

Introduction to High Performance Computing and COSMA

Alastair Basden

and the COSMA team

(Peter Draper, Richard Regan, Paul Walker, Mark Lovell, John Helly, and others)

cosma-support@durham.ac.uk

How to get help:

- cosma-support@durham.ac.uk
- Visit OCW220
- Alastair Basden
 - Technical Manager of COSMA
- Peter Draper
 - All-round software/hardware genius
- Richard Regan
 - DiRAC training manager, COSMA support
- Paul Walker, Mark Lovell
 - COSMA support
- John Helly
 - Cosmology software expertise, data storage expert

High performance computing

- What is HPC?
 - Use of parallel processing to run large applications
 - Typically applied to systems >100 TFLOPS
 - Aggregated computing power
 - More than can be obtained from a desktop
 - Used for solving large problems
- Cloud computing is not usually HPC
 - Typically only a single computer in the cloud is used



The TOP500

- Prestigious list of the World's most powerful supercomputers, released 6-monthly

- Includes some UK sites

- Also:

- Green TOP500

- Best performance/Watt

- I/O TOP 500

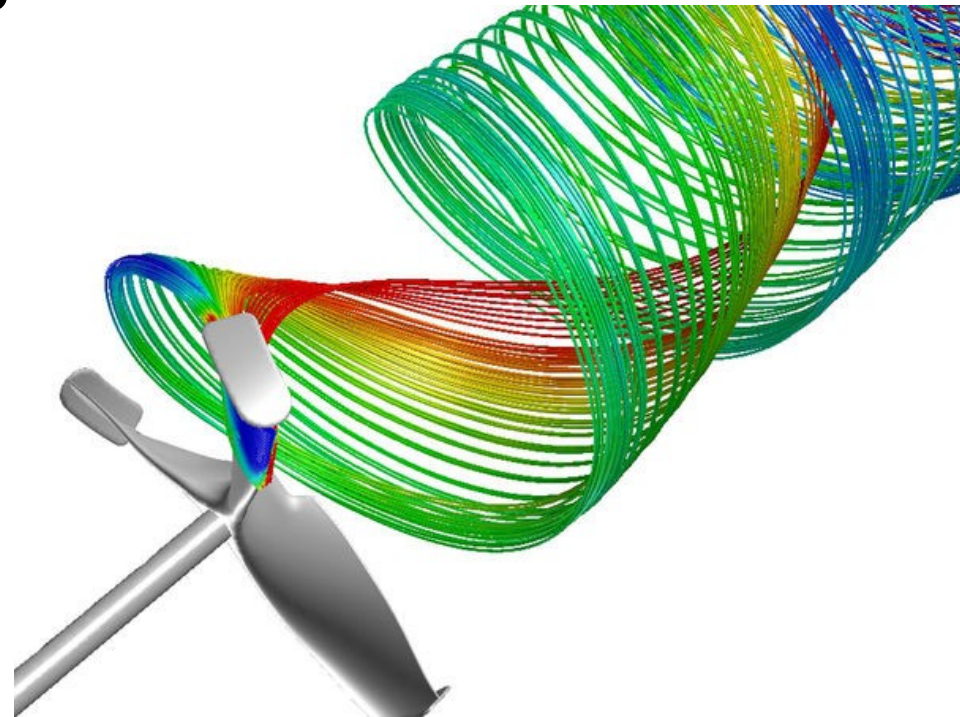
- Best I/O



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,299,072	415,530.0	513,854.7	28,335
2	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
6	HPC5 - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, Dell EMC Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252
7	Selene - DGX A100 SuperPOD. AMD EPYC 7742 64C 2.25GHz.	272,800	27,580.0	34,568.6	1,344

HPC use cases

- Galaxy/cosmology simulation
- Weather simulation
- Artificial intelligence
- Fluid dynamics modelling
- Materials modelling
- Genetics
- etc



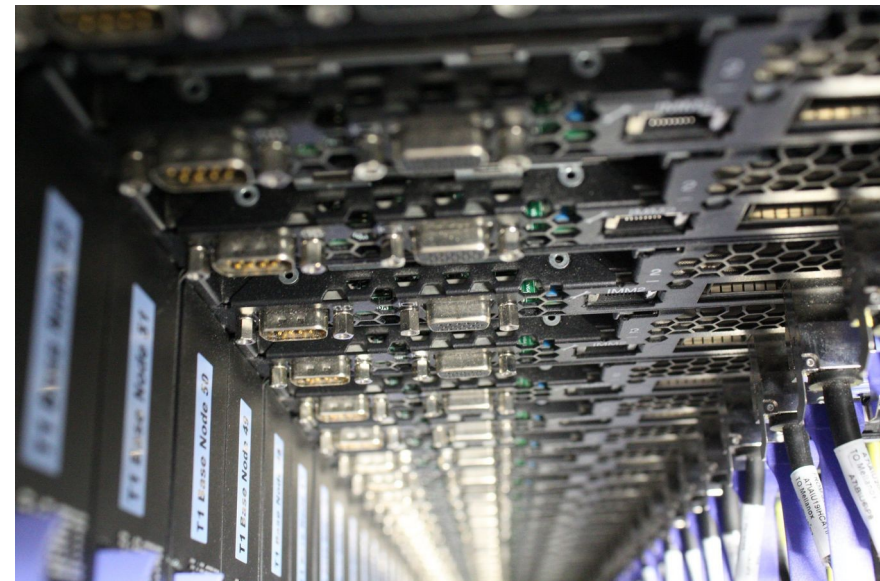
HPC in the UK

DiRAC

- Recently announced:
 - Bristol Isambard AI system
 - Pre-Exaflop Edinburgh system
 - Cambridge DAWN system
- Archer2 - Edinburgh, UK national facility (EPSRC)
- DiRAC, UK national facility (STFC)
 - Of which COSMA is part
- Met Office
- ECMWF
- Tier-2 systems
 - E.g. BEDE (at Durham)
- PRACE (European collaboration)
- Tier-3 systems
 - Local HPC resources, e.g. Hamilton

DiRAC

- A tier 1 HPC facility for the STFC theory community
 - Astronomy, cosmology, nuclear and particle physics
- 4 DiRAC sites:
 - Cambridge
 - Leicester
 - Durham
 - Memory intensive system – lots of memory per core
 - (note, memory is expensive, which makes COSMA unique)
 - Edinburgh
 - Extreme scaling system – maximum compute power
- DiRAC 2 (COSMA5) in 2012
- DiRAC 2.5 (COSMA6) in 2016 (retired in 2023)
- DiRAC 2.5x (COSMA7) in 2018 and 2019
- DiRAC 3 (COSMA8) in 2021 and 2023



Components of HPC

- Login nodes
 - For user interaction
- Compute nodes
 - For doing the work
- High performance fabric
 - For linking compute nodes and storage
- Storage
 - For storing data
- Facilities
 - A data centre, cooling, power
- User management
 - Personal details, etc
- Admin nodes
 - Job submission
 - Software modules
 - Logging
 - Authentication
 - etc



The COSmology MACHine

- The Durham DiRAC node
 - Now in its 8th generation
 - COSMA5 and 7 also available for use
 - And DINE (experimental 24-node cluster)



COSMA5

- A Durham-only facility
 - Including collaborators
 - Was a DiRAC facility until 2018
 - Now just ICC (and collaborators)
 - If you are part of a DiRAC project, please try to use COSMA7/8
- Arrived in 2012
- ~300 nodes
 - 16 cores per node (2 CPUs)
 - 128GB RAM per node
 - Diskless
- 2x login nodes: login5a, login5b



COSMA7

- A DiRAC facility
- Arrived in 2018 and 2019
- 452 nodes
 - (originally 147 nodes, April 2018, 300 nodes by end 2018, 452 nodes March 2019)
 - Each 2x 14 core Xeon CPUs (circa 2018)
 - 512GB RAM (18GB/core)
 - Originally 768GB RAM
- Also, mad01:
 - 3TB RAM, 28 cores
- mad02:
 - 1.5TB RAM, 4x CPUs,
 - 28 cores each (112 cores)
- mad03:
 - 6TB RAM, 56 cores
- login7a, login7b, login7c



DIRAC3: COSMA8

- £5M arrived in late 2020
 - System installed mid 2021
- £3m in late 2022
 - Operational in summer 2023
- Conventional CPU architecture
 - not GPU/FPGA
- AMD Rome/Milan processors
- Memory intensive
 - 528TB RAM
- 528 nodes each with 128 processors and 1TB RAM
- Some “fat” nodes: 4TB RAM
- Some GPU nodes (NVIDIA and AMD)



