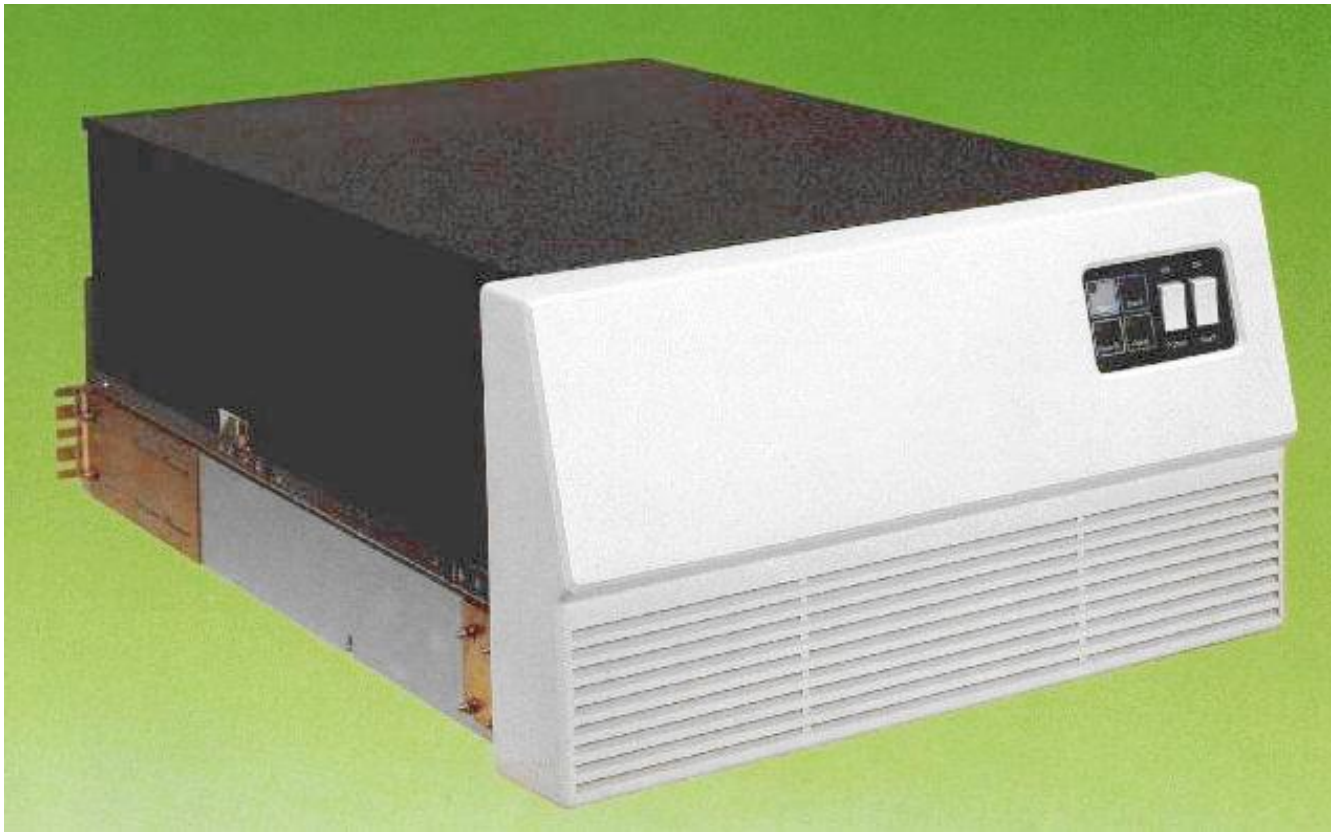
Inspiration

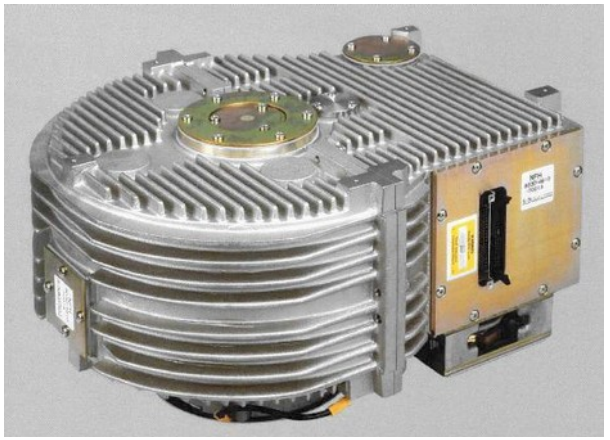
Old Habits Die Hard

Quick History Lesson

5 MB
$3200/Month
1956

Fujitsu Eagle





470 MB, $10K, 600W

RAID now enters, stage left……

This is where the whole idea about RAID got started.

