# Don’t lose my data

35+ years of storage war stories

Steven Ellis - Red Hat

LINUXCONFAU

# Agenda

A little bit of history

- Including some technology overviews

War Stories

- Names have been avoided to protect the “innocent”

A couple of tips and tricks along the way

# Where to begin?


LINUXCONFAU

https://commons.wikimedia.org/wiki/File:Bhimbetka_Cave_Paintings.jpg

https://pxhere.com/en/photo/1411286

# Common Disk Connection Protocols

| | |
|---|---|
| SAS | 12Gbps Serial Attached SCSI - Enterprise drives based on SCSI command set |
| NL-SAS | 12Gbps SAS Controller with SATA based media, disk and performance |
| SATA | Upto 6Gbps SATA - Mostly consumer based on Parallel ATA command set |
| SCSI | Physical Interface and command set |
| IDE | 8.3 MB/s - 133 MB/s - Parallel ATA command set |

# Enterprise Direct Connection Protocols

| | |
|---|---|
| SAS | 12Gbps Serial Attached SCSI - Enterprise drives based on SCSI command set |
| SATA | 6Gbps SATA - Mostly consumer based on Parallel ATA command set |
| PCIe | 31.5 Gbps PCIe 3.0 x4 - typically uses NVMe command set |
| M.2 | Form factor that can use SATA or PCIe (NVMe) |

# Data Protection Techniques

| | |
|---|---|
| RAID | Parity and striping across block devices to create sets of redundancy |
| EC | Erasure coding saves data in fragments with parity across different locations |
| Mirror | Storage array level synchronous and asynchronous mirroring of data (DR/BC) |
| Multipath | Redundant network paths from host to storage (dual HBA/NIC at host) |
| Cache | Battery or super-capacitor backed up cache |

# Block Storage Network - SAN

## Storage Array

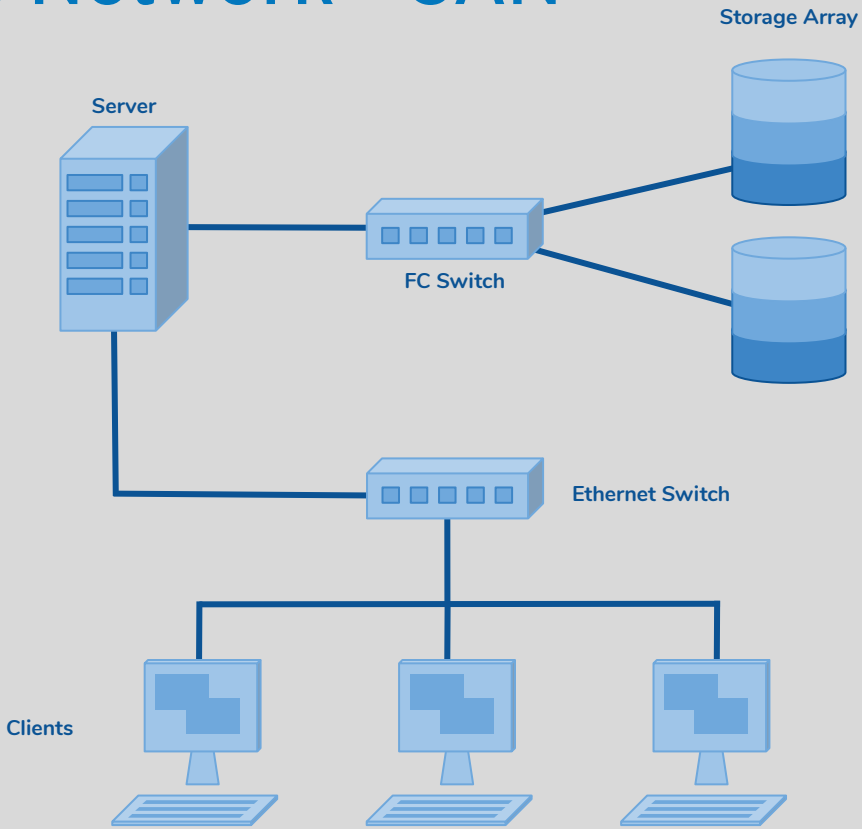
Modular or monolithic, Intelligent controllers

