CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# CRUK cluster introduction

## Using the Cambridge Institute's High Performance Computing

UNIVERSITY OF CAMBRIDGE

# Overview

This brief course will give you two things:

1. A refresher on unix and an introduction to cluster computing

2. Basic instruction on using our scheduler

3. Some performance hints

It *won't* make you an expert on parallel computing and HPC, but will let you get to work.

This course has a practical component, for which you will need an ssh client and cluster account.

# Session I

| | |
|---|---|
| 1 | Unix refresher |
| 2 | Cluster introduction |
| 3 | Practical – unix processes |

# Session II

| | |
|---|---|
| 4 | Using the scheduler |
| 5 | Practical – job submission |

# Session III

| | |
|---|---|
| 6 | Some performance hints |

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# Unix refresher

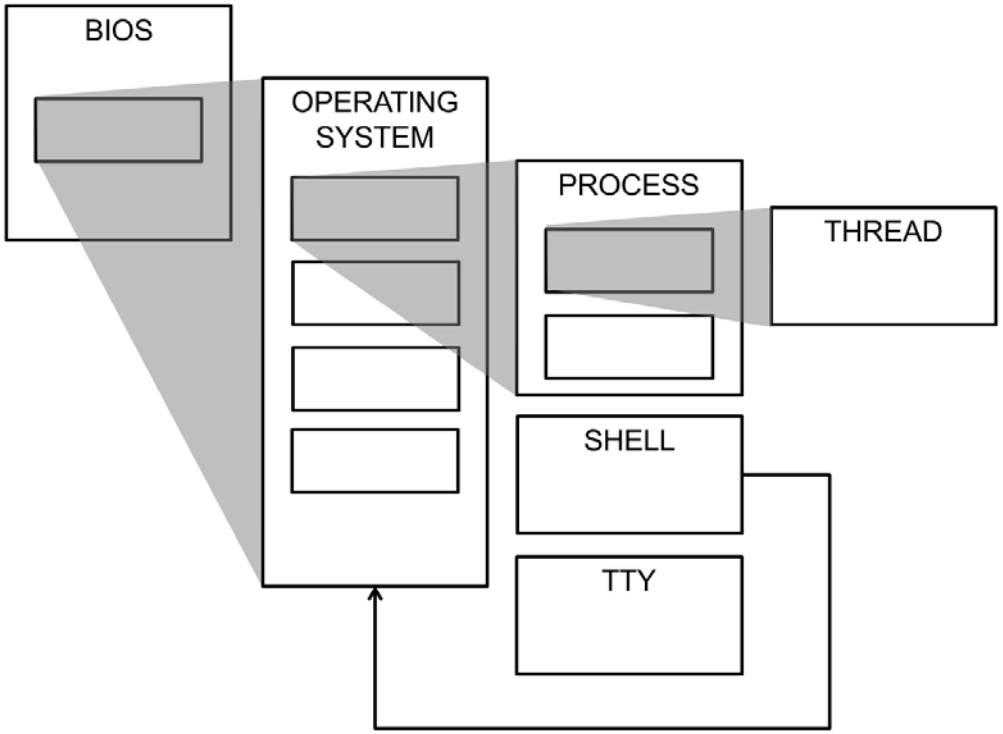
## (we have a course if this is all new…)

