

CRUK cluster introduction

Using the Cambridge Institute's High Performance Computing



Overview

This brief course will give you two things:

- I. A refresher on unix and an introduction to cluster computing
- 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

- I 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



Unix refresher

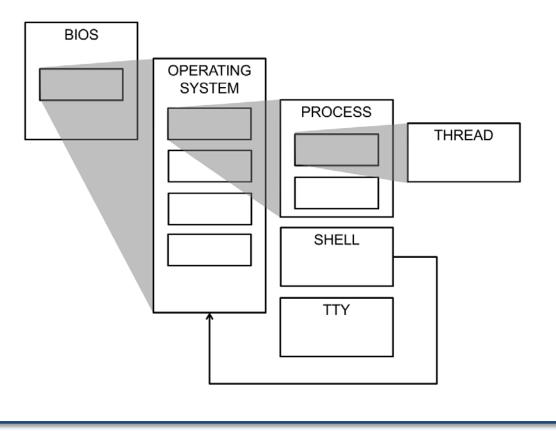
(we have a course if this is all new...)



Operating Systems and Processes

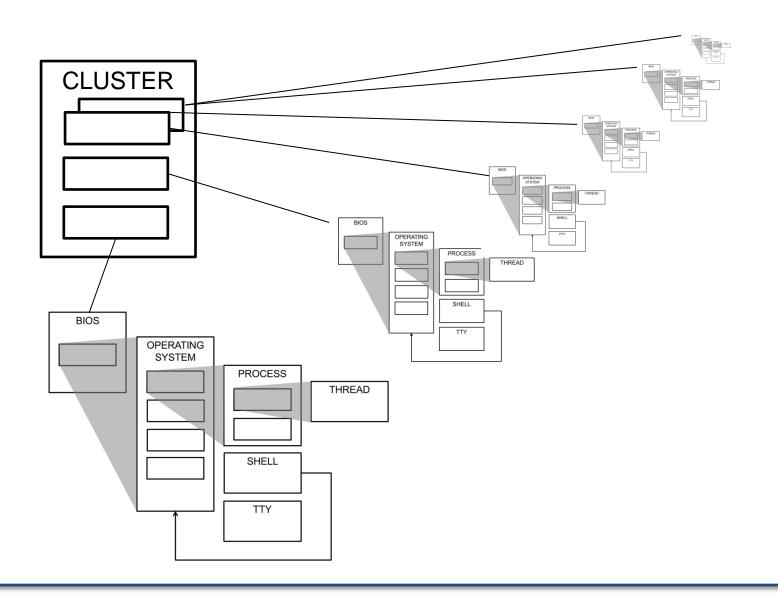
'Unix' or 'linux' (or 'UNIX') is our *operating system* – the program that controls the processes and their access to the network, screen, etc.

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.



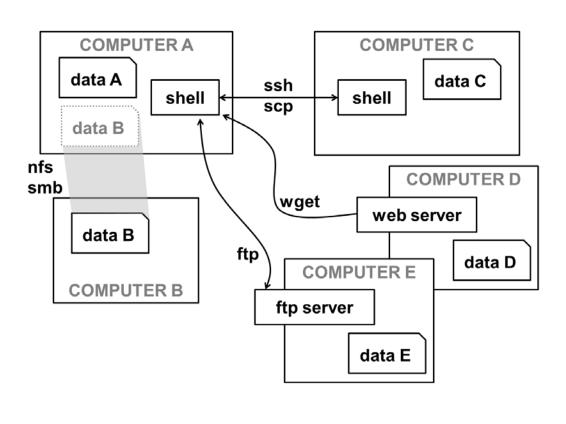
A cluster is just many computers together...

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



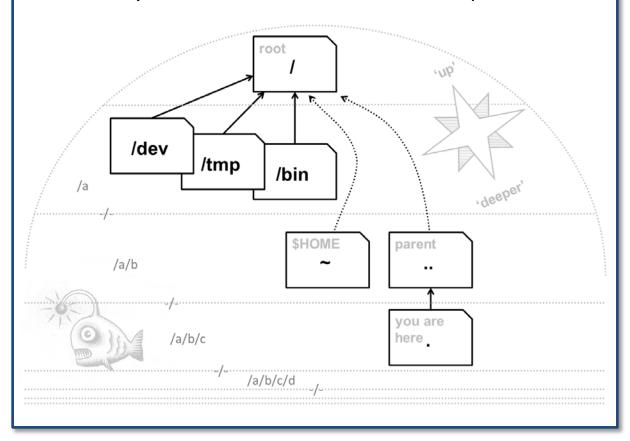
Moving data, or yourself

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.