Redundant multi pathing, dedicated network

FC, iSCSI, FCoE, NVMe, IB

Block Characteristics:

- Very fast
- Data written in blocks
- Not human friendly
- Complex addressing



# File Storage Network - NAS

## Storage Array (Filer)

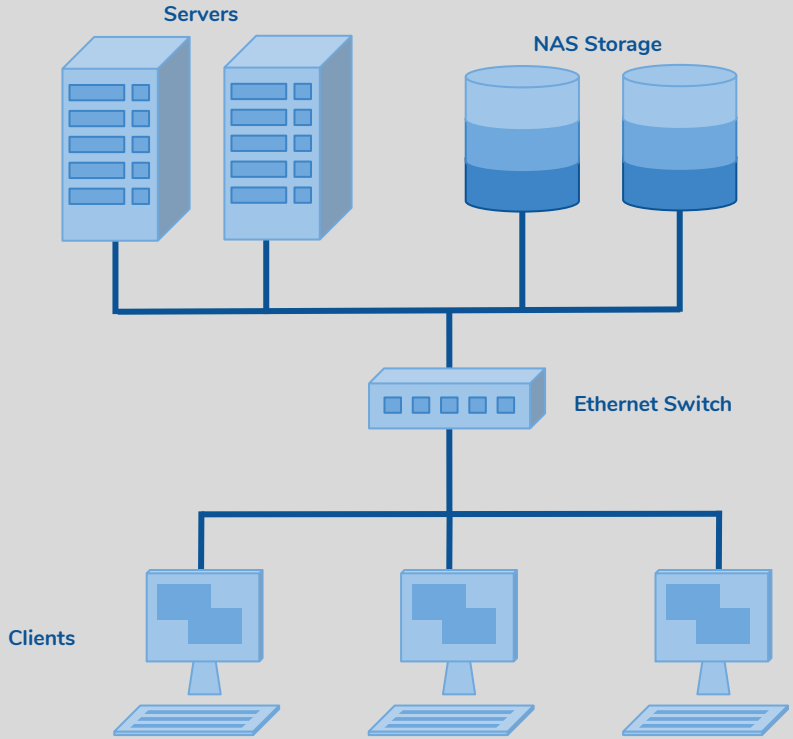
Usually Modular, Intelligent controllers

Redundancy via dedicated or existing ethernet network

NFS, CIFS

File Characteristics:

- Pretty fast
- Data written in files
- Human friendly paths
- Standard networking



