Edinburgh Compute and Data Facility

Dr Orlando Richards

orlando.richards@ed.ac.uk

ECDF Systems Team

Information Services

University of Edinburgh

13th June 2007

edikt

e-science data information & knowledge transformation

Edinburgh Compute and Data Facility

- Introduction
- Hardware
- Software
- Environment
- GridEngine
- Output
- GPFS
- Cluster Filesystems
- Basic Concepts
- Inter Cluster GPFS
- User Administration

Introduction

- Edinburgh ***Compute*** and ***Data*** Facility
- Compute – commodity ”Beowulf” cluster – named ***Eddie***
- Data – currently SAN based:
  - 156.8TB Sun StorEdge 6320
    - High performance fibre channel disk
    - ~4 years old
  - 18TB IBM DS4700
    - SATA disk
    - Attached to Eddie
  - Future expansion
    - Currently in procurement process for SATA disk

Hardware

Eddie

- Based on IBM x3550 servers
- 2 x 3.0GHz Intel Xeon Woodcrest dual core processors
- 8GB RAM per node (2GB per processor core)
- 2 x Gigabit Ethernet network
- 10 x IBM x3650 infrastructure servers
- InfiniPath / Infiniband network on 60 nodes

edikt
e-science data information & knowledge transformation

Software (1)

- Operating System: Scientific Linux 4 (64-bit)
  - Can be imaged with different operating systems if required
  - Based on RedHat Enterprise Linux 4
    - completely compatible
  - Additional packages commonly needed for scientific computing
- Compilers:
  - GNU C, C++, Fortran
  - Intel C, C++, Fortran
- Parallel Environments:
  - OpenMPI, OpenMP, LAM-MPI, MPICH, MPICH2, MVAPICH

Software (2)

- Allinea Software:
  - DDT – Distributed Debugging Tool
  - OPT – Optimization and Profiling Tool
    - http://www.allinea.com
- User space software
  - Installed by users of system
  - Managed by users of system
  - Help and advice is available from Information Services

Environment

- Environment is set up using ”modules”
  - `module available`
    - lists all available modules
  - `module add intel/cce/9.1.046`
    - loads Intel compiler module
- This sets environment variables, including:
  - `$PATH`
  - `$MANPATH`
  - `$LD_LIBRARY_PATH`

GridEngine

- Queuing and scheduling system
- Queues up submitted jobs and executes them in priority order
- Priority is determined by prior usage and target share of the system
- Key commands:
  - `qsub -l h_rt=01:00:00` – submit a job to the queue
  - `qstat` – see the status of jobs in the queue or running
  - `qacct` – see the details of a finished job, view total usage
  - `qmon` – loads a GUI that can do all of the above, and more

- ***Demo***

Output

- stdout and stderr (console output and error messages) by default go to the home directory
- Local scratch space is available:
  - `/local/scratch/`
  - this space is wiped after every job
- Jobs can read from and write to the group work space:
  - `/exports/work/group/`
  - This is a ***GPFS*** filesystem

Questions

GPFS

Ewan Roche

e.roche@ed.ac.uk

ECDF Systems Team

Information Services

University of Edinburgh

13th June 2007

edikt

e-science data information & knowledge transformation

Filesystems for Clusters

- We expect ~500 to 1500 simultaneous jobs to be running
  - Potential for huge I/O bottleneck
  - 1000 X 10Mbit/s = 10 Gigabits per Second
  - Not possible for a single server to cope
  - Network Filesystems (NFS/AFS) not able to cope (although some vendors suggested NFS.....)

GPFS

- General Parallel File System from IBM
  - True parallel filesystem with fine grained locking
  - Allows near linear increase in performance with increased parallelism
  - Scales to very large filesystems and numbers of nodes
  - Relies on standard IP networking for most of the data transfer
  - Kernel module presents a POSIX filesystem