UNIVERSITY OF CAMBRIDGE

# Operating Systems and Processes

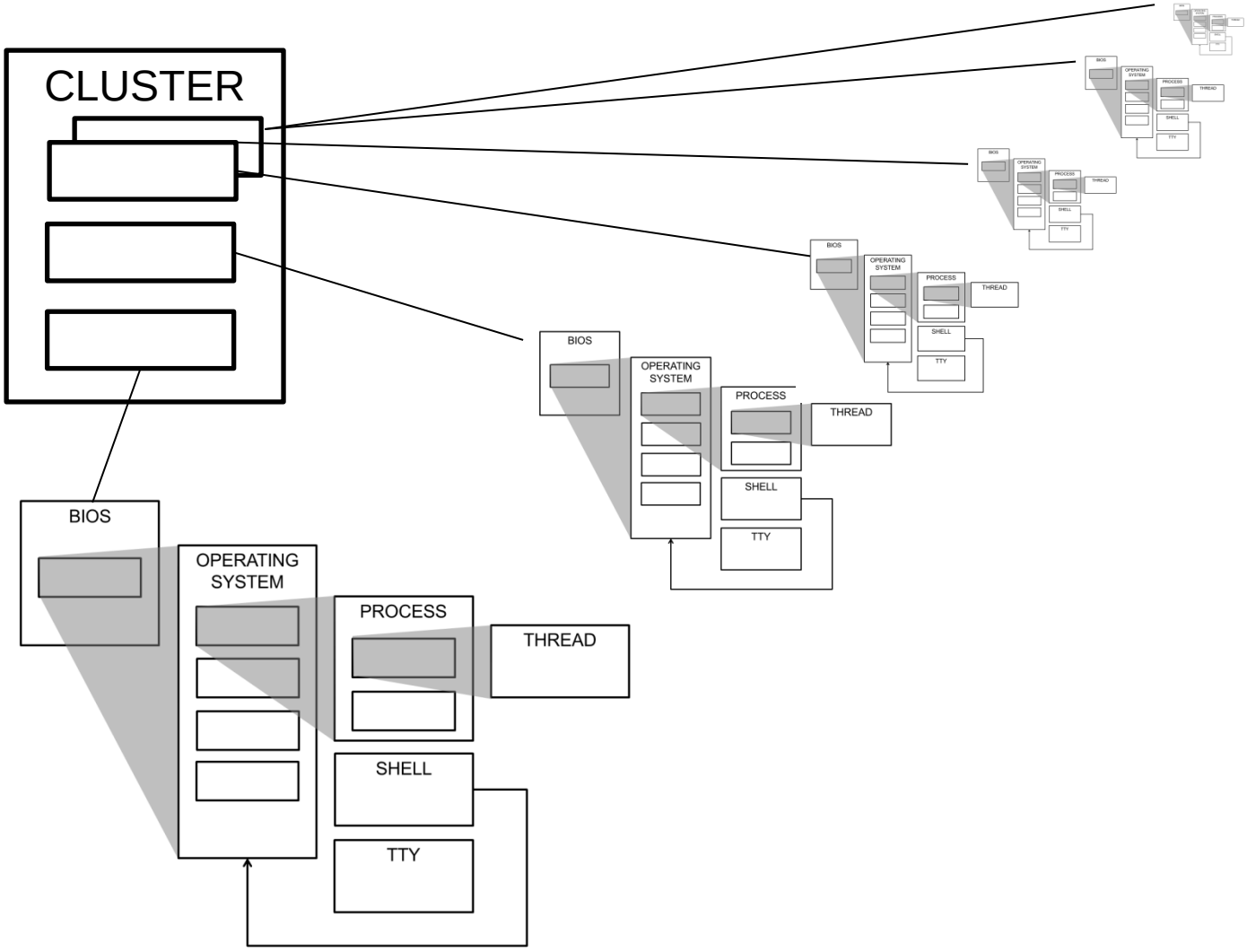
'Unix' or 'linux' (or 'UNIX') is our *operating system* – the program that controls the processes and their access to the network, screen, etc.

The shell is a *process* – it happens to be one that can see its own OS, which is one of the reasons it's so useful.



