The University *of* Edinburgh

Information Services

ECDF

Edinburgh Compute & Data Facility

Services and technologies for Data Intensive Research

Dr Orlando Richards, ECDF Systems Team manager orlando.richards@ed.ac.uk

- Intro to ECDF
- Version Control
- General Data Store
- Compute
- High Performance Computing Storage
- GPFS

ECDF

- What is ECDF?
  - Edinburgh
  - Facility
  - Compute
  - Data
- Formed 2006,
  - cluster service 2007 (eddie)
  - “proper” data service 2010 (ecdfnas)

Primary Service List

- Compute cluster (eddie)*
- General Academic File Store (ecdfnas)*
- Version Control Service (svn, SourcEd)*
- Middleware services (portals, Grid, ...)
- Consultancy services
- User support

Version Control Service

- Subversion (svn)
- Massively resilient

  > 4 copies of every version of every file!

  Active/active service front ends
- Good for:
  - Source code revision control
  - Collaborative editing
  - “Lab book” style configuration recording
  - “Golden copy” storage of source data

ecdfnas

- “General Purpose Academic File Store”
- Built to be somewhere to keep data
- Massively resilient, high performance, fully integrated, enterprise grade, etc, etc
- Scalable (started at 40TB, now 191TB, in a month – 450TB, in 6 months – 750TB-1PB)
- Standard access methods
  - CIFS (samba)
  - NFS
  - SSHFS

eddie

- Linux “Beowulf” cluster
  - 2992 “Westmere” Intel Xeon CPU cores
  - 7568GB RAM
  - Gigabit ethernet
  - QDR infiniband (816 cores)
- Batch processing (log in, qsub, get results)
- HPC storage...

HPC Storage

- A question of scale...
  - Desktop machine had one processor, and a 7200rpm SATA hard drive
  - One user, typically one job
  - Multitasking will kill it
  - We have 3000 tasks
    - 3000 hard drives??

Everything-ity

- **Performance (ability)**
- **Reliability**
- Manageability
- Integrity
- Usability
- Flexibility
- Security

Performance

- Measured in:
  - Random I/O Operations Per Second (IOPS)
  - MegaBytes per Second throughput (MB/s)
  - Metadata Operations Per Second (MOPs?)
  - User satisfaction
- Needs to be:
  - Fast
  - Responsive
  - Consistent

Reliability

- Downtime is bad
  - Typical task lasts 48 hours
    - One second complete interruption to service gives approx average 24 hours downtime
  - Recovery takes effort
    - Identifying and fixing fault
    - Identifying and re-running failed tasks
    - Checking for consistency
- Fault tolerance, performance and integrity

Manageability

- Single point of management
- Routine maintenance without service disruption
- Single namespace
- Fault tolerance

Integrity

- Multi-process environment requires file locking
- Successful writes must be acknowledged
- Operations should be “atomic”

Usability

- Interactive performance must be fast
- Performance must be consistent
- Interface should be familiar
- Interface should be *open*

Flexibility

- Should be able to accommodate almost all use cases, current and future
- Should be scalable, upgradable, replaceable

Security

- Must be able to restrict access
- Rigorous authentication and authorisation
- Protection against attack

