Introduction to High Performance Computing and COSMA

Alastair Basden
and the COSMA team
(Billy Carlyle, Mark Lovell, Richard Regan, Paul Walker,
John Helly, Sarah Johnston, Gokmen Kilic, and others)

cosma-support@durham.ac.uk

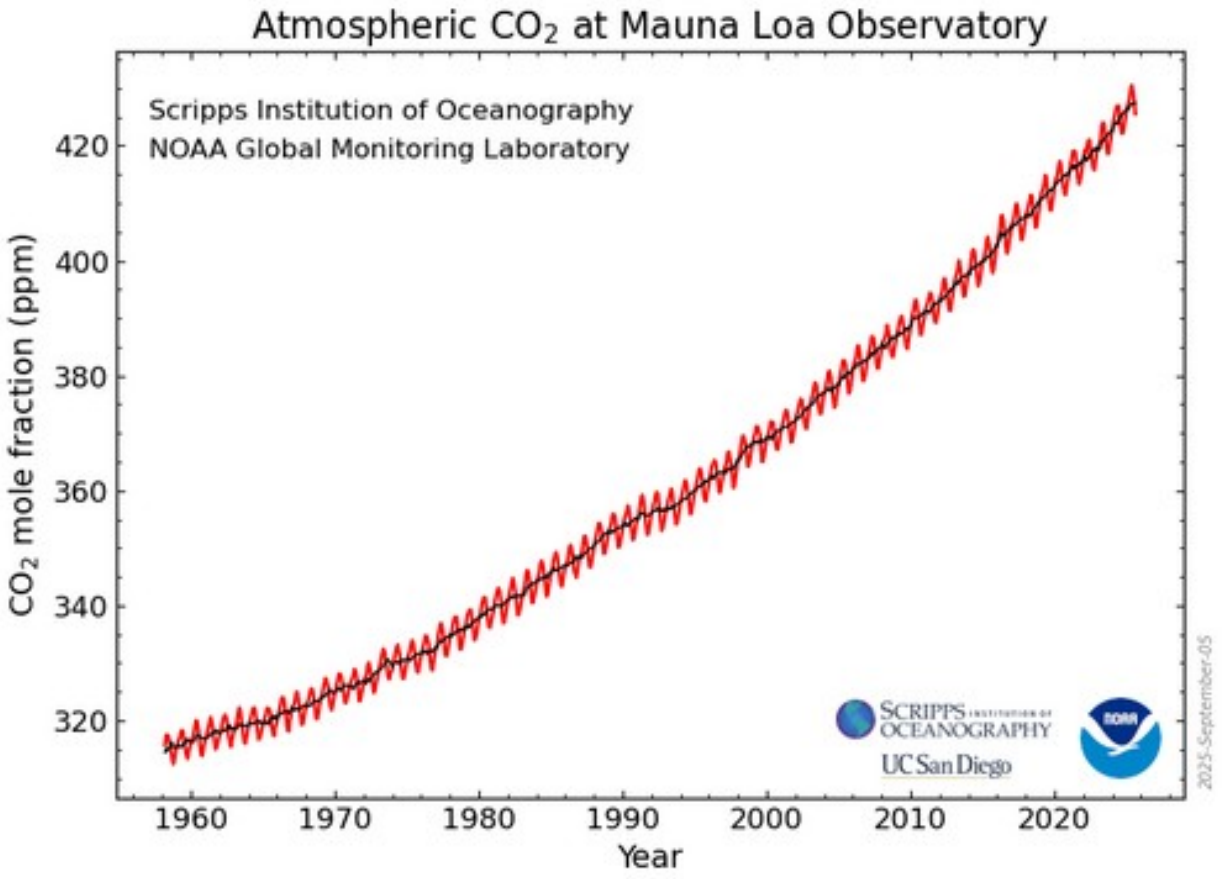
About me

- Spent a long time in astronomical instrumentation (CfAI)
  - Adaptive Optics and telescopes
- COSMA manager since 2018

How to get help:

- cosma-support@durham.ac.uk
- Visit Ph317, Ph321, OCW220 (or other offices)
- Alastair Basden
  - Director of COSMA
- Richard Regan
  - DiRAC training manager, COSMA support
- Paul Walker, Mark Lovell, Billy Carlyle
  - COSMA support
- John Helly
  - Cosmology software expertise, data storage expert
- Sarah Johnston, Gokmen Kilic
  - DiRAC Research Software Engineers (RSEs)
- Catch us at the CommuniTea event on Thursdays at 11am

High importance: Climate change



Institute and Faculty of Actuaries

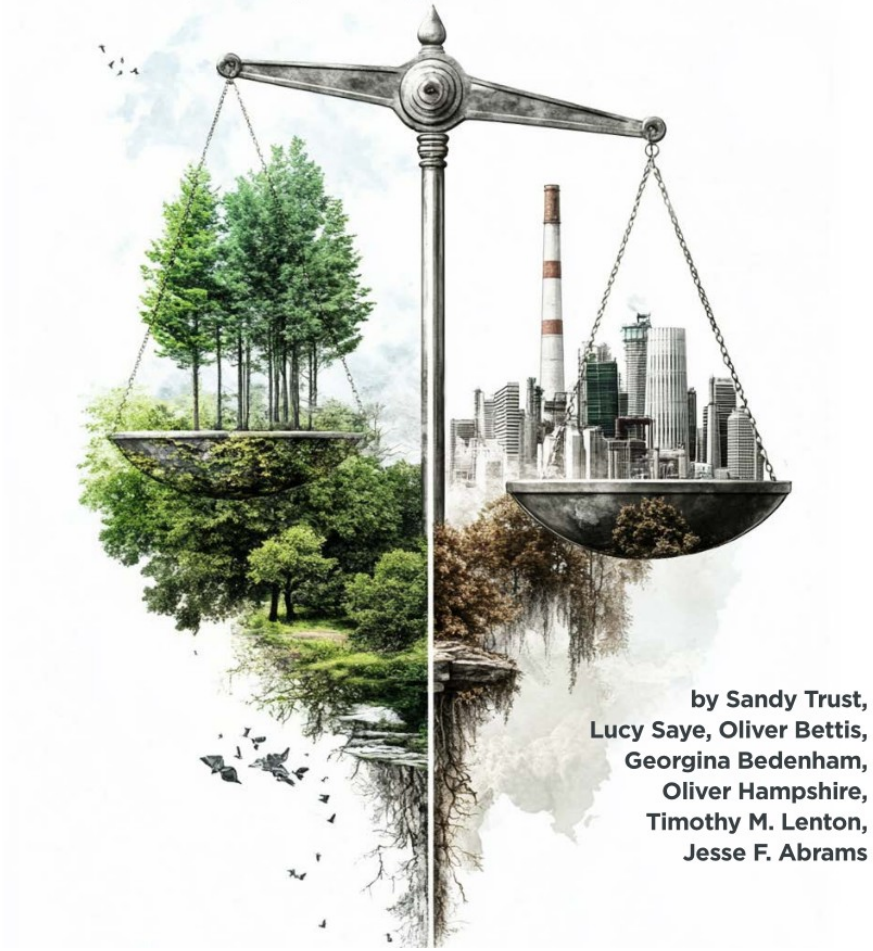
University of Exeter

Collaborative insights in the public interest

January 2025

Planetary Solvency – finding our balance with nature

Global risk management for human prosperity



by Sandy Trust, Lucy Saye, Oliver Bettis, Georgina Bedenham, Oliver Hampshire, Timothy M. Lenton, Jesse F. Abrams

Appendix I: Risk impact matrix

The table below shows the Planetary Solvency risk impact and likelihood matrix utilised for the illustrative Planetary Solvency outputs contained in previous sections.

Figure 12: Planetary solvency risk impact and likelihood definitions (illustrative)

Rating	Financial impact	Non-financial impact			
	GDP losses	Human mortality	Climate	Nature	Societal
Extreme	≥50%	≥50% > 4 billion deaths	3°C or more by 2050. Multiple climate tipping points triggered, tipping cascade.	Breakdown of several critical ecosystem services and Earth systems. High level of extinction of higher order life on Earth.	Significant socio-political fragmentation worldwide and/or state failure with rapid, enduring, and significant loss of capital and systems identity. Frequent large scale mortality events.
Catastrophic	≥25%	≥25% >2 billion deaths	2°C or more by 2050. High number of climate tipping points triggered, partial tipping cascade.	Breakdown of some critical ecosystem services and Earth systems. Major extinction events in multiple geographies. Ocean circulation severely impacted.	Severe socio-political fragmentation in many regions, low lying regions lost. Heat and water stress drive involuntary mass migration of billions. Catastrophic mortality events from disease, malnutrition, thirst and conflict.
Decimation	≥10% >$10 trillion annual losses	≥10% > 800 million deaths	Global warming limited to 2°C by 2050. Several climate tipping points triggered.	Severe reduction in several critical ecosystem services. Major extinction events in some geographies. Frequent global food and water crises.	Severe socio-political fragmentation in regions exposed to climate and/or nature impacts. Failure of vulnerable states and mass mortality events in impacted areas.
Severe	≥5% >$5 trillion annual losses	≥5% > 400 million deaths	Global warming limited to 1.5°C by 2050 following overshoot. Some proximate climate tipping points triggered.	Some impacts to critical ecosystem services. Ongoing species extinction. Regular global food and water crises.	Some socio-political fragmentation in most vulnerable states, where adaptation has been limited. Fragile states exposed to climate risks see mass migration and mortality events from heat, water stress and weather events.
Limited	≥1% >$1 trillion annual losses	≥1% > 80 million deaths	Global warming below 1.5°C by 2050, with limited overshoot. Climate tipping points largely avoided.	Mass extinction avoided and ecosystem services largely functional. Occasional global food crisis and widespread water crises.	Ongoing significant climate impacts with many hundreds of billion dollar + loss events annually and associated mortality and socio-political stress.