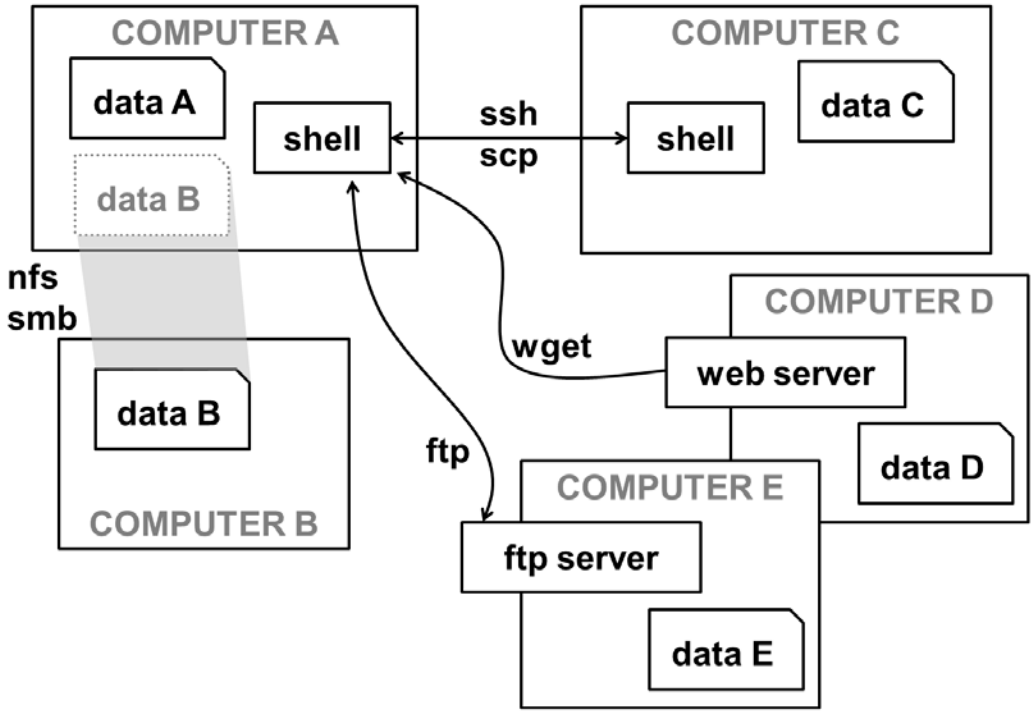
# A cluster is just many computers together...

Each one with its own OS, processes, and shell environments.



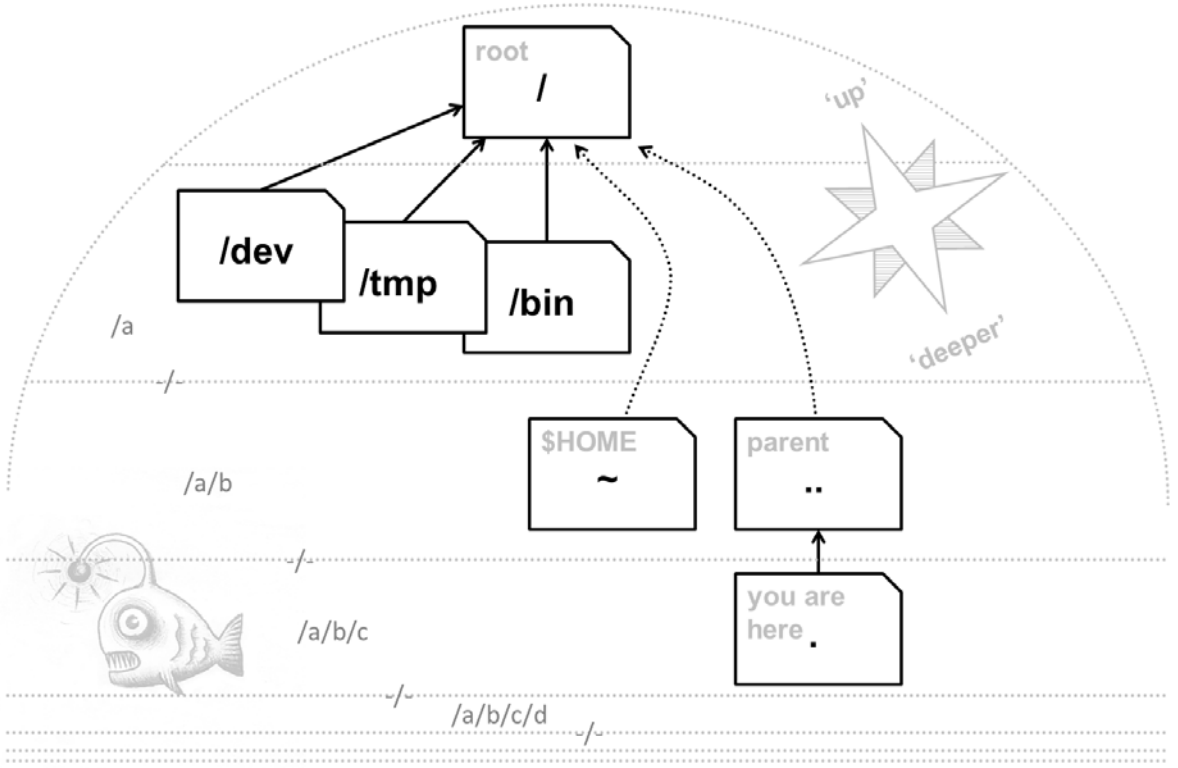
# Moving data, or yourself

Most of the ways of moving data around the internet were developed for Unix first. You also have the option of going to where the data is, with a remote shell.