Shugart (Seagate) ST-506

5 MB
$1500
1980

HGST “King Cobra” C15K600

$670, 600GB, 7.5W

HGST Ultrastar He^{12}

$670 12TB, 9.8W

What is this RAID stuff anyway?

Quick RAID History

UC Berkley

Also the home of vi, csh, UNIX TCP/IP, BSD UNIX and Bill Joy!

David Patterson, Garth Gibson, and Randy Katz

Mid-80s

Redundant Array of Inexpensive Disks

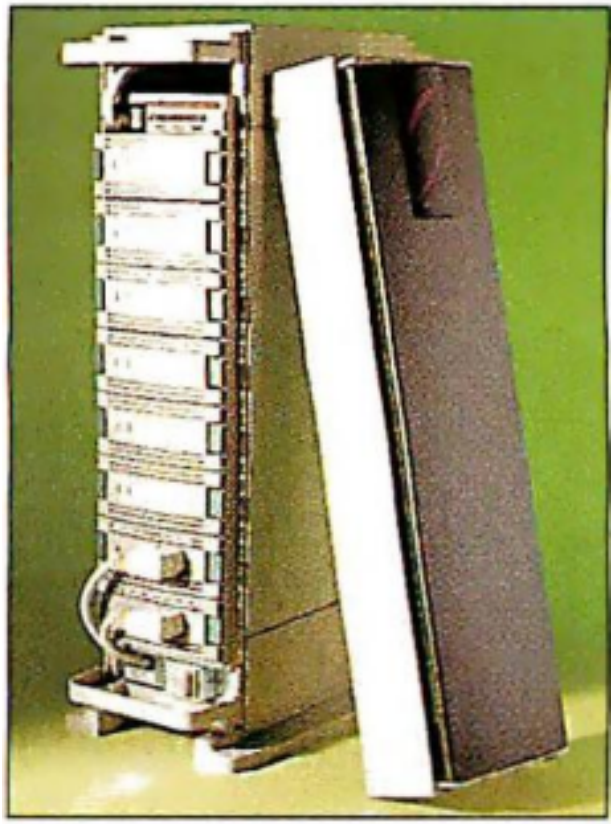
Now “Independent” Disks

IBM can also claim invention of RAID

Norman Ken Ouchi – RAID 4

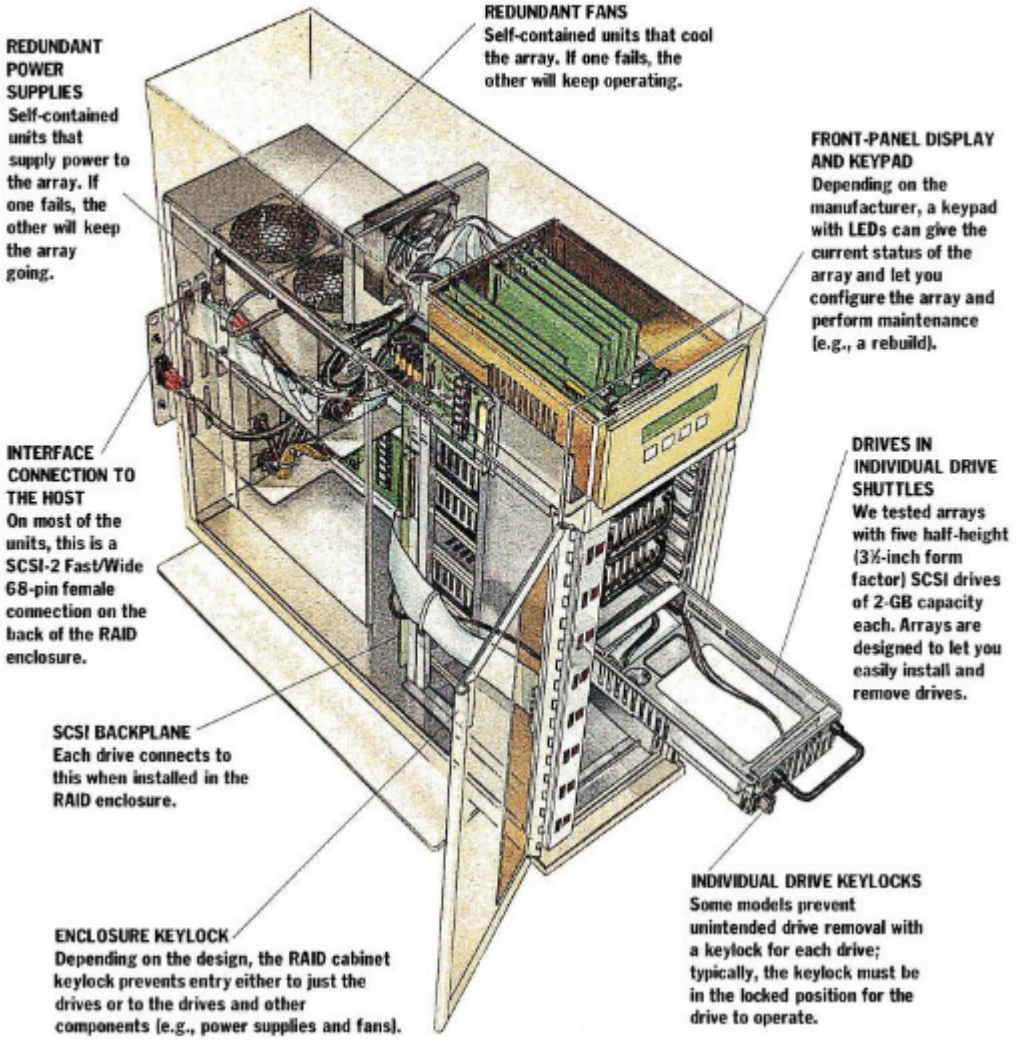
Clark, et al. - Patent on RAID 5 (1986)

Early RAID Systems



Digital StorageWorks RAID Array 230 Subsystem

A Pillar of Reliability



RAID Terminology

RAID-0

Striping; Super Important and widely used. **No Redundancy!**

RAID-1

Mirroring; Super important and widely used.

RAID-10

A stripe of mirrors. Super important and widely used.

N number of devices are lost capacity-wise.

RAID-2

Never Used

RAID-3 and RAID-4

Rarely used

RAID Terminology

RAID-5

Parity spread across N+1 devices; Can survive 1 device failure.