The GPFS Cluster

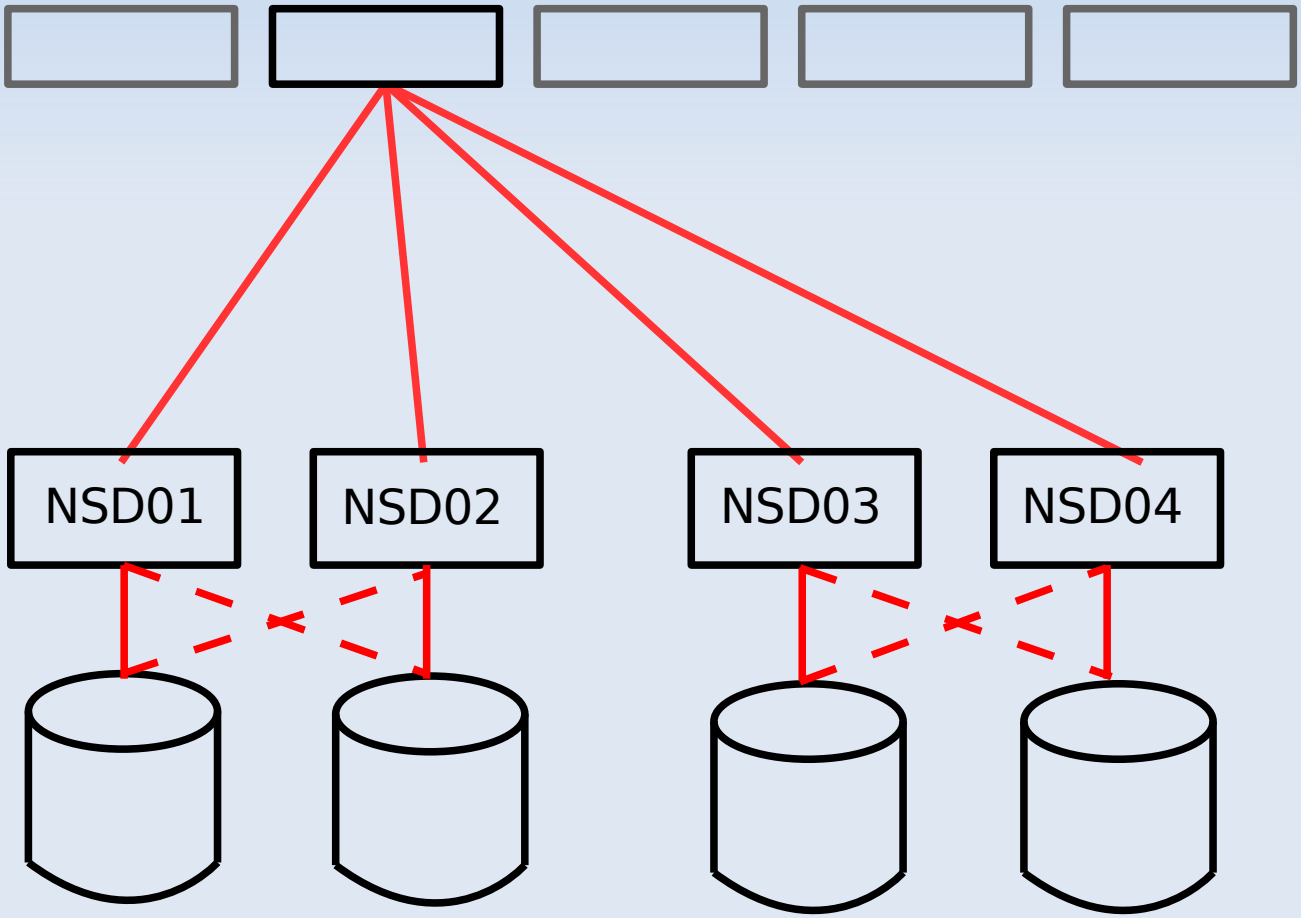
- All nodes are in the GPFS cluster but some are special.
- NSD Nodes
  - Connected directly to disk (via Fibre Channel)
  - Primary NSD does all transfer to a particular disk but a secondary (backup) can be specified.
- Quorum Nodes
  - Make decisions on integrity of FS when decisions in doubt. Maximum of 8.

