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

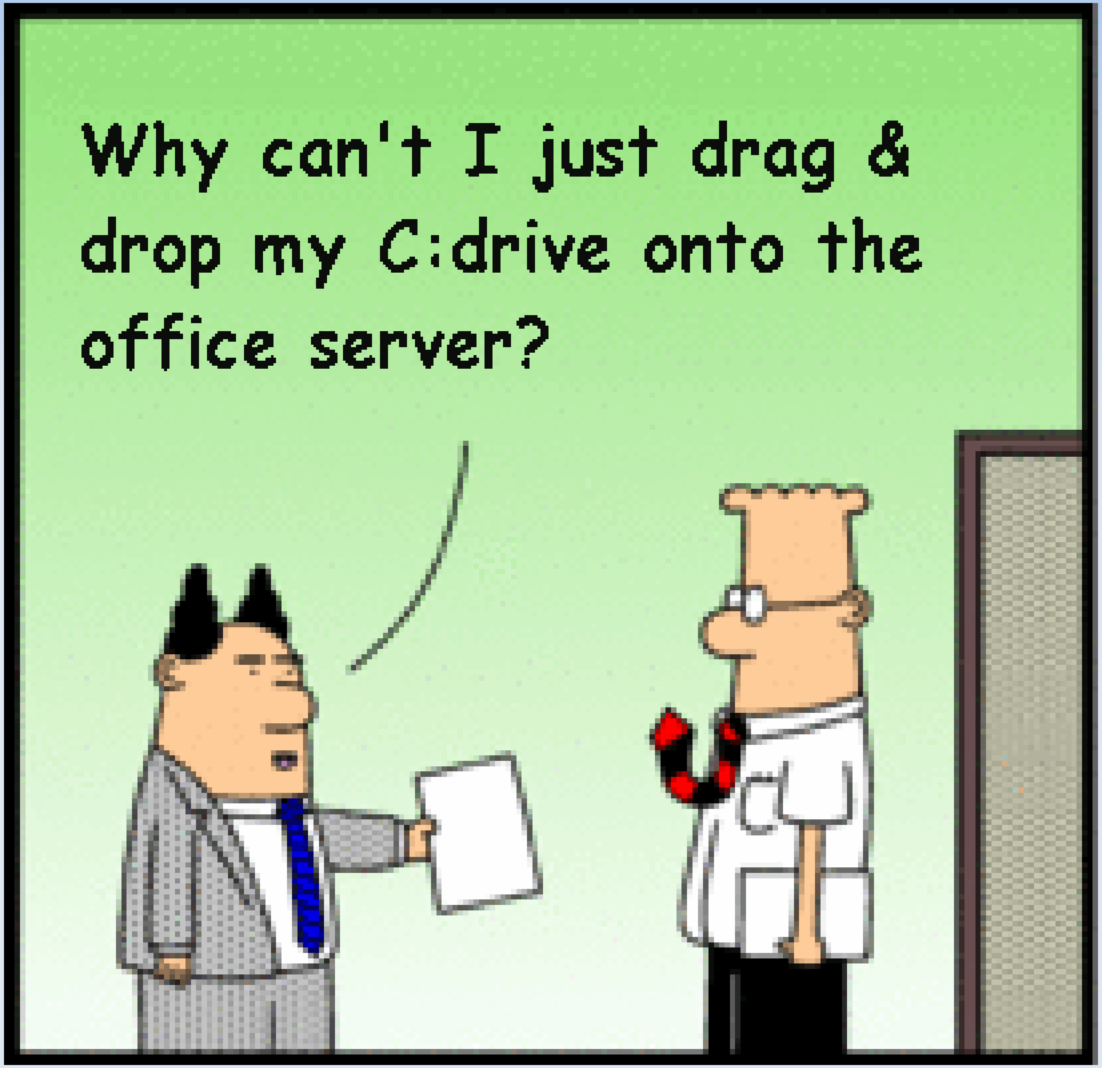
Raid Types

- 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]1
```

- 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!

OpenMedia

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!!

Analysis

- Create a large test file and checksum test

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

- Check the raid set