Can be implemented in both Hardware and Software

Single device capacity is lost

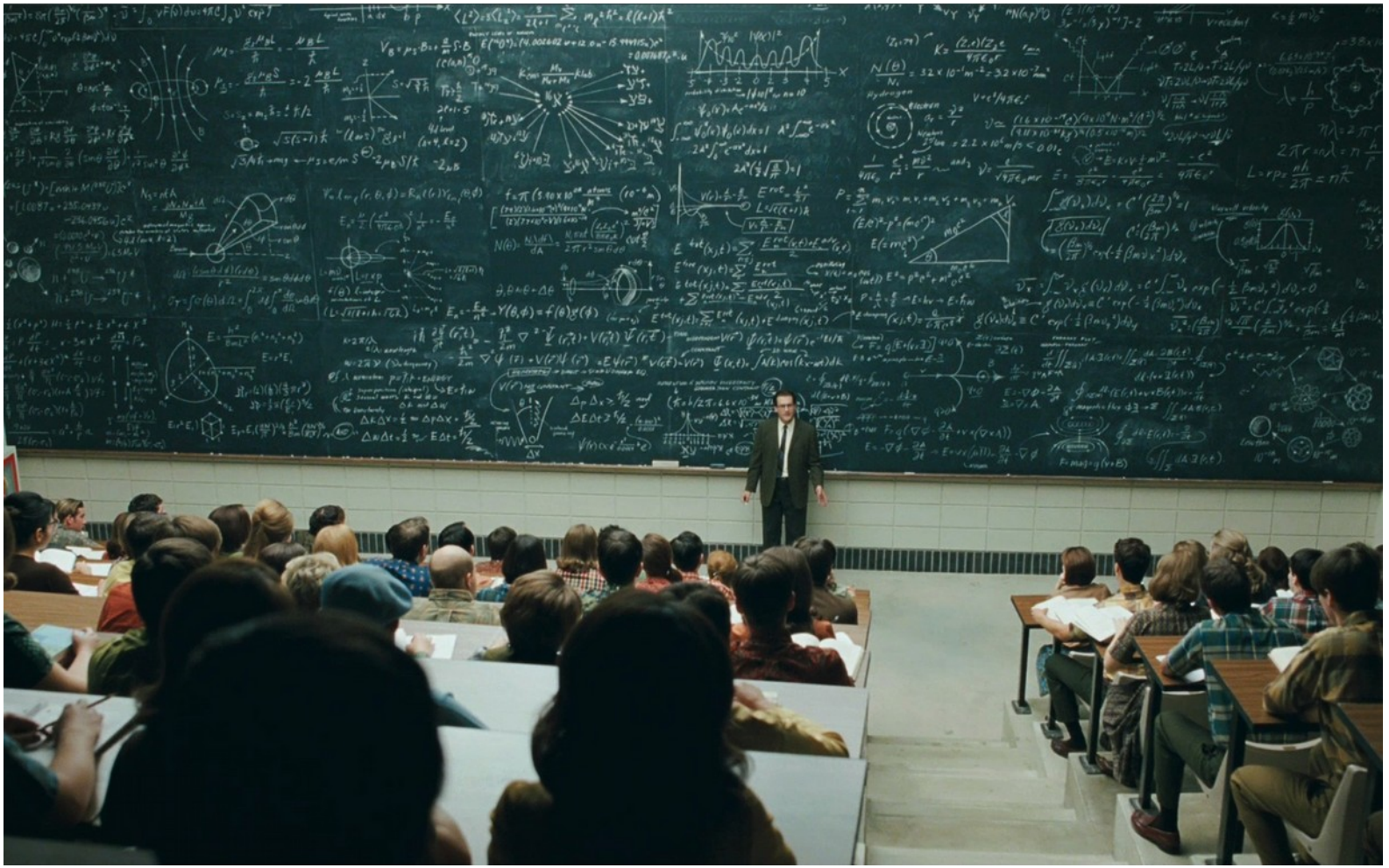
RAID-6

Parity spread across N+2 devices; Can survive 2 device failures.

Can be implemented in both Hardware and Software

Two device capacity is lost

So what is the problem?



Device failure means RAID Rebuild!

Not Really a big deal with sub-TB hard drives

We will see that data shortly

Became more Dangerous and Painful at 1TB

Solution – RAID 6! (well sorta..)

However, with 10TB devices (and beyond)...

Monster Problem!

As we shall see....

Methodology

Common Servers

Lenovo Broadwell based (Lenovo x3650 M5, 2U, 2 Socket)

CentOS 7.3 (.514 kernel)

Avago (LSI) RAID Adapter "Flatwoods" (mostly)

RAID-5 Array

5 Devices in RAID 5, with a hot spare (in most cases)

(and couple of interesting Software RAID Scenarios)

Common Load

Flexible I/O Tester "fio"

60/40 Random Read/Write

Queue Depth = 32 per job (20 jobs)