# Object Storage - Cloud

## Content Addressable Storage

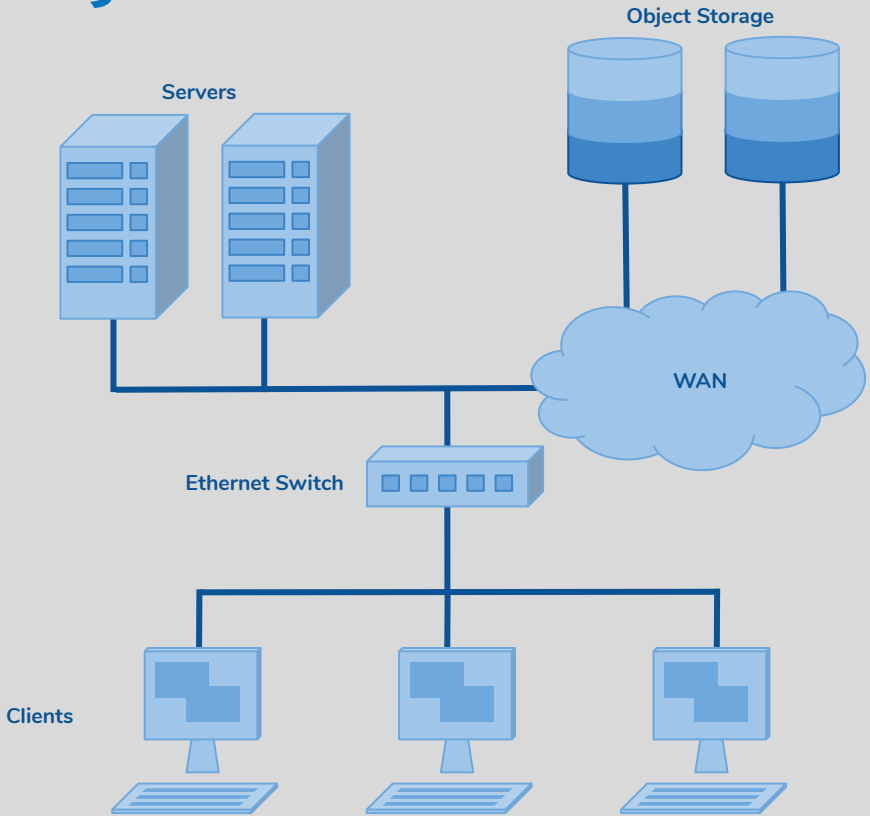
Usually massively scale-out

Redundancy via multiple data copies

REST API (http), Swift, S3

Object Characteristics:

- Not usually fast
- Data written in objects
- Eventually consistent
- Human friendly metadata
- Cloud networking



# Data Protection Techniques

| RAID | Redundant Array of Independent Disks |
|---|---|

RAID 0 - striping only, no protection

Performance only

RAID 1 - exact mirroring

Least capacity

RAID 5 - 5D+1P, parity blks striped

1 disk fail

RAID 6 - 4D+2P, parity blks striped

2 disk fail

# Data Protection Techniques

| EC | Erasure Coding - K + M |
|---|---|

EC 4+2

2 disk fail

EC 8+3

3 disk fail

# Data Optimisation/Reduction

| | |
|---|---|
| Compression | Algo to reduce redundant blocks, whitespace etc (unstructured data) |
| Deduplication | Avoids storing duplicate blocks (CPU intensive, rehydration, mapping) |

**Storage Vendor Terms**

- RAW Storage
- Usable storage
- Effective storage

- Decimal GB vs GiB

# Can I fsck that for you?

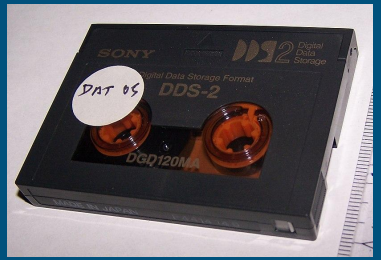
LINUXCONFAU

# 35+ years ago


ENERGY EFFICIENT AND
ANTI-RESONANCE
CASSETTE MECHANISM
B
XLII
POSITION HIGH
90
maxell

# Beware aging / legacy storage

- Bit Rot
- Disk Rot
- Flash failure
- Mould
- Old Interfaces (IDE / SCSI)
- Old Tape formats
- Hardware failure
- Old Filesystems




# Data recovery tips

Isopropyl Alcohol + lint free cloths

- Dust / oils can kill an optical drive
-

USB based dongles to reduce reboots

- CD/DVD Drive
- Floppy Drives
- IDE Drive
- SATA

NAS / SAN / External USB for initial archive

- I prefer locally attached USB-3 drives

Linux tools

- SystemRescueCd / UltimateBootCD
  - http://www.system-rescue-cd.org/
  - https://www.ultimatebootcd.com/
- ddrescue
- lzop
  - Fast lightweight compression
- testdisk / photorec
  - Recovery of filesystems and individual files off failed media

# Original Media Has Failed

USB based dongles don't always behave well with failed Hard Drives

- Time to dig out / borrow some old hardware
- Boot original hardware with a USB Live OS Image