```
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

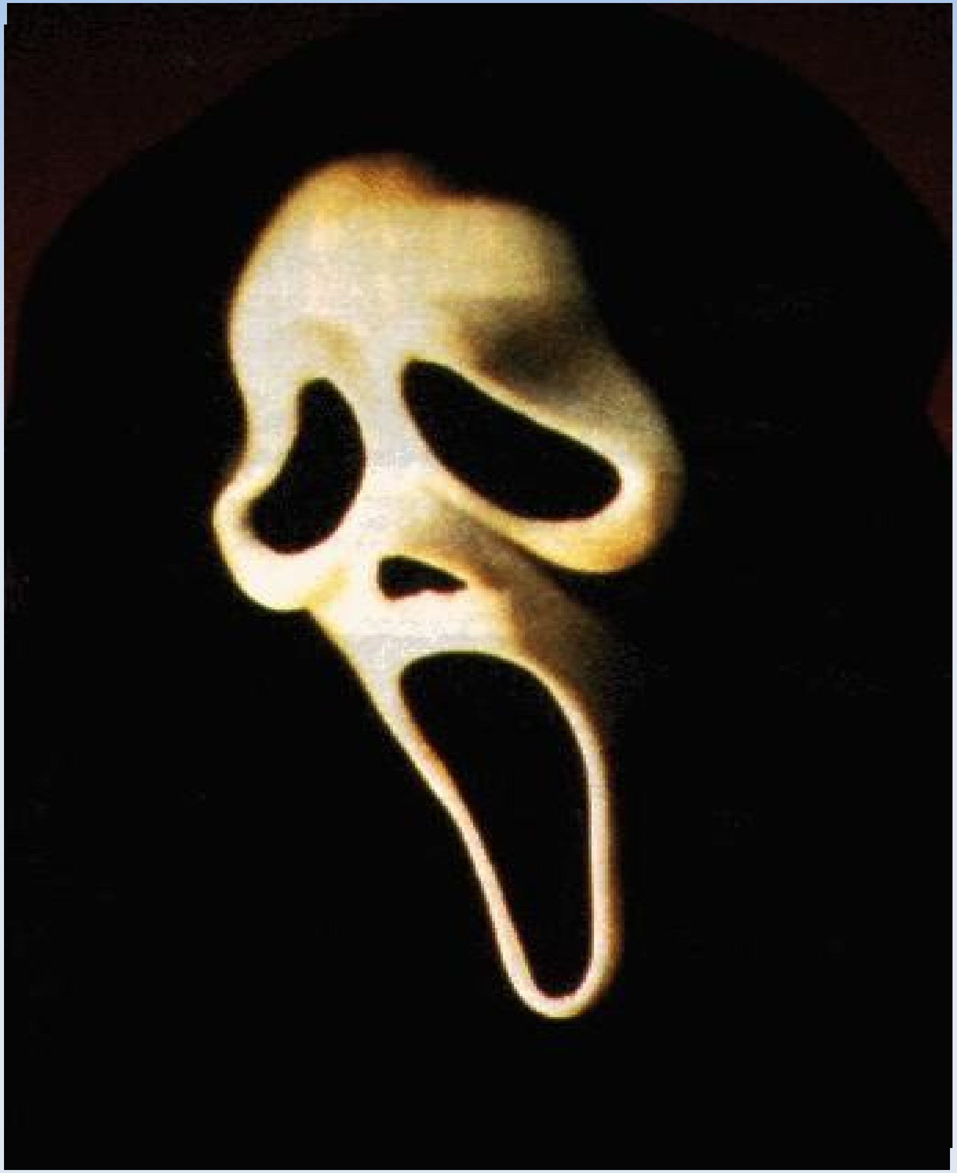
```
cat /sys/devices/virtual/block/md3/md/mismatch_cnt
  311424
```

- Try to fix it

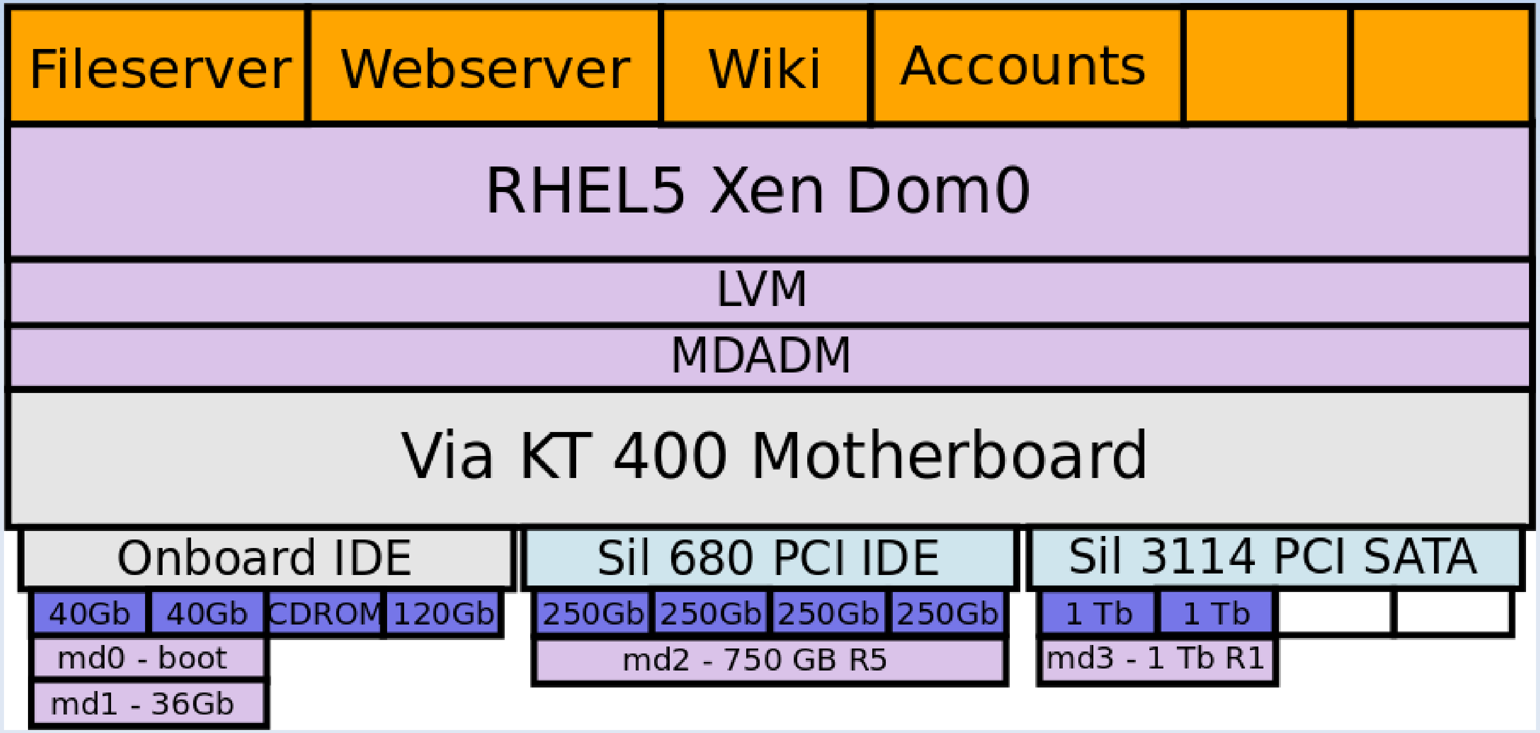
```
  echo repair >
    /sys/devices/virtual/block/md3/md/sync_action
```