Methodology

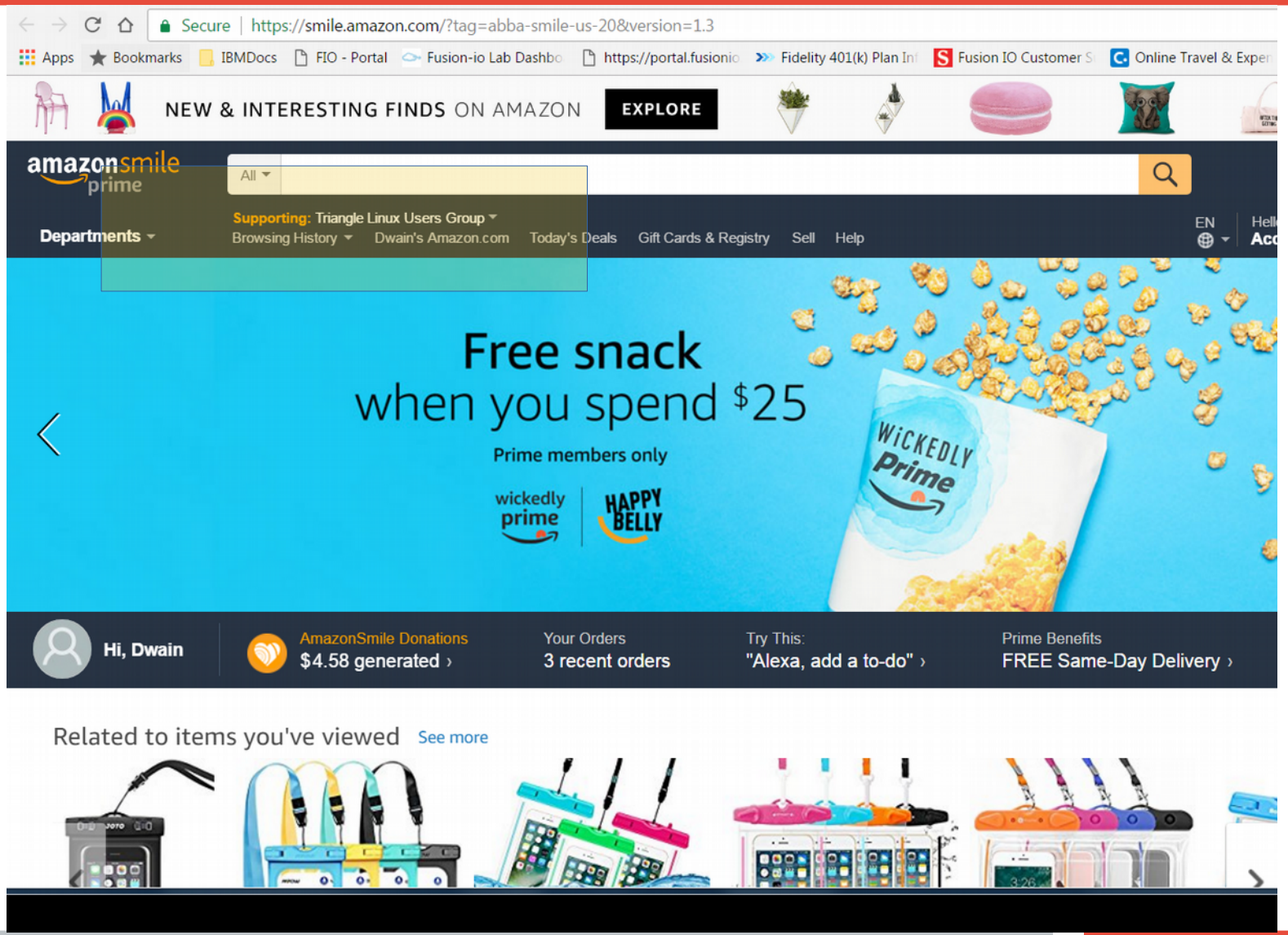
Measuring

- IOPS with No Load
- IOPS under Load
- RAID Rebuild time with No Load
- RAID Rebuild time under Load

And Now a Word from Our Sponser

YOU!