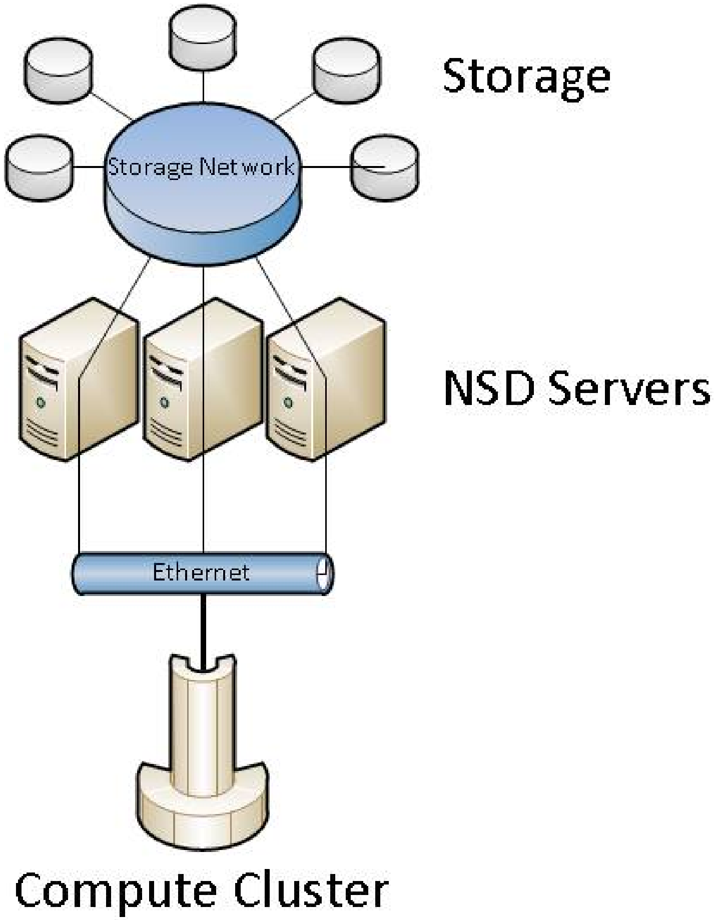
GPFS

- IBM's General Parallel File System

  www.ibm.com/systems/software/gpfs/
- Everything-ity
- See also: Lustre, CXFS, pNFS

GPFS Overview

- Based on shared disk accessed simultaneously by multiple servers
- Servers can communicate with the disks *through* the "NSD" servers
- Brokerage service



Everything-ity

- Performance (ability) - “scale-out”
- Reliability – inherently parallel
- Manageability – single namespace
- Integrity – full locking, atomic operations
- Usability – standard filesystem presentation
- Flexibility – fully posix compliant
- Security – access control lists, client node SSL authentication

The Disks

- The disk system is typically the bottleneck
- One SATA drive gives ~80 random IOPS (raw) – we need ~30,000 IOPS
- In the past – build to capacity, and the perf
for p
free

Year	Typical compute node performance SPECfp2006 Rate (Base)	Typical Hard Drive performance (4k Random IOPS)	Typical Hard Drive Capacity (GB)
2007	43	76	500
2011	210	76	2000

What we use:

- 2x IBM DS5300 SAN Disk systems
  - 114x 15k rpm Fibre Channel drives
  - 9x 73GB SLC SSD
- Sun StorageTek 6540
  - 110x 7200 rpm 750GB SATA drives
- Sun StorageTek 6540
  - 160x 7200 rpm 1000GB SATA drives

Storage Servers

- Need to move data fast – high bandwidth, low latency
- We have eight IBM X3650 servers, with:
  - 2x E5620 2.4GHz Intel Westmere CPUs
  - 48GB RAM
  - Dual-port 10GE ethernet adapter
  - Dual-port 8GB Fibre Channel HBA

GPFS Setup - Disks

- Tier 0 disk (SSD) holds metadata
  - Filesystem structure, folders, etc
- Tier 1 disk (15k rpm drives) holds “live” data
  - All new files are written here
- Tier 2 disk (7200 rpm drives) hold “bulk” data
  - Data which has aged
  - Large sequential files
- All disks configured in RAID5 pools, with

GPFS Setup - Filesystem

- 512kB Filesystem Block Size
  - Aligned with disk RAID pool stripe width
- 16kB sub-block size
- Metadata replication
- Capacity: 163TB
  - 74TB Tier 1 disk (39,312 RAW IOPS)
  - 88TB Tier 2 disk (20,520 RAW IOPS)
  - 950GB Tier 0 disk (n/a)

Performance

- As part of the acceptance test benchmarks:
  - 2.6 GByte/sec (read or write)
  - 28,000 random read 4k IOPS
  - 11,000 random write 4k IOPS
- Anecdotally – “very fast”

Efficient I/O On Eddie

- We give some advice:
  - Don't do lots of small I/O – save it up for a big sequential operation.
    - Write performance with 4k operations: 0.02GB/sec
    - Write performance with 512k operations: 1.53GB/sec
    - Read performance: 0.05GB/sec vs 3.3GB/sec
  - Store data in large files
    - Fewer metadata operations per data operation
    - “Easier” data management
    - Faster interactive response

Summary

- For big workloads, build big performance
- Sequential is good
- Use ECDF – for all your data needs!