- Wait
- Wait
- Wait some more
- Test the file system

Data Corruption!!!



Reduce the Problem



Fileserver

Webserver

Wiki

Accounts

RHEL5 Xen Dom0

LVM

MDADM

Via KT 400 Motherboard

Onboard IDE

40Gb

40Gb

CDROM

120Gb

md0 - boot

md1 - 36Gb

Sil 680 PCI IDE

250Gb

250Gb

250Gb

250Gb

md2 - 750 GB R5

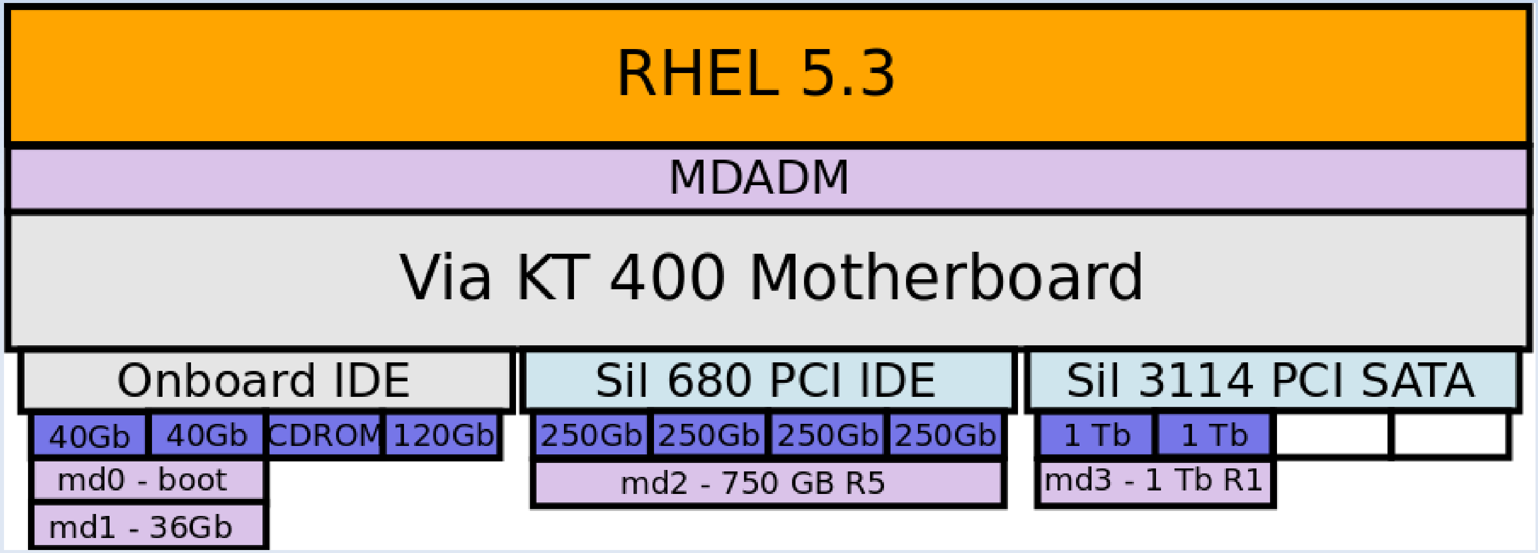
Sil 3114 PCI SATA

1 Tb

1 Tb

md3 - 1 Tb R1

Reduce the Problem



