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 He¹²

$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
Early RAID Systems

Digital StorageWorks RAID Array 230 Subsystem
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 Sponsor
YOU!
Easy Way to Sponsor

Free snack when you spend $25
Prime members only

Related to items you've viewed
## Collected Data

### RAID 5 Rebuild Times

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

| 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 a 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 architecture)
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)
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

<table>
<thead>
<tr>
<th>Drive</th>
<th>Server</th>
<th>RAID Adapter</th>
<th>RAID Array Size</th>
<th>Rebuild time Idle (hours)</th>
<th>Rebuild Time under Load (hours)</th>
<th>Normal Read IOPS</th>
<th>Normal Write IOPS</th>
<th>Rebuild Read IOPS</th>
<th>Rebuild Write IOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGST 10TB 12G SAS (Libra He10)</td>
<td>x3650 M5</td>
<td>LSI Avago M5210 RAID 10 4x2</td>
<td>40TB</td>
<td>16</td>
<td>1344</td>
<td>607</td>
<td>405</td>
<td>479</td>
<td>315</td>
</tr>
</tbody>
</table>
Where next?
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

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