Likelihood of the risk occurring over a certain timeframe	Extremely Unlikely	Highly Unlikely	Unlikely	Possible	Likely	Highly Likely
	<1%	1-10%	10-40%	40-60%	60-90%	≥90%

So...

- Please consider the implications when running your calculations, simulations, AI inferencing, etc!
  - Don’t waste compute cycles
  - Try to make codes efficient
  - Make good use of the calculations you run!

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
  - Though can be (but expensive)
- Related to HTC
  - High throughput computing



The TOP500

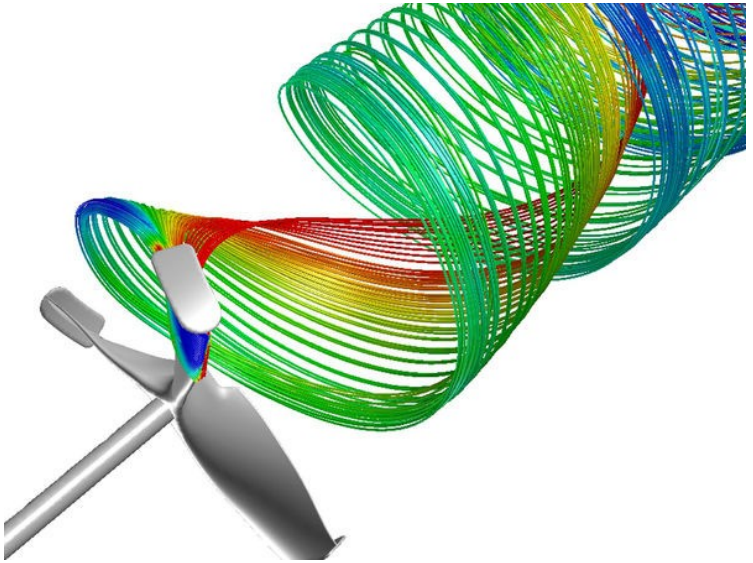
- Prestigious list of the World's most powerful supercomputers, released 6-monthly
  - Includes some UK sites
  - Becoming less relevant
- Also:
  - Green TOP500
    - Best performance/Watt
  - I/O TOP 500

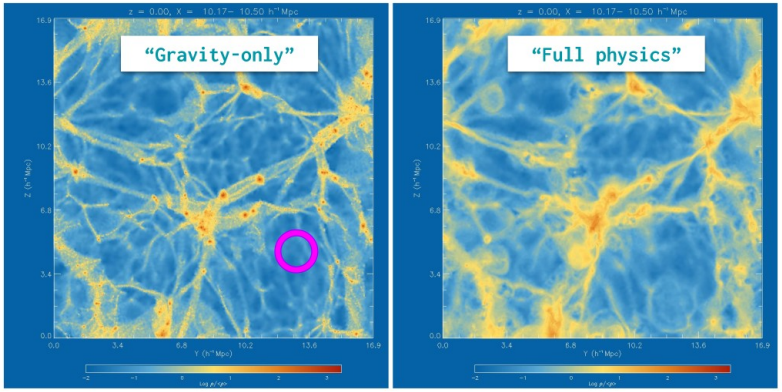


Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**LineShine** - LingKun, LX2 304C 1.55GHz, LingQi, Kylin OS, Shenzhen Cloud Computing Center Co., Ltd. National Supercomputing Centre in Shenzhen (NSCS) China	13,789,440	2,198.40	2,735.82	42,220
2	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,340,000	1,809.00	2,821.10	29,685
3	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
4	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
5	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, Bull EuroHPC/FZJ Germany	4,801,344	1,000.00	1,226.28	15,794
6	**HPC7** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, RHEL 9, HPE Eni S.p.A. Italy	3,461,472	571.50	861.13	8,735
7	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
8	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, HPE Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461
9	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
10	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124

HPC use cases

- Galaxy/cosmology simulation
- Weather simulation
  - And climate change modelling
- Artificial intelligence
- Fluid dynamics modelling
- Materials modelling
- Genetics
- Biomolecular simulation
- Medical - COVID, digital twins...
- etc



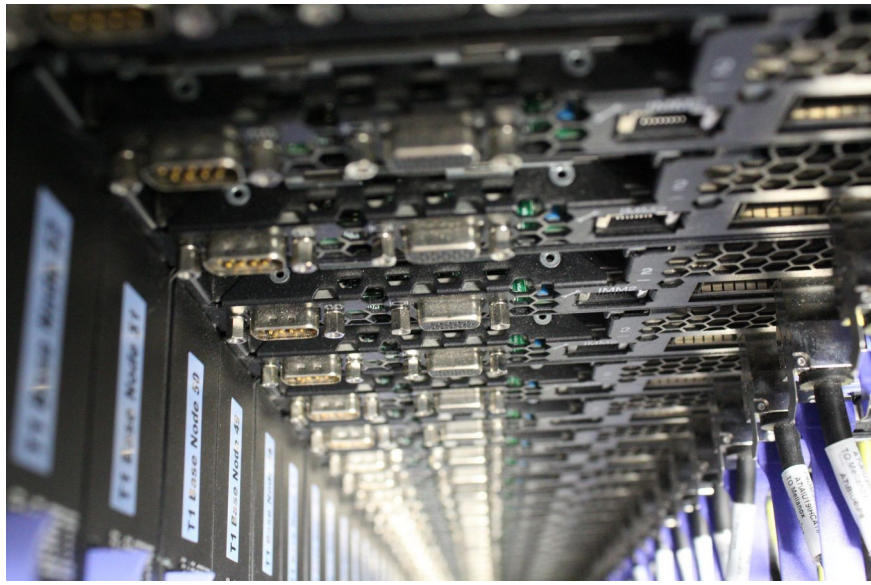


HPC in the UK

- Tier-1 systems (National)
  - AI Research Resource (AIRR):
    - Bristol Isambard AI system
    - Cambridge DAWN system
  - Archer2 - Edinburgh, UK national facility (EPSRC)
    - Retiring next month
  - DiRAC, UK national facility (STFC)
    - Of which COSMA is part
  - National Compute Resources (NCRs)
    - 6x systems for different workloads
  - Met Office
  - ECMWF
  - Next National Supercomputing Service (NNSS)
    - Coming in ~2028
- Tier-2 systems (Regional)
  - E.g. BEDE (at Durham), and a number of others
- PRACE (European collaboration)
- Tier-3 systems (Local)
  - 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 (refreshed 2025)
- DiRAC 2.5 (COSMA6) in 2016 (retired in 2023)
- DiRAC 2.5x (COSMA7) in 2018 and 2019
- DiRAC 3 (COSMA8) in 2020, 2021 and 2023



Science & Technology Facilities Council

DiRAC Distributed Research utilizing Advanced Computing

Components of HPC



- Login nodes
  - For user interaction
- Compute nodes
  - For doing the work
- Test nodes
  - For benchmarking, profiling, etc
- High performance network fabric
  - For linking compute nodes and storage
- Storage (i.e. lots of disks working together)
  - 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 (COSMA8)
  - COSMA5 and 7 also available for use
  - And DINE/DINE2 (experimental clusters)



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
- Originally ~300 old nodes (160)
  - 16 cores per node (2 CPUs)
  - 128GB RAM per node
- Now 8 new nodes
  - 256 cores per node
  - 1.5TB RAM
  - Some immersed in an oil tank!
- 1x login nodes: login5c



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)
  - 28 cores per node: 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

- Prototype system in early 2020
- £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 cores and 1TB RAM
- Some “fat” nodes: 4TB RAM
- Some GPU nodes (NVIDIA, AMD and Intel)



DINE/DINE2

- 24-node system
  - With network accelerators (NVIDIA DPU)
  - 32 cores/node
  - 512GB RAM
- 8-node system
  - With GPUs
  - 64 cores per node
  - 2TB RAM/node
- Useful for code testing, and exploring new technologies
- Part of the HPC Hardware Lab