RHEL 5.3

MDADM

Via KT 400 Motherboard

Onboard IDE

40Gb 40Gb CDROM 120Gb

md0 - boot

md1 - 36Gb

Sil 680 PCI IDE

250Gb 250Gb 250Gb 250Gb

md2 - 750 GB R5

Sil 3114 PCI SATA

1 Tb 1 Tb

md3 - 1 Tb R1

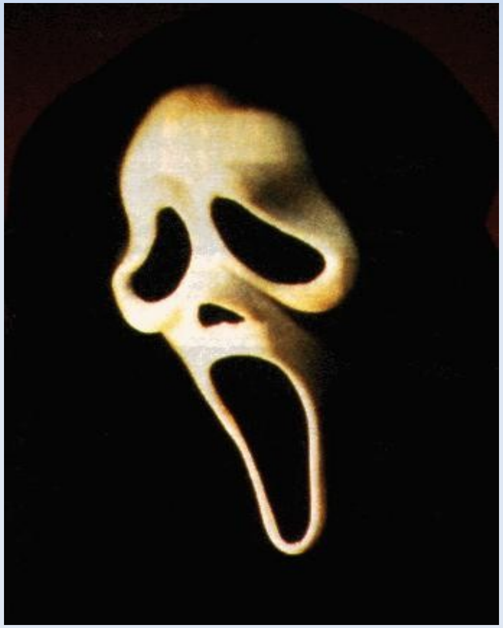
Research

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

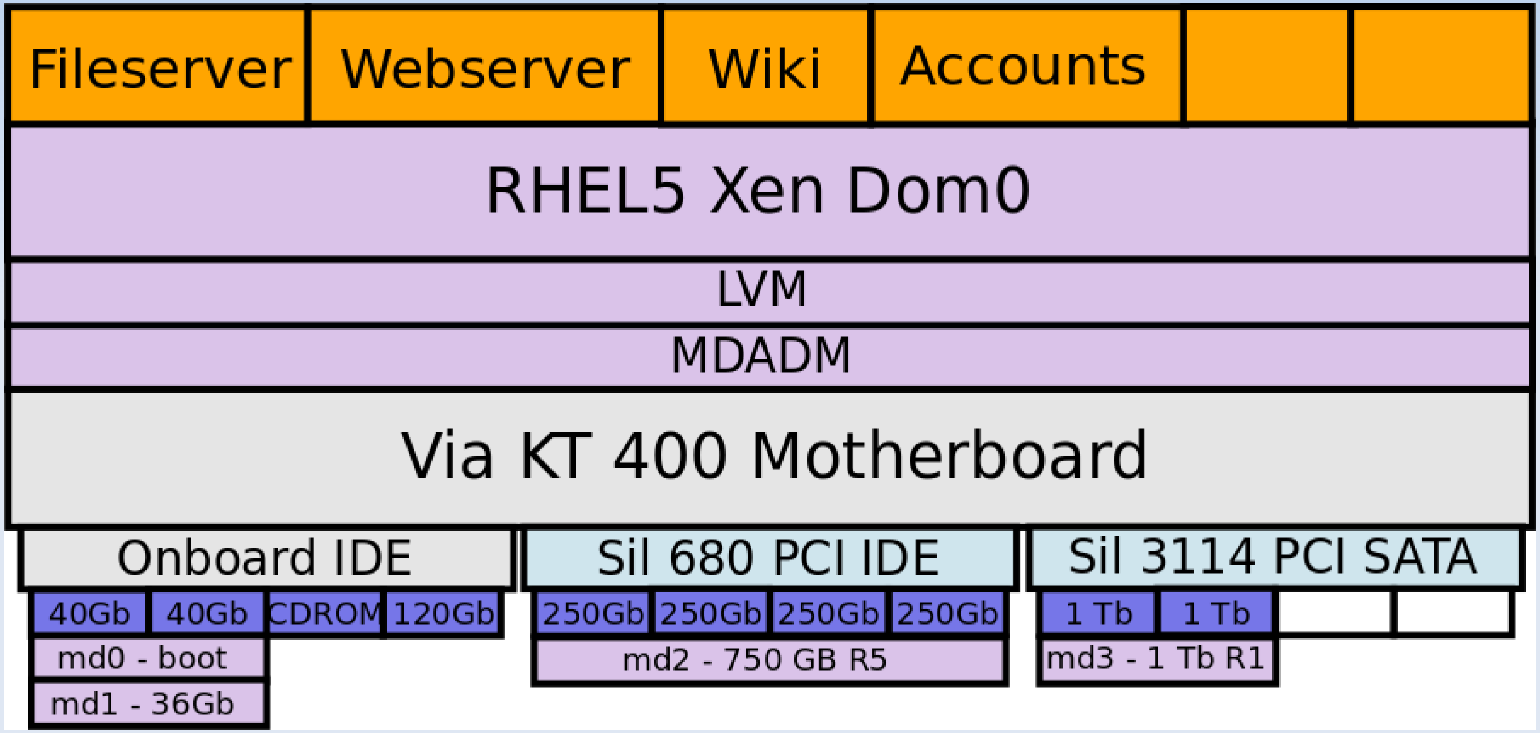
TLER

- 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?