DINE

- 24-node system
 - With NVIDIA DPU accelerators
 - 32 cores/node
 - 512GB RAM
- Useful for code testing, and exploring new paradigms

COSMA summary

	COSMA5	COSMA7	COSMA8
Nodes	300	452	528
Cores/node	16	28	128
Cores	4800	11200	67,584
Memory/node (GB)	128	512	1024
Memory/core (GB)	8	18	8
Total memory	38TB	230TB	528TB
Login nodes	2	3	2
Operating system	CentOS 7.9	CentOS 7.9	CentOS 7.9

COSMA network fabric

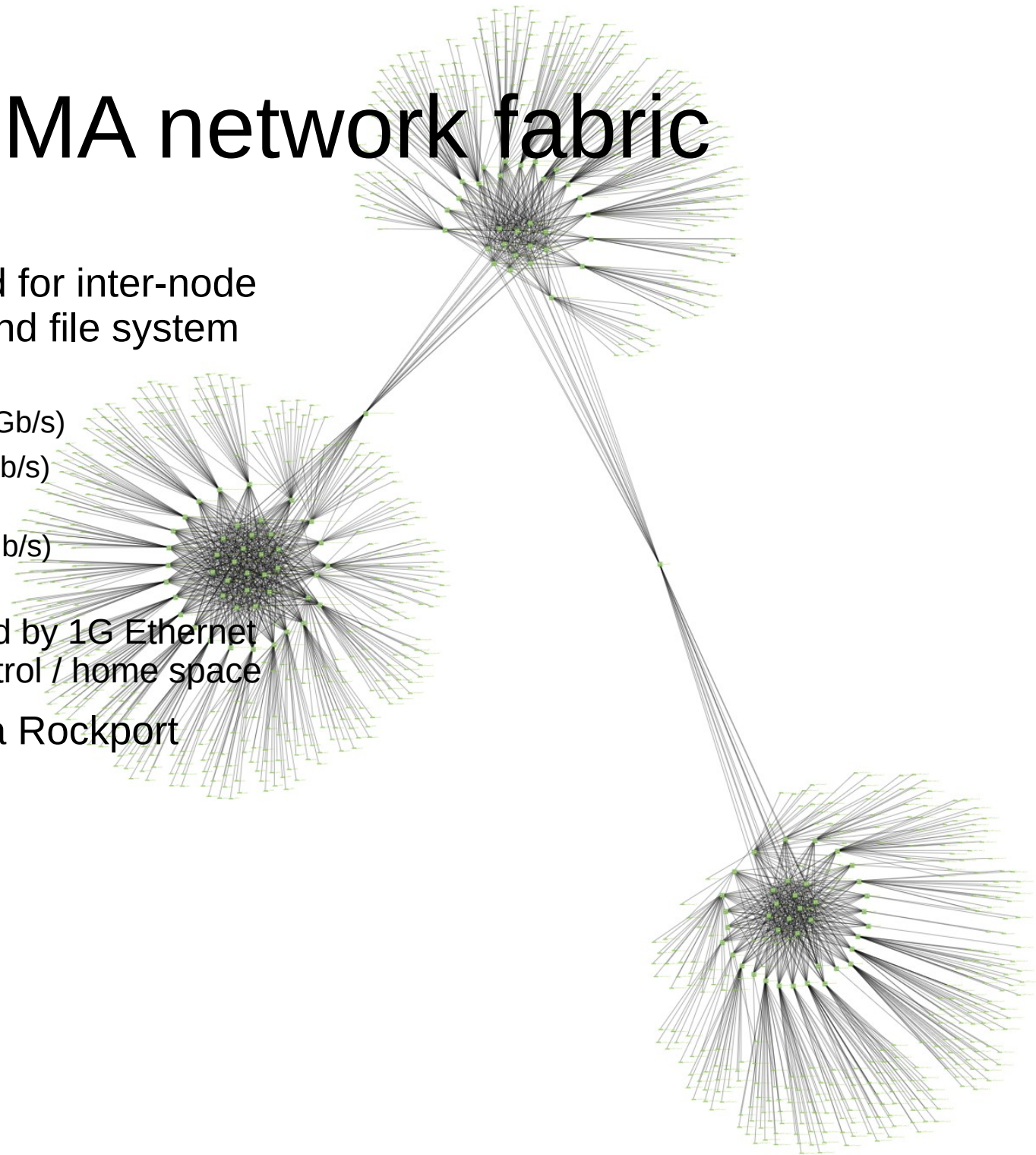
- COSMA uses Infiniband for inter-node communication (MPI) and file system access

- FDR10 for COSMA5 (40Gb/s)
- EDR for COSMA7 (100Gb/s)
 - 2:1 blocking ratio
- HDR for COSMA8 (200Gb/s)
 - non-blocking

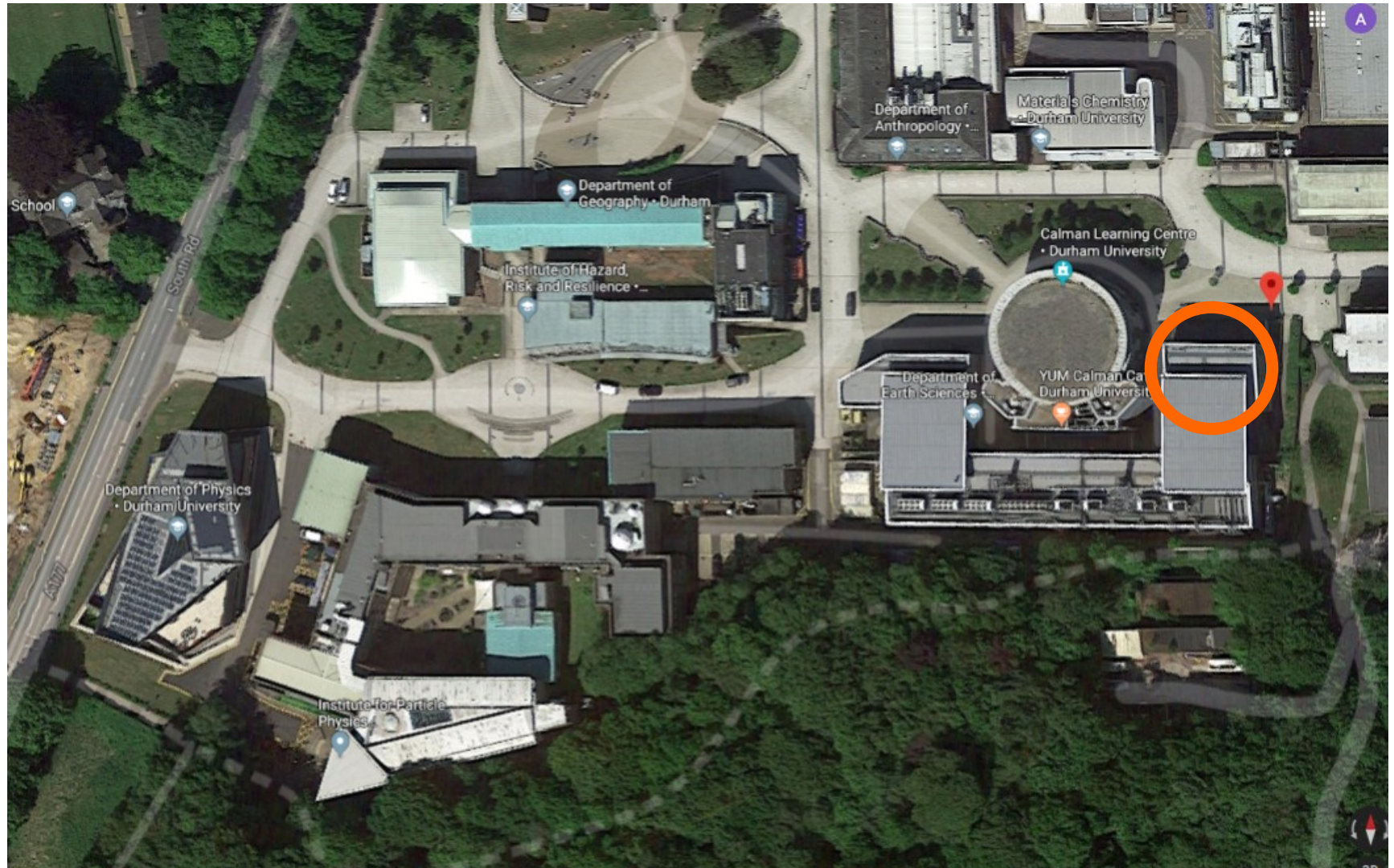
- All nodes also connected by 1G Ethernet for communication / control / home space

- Half of COSMA7 uses a Rockport fabric:

- 6D torus, switchless
- Experimental
- Ethernet, 100G

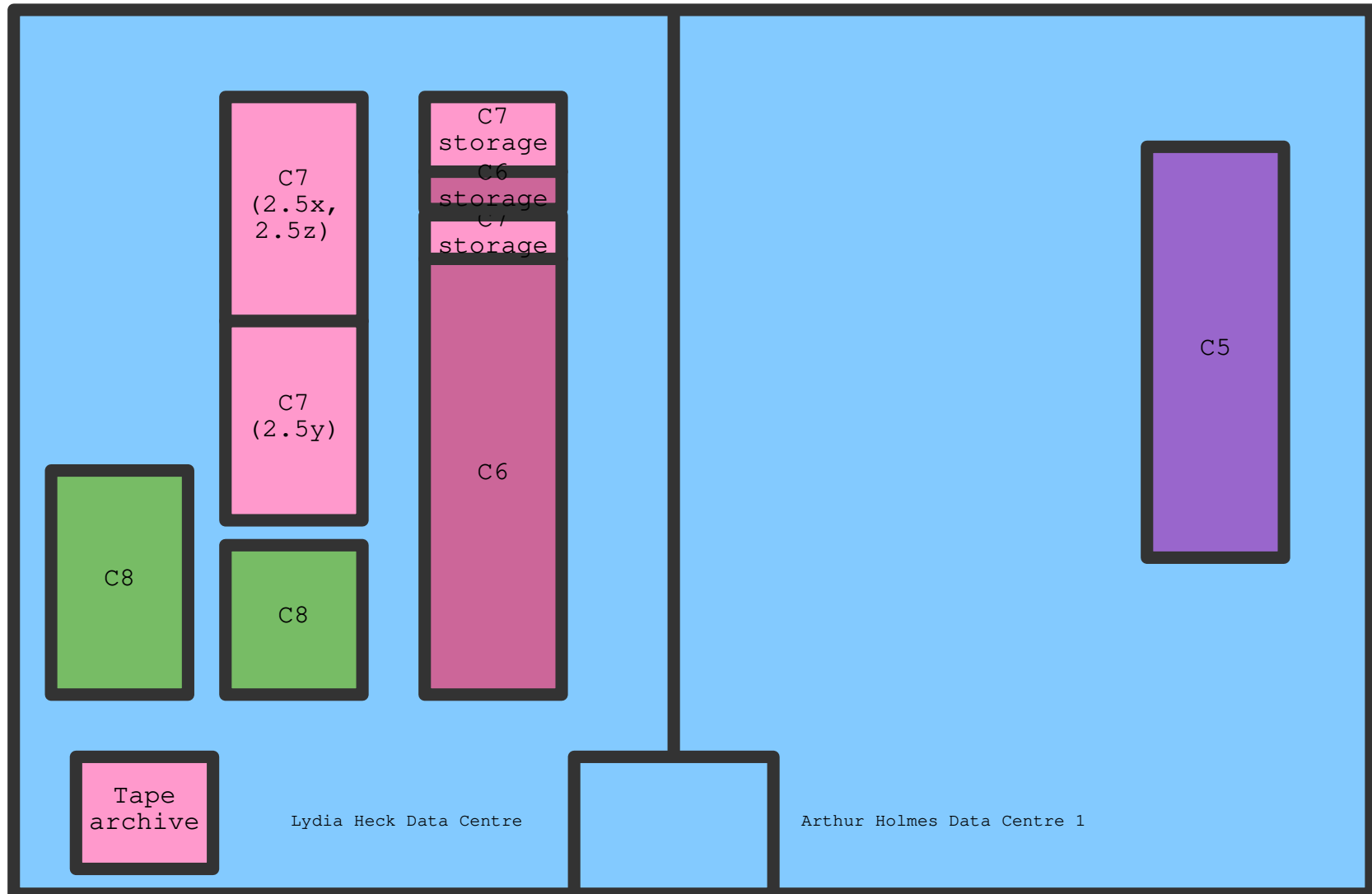


Arthur Holmes Data Centre



Arthur Holmes Data Centre

- Location of COSMA



Accessing COSMA

- First, you need a SAFE account
 - (Service Administration From EPCC)
 - (Used for all DiRAC facilities)
 - See <https://www.dur.ac.uk/icc/cosma/support/account>
 - In summary:
 - <https://safe.epcc.ed.ac.uk/dirac/>
 - Create an account (durham email, not personal email)
 - Upload an ssh key
 - Select your project
 - e.g. hpcicc or dp004 (ask your supervisor)
 - Select COSMA (not COSMOS)
 - Wait...



SAFE for DIRAC
Service Administration from EPCC



DIRAC SAFE Login

Welcome to the [DIRAC](#) SAFE. Through the SAFE, you can apply for an account on our high-performance computing systems, and perform other administrative tasks relating to your use of our machines.

Login using UKAMF or

Email or Dirac-global-id or Wiki name: *

Password *

Login [Forgot password?](#)

[Create an account](#)

As part of its normal function, when you log in the SAFE will install a temporary session cookie that will be removed when you log off or close your browser. If you do not wish this cookie to be set, disable cookies in your browser settings.

DIRAC SAFE Signup

This is the DIRAC SAFE.

Registration form

This form is to sign-up for a new SAFE account. If you already have an account then [login](#). If you have forgotten your password, then [recover your password](#).

Fields marked in **bold** ★ are mandatory.

We request CareerStage information for demographic analysis of our users.

Any SSH key you register here is a default for when new login accounts are requested. However this does not automatically mean that it will be installed when the account is created. See the individual system documentation for details of their policy on SSH keys.

All information supplied is held and processed in accordance with our Personal Data and Privacy Policy. You can find full details [here](#).

Email Address ★	<input type="text" value="name@example.com"/>
Your Nationality ★	<input type="text" value="United Kingdom"/>
Title (Mr,Mrs,Dr etc.)	<input type="text"/>
First Name ★	<input type="text"/>
Last Name ★	<input type="text"/>
Institution for reporting ★	<input type="text" value="Anglia Ruskin University"/>
Department	<input type="text"/>
Phone number (include International code e.g. +44 for UK)	<input type="text"/> + followed by numbers and spaces
Opt out of user Emails ★	<input type="checkbox"/>
Address Line 1	<input type="text"/>
Address Line 2	<input type="text"/>
Address Line 3	<input type="text"/>
Address Line 4	<input type="text"/>
Town/City	<input type="text"/>
Postcode	<input type="text"/>
Country	<input type="text" value="Not Selected"/>
SSH Public key	<input type="text"/> <input type="button" value="Browse..."/> No file selected.
Career stage	<input type="text" value="Not Selected"/>
HPC experience ★	<input type="text" value="No HPC experience"/>

Register

DIRAC SAFE Signup

This is the DIRAC SAFE.

Registration form

This form is to sign-up for a new SAFE account. If you already have an account then [login](#). If you have forgotten your password, then [recover your password](#).

Fields marked in **bold** ★ are mandatory.

We request CareerStage information for demographic analysis of our users.

Any SSH key you register here is a default for when new login accounts are requested. However this does not automatically mean that it will be installed when the account is created. See the individual system documentation for details of their policy on SSH keys.

All information supplied is held and processed in accordance with our Personal Data and Privacy Policy. You can find full details [here](#).

Email Address ★	<input type="text" value="diracsafe@mailinator.com"/>
Your Nationality ★	<input type="text" value="United Kingdom"/>
Title (Mr,Mrs,Dr etc.)	<input type="text" value="Dr"/>
First Name ★	<input type="text" value="Dirac"/>
Last Name ★	<input type="text" value="User"/>
Institution for reporting ★	<input type="text" value="University of Durham"/>
Department	<input type="text"/>
Phone number (include International code e.g. +44 for UK)	<input type="text"/> + followed by numbers and spaces
Opt out of user Emails ★	<input type="checkbox"/>
Address Line 1	<input type="text"/>
Address Line 2	<input type="text"/>
Address Line 3	<input type="text"/>
Address Line 4	<input type="text"/>
Town/City	<input type="text"/>
Postcode	<input type="text"/>
Country	<input type="text" value="Not Selected"/>
SSH Public key	<input type="text"/> <input type="button" value="Browse..."/> id_ed25519.pub
Career stage	<input type="text" value="Not Selected"/>
HPC experience ★	<input type="text" value="No HPC experience"/>

Register



SAFE for DIRAC services
Service Administration by EPCC



User Access Agreement

Please read our acceptable use policy.