HPC Hardware Lab

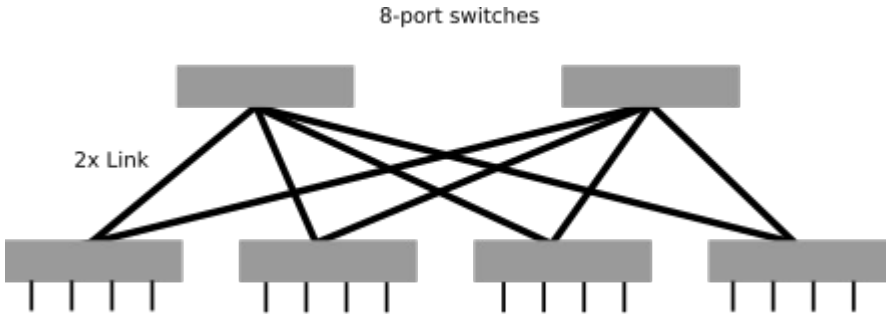
- A suite of new HPC technologies
  - Regularly updated
- Latest CPUs, GPUs, networking, storage
  - Ideal for porting and testing code performance

COSMA summary

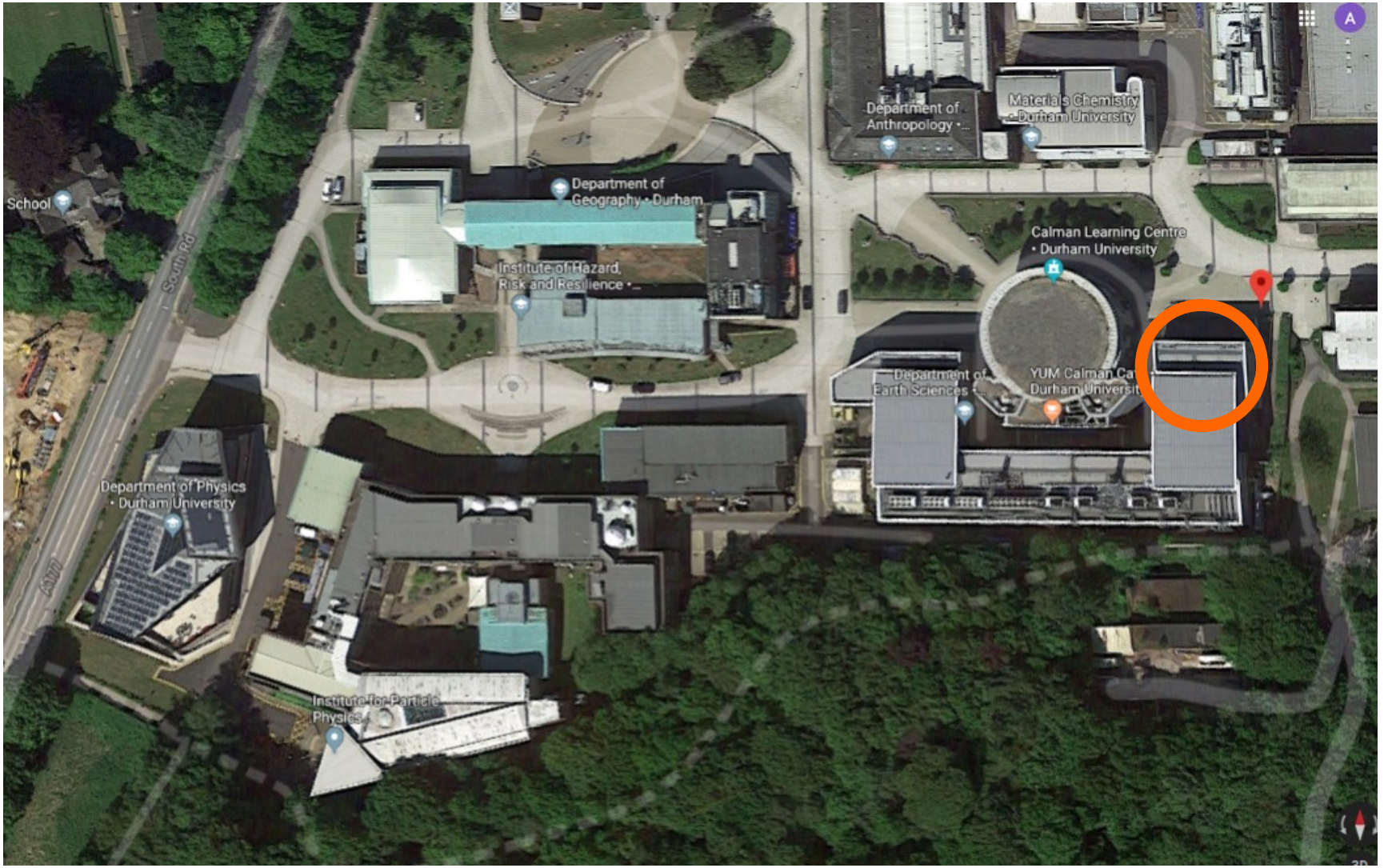
	COSMA5	COSMA7	COSMA8
Nodes	8 (was ~420)	452	528
Cores/node	256 (was 16)	28	128
Cores	2048	11,200	67,584
Memory/node (GB)	1536	512	1024
Memory/core (GB)	6	18	8
Total memory	12TB	230TB	528TB
Login nodes	1	3	2
Operating system	Rocky 9	Rocky 9	Rocky 9

COSMA network fabric

- A fast network is essential for HPC
  - Low latency also important
- COSMA uses Infiniband for inter-node communication (MPI) and file system access
    - Was FDR10 for COSMA5 (40Gb/s)
      - Now EDR
    - 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
  - A fat tree topology
- Half of COSMA7 uses a Rockport fabric:
  - 6D torus, switchless
  - Ethernet, 100G