Fileserver | Webserver | Wiki | Accounts

RHEL5 Xen Dom0

LVM

MDADM

Via KT 400 Motherboard

Onboard IDE

40Gb | 40Gb | CDROM | 120Gb

md0 - boot

md1 - 36Gb

Sil 680 PCI IDE

250Gb | 250Gb | 250Gb | 250Gb

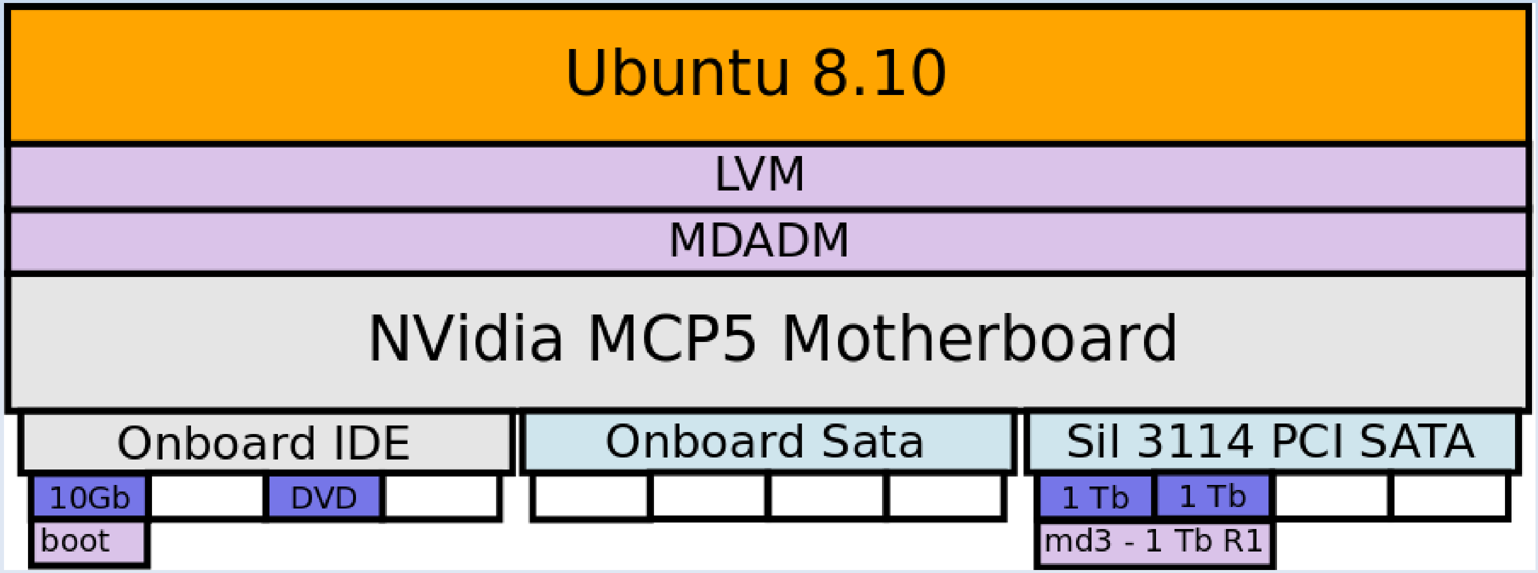
md2 - 750 GB R5

Sil 3114 PCI SATA

1 Tb | 1 Tb

md3 - 1 Tb R1

Test Hardware



Ubuntu 8.10

LVM

MDADM

NVidia MCP5 Motherboard

Onboard IDE

10Gb

boot