# Navigation concepts

You need to be able to navigate without a GUI.
Fortunately some things are always in the same place.
Unix file systems are trees, *with the roots at the top*.



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

# CRUK cluster

## HPC @ CRUK CI

UNIVERSITY OF CAMBRIDGE

## Key Cluster-related Staff

Peter Maccallum (Head of IT & SC)

Marc O'Brien (Technical Architect)

Nigel Berryman (Snr. Systems Administrator)

Simon Bell (Systems Administrator)
Luis Huang (Systems Administrator)
Beauty Bapiro (Systems Administrator)

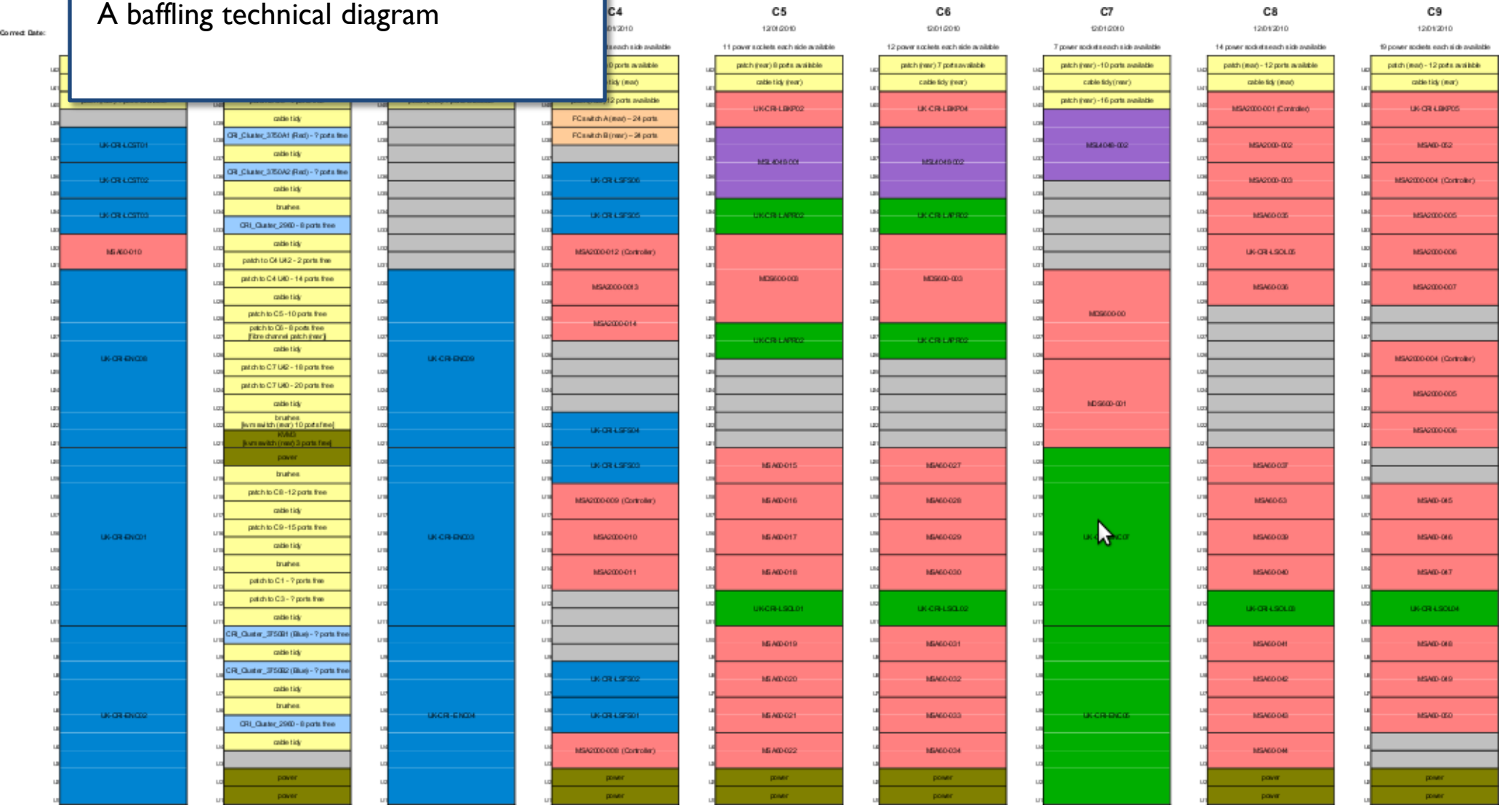