Always create a full copy of the original media

- ddrescue is your friend
- perform data recovery with a snapshot/copy of the backup
- Fail back to testdisk/photorec

# HDDs > SSDs

TRIM is critical to Flash Storage performance

- Allows for elegant wear leveling

Makes it nearly impossible to recover “deleted data”

- On a HDD a deleted file is “often” just unlinked from the filesystem

# Raid is not a backup mechanism

Raid 0/1/10/5/6 can be implemented via

- Hardware raid controllers
  - Proprietary or in kernel drivers
- “Fake Raid”
  - Really a software driver - dm-raid
- Software Raid
  - mdadm or LVM based

# Going Mad with MDADM Pt 1

Original talk from Sys Admin Miniconf - LCA 2010 in Wellington

Pinpointing the issue

- RAID / HBA Adapter
  - Firmware Issues
    - updates that can trash a Raid array
    - Raid metadata incompatible with different firmware versions
  - Legacy Adapter
  - Conflict with motherboard chipset

MDADM can be your friend

- running XFS also helps

# Going Mad with MDADM Pt 2

Problem - Hardware Raid Controller Failure

- No spare compatible hardware
  - Trade Me or Ebay was the only option for parts
- Installed a SATA/SAS HBA into a generic **modern** Linux box
  - Raid metadata was detected by dm-raid
  - Raid array assembled into a running state
  - Data recovered onto replacement hardware

# Going Mad with MDADM Pt 3

The Problem - recovering  failed a RAID 5 array

Software Raid-5 set via mdadm

- 4 x 3TB Drives
- Marvel 88SE9230 PCI-e SATA HBA
- DMA errors under high I/O
  - Or during weekly raid consistency check
- 2 Drives were removed from Raid Set

BZs

- https://bugs.launchpad.net/ubuntu/+source/linux/+bug/1810239
- https://bugzilla.redhat.com/show_bug.cgi?id=1337313

Solution

1 - Change HBA

2 - Check the Raid Set

```
mdadm --detail /dev/md2
```

3 - Confirm Event of 3 disks is close enough

```
mdadm --examine /dev/sd[abd]2 | \
grep Event
```

4 - Force start a degraded array and cross fingers

4 - Consistency check on LVM and filesystems

LINUXCONFAU

# Go fsck

LINUXCONFAU

# How fsck is your storage

Sanity check for

- HDD
- SSD
- SAN

hdparm -t

Additional tool for flash storage

f3read

f3write

# Cluster fsck

# I need HA storage for/because

Beware

- there are dragons ahead
- may you live in interesting time
- life is to short to build a cluster of two

What is the use case

- RTO / RPO
- Workload performance requirements
- Any latency issues

HA / DR / Backup

- HA isn't a backup mechanism
- DR with high or variable latency
- Fail over / Fail back

Cluster of two

- Is a problem waiting to happen
- 3rd quorum node / arbiter is critical

# HA NFS

Approaches

- Active/Active
  - Requires a cluster aware file system
    - gpfs / gfs2
- Active / Passive
  - Shared storage over FC/iSCSI
  - Partition is only mounted on a single node
  - Pacemaker + VIP

Issues

- Application services are latency sensitive
  - Requirement sub 5ms
  - NFS failover is >= 30 seconds
- Scale
  - 2 node cluster couldn't cope with workload
  - Had to scale to 3 nodes with considerable added complexity
- No Live migration
  - Environment was virtualised