DVD

Onboard Sata

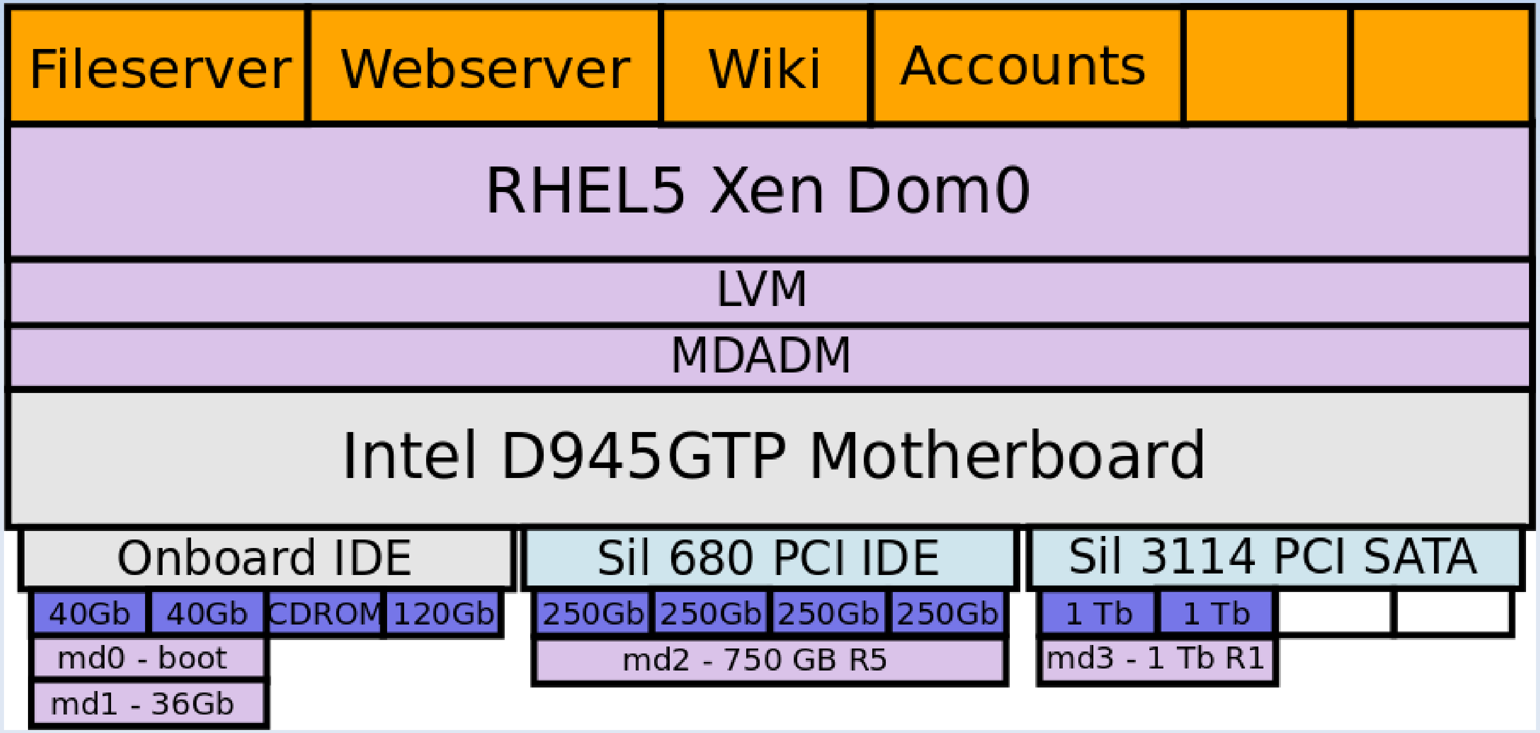
Sil 3114 PCI SATA

1 Tb

1 Tb

md3 - 1 Tb R1

Alternative Motherboard



Fileserver | Webserver | Wiki | Accounts

RHEL5 Xen Dom0

LVM

MDADM

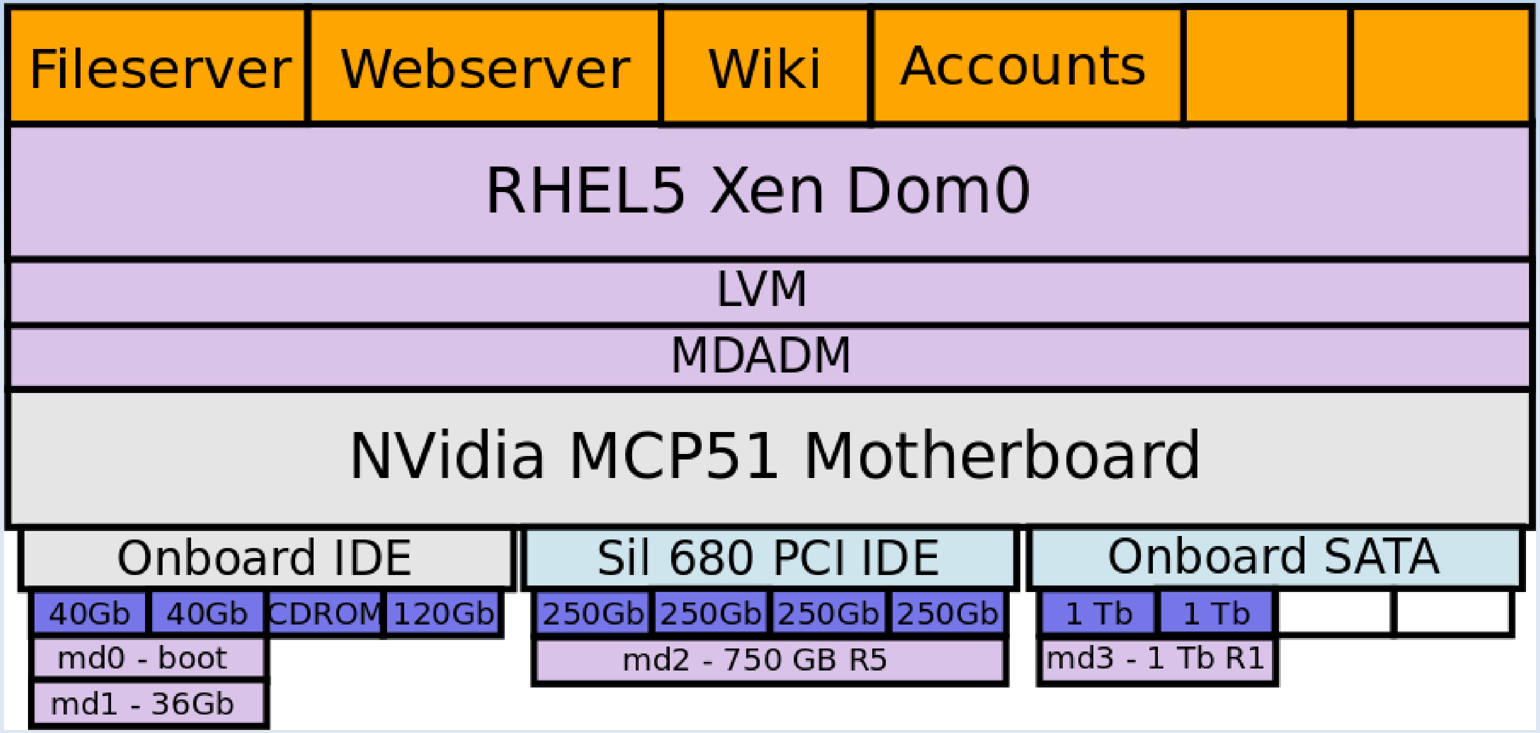
Intel D945GTP Motherboard

Onboard IDE: 40Gb | 40Gb | CDROM | 120Gb
md0 - boot
md1 - 36Gb

Sil 680 PCI IDE: 250Gb | 250Gb | 250Gb | 250Gb
md2 - 750 GB R5