## CRI IT & SC Help Desk

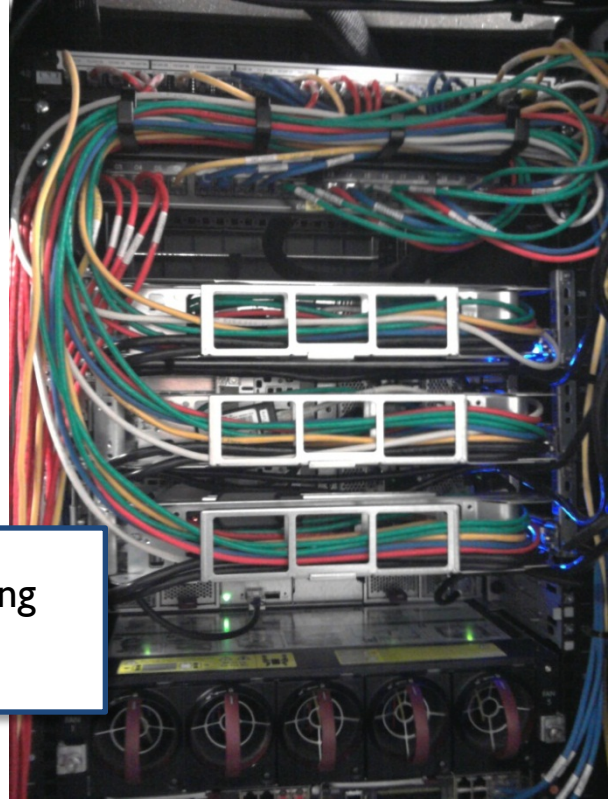
Use the usual helpdesk route for day-to-day problems – if an issue is affecting you it may be affecting many people.

helpdesk-it@cruk.cam.ac.uk

01223 769600

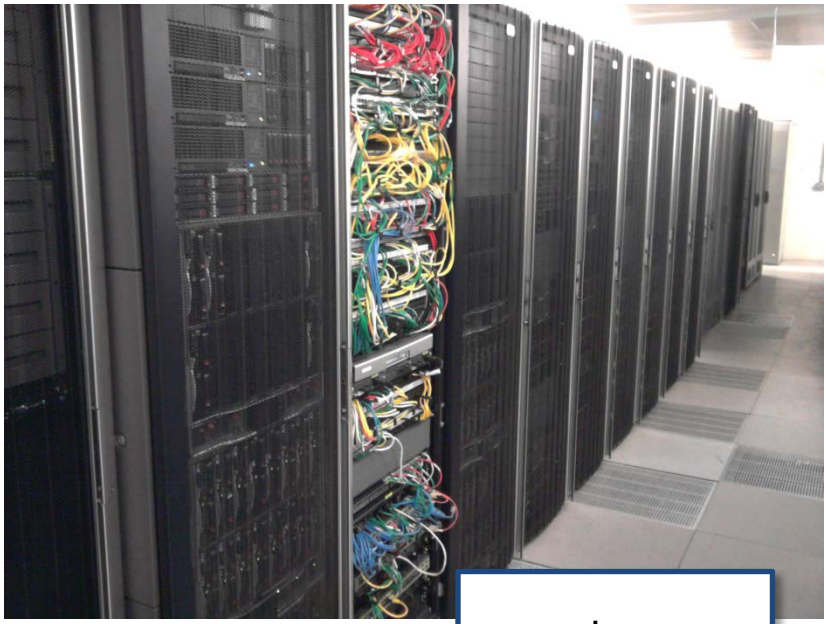
A baffling technical diagram





Networking

Compute blades



…and storage

## CRI HPC Cluster Specifications (March 2011)

· 70 computer nodes
- 560 cores
- 2 x 2.66GHz Quad-core Intel CPU
· 16, 24, 32 or 48GB per compute node
· 48TB Lustre parallel file-system
· Job scheduler (Platform LSF)
· Red Hat Enterprise Linux 5.4