Easy Way to Sponser



Collected Data

RAID 5 Rebuild Times

Drive	RAID Array Size	Rebuild time Idle (hours)	Rebuild Time under Load (hours)	Normal Read IOPS	Normal Write IOPS	Rebuild Read IOPS	Rebuild Write IOPS
500GB 7200 6G SAS	2TB	1.5	134	265	170	170	125
HGST King Cobra F 15K 300G 12G SAS	1.2TB	0.7	54	564	375	434	284
HGST Cobra F 10K 600GB 12G SAS	2.4TB	1.5	58	514	343	350	217
HGST 10TB 12G SAS (Libra He10)	40TB	77	4200 (extrapolated)	313	209	208	127
CloudSpeed II 1.92TB SATA	7.7TB	2	18	33.7K	22.5K	12.8K	8.6K
Optimus II Max 3.84TB 6G SAS	15.4TB	5.5	14.5	29.4K	19.6K	18.4K	12.2K
Optimus II Ascend 800GB 6G SAS	3.2TB	0.5	6	33.7K	22.5K	15.8K	10.8K
Bear Cove 10DWPD 800G 12G SAS R100 (14W)	3.2TB	0.5	6	33.4K	22.3K	16.7K	11.2K
Bear Cove 10DWPD 800G 12G SAS R100 (9W)	3.2TB	0.5	6	32.9K	21.1K	16.8K	11.3K
Fusion ioMemory SX350 3.2TB PCIe	12.8TB	5	122	49.6K	33.5K	16.8K	12K
Fusion ioMemory SX350 3.2TB PCIe (Thread=32)	12.8TB	1	25	182K	121K	144K	95.7K
HGST SN-150 1.6TB NVMe	6.4TB	1	83	134.7K	89.8K	44.4K	28.5K
HGST SN-150 1.6TB NVMe (Threaded=16)	6.4TB	0.5	4	164K	109K	125K	81.9K
Fusion ioMemory SX350 3.2TB PCIe	12.8TB			296K	197K		
Fusion ioMemory SX350 3.2TB PCIe	16TB			330K	220K		
Fusion ioMemory SX350 3.2TB PCIe	3.2TB			154K	103K		

Consequences!

RAID-5(6) Rebuild times on current “Capacity” (10,12 TB) drives are enormous!

- **4200 Hours ≈ 5 ½ Months**
- **Staggering!!**

Devices are stressed even more during rebuild

- Increased chance of additional device(s) failing

Relatively slow devices now run even slower!

Is there Better Way?

Absolutely!

Application Redundancy

Let your application take care of Redundancy

- **MySQL Master-Slave Replication**
- **Oracle Data Guard**
- **Microsoft SQLserver AlwaysOn Application Cluster**
- **SAP Hana**
- **Hadoop (in the base architecure)**
- **OpenStack and Ceph**

Not only protects against storage failure, but system failure as well

Erasure Coding

RAID-6 is a primitive Erasure Code

Tahoe-LAFS

Ceph – Block and Object

Hadoop

Swift – and other Object Storage Solutions

HGST ActiveScale – S3

API (ie Reed-Solomon, OpenRQ)

Software Defined Storage

Ceph

Swift

SUSE Enterprise Storage

VMware VSAN

Microsoft Storage Spaces Direct

DataCore

Nexenta

Nutanix

(and a score of others)

Remember the Revolution….

Flash Storage

- UBER
  - Typically an order of magnitude (or two!) better than spinners
- No Moving Parts
- Built-in Resiliency

Tools

Fio

- The Flexible I/O Tester
- Small learning curve yields great results
- Very script-able

Tips

- Remember to “Pre-Condition” (especially Flash devices)
- Watch your Queue Depth
- Use the right “io engine”
- **Beware – power tools can injure!**

Fio sample script

```
[global]
readwrite=write
rwmixread=0
blocksize=4M
ioengine=libaio
thread=0
size=100%
iodepth=16
group_reporting=1
description=fio PRECONDITION sequential 4M complete write

[/dev/sda]
filename=/dev/sda
cpus_allowed=0-19
```

More Tools

MegaRAID Storage Manager

Linux md RAID tools

```
cat /proc/mdstat
mdadm –misc –detail /dev/mdYYY
dmesg -H -w
```