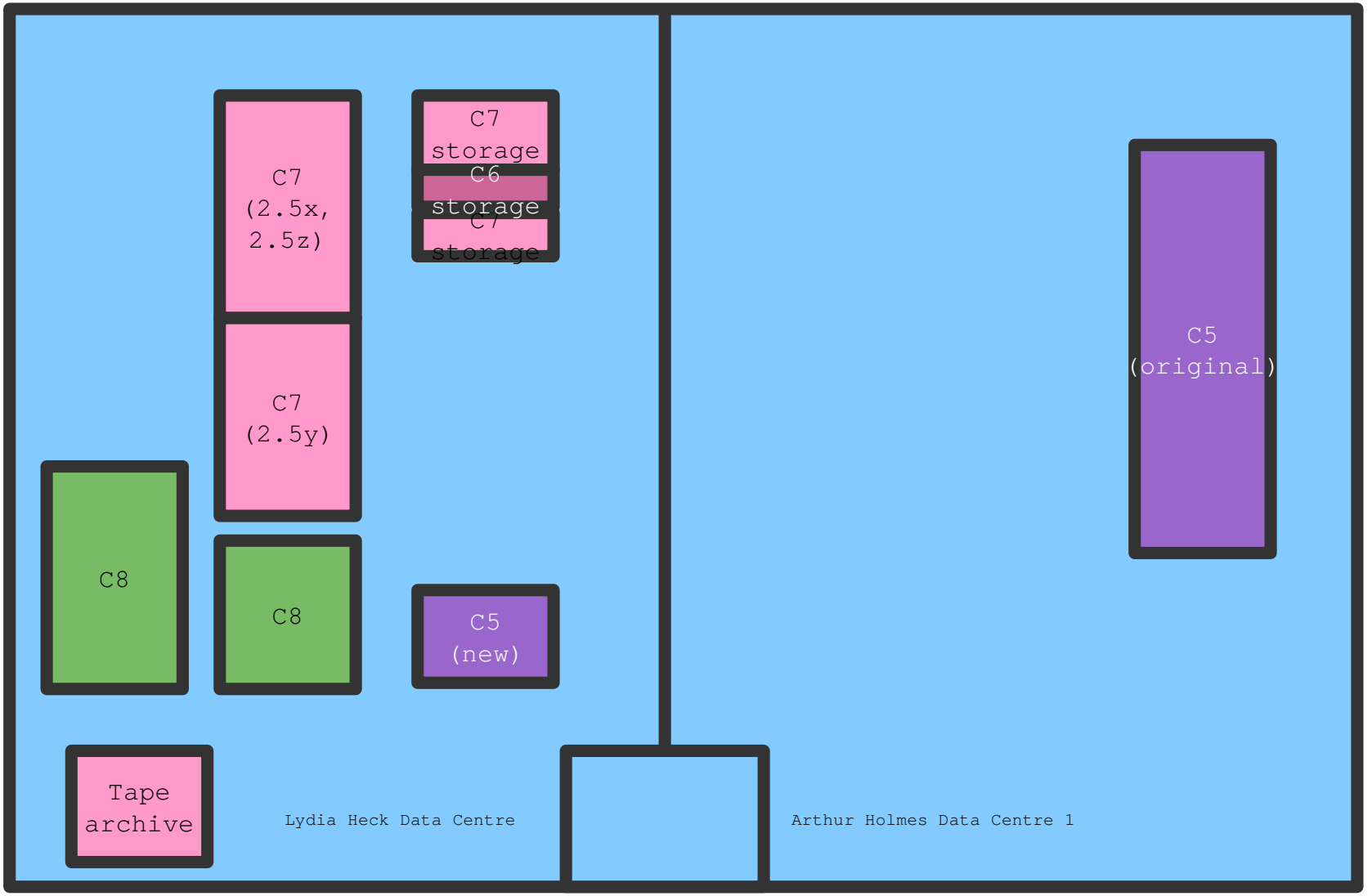


Arthur Holmes Data Centre



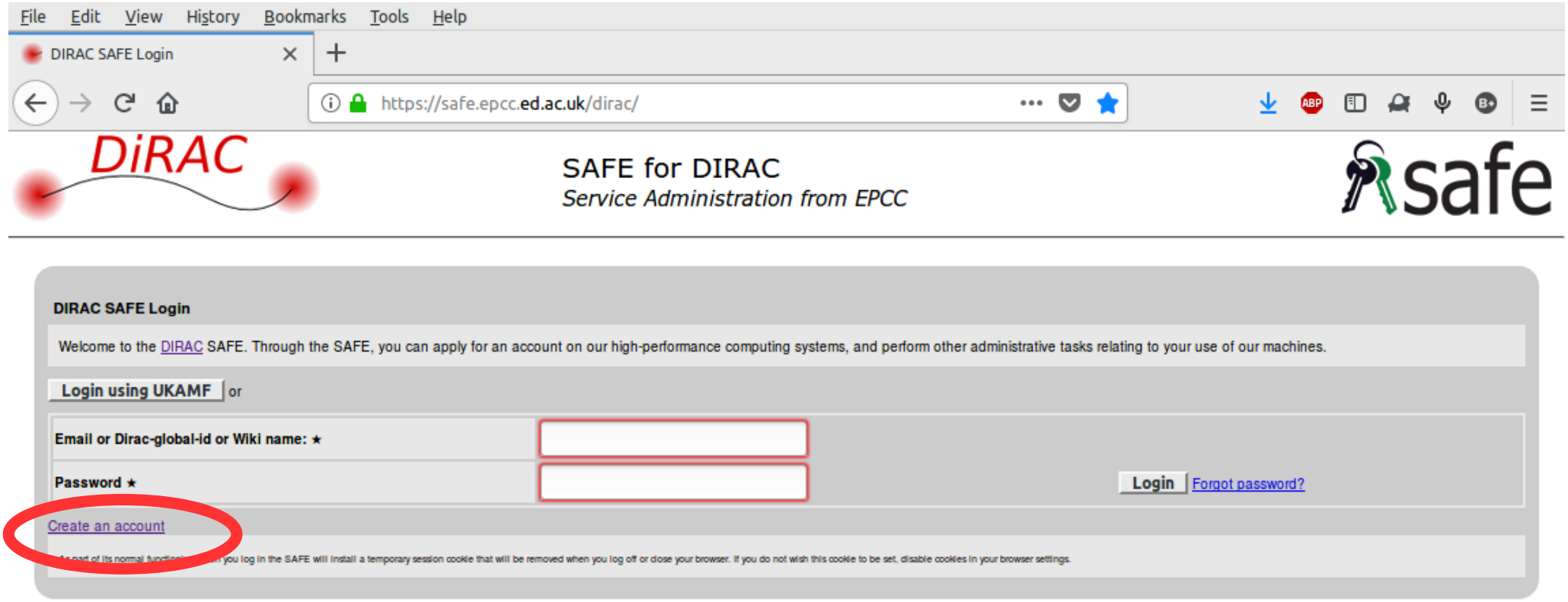
Arthur Holmes Data Centre

- Location of COSMA



Accessing COSMA

- First, you need a SAFE account
  - (Service Administration From EPCC)
    - Used for all DiRAC facilities
  - See https://cosma.readthedocs.io/
    - → Creating an 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...


File Edit View History Bookmarks Tools Help
DIRAC SAFE Login
https://safe.epcc.ed.ac.uk/dirac/
DiRAC
SAFE for DIRAC
Service Administration from EPCC
safe
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
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 for DIRAC services

Service Administration by EPCC

safe by epcc

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

Register

IDiRAC

SAFE for DIRAC services
Service Administration by EPCC

safe by epcc

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

Register

DiRAC

SAFE for DIRAC services
Service Administration by EPCC

safe by epcc

User Access Agreement

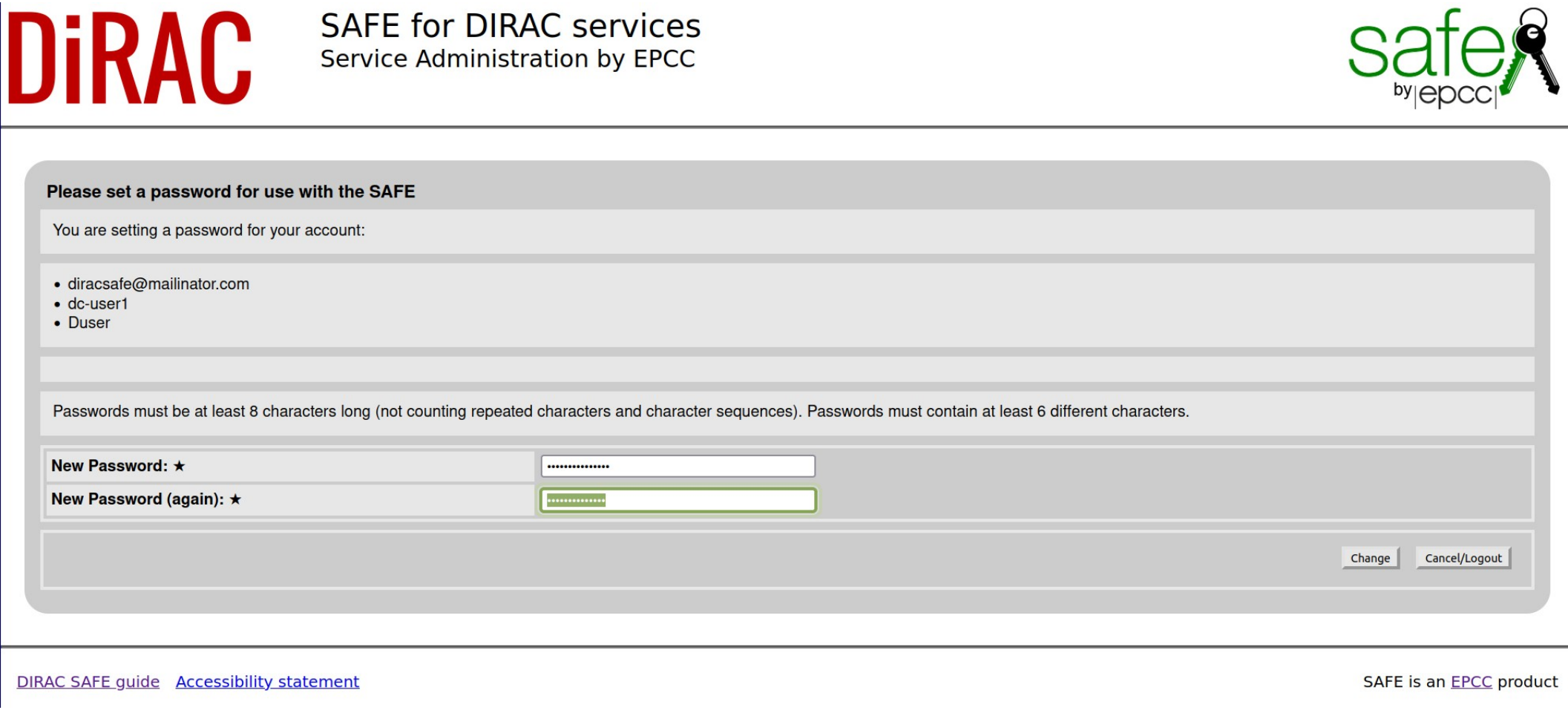
Please read our acceptable use policy.

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...



DiRAC

SAFE for DIRAC services

Service Administration by EPCC

safe 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

DiRAC

SAFE for DIRAC services

Service Administration by EPCC

safe by epcc

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 like this it indicates a pending request that needs your attention.

DiRAC

SAFE for DIRAC services
Service Administration by EPCC

safe 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 ★ hpcicc

Next

DiRAC

SAFE for DIRAC services
Service Administration by EPCC

safe by epcc

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
◉	cosma: The Durham COSMA machine		The Durham COSMA machine	• 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

SAFE for DIRAC services
Service Administration by EPCC

safe by epcc

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.