CRUK cluster HPC @ CRUK CI



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)

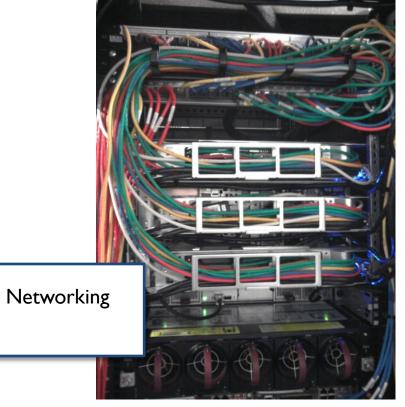
CRI IT & SC Help Desk

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

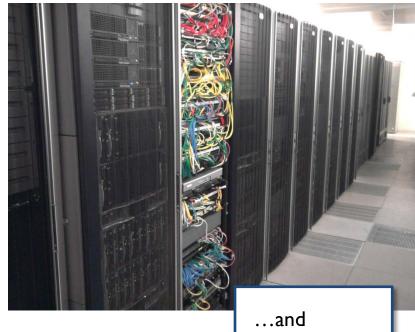
helpdesk-it@cruk.cam.ac.uk

01223 769600

A baffling technical diagram			C4	C5	C6	C7	C8	C9 1201/2010
	_	_	taeach side availathe	11 power societs each side available	12 power societs each side available	7 power socket seach side available	14 power socket seach side available	19 power sockets each side
			O porta avaliable	patch (rear) il pota avsibble	patch (var) 7 pot savalable	patch (ver) -10 ports available	paich (war) - 12 ports available	patch (mar) - 12 pots av
			tick (mar)	cable tidy (near)	cable fidy (rear)	cubie fidy(rear)	cable fidy (man)	cable tidy (mar)
			12 corts available	LET COMPANY	LET	prich (ver) -16 ports available	LM	LET
	cable tidy		FCsaltch A(mar) - 24 ports	UKCR-LBIP02	UK CR-LBIPOI	Ott heruitan), to but a service	MSACCOCOT (Controller)	UK OR LEKPOS
UK-CRILOSTO1	CR _Chater_3/50At (Red) - 7 pots the	2	FC saitch B (rear) – 24 ports	Lise	List	100	LTH	130
	code tidy		1.00			MSL4040-002	MSAG000-002	MSAID-052
UK-OR-LOSTIO2	CR_Chater_3750A2 (Red) - 7 pots time		100	163L4049 CO1	1652404B 002	100	LOW COMPANY OF THE PARK OF THE	
	code tidy		UK OR LSPSON	17			MSACOCO-COC	MSA2000004 (Corti
UK OR LOSTOS	bruhes (c)		un en			LD4	LD.	
	CRI_Cluster_2960 - 6 ports free		UNIOR LISTEDS	UKCR LARRO	UK CRI LAPRO		MSAGO COS	M2V2000002
NE-AGO-010	cable tidy		lan	1.00		LID.		
	um patch to Of UR2 - 2 ports free	21	MSA2000-012 (Controller)	un	un	LOT	UH-GRILSOLES	MSAG000006
UK-CREDYCON	patch to C4 UID - 14 parts free		MSAZ0000013	M09600-000	M29600-000	Long	MSAGO COS	MSAC000007
	code tidy		LON	UN	un	Los	LON	SEAGUIOU/
	patch to CS-10 ports free		LON	120	GH CO	M29600-00	Low	law .
	patch to C6 - 6 posts free Lor [Fibre channel patch (next)] to	UK-CR-ENIZO	MSACCOD-014	UKCR LARRO	UK CR LAPRO2	Loo	Las	ue
	cable tidy		LON	UNCHE DIFFECE	Of Circle Davies	Loui	LON	MSA2000-004 (Cort
	patch to C7 UK2 - 18 parts free		LOI	L21	un	LON	Lan	REACTED-CON (CON
	patch to C7 UID - 20 parts free	24	LON	LEA	iai	LON	LON	MSACODOOS
	Loo Cable tidy	2	Lon	us	un	MD 5600-001	Lan	un REACTOOLS
	brushes uzo [iwws.saitch (nor) 10 pot stime] uz		LIST COLUMN TO A STREET	uo	LID	Loo	LOD	MSA2000-006
	uzn jium svitch (nex) 3 parts finej uz	21	UK-CR LSFSO4	un	un	Lan	Lon	UT)
	Long power Lo		UNI CREASESSO	UII NG A00-015	UN NEACOCOT	Last	LOI MSAGO COF	un
uk at ekant	ure bruhes u	79 79 75 75 18 UK CIR-ENCOO	Uni NESAZODO COD (Controlar)	MG AGO O16 MG AGO O17	LIN MEACO-029 LIN MEACO-029 LIN MEACO-029	U W	ш	Lim
	patch to CB -12 ports free					un un un un	UN MESAGO COD UN MESAGO COD	LIN MSAKO-ONS LIN MSAKO-ONG
	ure cable tidy u		un					
	patch to C9-15 parts free		UN MSACODO10					
	um cable tidy u		UII					
	us brahes u	N	UNI MSAGODOH1	INS ARD-018	MSA60400	UNI	UN MISAGO DAD	INSARD-OUT
	uro patich to C1 - 7 ports thee	0	un	LID	Lo	un	ив	LID
	uro petich to C3 - 7 ports thee u	0	u o	UNICR-LIQUOT	UK CR-LSQ.02	ura	UNIOR LISOLO	UK-GR LSOLDA
	UTI Codie tidy U		ит	List Control of the C	Line Company	un	un	LITE
	Uni CR_Chater_375081 (Blue) - ? parts free		UTI	MS A00-019	MSAGOGOI	UTB	MSAGOOH	MSAID-068
UK OR ENION	cable tidy	us us us us uscer-cnoss	List Control of the C	u	u	to the same of the	List Control of the C	и
	CR_Cluster_375082 (Blue) - 7 ports free		us on careas	MS A60-020	MSA60632	to the same of the	MSAGO D42	MSAID-019
	cottle tidy			U	u	Lor	LO COMPANY	ur .
	Le brushes .		UK-CRILSESOI	MEA(0-021	MSAGOGGG	uk ca-picos	MSAGOOKI	MSAGO-GGO
	CRI_Cluster_2960 - 6 ports free		LI CONTRACTOR OF THE CONTRACTO	u	u	List Control of the C	Lis	ш
	Lid Cable Hidy	4	MSAG000-008 (Cortroller)	16 A00-022	MEAGOGOA	ш	MSACOOM	La .
	LO LO	0	La Company	u en	u	LO	LO COMPANY	u .
	pover		power	(3) power	u power	LO	pover	Domes.