Before accepting the Terms and Conditions, please note: you can change any of the details you have input by clicking your browser's BACK button and then editing them. You can also change them later by returning to this website. No specific copy of these Terms and Conditions will be filed under your name, but you can look at them at any time by going to

https://safe.epcc.ed.ac.uk/dirac/safe_acceptable_use.jsp.

These Terms and Conditions are offered only in English.

You may read the terms of the agreement [here](#).

I accept the Terms and Conditions of Access

After email arrives, click the link to get here...



SAFE for DIRAC services
Service Administration by EPCC



Please set a password for use with the SAFE

You are setting a password for your account:

- diracsafe@mailinator.com
- dc-user1
- Duser

Passwords must be at least 8 characters long (not counting repeated characters and character sequences). Passwords must contain at least 6 different characters.

New Password: ★

New Password (again): ★

Change

Cancel/Logout

[Your details](#)[Service information](#)[Login accounts](#)[Help and Support](#)

DIRAC

SAFE for DIRAC services
Service Administration by EPCC



Welcome to the DIRAC Administration Website

You can use this site to view your project details, current budgets and resource allocations and see usage reports.
If you are a project PI or manager you can change time and resource allocations.

If you have any problems using the system, please submit a support query.

Regards,

The DIRAC Support Team

[Continue](#)



SAFE for DIRAC services

This is the DIRAC SAFE. It is a web-site used to administer the DIRAC HPC services.

You are currently recorded as a new user of the SAFE and will see a restricted view of the available information until you have been accepted into a project. To join a project you should apply for a login account on the HPC service using [this link](#) or via the **Login accounts** menu. You will need to select the project you are applying for. Once your login account has been approved the login account will be created.

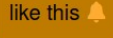
All of the functions of the SAFE are available from the menus at the top of the page.

Use the **Your details** menu to view and update the information we hold about you or to change the settings of your SAFE account.

Use the **Service Information** menu to view information about the service or to generate reports from our database.

Use the **Login Accounts** menu to apply for an account on DIRAC machines. All accounts need to be part of a funded project so this request will need to be approved by a manager of the project you select before the account can be created. If you already have an account you can also use this menu to view the status of, and make changes to, your accounts.

The **Help and Support** menu gives access to the help-desk and contains links to SAFE guides and documentation.

If a menu looks  it indicates a pending request that needs your attention.

[Your details](#)[Service information](#)[Login accounts](#)[Help and Support](#)

DIRAC

SAFE for DIRAC services

Service Administration by EPCC



DIRAC Login account Request

This form is for requesting new login accounts. If you wish to add additional access to an existing account, select that account from the navigational menu at the top of the page to see the available options.

Please note that when you apply to join a project some of the personal data (such as your name and email address) that we hold about you will be shared with managers of the project to allow them to process the application. If your application is approved then the project managers will continue to have access to this data while you remain a member to allow them to manage their project effectively.

Project ★



DIRAC Login account Request

Project: hpcicc - Durham

Your personal details stored in the SAFE will also be accessible to the operators of that service if this request is approved

New account policies

If a check-box does not appear beside a machine then the project you selected is allowed to use the machine but one of the policies that apply to the machine is preventing you from applying.

A cross will be marked against the policy that is preventing you from applying.

You would also be able to enable access to this machine by updating your account to meet any policy marked with an arrow.

Select a machine for the login account

Select	Machine	Type	Description	Policies
<input checked="" type="radio"/>	cosma: The Durham COSMA machine		The Durham COSMA machine	<ul style="list-style-type: none"> Accounts named after dirac-id ✓ Only one account per person is allowed ✓ Users must have a public key registered to use the machine ✓

[Next](#)



DIRAC Login account Request

This form is for requesting new login accounts. To request additional access for an existing account, select it from the navigation menu at the top of the page

Your username will be visible to other users on the system

This machine support ssh key authentication. You can upload a public key to use here.

A SSH public key is required to use this machine.

Your default key will automatically be added if no other key is specified

Requested username ★

dc-user1

SSH public key

ssh-rsa AAXXYZ...

Browse...

id_ed25519.pub

Request

Then wait...

- While the account is first authorised...
- And then created...
- Finally, you will receive an email!



Your details

Service information

Login accounts

Help and Support

DIRAC

SAFE for DIRAC services
Service Administration by EPCC



New Account requested

Your account dc-user1@cosma has been requested, the project's principal investigator or a project manager will be informed of your request and you will be contacted once your machine account has been activated.

Accessing COSMA...

- Login nodes
 - This is where you do your work!
 - Prepare scripts
 - Edit code
 - Compile code
 - Inspect results
 - A shared facility
 - Usually with extra RAM
 - Try not to run jobs here
 - Submit them to a “queue” instead
 - Via ssh:
 - ssh USERNAME@login.cosma.dur.ac.uk
 - ssh USERNAME@login5.cosma.dur.ac.uk
 - ssh USERNAME@login6.cosma.dur.ac.uk
 - ssh USERNAME@login7.cosma.dur.ac.uk
 - ssh USERNAME@login8.cosma.dur.ac.uk



Credit: fotolia.com

#163700234

ssh on COSMA

- Authentication requires an SSH key:
 - A key has 2 parts
 - Private part (keep this very safe!)
 - Please put in a passphrase when you generate it
 - A public part (give this to COSMA – or anything else)
 - Uses “public key cryptography”
 - When you try to connect:
 - COSMA will use the public key to generate a “challenge” - an encrypted message
 - Only the private key can decode this
 - Your computer then sends the correct response to COSMA
 - Access is then granted
- And a password
 - Which you can change with the passwd command

Generating an ssh key

- `ssh-keygen -t ed25519 -C "unique name to identify this key."`
 - This will ask for a passphrase
 - Please use one – this protects your private key
 - This will create (by default):
 - `id_ed25519` (private key – keep this safe, like a physical key)
 - `id_ed25519.pub` (public key – upload this to SAFE)
 - We will then add your `id_ed25519.pub` key to the authorized keys file in COSMA
 - You can use the same public key on any other servers that you use
 - e.g. `mira.dur.ac.uk`, `hamilton.dur.ac.uk`, your desktop, etc.
- Instructions at www.dur.ac.uk/icc/cosma/support

ssh key example

```
ssh-keygen -t ed25519 -C cosmakey
Generating public/private ed25519 key pair.
Enter file in which to save the key (/home/ali/.ssh/id_ed25519):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/ali/.ssh/id_ed25519
Your public key has been saved in /home/ali/.ssh/id_ed25519.pub
The key fingerprint is:
SHA256:zSCTo1ET81TIZb2DhUByq37pEQkwpOM/M2UQ6Yqx6Uw cosmakey
The key's randomart image is:
```

```
+--[ED25519 256]--+
|
|  . = . * + = = + o
|  .. = Ooo. o
|  o.o * + o .
|  .. .. + * = . o
|  = ... + S o .
| +E.. + o
| + = . +
| o + o .
|
|  .
+-----[SHA256]-----+
```

Could have different files for different keys

A passphrase (password) was entered here

Upload this file to SAFE

After login...

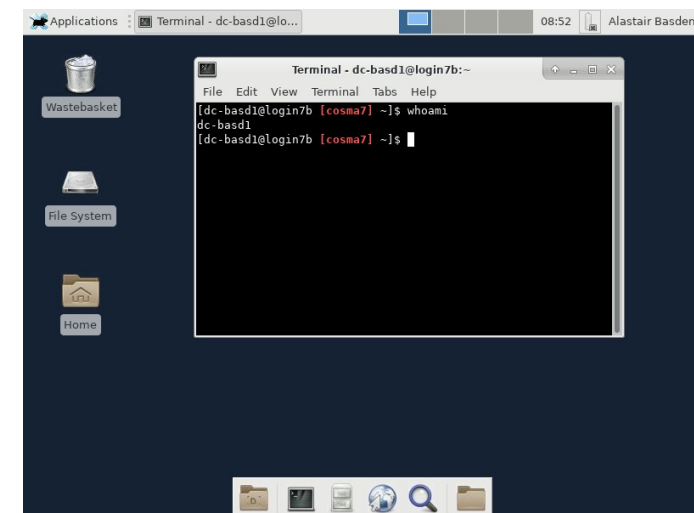
- Your terminal will now be redirected to COSMA
 - Anything you type will be interpreted by COSMA (not your PC)
 - Use common Unix commands
 - e.g. ls, pwd, mkdir, cp, etc
 - Start text editors
 - e.g. vi, emacs etc
- If you want to access COSMA from other computers, upload the public key to SAFE

Some useful commands

- `id`
 - Gives you User ID, and the groups you are in
- `finger USERNAME`
 - Information about a user
- `whoami`
 - Your USERNAME
- `w` and `who`
 - see who else is logged on at the moment
- `top` and `htop`
 - see who is hogging resources!
- And don't forget ***tab-completion***

