Going Mad with MDADM

The joy and pain of using Software Raid

System Administrators Mini Conf
Linux.Conf.AU 2010
Wellington New Zealand

Steven Ellis

Technical Director OpenMedia Limited,
Director Global Engineering Bulletin.net
Raid Types

- Hardware Raid Controllers
  - 3ware
  - Adaptec
  - LSI Logic

- Hardware / Bios assisted “fakeraid” – dmraid
  - Intel
  - Highpoint
  - LSI Logic
  - NVidia
  - Promise
  - Silicon Image

- Linux Software Raid - mdadm
Why MDADM?

- Low Cost Solution
  - Any type of HD
  - Any Controller

- Portable
  - Not tied to a particular HW Controller

- Performance
  - Raid 1 has adequate performance on a modern CPU
The need for Storage

Why can't I just drag & drop my C: drive onto the office server?
The Dream
The Budget
Cold Hard Reality

- SMB Office Server
  - Consumer grade motherboard
    - VIA KT400, Athlon XP CPU
  - On-board IDE – Full
  - SIL 680 PCI IDE Card – 4 disk
    Raid 5 set
  - No SATA Ports
  - One spare PCI Slot
  - No PCIe Slots
The Upgrade

- **Hardware**
  - SIL 3114 4 Port PCI SATA Card
  - 2 x 1TB WD WD10EACS

- **Test Environment**
  - Different motherboard chipset
  - Different CPU
  - Different OS (Ubuntu vs RHEL5)
Set-up the Array

- Smart Test the HDs
  - `smartctl -t long /dev/sda`
  - `smartctl -t long /dev/sdb`

- Create a single partition with type fd

- Build raid set with mdadm
  - `mdadm --create /dev/md3 --level=1 \ --raid-devices=2 /dev/sd[ab]`

- Migrate some of the production data

- Stress test
Go Live

- Move the disks to production server
- Confirm no issues with filesystem on new raid set
- Complete data migration
- Assign new volumes for production use.
  - Retain old volumes for the next couple of weeks
Data Corruption!
Troubleshooting

- The obvious
- Check Filesystem(s)
- Smart Check the Hard Drives
- Avoid production impact
  - Move back to test environment
- Can't reproduce the problem
- Back to Production
Data Corruption!!
Create a large test file and checksum test

```bash
dd if=/dev/urandom of=testfile bs=1M count=2048
md5sum testfile;
  628e063d881169bd75d4d59517067689  testfile
md5sum testfile;
  ef9bad771d7e50cf8a67b0016867ff2b  testfile
```

Check the raid set

```bash
cat /proc/mdstat
  Personalities : [raid1] [raid6] [raid5] [raid4]
md3 : active raid1 sdb1[0] sda1[1]
    976759936 blocks [2/2] [UU]
```
MDADM Checks

- Force a check on the raid array
  
  `echo check > /sys/devices/virtual/block/md3/md/sync_action`

- This might take some time
  
  `cat /proc/mdstat`
  
  Personalities : [linear] [multipath] [raid0] [raid1] [raid6] [raid5] [raid4] [raid10]
  
  md3 : active raid1 sda1[0] sdb1[1]
  
  976759936 blocks [2/2] [UU]
  
  [>....................] check = 0.0%
  
  (487104/976759936) finish=200.3min speed=81184K/sec
Fix the errors

- High mismatch count
  
  ```
  cat /sys/devices/virtual/block/md3/md/mismatch_cnt
  311424
  ```

- Try to fix it
  
  ```
  echo repair > /sys/devices/virtual/block/md3/md/sync_action
  ```

- Wait
- Wait
- Wait some more
- Test the file system
Data Corruption!!!
Reduce the Problem

<table>
<thead>
<tr>
<th>Filesver</th>
<th>Webserver</th>
<th>Wiki</th>
<th>Accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RHEL5 Xen Dom0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LVM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MDADM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Via KT 400 Motherboard

<table>
<thead>
<tr>
<th>Onboard IDE</th>
<th>Sil 680 PCI IDE</th>
<th>Sil 3114 PCI SATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40Gb 40Gb CDROM 120Gb</td>
<td>250Gb 250Gb 250Gb 250Gb</td>
<td>1Tb 1Tb</td>
</tr>
<tr>
<td>md0 - boot</td>
<td>md2 - 750 GB R5</td>
<td>md3 - 1 Tb R1</td>
</tr>
</tbody>
</table>
Reduce the Problem