## CRI HPC Cluster Specifications (as of April 2011)

· 134 computer nodes
- 1280 cores
- 2 x 2.66GHz Quad and Six-core Intel CPU
· 16, 24, 32, 48GB and 96GB per compute node
· 96TB Lustre parallel file-system
· Job scheduler (Platform LSF)
· Red Hat Enterprise Linux 5.4

# Other ways of computing: Dedicated Servers

Database servers
- BioInformatics database servers
- Many 10s TB storage

Group Servers
- Dedicated server for individual research groups

Virtual Servers
- VMware cluster (virtual servers)
- Ideal for non-CPU intensive applications, such as web servers

# Accessing the HPC cluster

Login to the head node via secure shell (ssh)

```
user@laptop:~> ssh uk-cri-lcst01
Last login: Wed Nov 24 13:15:12 2010 from 143.65.50.119
.--------------------------------------------------------------------,
| CRUK Cambridge Research Institute - HPC Cluster |
|====================================================================|
| Please report any issues to CRI IT Helpdesk <cri.it@cancer.org.uk> |
`--------------------------------------------------------------------'
Disk quotas for group xxlab (gid 1234):
file-system kbytes quota limit grace files quota limit
grace
/lustre 128 900000000 1000000000 - 31 0 0 -
Disk quotas for user xxuser01 (uid 87654321):
Filesystem blocks quota limit grace files quota limit grace
uk-cri-lcst03:/home
*10987650 10000000 10485760 21824 0 0
[user@uk-cri-lcst01 ~]$
```

# /home Directory

· Relatively small 670GB file systems
- Intended for source code and common user applications & libraries
· User quota set to 10GB
- Receive automated email if user exceed soft limit
- 7 days grace period

Mounted on head node and all compute nodes
- Read-only on compute nodes
· Daily backed up

```
[user@uk-cri-lcst01~ ]$ quota
Disk quotas for user xxuser01 (uid 87654321):
Filesystem blocks quota limit grace files quota limit grace
uk-cri-lcst03:/home
*10987650 10000000 10485760 21824 0 0
```

# Other Available File Systems

Group File Systems
- 30+ TB
- NFS mounted on head node as "/group-dirs"
- Quoted per group

Archive!

# Application File System Hierarchy

· Standard Linux applications in /apps/usr
- Based on "Gentoo Prefix" project
( http://www.gentoo.org/proj/en/gentoo-alt/prefix/ )
- Complete Independence from underlying operating system
(currently Red Hat Enterprise Linux)
- Access through commonly used /usr/local via symbolic
links for compatibility
- Mounted on all compute nodes

```
$ ls -ld /usr/local/{bin,lib,include}
lrwxrwxrwx 1 root root 13 Aug 4 20:04 /usr/local/bin -> /apps/usr/bin
lrwxrwxrwx 1 root root 17 Aug 4 20:05 /usr/local/include ->
/apps/usr/include
lrwxrwxrwx 1 root root 13 Aug 4 20:04 /usr/local/lib -> /apps/usr/lib
```

# Supported Applications

· Introduction to the "Environment Module Package"

- User environment can be modified on a per-module basis
- Set, append, remove and delete environment variables such as PATH, MANPATH & LD_LIBRARY_PATH

```
$ module avail
------------ /apps/usr/pkg/Modules/modulefiles ------------
ImageMagick/6.3.7.9 R/2.7.1 java/1.6.0.21
ImageMagick/6.5.2.9 R/2.8.1 matlab/2008b
ImageMagick/6.5.8.8 R/default-2.8.1 modules
ImageMagick/default-6.5.8.8 dot muscle/3.6
MEME/4.3.0 gcc/3.4.6 openmpi/1.4.3
R/2.10.0 gcc/4.1.2 ruby/1.8.7
R/2.10.1 gcc/4.2.4 ruby/1.9.1
R/2.11.1 gcc/4.3.2 use.own
R/2.12.0 gcc/4.4.3 vienna-rna/1.8.4
R/2.7.0 gcc/default-4.4.3
```