Graphical access

- If you want graphical access:
 - x2go is the best option
 - Install an x2go client, and use it to connect to a login node.
 - Alternatively, use X11 forwarding:
 - `ssh -X USER@login.cosma.dur.ac.uk`
 - (or `-Y` if that doesn't work well – less secure)
 - You can then start remote programs which will display windows locally, e.g.
 - text editors
 - plots of your data
 - movies etc are not recommended unless you have an excellent connection
- Note: X11-forwarding is bandwidth heavy
 - Requires a good network connection to be useable
 - You might struggle from home if on a 1MBit connection
 - To edit files over a poor connection, use a terminal based editor
 - `vi`, `emacs -nw`, `nano`
- x2go is faster

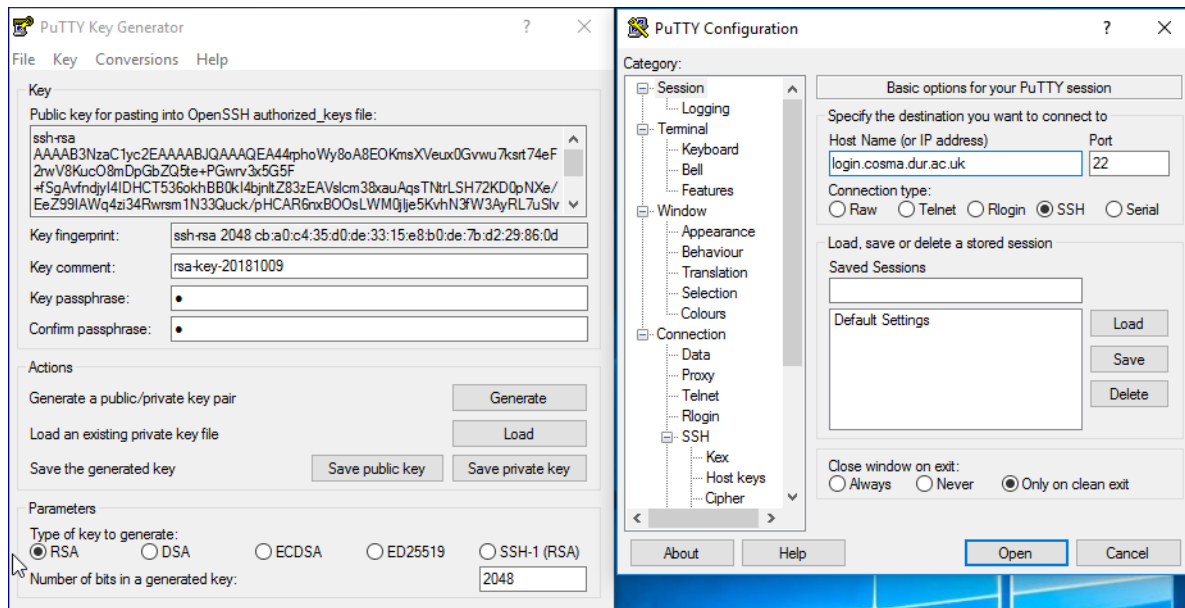


COSMA passwords

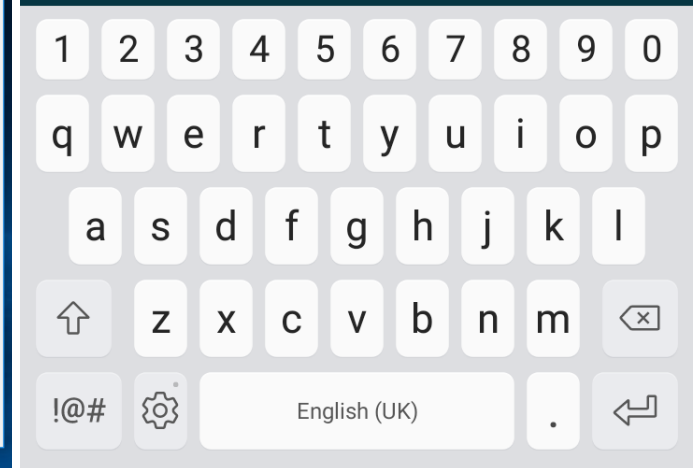
- Are also used to authenticate with the cosma websites
 - (to access usage statistics, Jupyter hub, etc)
- These are different from your ssh key passphrase
 - We can reset your password
 - If you lose your ssh passphrase, you will need to generate a new key (and upload to SAFE)

ssh from other platforms

- On Windows, best to use the cmd prompt.
 - Can also use apps like e.g. Putty
 - You will need to run puttygen first to create the ssh key pair
 - Or mobaXterm, or others...
 - Or WSL (Windows Subsystem for Linux)
- On Android, use e.g. JuiceSSH
 - You will need to generate an ssh key pair (using the app)



```
Last login: Tue Oct 9 10:19:27 2018 from host-92-14-183-32.as43234.net
[dc-basd1@cosma-m [cosma7] ~]$ ls
AdaJavaUI.bin          test15113.txt
atempo.lic            test15578.txt
bsubSerial.bsub       test18216.txt
bsubSerial.bsub~     test20165.txt
dead.letter           test22149.txt
Desktop               test29386.txt
Documents             test31771.txt
Downloads             test5379.txt
getHomeDirs.py        test5436.txt
getHomeDirs.py~     test5504.txt
installADA.exe        test5959.txt
installADA.gz         test7637.txt
installADA.tgz        test8431.txt
installer.properties test8822.txt
local                 test_err
makeLargeFiles.py     test.out
makeLargeFiles.py~   testScript.py
module.txt            testScript.py~
Music                 tmp2.txt
parallel_tasks        tmp3.txt
Pictures              tmp3.txt~
Public                tmp4.txt
root@console51        tmp5.txt
rsync-3.1.0           tmp5.txt~
rsync-3.1.0.patch     tmpTest.txt
rsync-3.1.0.tar.gz    tmp.txt
rsyncData.py          tmp.txt~
rsyncData.py~        userdirs.txt
setup_ada_admin_x86_64.exe users.txt
Templates             test~
test11062.txt         valgrind-3.13.0.tar.bz2
test12984.txt         Videos
[dc-basd1@cosma-m [cosma7] ~]$
```



Typical workflow

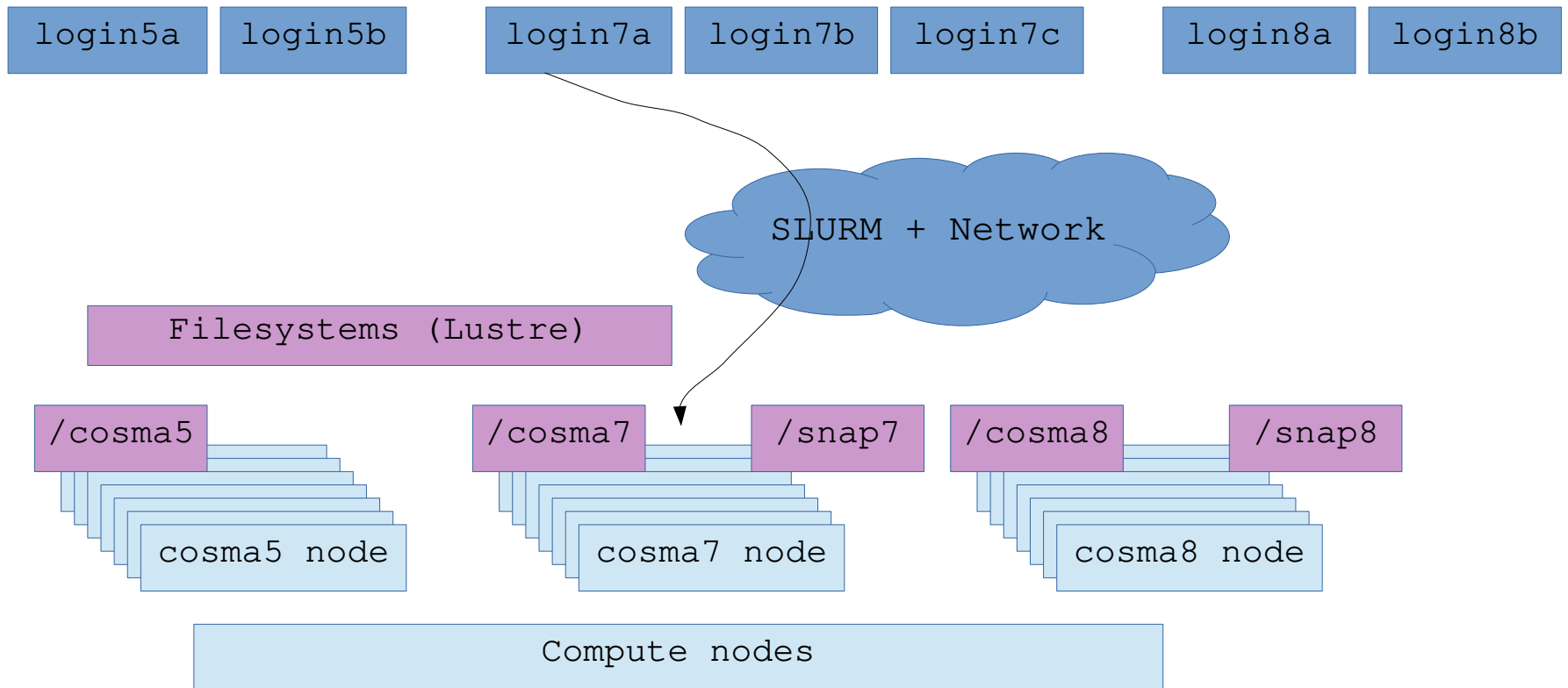
- Log in to a login node
- Edit files and scripts
 - If you have a particular editor that isn't available, or are on a slow connection and want a graphical editor:
 - consider using sshfs to mount your COSMA file system:
 - `cd && mkdir mnt`
 - `sshfs USER@login.cosma.dur.ac.uk:/cosma/home/PROJECT/USER ~/mnt`
 - `ls mnt/` locally will then show your COSMA homepage files
 - Useful options for sshfs include `-o reconnect,ServerAliveInterval=15,ServerAliveCountMax=3`
- Submit jobs to the batch queue (see later)
- Monitor jobs if necessary
- If you need to access an internal webpage on COSMA, e.g. jupyter hub, either start firefox over x2go, or forward the ssh connection from your desktop:
 - `ssh -D 1234 USER@login7a.cosma.dur.ac.uk`
 - `chromium-browser --proxy-server="socks5://localhost:1234" https://webpage.on.cosma`
 - `ssh -L 1234:localhost:80 USER@login7a.cosma.dur.ac.uk`
 - And then visit `http://localhost:1234`