### RHEL 5.3

<table>
<thead>
<tr>
<th>MDADM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Via KT 400 Motherboard</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component Type</th>
<th>Storage Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onboard IDE</td>
<td>40Gb 40Gb CDROM 120Gb md0 - boot md1 - 36Gb</td>
</tr>
<tr>
<td>Sil 680 PCI IDE</td>
<td>250Gb 250Gb 250Gb 250Gb md2 - 750 GB R5</td>
</tr>
<tr>
<td>Sil 3114 PCI SATA</td>
<td>1 Tb 1 Tb md3 - 1 Tb R1</td>
</tr>
</tbody>
</table>

OpenMedia
Research

- Kernel Mailing List
- Linux Sata Drivers
- Linux Raid Mailing List
- WD Hard Drive Issues
TLER

- Time Limited Error Recovery
- Only enabled on WD Raid/Enterprise series drives.
- Can be enabled on Green Drives
- Google for WDTLER.EXE
Surprise Surprise
## Hardware conflicts?

<table>
<thead>
<tr>
<th>Fileserver</th>
<th>Websserver</th>
<th>Wiki</th>
<th>Accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHEL5 Xen Dom0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDADM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Via KT 400 Motherboard

<table>
<thead>
<tr>
<th>Onboard IDE</th>
<th>Sil 680 PCI IDE</th>
<th>Sil 3114 PCI SATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40Gb 40Gb CDROM 120Gb</td>
<td>250Gb 250Gb 250Gb 250Gb</td>
<td>1 Tb 1 Tb</td>
</tr>
<tr>
<td>md0 - boot</td>
<td>md2 - 750 GB R5</td>
<td>md3 - 1 Tb R1</td>
</tr>
<tr>
<td>md1 - 36Gb</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Test Hardware

Ubuntu 8.10

LVM

MDADM

NVidia MCP5 Motherboard

Onboard IDE

10Gb DVD boot

Onboard Sata

Sil 3114 PCI SATA

1 Tb 1 Tb

md3 - 1 Tb R1
### Alternative Motherboard

<table>
<thead>
<tr>
<th>Fileserver</th>
<th>Webservice</th>
<th>Wiki</th>
<th>Accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHEL5 Xen Dom0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDADM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Intel D945GTP Motherboard**

<table>
<thead>
<tr>
<th>Onboard IDE</th>
<th>Sil 680 PCI IDE</th>
<th>Sil 3114 PCI SATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40Gb</td>
<td>250Gb</td>
<td>1 Tb</td>
</tr>
<tr>
<td>40Gb</td>
<td>250Gb</td>
<td>1 Tb</td>
</tr>
<tr>
<td>CDROM</td>
<td>250Gb</td>
<td>md2 - 750 GB R5</td>
</tr>
<tr>
<td>120Gb</td>
<td>250Gb</td>
<td>md3 - 1 Tb R1</td>
</tr>
<tr>
<td>md0 - boot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>md1 - 36Gb</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Final Solution

<table>
<thead>
<tr>
<th>Fileserver</th>
<th>Webserver</th>
<th>Wiki</th>
<th>Accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHEL5 Xen Dom0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDADM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NVidia MCP51 Motherboard

- **Onboard IDE**
  - 40Gb
  - 40Gb
  - CDROM
  - 120Gb
  - md0 - boot
  - md1 - 36Gb

- **Sil 680 PCI IDE**
  - 250Gb
  - 250Gb
  - 250Gb
  - 250Gb
  - md2 - 750 GB R5

- **Onboard SATA**
  - 1 Tb
  - 1 Tb
  - md3 - 1 Tb R1
We Celebrate
What Did We Learn

- Some Hardware sucks
- How to troubleshoot Software Raid
- Patience
- Virtualisation Rocks
- Have a better test environment
Links and References

- **TLER Background**
  

- **Debian Thread on debugging mdadm**
  
  [http://marc.info/?l=debian-user&m=123115382721512&w=2](http://marc.info/?l=debian-user&m=123115382721512&w=2)

- **Linux Raid Page at Linux Foundation**
  
  [http://www.linuxfoundation.org/collaborate/workgroups/linux-raid](http://www.linuxfoundation.org/collaborate/workgroups/linux-raid)

- **Linux Raid Mailing List**
  