# Module Environment Package

Show application information and environment changes

```
$ module show R/2.11.1
-------------------------------------------------------------------
/apps/usr/pkg/Modules/modulefiles/R/2.11.1:
module-whatis R is a language and environment for statistical
computing and graphics
conflict R
prepend-path PATH /apps/usr/pkg/R/2.11.1/bin \
/apps/usr/pkg/imagemagick/6.5.8.8/bin
prepend-path MANPATH /apps/usr/pkg/R/2.11.1/share/man
prepend-path LD_LIBRARY_PATH /apps/usr/pkg/imagemagick/6.5.8.8/lib
-------------------------------------------------------------------
```

# Example "Module" in action...

· Set per command shell

```
$ module load R/2.10.1
$ which R
/apps/usr/pkg/R/2.10.1/bin/R
$ module load muscle
$ module list
Currently Loaded Modulefiles:
1) R/2.10.1 2) muscle/3.6
$ module switch R/2.8.1
$ which R
/apps/usr/pkg/R/2.8.1/bin/R
$ module unload R
$ module list
Currently Loaded Modulefiles:
1) muscle/3.6
$ which R
/usr/bin/which: no R in (/apps/usr/pkg/muscle/3.6/bin:....)
```

# Transfer data into and out of the cluster

Use SSH Secure File Transfer (SFTP) to transfer data

Bulk transfers using rsync (delta-transfer algorithm)

```
user@laptop:~> scp -r data/ uk-cri-lcst01:/lustre/xxlab/
file01.dat 100% 100KB 100.0KB/s 00:00
file02.dat 100% 100KB 100.0KB/s 00:00
file03.dat 100% 100KB 100.0KB/s 00:00

user@laptop:~> rsync -av data/ uk-cri-lcst01:/lustre/xxlab/data/
sending incremental file list
file01.dat
file02.dat
file03.dat
sent 307464 bytes received 72 bytes 615072.00 bytes/sec
total size is 307200 speedup is 1.00
```

# Cluster Storage

1. What's wrong with NFS or CIFS?
2. The Lustre parallel file-system
3. Working with Lustre

## The Problem

Extreme I/O demand on storage
- HPC cluster can have 10s to 1000s compute nodes
- x Many users
- x 1000s jobs
- + Millions of small and large files

Required shared filesystem
Breaks most filesystems !!!