COSMA login nodes

- Currently:
 - COSMA5 has 2 nodes: login5a, login5b
 - COSMA7 has 3 nodes: login7a, login7b, login7c
 - COSMA8 has 2 nodes: login8a, login8b
- All login nodes offer access to all facilities
 - File space/storage, job queues, libraries, tools, compilers, etc
- In general, please use COSMA7 or 8 for DiRAC projects, and use COSMA5 for internal Durham projects
 - Ask (us or supervisor) if not sure
- `login.cosma.dur.ac.uk` and `login5.cosma.dur.ac.uk` (round-robin allocation to 5a, 5b)
- `login7.cosma.dur.ac.uk` (round-robin allocation to 7a, 7c)
- `login8.cosma.dur.ac.uk` (round-robin allocation to 8a, 8b)

Graphical COSMA

`login[5,6,7,8].cosma.dur.ac.uk`



HPC storage

- 3 main types of storage:
 - Home space (10GB)
 - /cosma/home/
 - Backed up nightly
 - Bulk storage / data space (5-10TB)
 - e.g. /cosma7, /cosma8
 - Scratch storage (unlimited)
 - e.g. /snap7, /snap8
 - No redundancy
 - You may lose data
 - Use with care

COSMA file system

- Your home folder will be in
 - /cosma/home/PROJECT/USERNAME
 - PROJECT is probably durham or dp004
 - When you first log in, typing “pwd” will show you where you are
 - You have a 10GB quota
- You will also have data space at some of:
 - /cosma5/data/PROJECT/USERNAME (10TB/2.4PB)
 - /cosma7/data/PROJECT/USERNAME (5TB/3.5PB)
 - /snap7/scratch/PROJECT/USERNAME (Unlimited/440TB)
 - /cosma8/data/PROJECT/USERNAME (10TB/14PB)
 - /snap8/scratch/PROJECT/USERNAME (Unlimited/1PB)
- Data space is optimally connected to the corresponding COSMA
 - Reading /cosma5/ from COSMA5 will be faster than from COSMA7
 - /snap7, /snap8 space is temporary storage to use within a single run or for a short time period
 - e.g. for restart points
 - Fast SSDs, at one point was the fastest storage in Europe
 - /cosma5 is not available from the COSMA6/7 compute nodes, and vice versa
- /cosma/local/ is where tools and libraries are located

File system quotas

- quota (for home space):

```
Disk quotas for user dc-basdl (uid 20957):
  Filesystem      blocks quota  limit  grace  files  quota  limit  grace
172.17.170.16:/export/vol1 2256360 10485760 30000000 1208 2000000 2200000
```

- c7quota (for /cosma7, on a C7 login node):

```
Quota for dc-basdl
Filesystem      usage      quota      limit      files      quota      limit
-----
/madfs          0MB         0MB         0MB         0           0           0
/cosma7         0.00390625MB 0MB         0MB         1           0           0
/snap7          0MB         0MB         0MB         0           0           0
```

- c5quota (for /cosma5, on a C5 login node):

```
Quota for dc-basdl
Filesystem      usage      quota      limit      files      quota      limit
-----
/gpfs           0GB         0GB         0GB         0           0           0
/cosma5         33.4439GB  5120GB      5200GB      23          0           0
```

- c6quota (for /cosma6, on a C6 node):

```
Quota for dc-basdl
Filesystem      usage      quota      limit      files      quota      limit
-----
/cosma7         0.00390625MB 0MB         0MB         1           0           0
/cosma6         31.9992GB  0MB         0MB         17          0           0
```

- Running “quota” should tell you all this information

Files

- Your quota is not just restricted to total storage used
 - The number of files is also important
 - Each file uses a single “inode”
 - Metadata about a file, e.g. name, creation time, access rights, etc
 - Total number of inodes is limited
- Lots of small files is BAD
 - If you need this, please consider tarring up files, or concatenating them during writing
 - Your codes will run better if you do your I/O properly

Over-quota

- If you go over quota:
 - You will be sent a daily email reminder
 - You have a soft limit and a hard limit
 - You can go over your soft limit for up to 7 days
 - You cannot go over your hard limit
 - After 7 days, you will not be able to write files until you get back under quota
 - If you really need more quota, try asking!

Group/project quotas

- Groups/projects also have quotas
 - If a project goes over quota, no one will be able to write files
 - To check group quotas:
 - `c7quota -g durham`
 - Worth noting: If you are within your quota, but cannot write files, check the group quota

What to put where...