Take Time to Tune your md Array

Threads

$ sudo echo 16 > /sys/block/md0/md/group_thread_cnt

Speed Limits

dev.raid.speed_limit_max = xxyyzz

Defaults to dev.raid.speed_limit_max = 200000

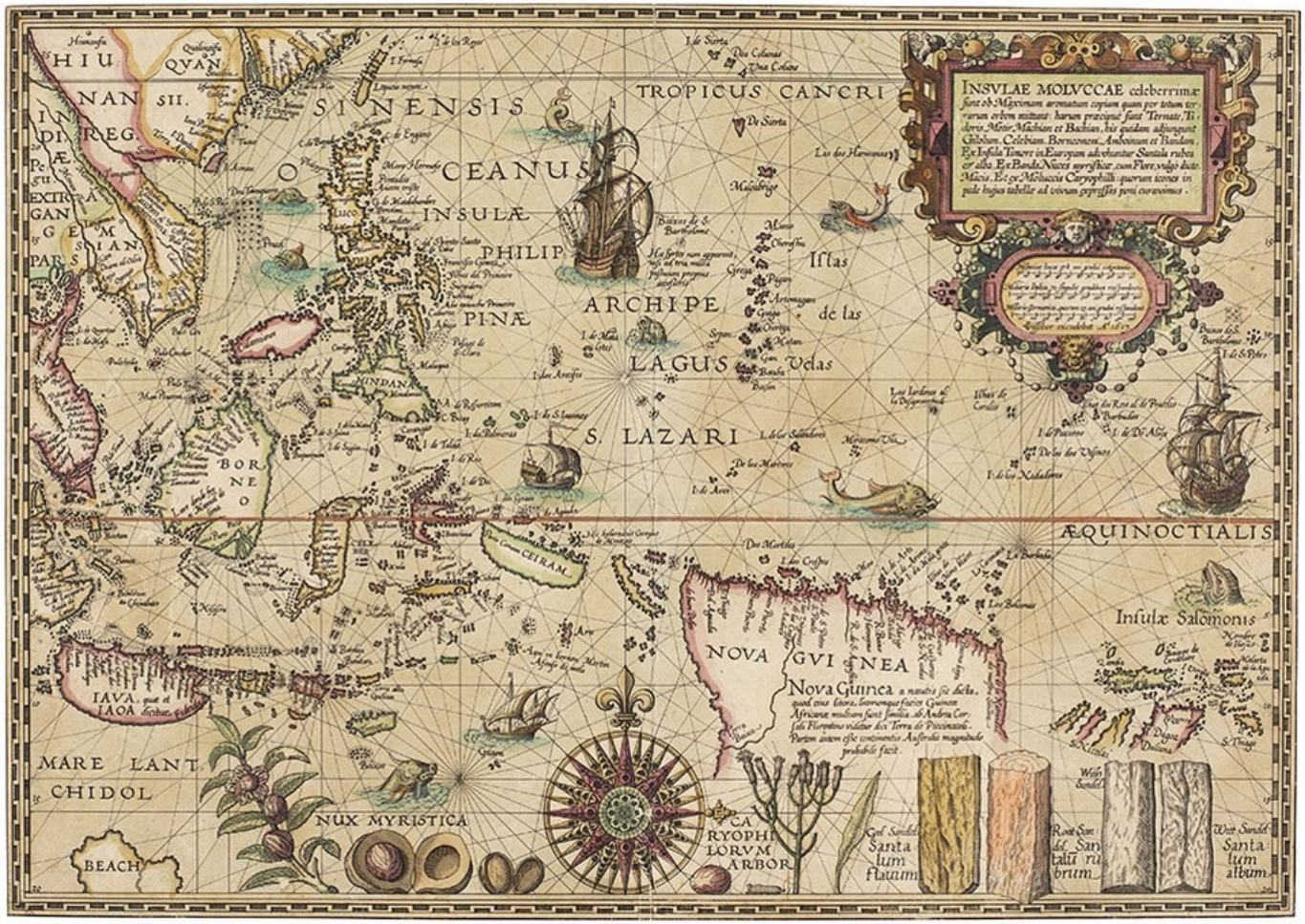
Things to Remember

- **RAID 0 and 1 (and 10) are still very viable**
  - Maybe not so much with RAID 10….
- **RAID 5 and 6 are still OK for Flash Devices**
  - Understand your Limitations!
  - The RAID Adapter will be your limiting factor
- **RAID 6 is likely OK for sub-TB Spinning Disk**
  - As long as you can get them!
- **RAID Hardware varies widely in performance!**
- **Capacity Hard Drives Require a different Data Resiliency Technique**
- **Using md Software RAID? Do not forget to tune!**

Maybe some concern with RAID 10...

RAID 5 Rebuild Times									
Drive	Server	RAID Adapter	RAID Array Size	Rebuild time Idle (hours)	Rebuild Time under Load (hours)	Normal Read IOPS	Normal Write IOPS	Rebuild Read IOPS	Rebuild Write IOPS
HGST 10TB 12G SAS (Libra He10)	x3650 M5	LSI Avago M5210 RAID 10 4x2	40TB	16	1344	607	405	479	315

Where next?