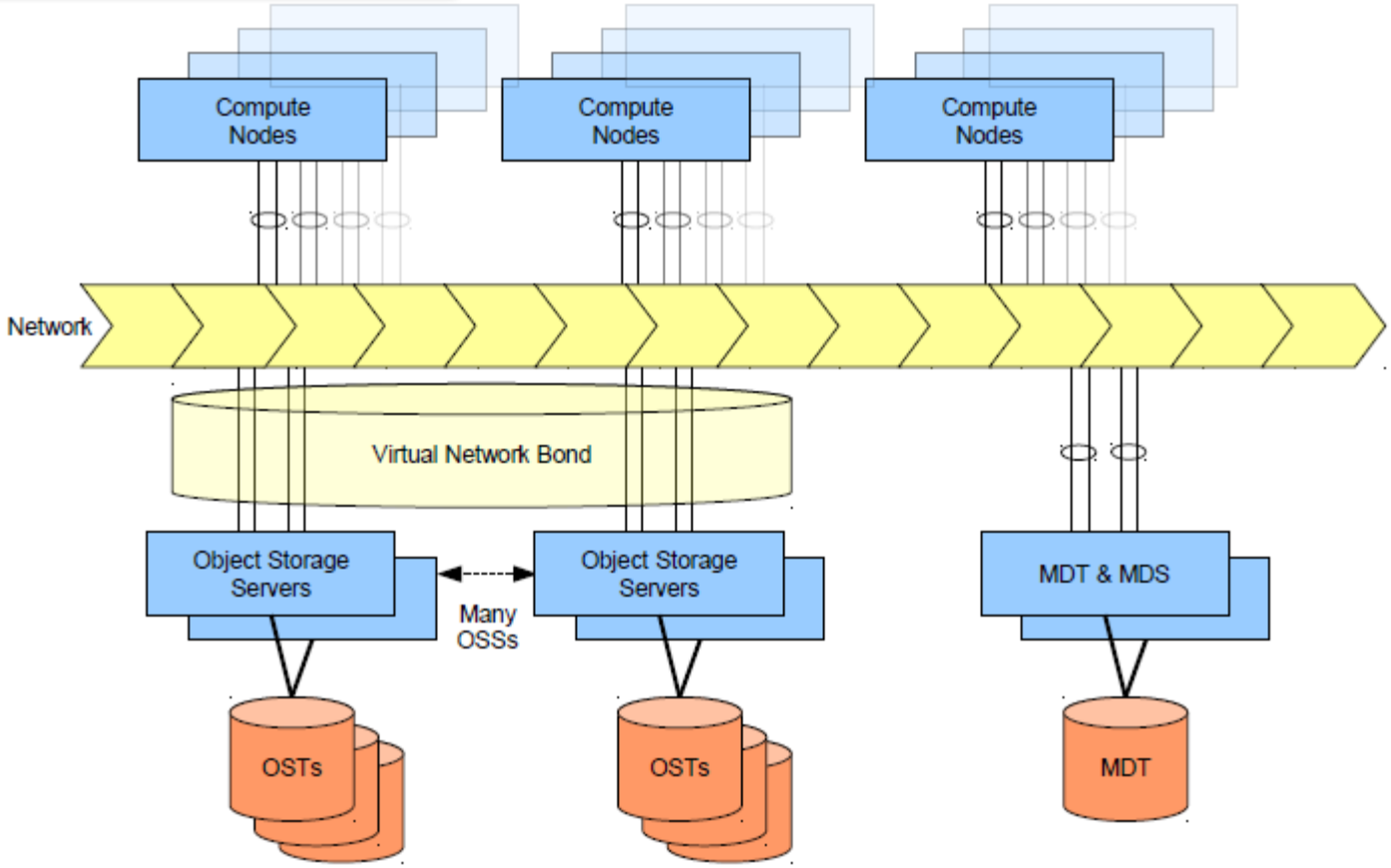
## Lustre: Parallel Filesystem

Lustre is a massively parallel distributed file system
- Deployed in 7 out of 10 most powerful supercomputers
- POSIX compliant

Lustre design paradigm concepts
- Separation of file meta-data and storage allocation
- Scalable data serving through parallel data striping
- Aggregate network bandwidth
- Distributed operation

.

# Lustre Architecture



# Lustre Quotas

· /lustre quotas are applied to each group
- Units in kilobytes (oddity)
- "quota" & "limit" of zero signifies no quota (as shown for user)

```
uk-cri-lcst01 ~ $ lfs quota /lustre
Disk quotas for user user321 (uid 442255):
Filesystem kbytes quota limit grace files quota limit grace
/lustre 275648748 0 0 - 984 0 0 -
Disk quotas for group xxlab (gid 987):
Filesystem kbytes quota limit grace files quota limit grace
/lustre 3471466428 3600000000 4000000000 - 681541 0 0 -
```

# Lustre Health Check

Check the status of each lustre
Display the usage of each distributed Object Storage Target component

```
uk-cri-lcst01 ~ $ lfs check servers
lfs check servers
lustre-MDT0000-mdc-ffff81041158cc00 active.
lustre-OST0000-osc-ffff81041158cc00 active.
...
lustre-OST0007-osc-ffff81041158cc00 active.

uk-cri-lcst01 ~ $ lfs df -h
UUID bytes Used Available Use% Mounted on
lustre-MDT0000_UUID 239.0G 856.7M 224.5G 0% /lustre[MDT:0]
lustre-OST0000_UUID 5.4T 2.2T 2.9T 40% /lustre[OST:0]
...
lustre-OST0007_UUID 5.4T 2.1T 3.0T 39% /lustre[OST:7]
filesystem summary: 42.9T 16.3T 26.6T 37% /lustre
```