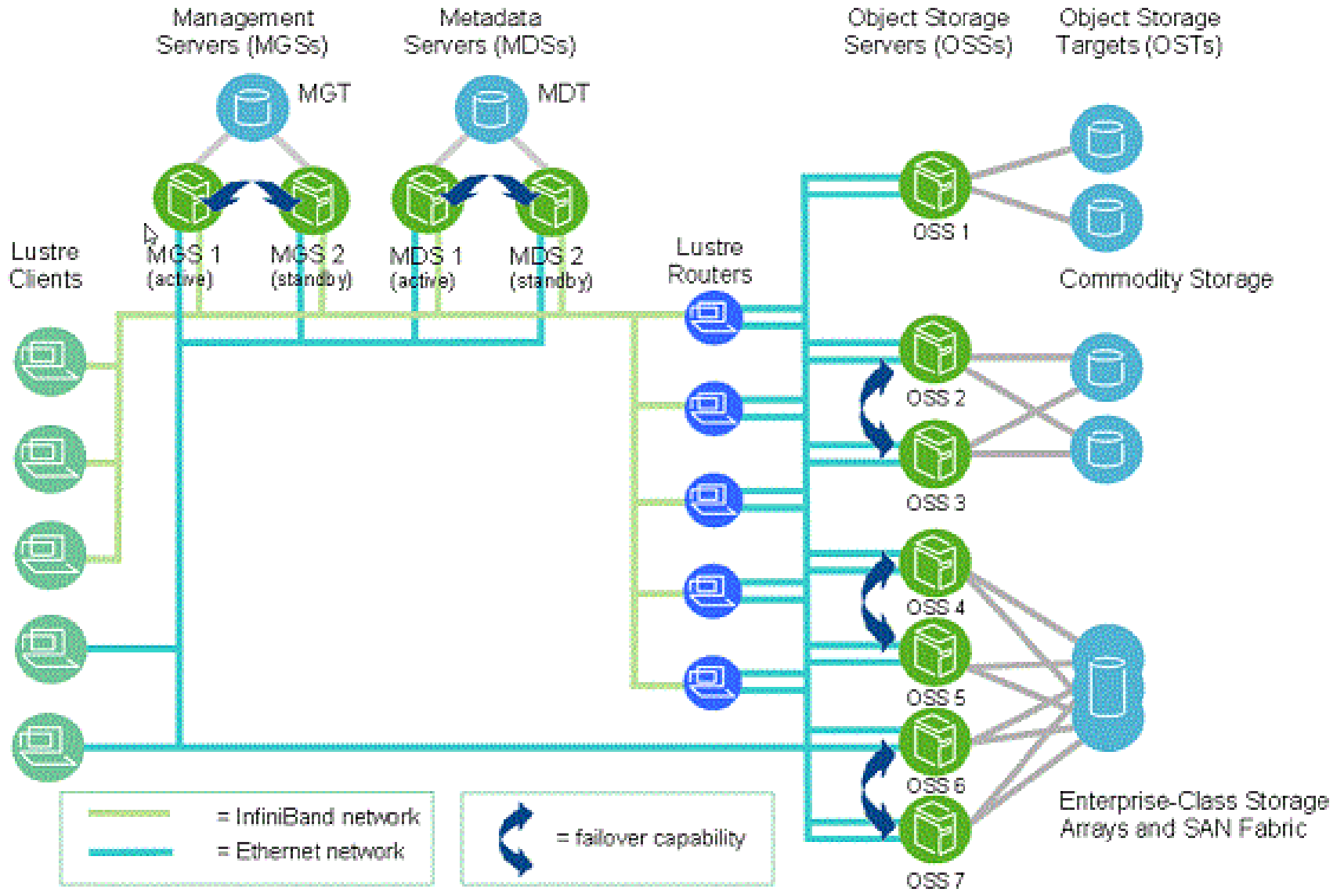
- Use `/cosma/home/` for scripts, code, self-compiled libraries, python modules etc.
 - This is backed up
 - No data files from runs
 - No log files from runs
- Use `/cosma[5,7,8]/data/` for input and output data produced by your runs
 - This can be archived to tape (upon request)
 - Not backed up, but does have redundancy
- Use `/snap[7,8]/scratch/` if running on COSMA7/8 for staging posts, restart files, checkpointing etc.
 - This is not backed up, no redundancy, purged 3x per year

Parallel file systems

- Storage of data across multiple servers
 - Data is distributed (striped) across these
- High performance access
 - Simultaneous reads/writes
- COSMA uses:
 - Lustre for /cosma5, /cosma7, /snap7, /cosma8, /snap8
 - NFS (not parallel) for /cosma/home and /cosma/local
- Getting file system access right is an important skill in HPC

The Lustre file system

- An object-based parallel file system
- Main components:
 - Metadata servers (MDS)
 - Metadata targets (MDT)
 - aka disks
 - Object Storage Servers (OSS)
 - Object Storage Targets (OST)
 - Lustre clients
 - e.g. login nodes and compute nodes

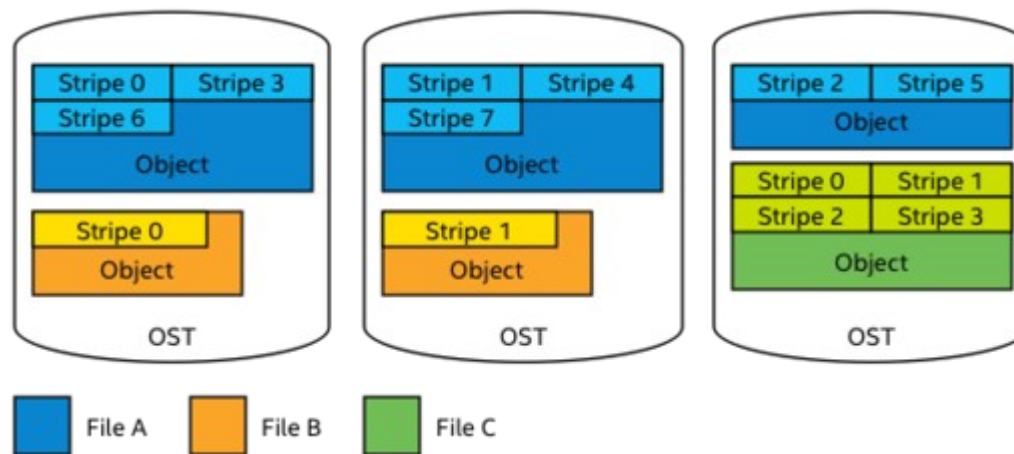


Lustre sequence

- To read a file:
 - Client requests information about the file from the MDS
 - Object identifiers and layout transferred from MDS (meta-data server) to client
 - Client can then directly interact with corresponding OSSs (object storage servers) where the object (data) is actually held
 - Client can then perform I/O in parallel across multiple OSSs without further communication with the MDS
 - Client then presents the file
- To write a file:
 - Client requires new file creation from an MDS
 - Client receives details about where to write the information
 - Client then writes data to the relevant OSSs
 - which store the data on the OSTs

Lustre striping

- Strength of parallel file systems comes from ability to stripe data across multiple targets (HDDs)
 - Capacity and bandwidth scale with the number of OSSs
 - Each Lustre file specifies its own stripe count and size
- With Lustre, once a file is created, its striping cannot be changed
 - i.e. the number of stripes
 - unless the file is recreated (overwritten) or migrated
- Striping is inherited from the directory that a file is in
- Changing striping is easy, and can improve performance



Writing large files

- If you are writing large files:

1) Set striping for the directory:

```
lfs setstripe -S 4M -c -1 /path/to/directory
```

All files here will then have this striping

2) “Touch” a file first:

```
lfs setstripe -S 4M -c -1 /path/to/new/file
```

(this creates an empty file)

When written to, this file will then have this striping

- For very large files, striping is essential if the file won't fit on a single HDD
- A C API also exists (`#include <lustre/lustreapi.h>`)
 - But don't reinvent the wheel
 - e.g. use parallel HDF5 to write files...

Size of the
individual data
blocks

Number of OSTs
to stripe across

File striping

- To check the current striping on a file or directory:

– `lfs getstrip /path/to/file/or/directory`

```
tmp3.txt
```

```
lmm_stripe_count: 2
```

```
lmm_stripe_size: 2097152
```

```
lmm_pattern: 1
```

```
lmm_layout_gen: 0
```

```
lmm_stripe_offset: 10
```

obdidx	objid	objid	group	
10	460960	0x708a0	0	
4	463821	0x713cd	0	

Hints for striping

- Good practice is to have dedicated directories with high striping for writing large files into
- Small files should be written with no striping
- With a file-per-process I/O pattern, best to use no striping
 - This will limit OST contention
- Accessing a single shared file with many processes, strip count is best if equalling the number of processes
 - Size and location of I/O operations can then be managed to allow stripe alignment with each process accessing a single OST
- Avoid patterns where a single process accesses all OSTs
- Open files as read-only where possible
- Try to avoid:
 - Multiple processes accessing the same small file
 - Use a single process to broadcast the information
 - Excessive use of stdout and stderr for parallel processes
- A good stripe size is something like 0.1-1GB
 - HDDs write at ~100 – 200MB/s
 - Feel free to investigate best sizes for your application

General Lustre hints

- Avoid using “ls -l” on large directories
 - File size is only stored on the OSSs
 - Use “/bin/ls” to see if a file exists (not ls)
 - Use “ls -l FILENAME” to get the size of a file
- Avoid having thousands of files in the same directory
- Avoid accessing small files under lustre
 - Either keep them in /cosma/home, or copy to /tmp before starting your job

Data management

- Important to consider data management
 - Long term storage
 - Access for other people (on and off COSMA)
- A tape library can be used to store data off-line
 - Freeing up space on the file system
- FAIR principles
 - Findable, accessible, interoperable, reusable
- Consider metadata
 - Data which explains the data!
- Consider version control
 - e.g. git
- Use tape for long-term archive

The Atempo tape library

- Uses tape for long-term storage
 - If you have data that you may want in the future
 - But don't need now
 - Frees up space on disk
- Ask cosma-support to transfer data

COSMA Modules

- COSMA uses a “Module” environment
 - If you need specific tools/libraries/compilers, load the corresponding module
 - All this does is sets the correct environment variables
- e.g.
 - `module load gnu_comp`
 - Will “load” the GNU gcc compiler module
 - (actually adds `/cosma/local/gnu_comp/.../.../bin` to `$PATH`)
 - `module load fftw`
 - Will load the FFTW libraries
 - (actually adds stuff to `$LDFLAGS`, `$LIBRARY_PATH`, `$CPATH`, `$CMAKE_INCLUDE_PATH`, etc)