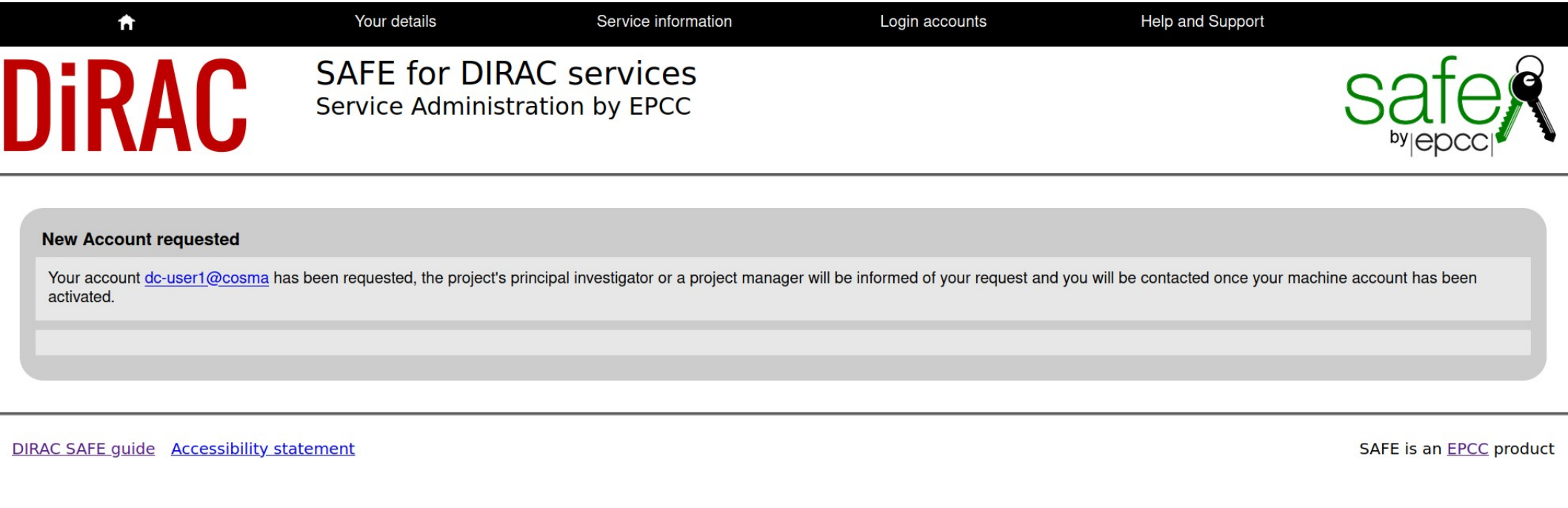
Your default key will automatically be added if no other key is specified

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!



Accessing COSMA...

- Login nodes
  - This is where you do your work!
    - Prepare scripts,
    - Edit code,
    - Compile code,
    - Inspect results,
    - Run small, quick tests
  - A shared facility
    - Usually with extra RAM
    - Try not to run larger 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@login7.cosma.dur.ac.uk
    - ssh USERNAME@login8.cosma.dur.ac.uk




```
ali@xps14:~$ ssh dc-basd1@login8a.cosma.dur.ac.uk
(dc-basd1@login8a.cosma.dur.ac.uk) Password:

Welcome to new-look COSMA.

There has been a major system update recently.  For further information, please
see

https://cosma.readthedocs.io/en/latest/clusterupgrade.html

Last login: Sun Oct 20 22:24:25 2024 from 80.45.116.86
Install package 'keychain' to provide command 'keychain'? [N/y] n

[dc-basd1@login8a ~]$
```

First time login

- People often get stuck here:
  - ssh -i /path/to/ssh/key USER@login.cosma.dur.ac.uk
    - First you will be asked for your ssh key passphrase
    - Then you will be asked for your COSMA password
      - This will be in the welcome email.
    - If you get this right, you will then be asked to change it
      - YOU NEED TO ENTER IT AGAIN
        - To start the reset process
      - And then enter a new password twice
        - >=10 characters, 2 character classes

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
  - This is different from your SAFE password and your ssh key passphrase

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 cosma.readthedocs.io

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...

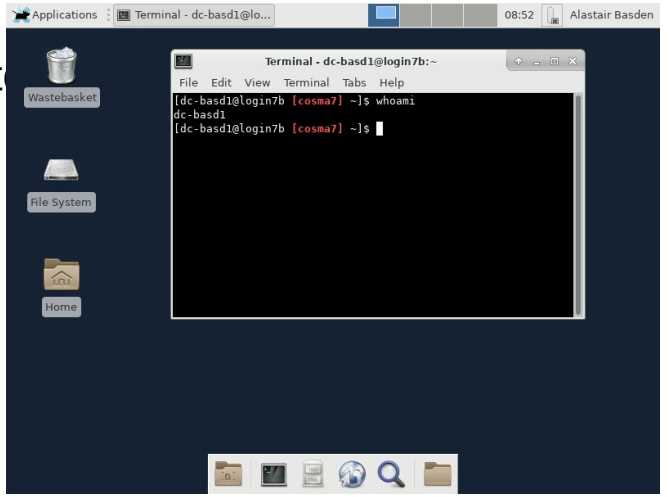
- Your terminal will now be redirected to COSMA
  - Anything you type will be interpreted by COSMA (not your PC)
  - Use common Unix commands
    - e.g. ls, pwd, mkdir, cp, etc
  - Start text editors
    - e.g. vi, emacs etc
- If you want to access COSMA from other computers, upload the public key to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

Some useful commands

- id
  - Gives you User ID, and the groups you are in
- whoami
  - Your USERNAME
- w and who
  - see who else is logged on at the moment
- top and htop
  - see who is hogging resources!
- cd, ls, rm, less, cat
- And don’t forget ***tab-completion***

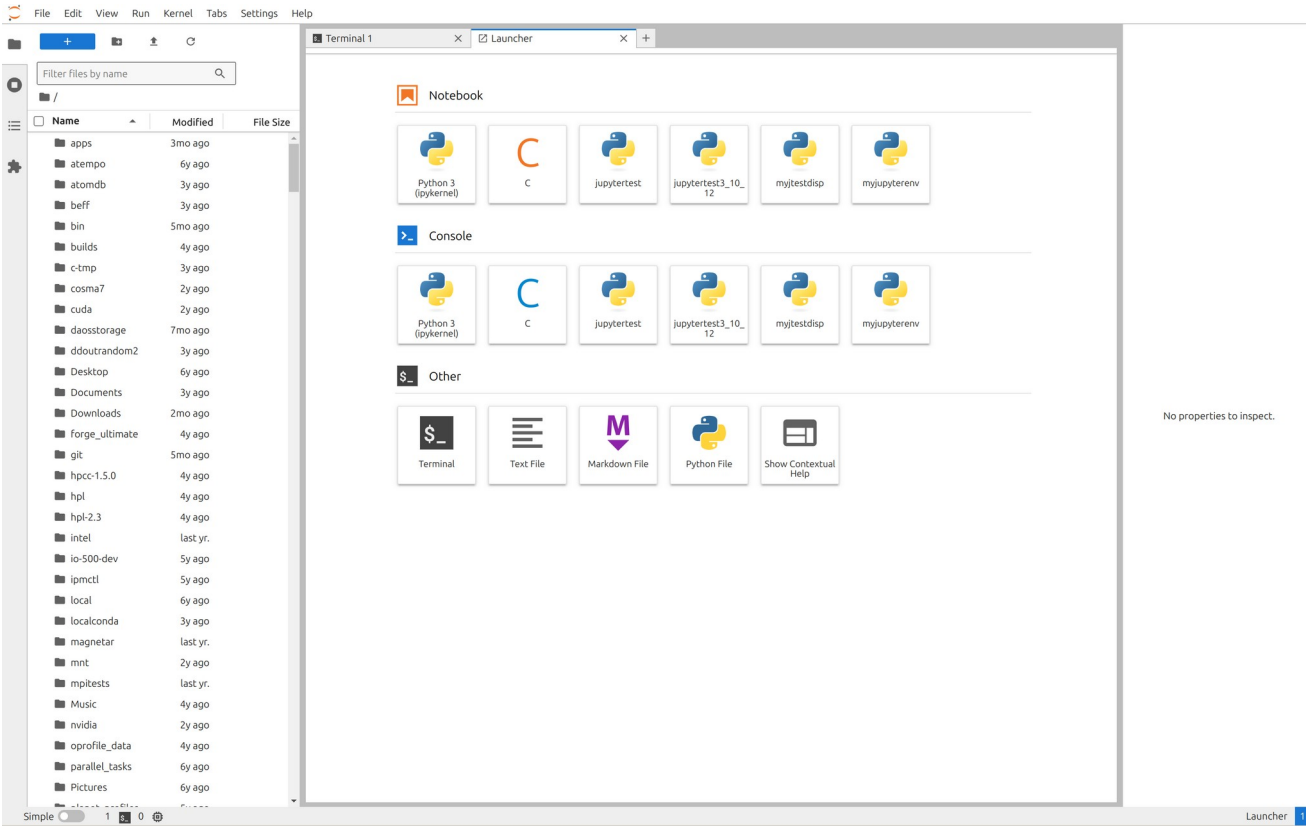
Graphical access

- If you want graphical access:
  - x2go is the best option
    - Install an x2go client, and use it to connect to a login node.
      - See cosma.readthedocs.io
  - 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
  - To edit files over a poor connection, use a terminal based edit
    - vi, emacs -nw, nano
- x2go is faster