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)

/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 "/groupdirs"
- 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/gentooalt/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 permodule basis
- Set, append, remove and delete environment variables such

as PATH, MANPATH & LD_LIBRARY_PATH

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

- I. 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

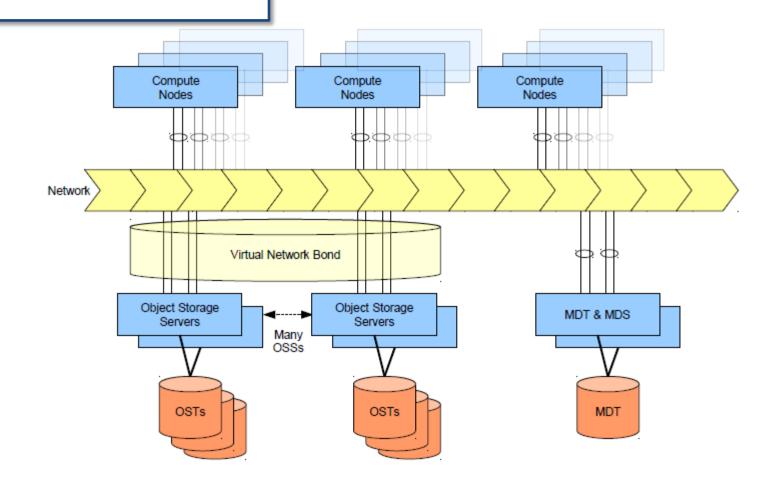
- Deployed in 7 out of 10 most powerful supercomputers
- POSIX compliant

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 ~ \$ 1fs 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
```