Module commands

- `module avail`
 - Too see available modules
 - Can search by appending a name, e.g.
 - `module avail ff` → will show all modules starting with ff
- `module list`
 - Lists currently loaded modules
- `module load MODULENAME`
- `module unload MODULENAME`
- `module purge`
 - Unloads all modules
- `module show MODULENAME`
 - Shows information about the module
- Commands can be shortened (e.g. `module av`)
 - Tab completion works

Module dependencies

- Some modules depend on others
- e.g. for FFTW, you need a compiler module and an MPI module loaded first
- Others conflict
- e.g. you cannot load both python2 and python3 modules
 - (python/2.7.15 and python/3.6.5)
 - *Note, if you load the python3 module, you need to use python3, rather than python (which would give you the old system python2)*

Compilers and MPI libraries

- gnu_comp/7.3.0, /8.2.0, /9.1.0, /10.x.x
- intel_comp/2017, /2018, /2019, /2020, /2021
- openmpi/3.0.1, /4.0.1, /4.1.4
- intel_mpi/2017, /2018, /2019, /2020
- hpcx-mt/2.2 → Openmpi optimised for infiniband
- If you need new modules, please ask us to add them

Message Passing Interface (MPI)

- Most HPC codes use MPI to communicate between nodes
 - Mostly transparent to a user
 - But some knowledge required for code development
 - See other lectures/courses

SLURM

- Job scheduling system
- Used to allocate resources (nodes) to users
- Monitoring of jobs
- Maintaining a fair work queue
- COSMA has several work queues or “partitions”
- Useful commands:
 - sinfo, squeue, sbatch, scontrol, scancel, showq, sprio, srun, squota

COSMA5 partitions

- cosma
 - Standard queue
- cosma-prince
 - For users who require a large number of nodes
 - (explicit permission required)
- cosma-analyse
 - For users of the analyse group
- cordelia
 - For single core jobs (i.e. not parallel jobs)

COSMA7 partitions

- `cosma7`
 - Standard queue
- `cosma7-pauper`
 - For users with no allocation left
- `cosma7-prince`
 - When a large number of nodes are required
- `cosma7-shm`, `cosma7-shm2`
 - Single node for large memory jobs or single core jobs
 - Not exclusive

COSMA8 partitions

- cosma8
- cosma8-pauper
- cosma8-prince
- cosma8-shm, cosma8-shm2, cosma8-shm3
- cosma8-rome
- cosma8-milan
- cosma8-serial

sinfo

- sinfo
 - Shows a list of the nodes allocated to the different queues
- sinfo -p cosma7
 - Show only cosma7 nodes
- See the man pages for further information:
 - man sinfo

PARTITION	AVAIL	TIMELIMIT	NODES	STATE	NODELIST
cosma7	up	3-00:00:00	1	down*	m7037
cosma7	up	3-00:00:00	2	drain	m[7064,7143]
cosma7	up	3-00:00:00	144	alloc	m[7001-7036,7038-7063,7144-7147]

squeue

- Shows the current state of the queues
- e.g. `squeue -p cosma`
 - Shows only cosma
 - Column ST shows state
 - Common states are:
 - R=running, PD=pending, CA=cancelled, CF=configuring (e.g. waiting for servers to book), CG=completing, DL=deadline (job terminated on deadline), NF=node fail, TO=timeout

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST (REASON)
86362	cosma	LWHalo	dc-rega4	R	14:56	36	m[5124-5136,5146-5156,5159-5170]
86282	cosma	L400_de	dc-pfeil	R	2:37:07	32	m[5184,5219-5242,5244-5250]
86281	cosma	L400_de	dc-pfeil	R	2:37:39	32	m[5185-5212,5214,5216-5218]
86363	cosma	MMHalo_N	dc-rega4	R	14:42	12	m[5123,5171-5181]
86376	cosma	RECAL-h0	rcrain	R	0:54	8	m[5182-5183,5251-5256]
86191	cosma	IC_Gen	arj	R	4:54:39	8	m[5137-5138,5140-5145]
86374	cosma	energy	likm	R	5:47	1	m5262
85285	cosma	cal_all	shliao	R	1-05:32:27	1	m5122

sbatch

- Use sbatch to submit a job:
 - sbatch /path/to/job/file
- A job file will contain the necessary information for SLURM
- Sample scripts in /cosma/home/sample-user

```
#!/bin/bash -l
#SBATCH -n 1                # Number of cores 1
#SBATCH -J job_name
#SBATCH --exclusive        # No sharing of node
#SBATCH -t 10              # Time limit of 10 minutes
#SBATCH -p cosma           # Use partition (queue) cosma
#SBATCH -A durham          # group durham for accounting purposes
#SBATCH -o std_%j.out      # Output file
#SBATCH -e stderr_%j.err   # Error file
#SBATCH --mail-type=END    # Notification when job ends (done or failed)
#SBATCH --mail-user=user@durham.ac.uk # Where to send emails

module load fftw
cd /path/to/my/code
./startMyJob
```

scontrol

- Can be used to see which groups are allowed to submit to a partition
- e.g. `scontrol show partition cosma7`

```
PartitionName=cosma7
  AllowGroups=ALL AllowAccounts=do004,dp004,dp019,dp034,dp104,dp105,ds007 AllowQos=ALL
  AllocNodes=ALL Default=NO QoS=N/A
  DefaultTime=NONE DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
  MaxNodes=UNLIMITED MaxTime=3-00:00:00 MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
  Nodes=m[7001-7147]
  PriorityJobFactor=50 PriorityTier=50 RootOnly=NO ReqResv=NO OverSubscribe=EXCLUSIVE
  OverTimeLimit=NONE PreemptMode=OFF
  State=UP TotalCPUs=4116 TotalNodes=147 SelectTypeParameters=NONE
  DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED
```

- `scontrol show job #JOB NUMBER`

scancel

- Cancel submitted jobs:
 - scancel jobID

showq and sprio

- `showq -l -o -p cosma7`
 - Shows running and waiting jobs, ordered by priority
- `sprio -l -p cosma7`
 - shows information about priorities

Scheduling

- SLURM priority calculation is complex
- Consider the case of a large job requiring many nodes, and many smaller jobs.
 - Small jobs can back-fill (i.e. use unused nodes)
 - But this would mean that there are never enough jobs for the large job
 - So a priority is calculated based on many things
 - If you have short jobs, please specify their estimated runtime accurately, so that they can be used to back-fill nodes while the large jobs are waiting to start
- Command `c[5,7,8]backfill` shows nodes currently available for small jobs

Compiling and optimising

- Use the login nodes for compiling code
- Optimisation flags can be used to improve code performance:
 - For gcc:
 - -O3 -march=native
 - For icc:
 - -O3 -xHOST
- Other flags also available (>1000!)
- But, e.g. if compiling for a COSMA5,8 queue on the COSMA7 login nodes, do not optimise too far
 - C5 nodes are older and illegal instructions may result
 - C8 nodes don't have AVX512 instructions
 - So, compile on the relevant login node!

CPU architectures

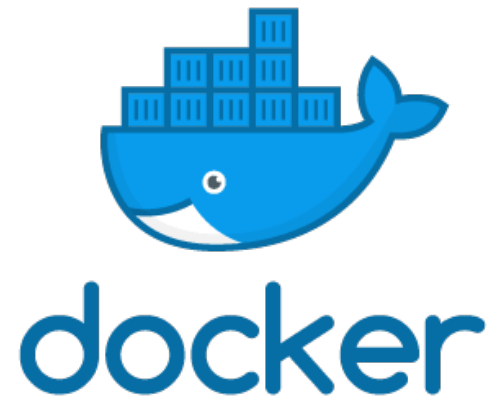
- COSMA 5: Intel Sandybridge
- COSMA 7: Intel Skylake
- COSMA8: AMD Rome
- Different architectures can run different CPU instructions
 - Though all have a very similar base set
 - Key difference: AVX-512 instructions on Skylake

Jupyter hub

- A Jupyter hub is available on a COSMA7/8 login node
 - But, this is not a good HPC paradigm and should really only be used for analysis of results
 - It is a shared resource, and only accessible on that server
- Can also be launched on compute nodes.
- See www.dur.ac.uk/icc/cosma/support/jupyter for how to get set up
- It will require your cosma username and password
 - Note, not your ssh key passphrase

Containers

- COSMA has a “singularity” module
 - Can be used with singularity or docker containers



Future updates

- DiRAC4
- Data Curation Service
 - Long-term storage of data
 - Including tape archive storage
 - Accessibility
 - Use metadata tagging to add meaning and context to the data