Jupyter Hub

- Several login nodes host a Jupyter Hub
- You first need to set up a ssh tunnel:
  - A bit like a VPN
  - ssh -L localhost:8443:localhost:443 USER@login8b.cosma...
- Then visit https://localhost:8443



Visual editors

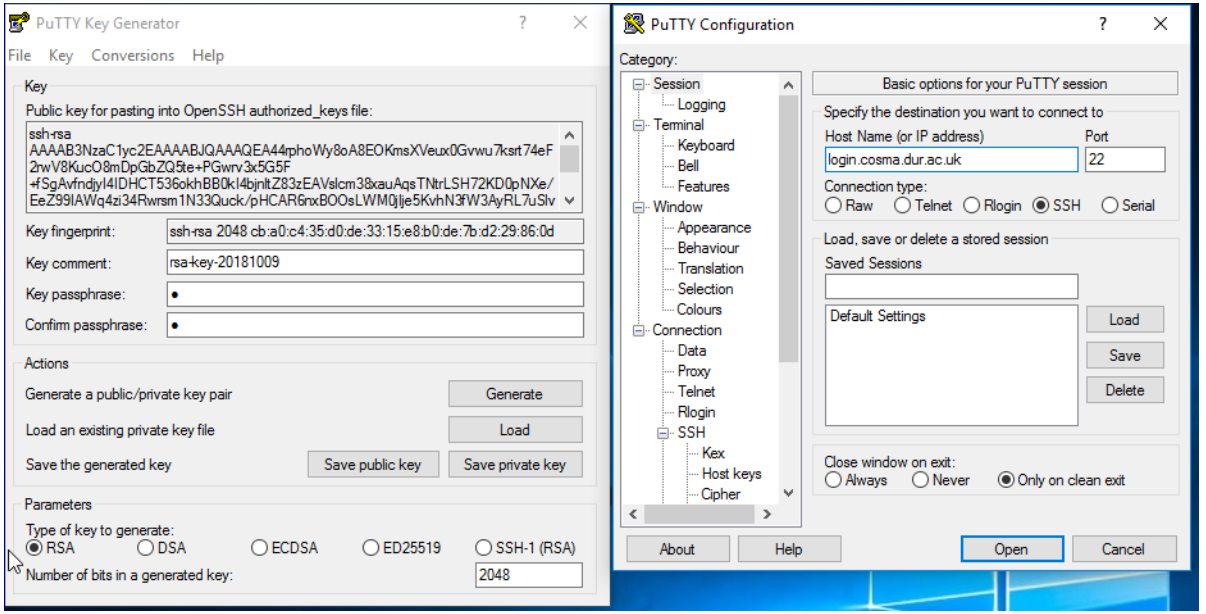
- Many people use e.g. VSCode
  - Powerful tool
  - Lots of extensions
    - Beware, some of these are nefarious
      - Think before you use them!
  - Lots of AI extensions
    - Allowing connection to Claude/OpenAI, etc
  - Think carefully about your setup
    - Many users get temporarily banned by failed password attempts (repeated attempts by VSCode)
    - See the help pages for details

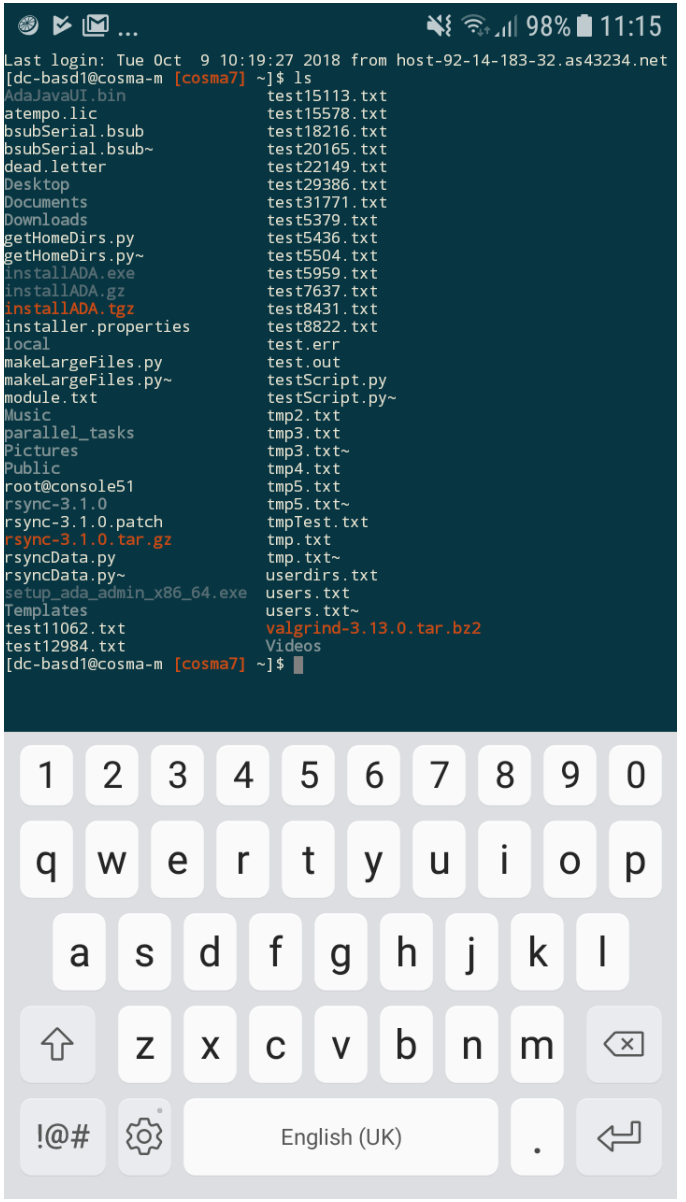
COSMA passwords

- Are also used to authenticate with the cosma websites
  - (to access usage statistics, Jupyter Hub, etc)
- These are different from your ssh key passphrase
  - You can reset your password on SAFE
  - 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 WSL (Windows Subsystem for Linux)
  - Which provides a Unix-like 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…
  - People have mixed success with these
- On Android, use e.g. JuiceSSH, Termux, etc
  - You will need to generate an ssh key pair (using the app)