TROPICUS CANCRI
SINENSIS
OCEANUS
INSULÆ
PHILIP
PINÆ
ARCHIPE
LAGUS
S. LAZARI
Islas
de las
Velas
ÆQUINOCTIALIS
INSVLAE MOLVCCAE celeberrimæ
BORNEO
IAVA
MARE
CHIDOL
NUX MYRISTICA
NOVA GVINEA
Nova Guinea
Insulæ Salomonis
Santalum album

Resources

https://archive.org/details/byte-magazine

(Sept 1995, page 248)

https://www.youtube.com/watch?v=V-WbdMPiM1A

Fujitsu Eagle Spinup!

http://queue.acm.org/detail.cfm?id=1670144

Triple-Parity RAID and Beyond (Adam Leventhal, Sun)

https://github.com/axboe/fio

Flexible I/O Tester (fio) (Jens Axboe)

https://en.wikipedia.org/wiki/RAID

https://raid.wiki.kernel.org/index.php/RAID_setup

Excellent md RAID tutorial

Thanks!!!

Dwain Sims

dsims@bayleafnc.org

Google Voice: 919-480-1774

Collected Data

RAID 5 Rebuild Times

Drive	Server	RAID Adapter	RAID Array Size	Rebuild time Idle (hours)	Rebuild Time under Load (hours)	Normal Read IOPS	Normal Write IOPS	Rebuild Read IOPS	Rebuild Write IOPS
500GB 7200 6G SAS	x3500 M2	MR10i	2TB	1.5	134	265	170	170	125
HGST King Cobra F 15K 300G 12G SAS	x3650 M5	LSI Avago M5210	1.2TB	0.7	54	564	375	434	284
HGST Cobra F 10K 600GB 12G SAS	x3650 M5	LSI Avago M5210	2.4TB	1.5	58	514	343	350	217
HGST 10TB 12G SAS (Libra He10)	x3650 M5	LSI Avago M5210	40TB	77	4200 (extrapolated)	313	209	208	127
HGST 10TB 12G SAS (Libra He10)	x3650 M5	LSI Avago M5210 RAID 10 4x2	40TB	16	1344	607	405	479	315
CloudSpeed II 1.92TB SATA	x3650 M5	LSI Avago M5210	7.7TB	2	18	33.7K	22.5K	12.8K	8.6K
Optimus II Max 3.84TB 6G SAS	x3650 M5	LSI Avago M5210	15.4TB	5.5	14.5	29.4K	19.6K	18.4K	12.2K
Optimus II Ascend 800GB 6G SAS	x3650 M5	LSI Avago M5210	3.2TB	0.5	6	33.7K	22.5K	15.8K	10.8K
Bear Cove 10DWPD 800G 12G SAS R100 (14W)	x3650 M5	LSI Avago M5210	3.2TB	0.5	6	33.4K	22.3K	16.7K	11.2K
Bear Cove 10DWPD 800G 12G SAS R100 (9W)	x3650 M5	LSI Avago M5210	3.2TB	0.5	6	32.9K	21.1K	16.8K	11.3K
Fusion ioMemory SX350 3.2TB PCIe	x3650 M5	Linux MD RAID	12.8TB	5	122	49.6K	33.5K	16.8K	12K
Fusion ioMemory SX350 3.2TB PCIe (Thread=32)	x3650 M5	Linux MD RAID	12.8TB	1	25	182K	121K	144K	95.7K
HGST SN-150 1.6TB NVMe		Linux MD RAID	6.4TB	1	83	134.7K	89.8K	44.4K	28.5K
HGST SN-150 1.6TB NVMe (Threaded=16)		Linux MD RAID	6.4TB	0.5	4	164K	109K	125K	81.9K
Fusion ioMemory SX350 3.2TB PCIe	x3650 M5	Linux MD RAID0 x4	12.8TB			296K	197K		
Fusion ioMemory SX350 3.2TB PCIe	x3650 M5	Linux MD RAID0 x5	16TB			330K	220K		
Fusion ioMemory SX350 3.2TB PCIe	x3650 M5	No RAID	3.2TB			154K	103K		