GPFS on Eddie

```
[eroche@frontend01]$ df -h
Filesystem                 Size  Used Avail Use% Mounted on
/dev/sda1                  5.0G  1.4G  3.4G  30% /
/dev/sda6                  203G  754M  192G   1% /local
/dev/sda3                  2.0G   53M  1.9G   3% /tmp
/dev/sda2                  5.0G  114M  4.6G   3% /var
192.168.195.254:/usr/local
                            16G  4.3G   11G  30% /usr/local
/dev/fs0                    24T  1.3T   23T   6% /exports/work
/dev/fs1                   5.5T   24G  5.5T   1% /exports/home
```

GPFS Basics

Accessing a file from a worker node

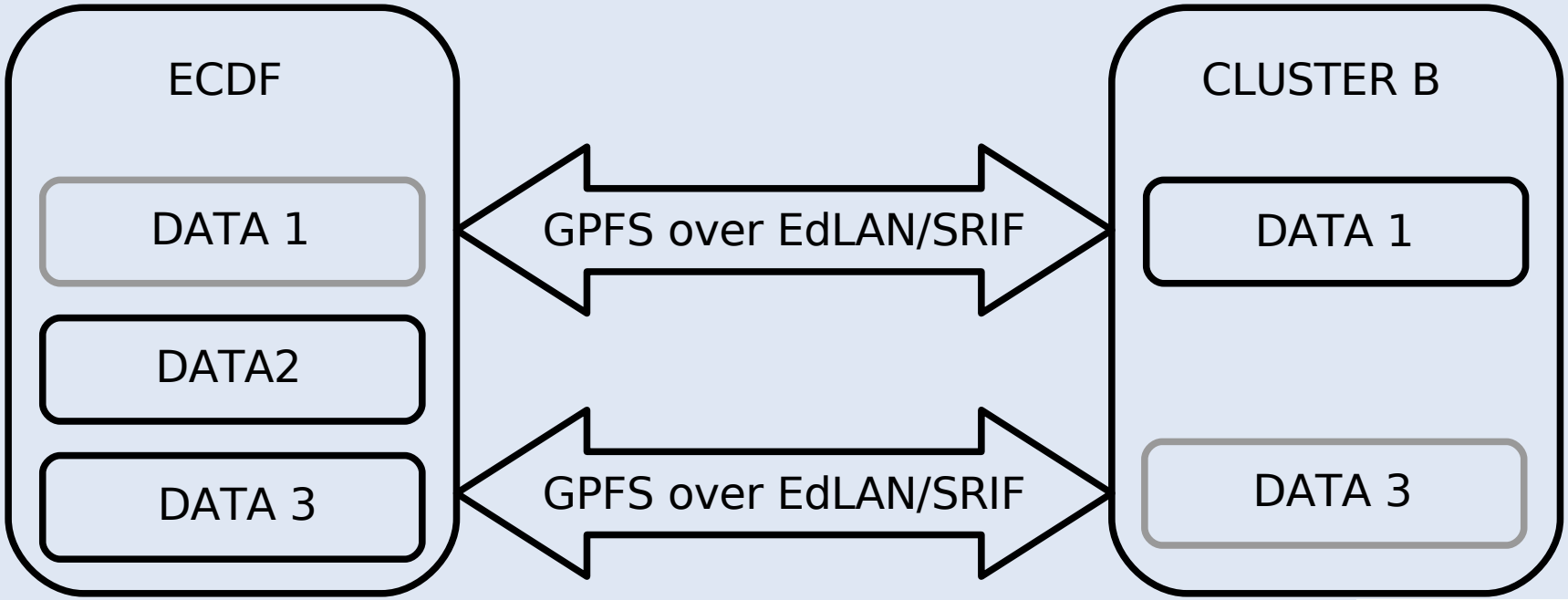


However.....

- Security
  - One can only share entire filesystems
  - Root on the cluster has access to all shared files
  - Traffic across the network not encrypted
- Scalability
  - Currently limited to 32 filesystems

Inter Cluster Communications

- There exists the capability to share filesystems between clusters via GPFS



User Administration

- Quotas can be imposed per filesystem for
  - Users e.g. 1GB for `/exports/home/eroche`
  - Groups
  - Filesets e.g. 1TB for `/exports/work/ug`
  - Possible to combine the above controls

- A fileset is like a filesystem within a filesystem and is a very useful organisational construct

Questions