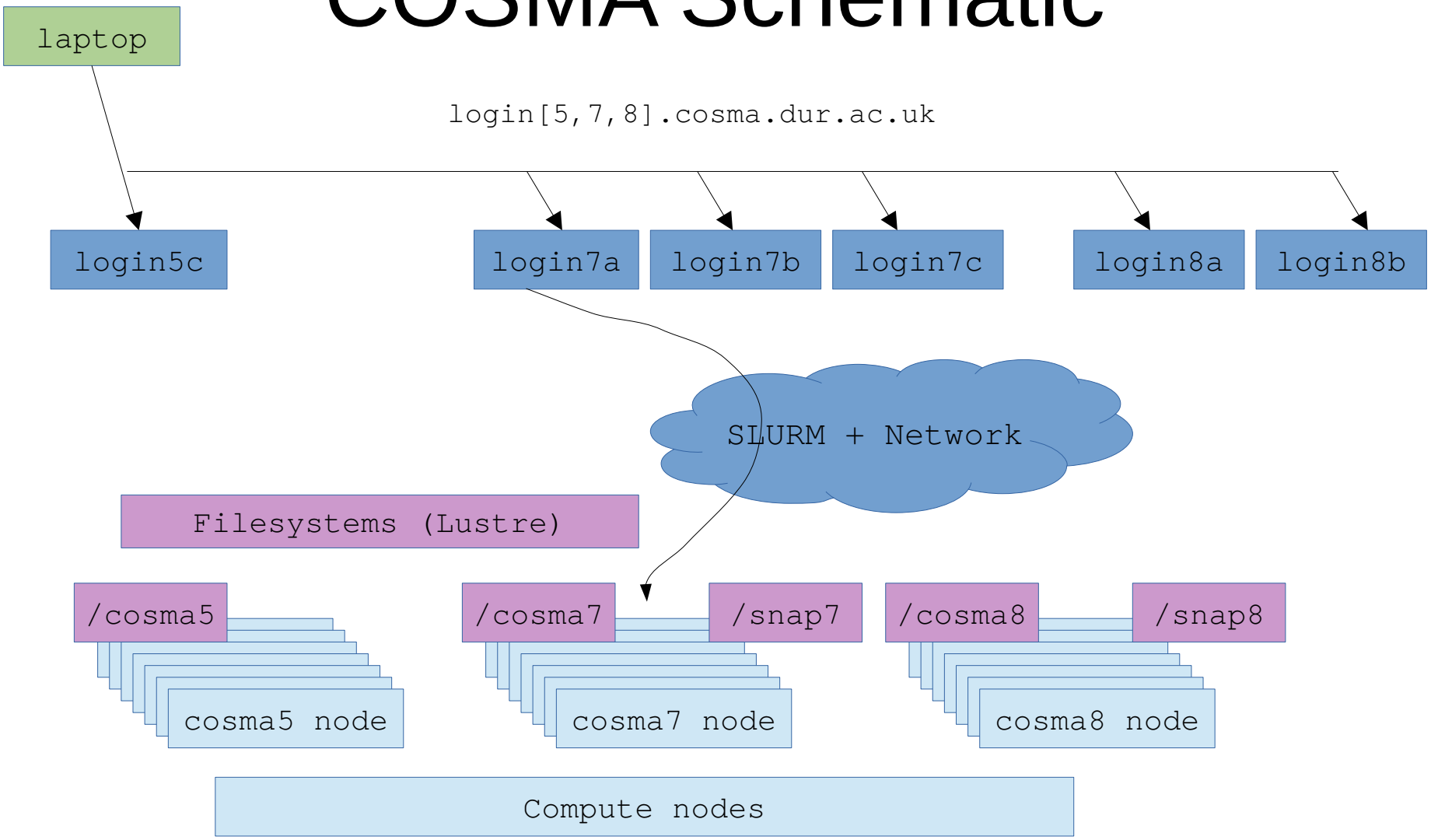
Typical workflow

- 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:
    - Become proficient with git, or
    - 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 homespace 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 -L localhost:1234:localhost:80 USER@login.cosma.dur.ac.uk
    - And then visit http://localhost:1234

COSMA login nodes

- Currently:
  - COSMA5 has 1 node: login5c
  - 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, login5.cosma.dur.ac.uk (login5c)
- login7.cosma.dur.ac.uk (round-robin allocation to 7a, 7c)
- login8.cosma.dur.ac.uk (round-robin allocation to 8a, 8b)

COSMA Schematic



HPC storage

- 4 main types of storage:
  - Home space (10GB)
    - /cosma/home/
      - Snapshotted hourly, nightly to tape
  - Application space (100G)
    - /cosma/apps
      - Snapshotted daily
  - Bulk storage / data space (5-10TB)
    - e.g. /cosma7, /cosma8
    - More if needed upon request
  - 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 can install applications (python libraries, venvs, code libraries, spack, tools etc) at:
  - /cosma/apps/PROJECT/USERNAME
- 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/18PB)
  - /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 typically available from the COSMA7 compute nodes, and vice versa
- /cosma/local/ is where system tools and libraries are located

What to put where...

- Use /cosma/home/ for scripts, code, config files
  - This is backed up
    - And snapshotted hourly (kept for 1 week)
  - No data files from runs
  - No log files from runs
- Use /cosma/apps for Python modules, self-compiled libraries, etc.
- 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

- Use the “quota” command:

```
For each file system, you have a soft quota (quota), and a hard quota (limit)
If you go over your soft quota, you will be warned, but still be able to
write files.  If you reach your hard quota, you will no longer be able to
write files.

######## /cosma5 quota ########

Disk quotas for usr dc-basd1 (uid 20957):
     Filesystem    used   quota   limit   grace   files   quota   limit   grace
        /cosma5  99.61G      5T      6T       -   12102  200000  220000       -
uid 20957 is using default block quota setting
uid 20957 is using default file quota setting

To see your /cosma6 quota, please run this on another login node

######## /cosma7 quota ########

Disk quotas for usr dc-basd1 (uid 20957):
     Filesystem    used   quota   limit   grace   files   quota   limit   grace
        /cosma7   2.76G      5T      6T       -      45  200000  220000       -
uid 20957 is using default file quota setting

######## /cosma8 quota ########

Disk quotas for usr dc-basd1 (uid 20957):
     Filesystem    used   quota   limit   grace   files   quota   limit   grace
        /cosma8  2.568T     20T     22T       -   26896  200000  220000       -
uid 20957 is using default file quota setting

######## /cosma/home (homespace) and /cosma/apps (application space) quota ########

Disk quotas for user dc-basd1 (uid 20957):
     Filesystem   space   quota   limit   grace   files   quota   limit   grace
172.17.170.16:/export/vol1
                36412M  40960M  48829M            354k    400k    540k
nfs.vastnet.pri.cosma.local:/apps
                 6421M    100G    110G            146k  10000k  11000k
nfs.vastnet.pri.cosma.local:/home
                39082M  51200M  61440M            266k    300k   1000k
```

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 not good for HPC
  - If you need this, please consider tarring up files, or concatenating them during writing
    - Or better, change the code
  - 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

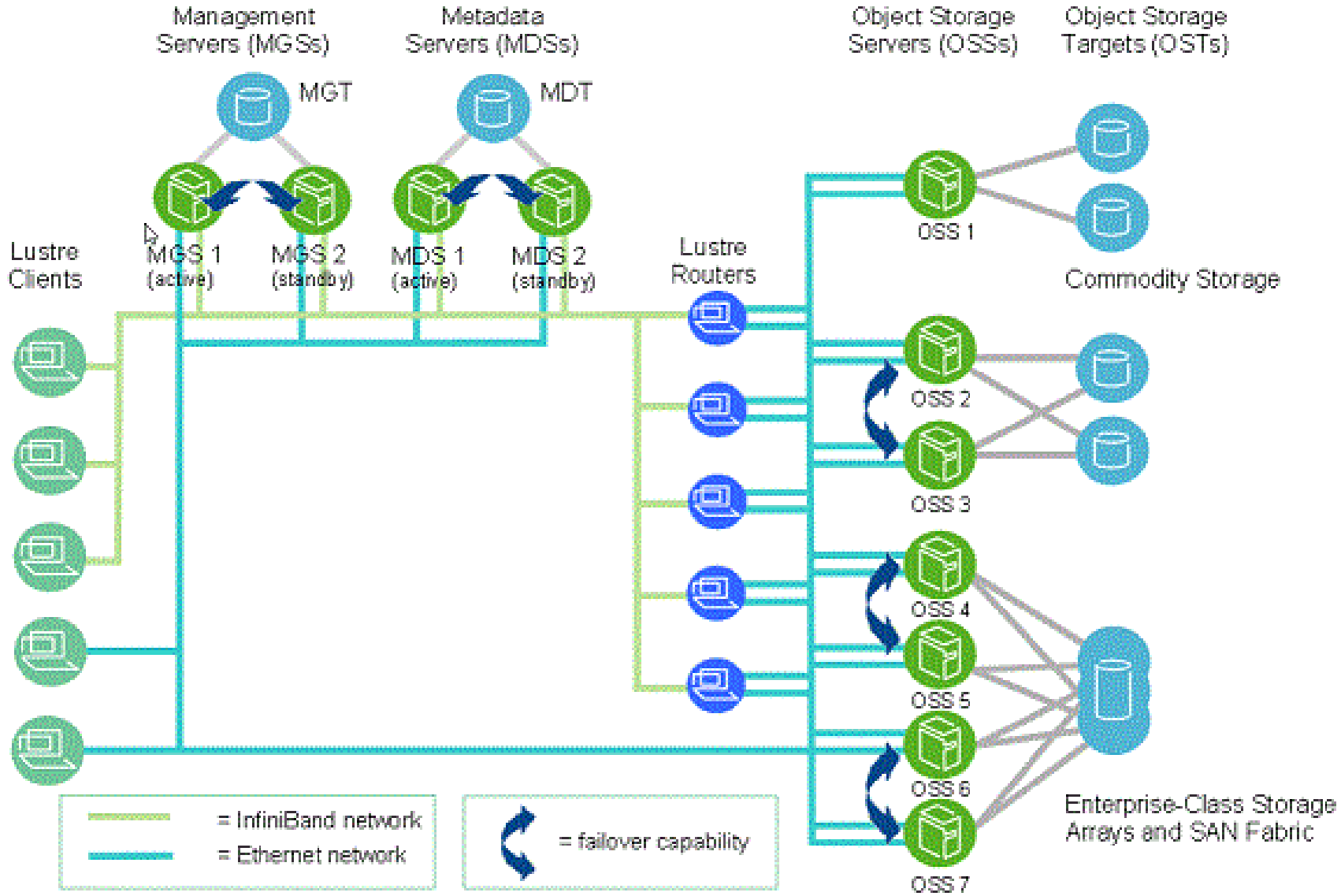
- Groups/projects also have quotas
  - If a project goes over quota, no one in that project will be able to write files
  - To check group quotas:
    - c7quota -g durham
    - lfs quota -g dp004 /cosma8 -h
  - Worth noting: If you are within your quota, but cannot write files, check the group quota