Sil 3114 PCI SATA: 1 Tb | 1 Tb
md3 - 1 Tb R1

Final Solution



Fileserver
Webserver
Wiki
Accounts

RHEL5 Xen Dom0

LVM

MDADM

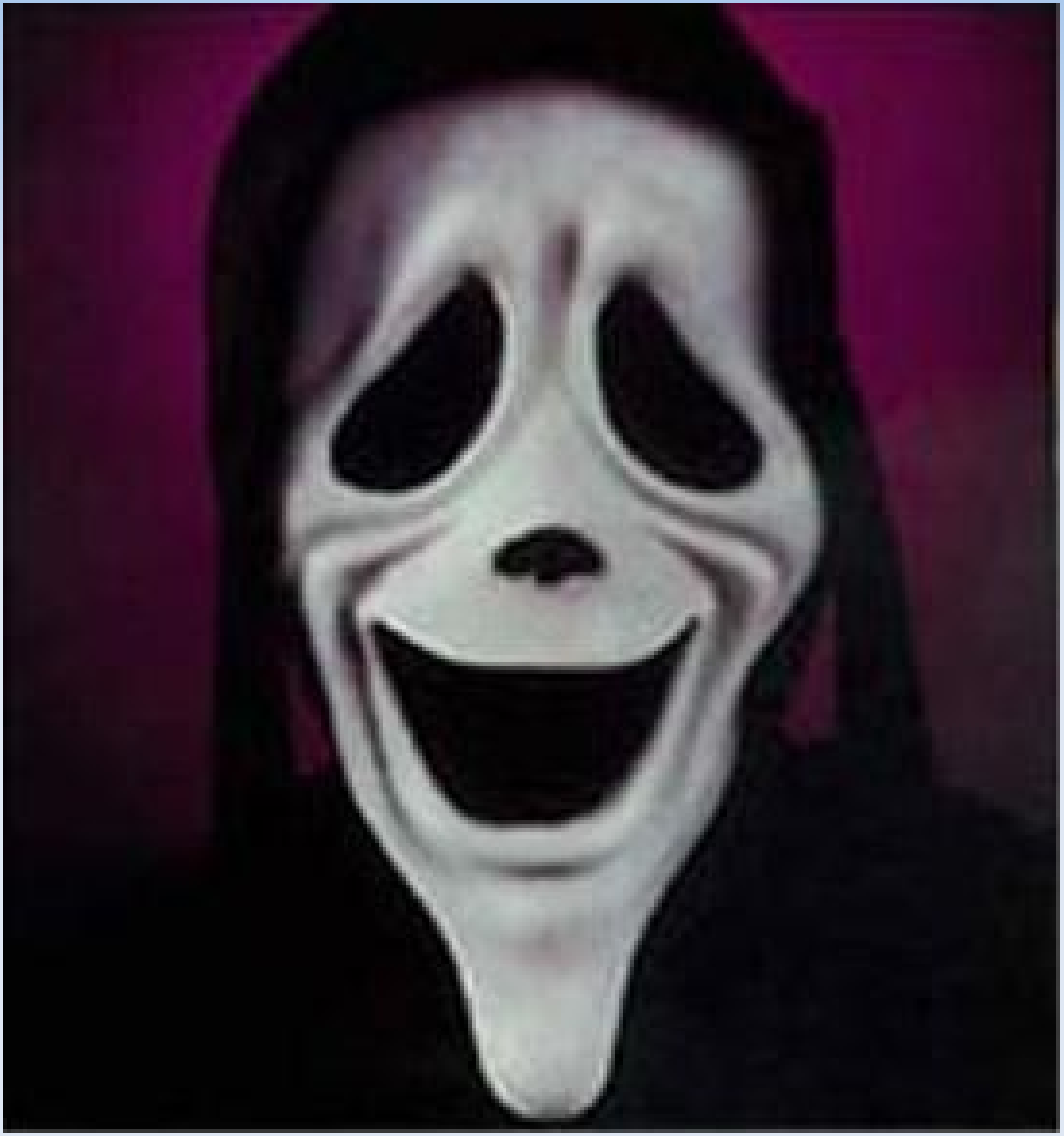
NVidia MCP51 Motherboard

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

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

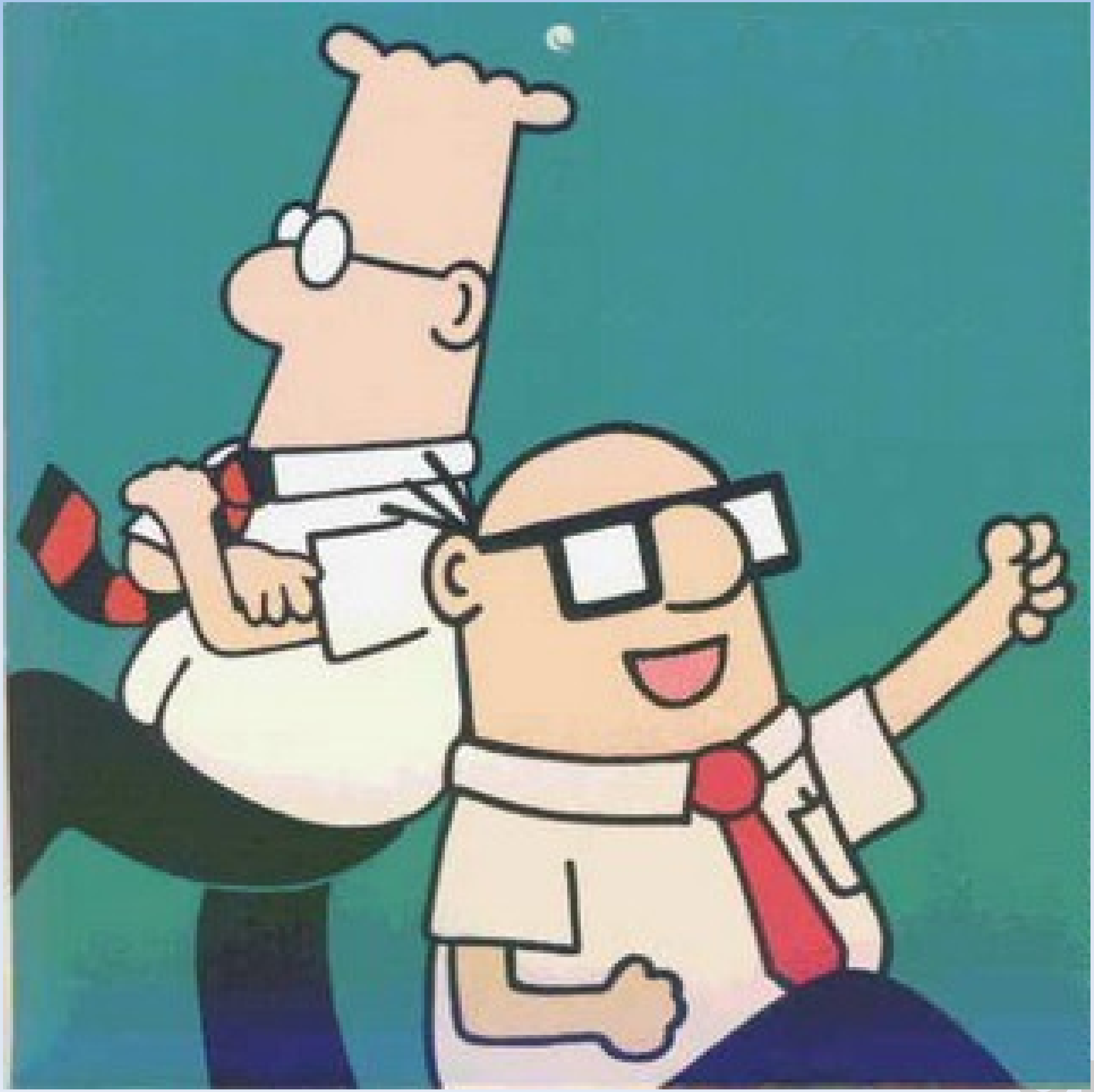
Onboard SATA
1 Tb
1 Tb
md3 - 1 Tb R1

We Win



OpenMedia

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 Backgound

  http://www.hardforum.com/archive/index.php/t-1191548.html
- Debian Thread on debugging mdadm

  http://marc.info/?l=debian-user&m=123115382721512&w=2
- Linux Raid Page at Linux Foundation

  http://www.linuxfoundation.org/collaborate/workgroups/linux-raid
- Linux Raid Mailing List

  http://vger.kernel.org/vger-lists.html#linux-raid

Questions