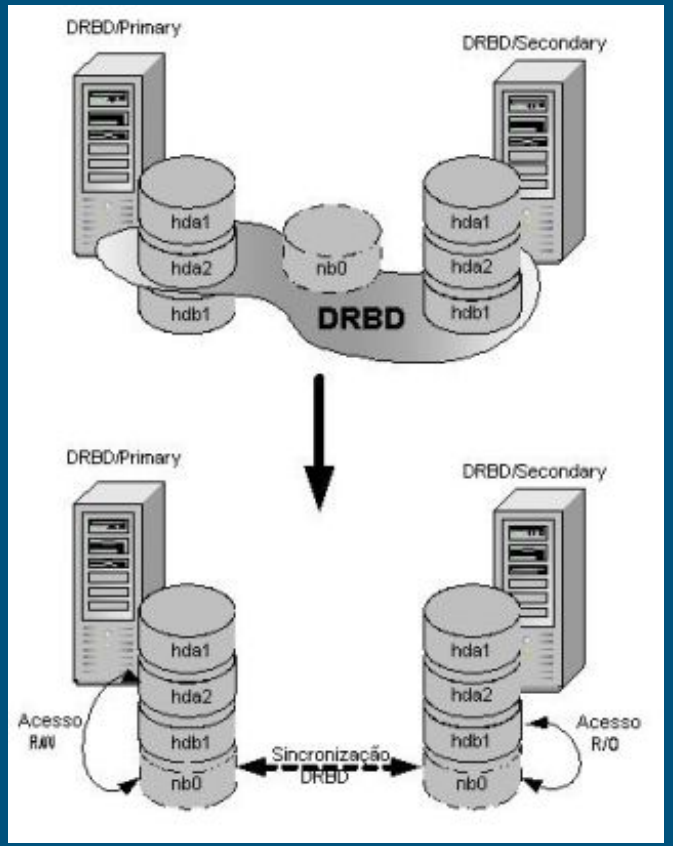
# DRBDont

DRBD

- Distributed Replicated Block Device

DRBDont

- Maintenance can be (was) painful
- Fail back issues
- Cluster of 2
- STONITH !!!!!



# Multiple Single points of failure

Understand the requirements

- And the existing infrastructure
- Especially any SAN arrays
    - And any associated  network infrastructure

Real cost of the solution

- What does an NFS head for the SAN cost
- vs project and operational cost of your “busy work”

Common statement is the existing storage infrastructure isn’t reliable or meet the RTO/RPO requirements of a project

- Secondary requirement is solution has to be virtualised

All Virtual infrastructure runs off the same SAN array

- But you need to meet a higher SLA than the array

# Software Defined Storage

Gluster

- Suits file centric workloads
- Simple to implement
- Can run virtually or on bare metal
- Scales elegantly
- Supports CIFS/NFS/pNFS + Gluster Fuse

Ceph


- Object / Block / File
- Focused on bare metal
- See my rook talk tomorrow for containers
- Vibrant community
- Replica 3 for performance
- Excellent EC implementation for scale

# What the fsck!

LINUXCONFAU

# Dust and Humidity

Existing machine room re-sized

- Shrunk to provide additional storage space
- New drywall installed
  - and sanded
- But they did install drop cloths over the racks

Outcome

- We had to vacuum out all the servers
- Almost every hard drive was replaced over next 9 months

Aircon unit leak

- Wet carpet in the machine room
- Temporary aircon couldn't deal with additional humidity from drying out carpets

Outcome

- Almost every hard drive failed over next 6 months

# Expect the unexpected

Corrupted LVM

- Multiple LUNs from SAN
- Combined into a single VG via LVM

Issue - FC LUNs had been allocated to 2 systems

- No partition table was present
- Unix team had re-formatted the LUN
- LUN was in the middle of a Linux LVM VG

**Recommendation** - Always create a partition table

Corrupted filesystem

- xfs filesystem consistency issues
- Rebooting host inconsistent behavior

Issue - Poor grouping of LUNs

- Virtual Guests hosted on KVM
- Direct LUNs mapped to Virtual Guests
- Guests mounted wrong /var or /data

**Recommendation**

- Unique LV UUIDs & mount by UUID

# Back to the future

# Everything old is new again

You will always need more storage

HDDs (rust) aren't dead (yet)

At some point TCO for Flash will drop below rust

EDSFF for hyperdense flash storage

Persistent Memory

Cloud players will continue to innovate

# Questions?

sellis@redhat.com
http://people.redhat.com/sellis

LINUXCONFAU