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, /cosma/apps 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



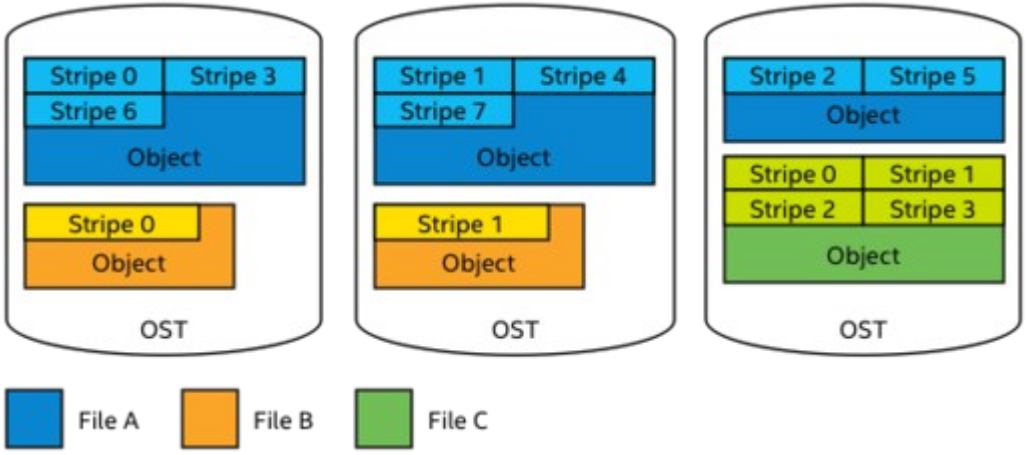
Credit: lustre.org

Lustre file operation 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



Credit: lustre.org

Writing large files

- If you are writing large files:

1) Set striping for the directory:

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

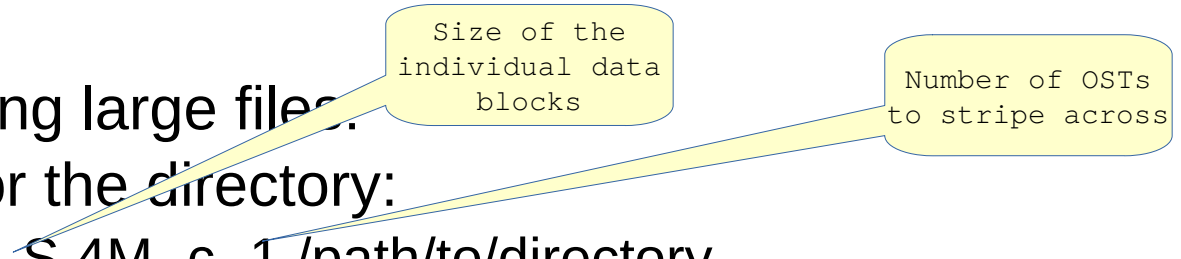
   All files here will then have this striping

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...

File striping

- To check the current striping on a file or directory:
  - lfs getstripe /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” 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
- Try to avoid accessing many 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
  - For long-term archive
    - e.g. post-publication
  - Ask if you want to transfer data...
- FAIR principles
  - Findable, accessible, interoperable, reusable
- Consider metadata
  - Data which explains the data!
- Consider version control
  - e.g. git
- There are various tools for sharing data with collaborators on and off COSMA

HPC Software installation/management

- A bit different from on your laptop
  - Many libraries are centrally managed
  - Usually performance optimised for the system
- You can also install your own
- Typically use the “Module” environment

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/10.5.2`
  - 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`
  - To 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 multiple versions of the same library (e.g. python)

Compilers and MPI libraries

- gnu_comp/VERSION
- intel_comp/VERSION
  - Follow the instructions once loaded
    - It suggests which modules you might want to load
- openmpi/VERSION
- intel_mpi/VERSION
  - Only for older versions
- aocc/VERSION

- 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 many work queues or “partitions”
- Useful commands:
  - sinfo, squeue, sbatch, scontrol, scancel, showq, sprio, srun, squota

COSMA5 partitions

- cosma5
  - Non-exclusive
    - If you don’t ask for all the cores on a node you will share it with others
      - Good for single core jobs
  - 2048 cores (8 nodes, 256 cores per node)
  - Often underused
    - So good for rapid turnaround on small jobs
  - 1.5TB per node – so good for large memory jobs

COSMA7 partitions

- cosma7
  - Standard queue: 224 nodes, 28 cores/node, 512GB
- cosma7-rp
  - Identical, but using a different network fabric
- cosma7-pauper
  - For users/projects with no allocation left
    - Allocations are arranged quarterly (every 3 months)
- cosma7-prince
  - When a large number of nodes are required
- cosma7-shm, cosma7-shm2
  - Single nodes 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

Summary of nodes

- cosma5: 256 cores/node, 1.5TB RAM
  - 100Gb/s
- cosma7: 28 cores/node, 512GB RAM
  - 100Gb/s
- cosma8: 128 cores/node, 1TB RAM
  - 200Gb/s

Others of interest

- bluefield1 partition:
  - ~16 nodes with 32 cores, 512GB
- dine2 partition:
  - 8 nodes with 64 cores, 2TB
- gracehopper
  - Single Grace CPU with Hopper GPU
- mi300x
  - Single node with 8x MI300X GPUs
- hardwarelab-gpu
- hardwarelab-cpu

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 cosma5
  - Shows only cosma5
  - Column ST shows state
    - Commen 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    cosma5   LWHalo dc-rega4 PD      14:56     36 m[5124-5136,5146-5156,5159-5170]
86282    cosma5  L400_de dc-pfei1  R    2:37:07     32 m[5184,5219-5242,5244-5250]
86281    cosma5  L400_de dc-pfei1  R    2:37:39     32 m[5185-5212,5214,5216-5218]
86363    cosma5 MMHalo_N dc-rega4  R      14:42     12 m[5123,5171-5181]
86376    cosma5 RECAL-h0   rcrain  R       0:54      8 m[5182-5183,5251-5256]
86191    cosma5   IC_Gen      arj  R    4:54:39      8 m[5137-5138,5140-5145]
86374    cosma5   energy     likm  R       5:47      1 m5262
85285    cosma5  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 cosma5                    # 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
```

srun

- Usually for testing
  - Allocates a node and moves your terminal prompt there
    - Within the Slurm environment – so that mpirun etc work
  - e.g.:
    - srun -p cosma8 -A dp004 -t 10:00:00 --pty /bin/bash

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 COSMA8 queue on the COSMA7 login nodes, do not optimise too far
  - C8 nodes don’t have AVX512 instructions
  - So, compile on the relevant login node!

CPU architectures

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

Python

- The installed Python has very few libraries
  - You should set up a virtual environment
  - Module load python/VERSION
  - Python -m venv /path/to/new/venv
  - Source /path/to/new/venv/bin/activate
  - Pip install whatever

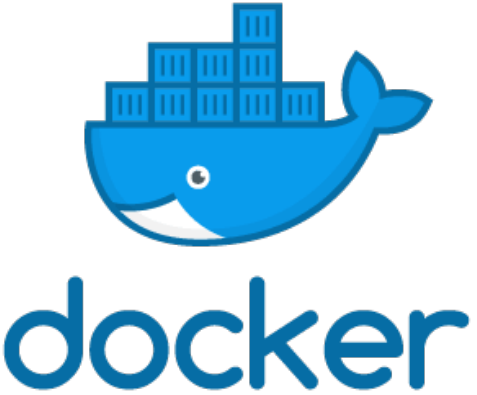
Jupyter hub

- A Jupyter hub is available on COSMA5/7/8 login nodes
  - And various other places
  - 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 cosma.rtfd.io for how to get set up
- It will require your cosma username and password
  - Note, not your ssh key passphrase

Containers

- COSMA has Apptainer installed
  - Can be used with apptainer or docker containers





Experimental hardware

- COSMA contains a HPC Hardware Lab:
  - New/novel kit for experimentation
  - This can be a good way of getting a publication
    - i.e. porting and running a code on a new GPU and reporting on performance
- Experiences with any large run may also be publishable
  - The tweaks required to get it running
  - Performance tuning
  - Profiling
  - etc

Cloud facilities

- COSMA includes a cloud platform:
  - Spin up your own VMs (with root access), slurm platforms, Jupyter Hubs, kubernetes, clusters
  - Etc
- Ask if this would be useful

Future updates

- DiRAC4
- Data Curation Service
  - Accessibility: Making data available for reuse
    - Maximise the extent of your research
    - Use metadata tagging to add meaning and context to the data
  - Long-term storage for projects
- Ongoing additions to the Hardware Lab