# Introduction to COSMA

Alastair Basden
and the COSMA team
(Peter Draper, Lydia Heck, Richard Regan and others)

cosma-support@durham.ac.uk

# High performance computing

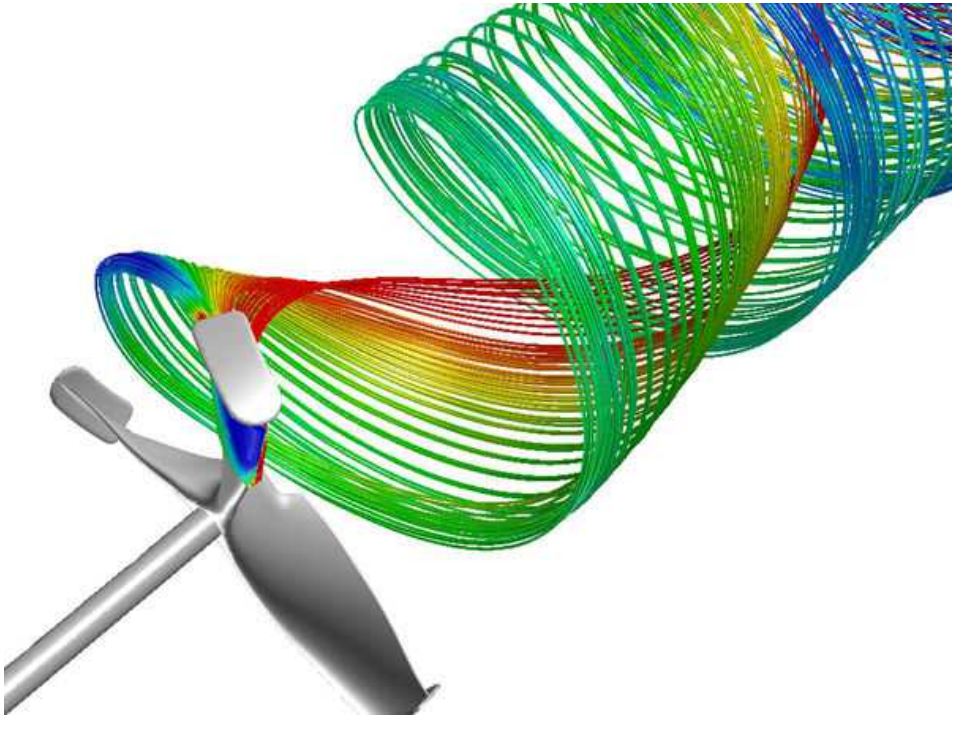
- What is HPC?
  - Use of parallel processing to run large applications
  - Typically applied to systems >10 TFLOPS
  - Aggregated computing power
    - More than can be obtained from a desktop
    - Used for solving large problems
- Cloud computing is not usually HPC
  - Typically only a single computer in the cloud is used

# The TOP500

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

# HPC use cases

- Galaxy simulation
- Weather simulation
- Artificial intelligence
- Fluid dynamics modelling
- etc



# DiRAC

- A tier 1 HPC facility for the STFC community
  - Astronomy, cosmology and particle physics
- 4 DiRAC sites:
  - Cambridge
    - GPU and Xeon Phi systems
  - Leicester
  - Durham
    - Memory intensive system – lots of memory per core
    - (note, since memory is expensive, COSMA typically has lower total compute power than other sites, but can perform simulations not possible elsewhere)
  - Edinburgh
    - Extreme scaling system – maximum compute power
- DiRAC 2 (At Durham = COSMA5) in 2012
- DiRAC 2.5 (At Durham = COSMA7) in 2018
- DiRAC 3 expected in 2020 or 2021

# The COSmology MAchine

- The Durham DiRAC node
  - Now in its 7th generation
  - COSMA5, 6 and 7 available for use

# COSMA5

- A Durham-only facility
  - Including collaborators
  - Was a DiRAC facility until 2018
  - If you are part of a DiRAC project, please try to use COSMA6/7
- Arrived in 2012
- ~300 nodes
  - 16 cores per node (2 CPUs)
  - 128GB RAM per node
  - Diskless
- 3x login nodes: cosma-a, -b and -c

# COSMA6

- A DiRAC facility
- Arrived in 2016
  - Second-hand, originally at STFC Hartree
- Approx 575 nodes
  - Identical to COSMA6 nodes
  - But also with disks
  - 9200 compute cores

# COSMA7

- A DiRAC facility
- Arrived in 2018
- Currently 147 nodes, to be 300 nodes by end October
  - ~400 nodes by March
  - Each 2x 14 core Xeon CPUs (circa 2018)
  - 768GB RAM (28GB/core)
- Also, mad01:
  - 3TB RAM
- mad02:
  - 0.75TB RAM, 4x CPUs,
    - 28 cores each (112 cores)

# DiRAC3: COSMA8

- £8M due in 2020-2021
- Likely to be conventional CPU architecture
  - (not GPU/FPGA)
- May include ARM or AMD processors
- Will again be memory intensive

# COSMA summary

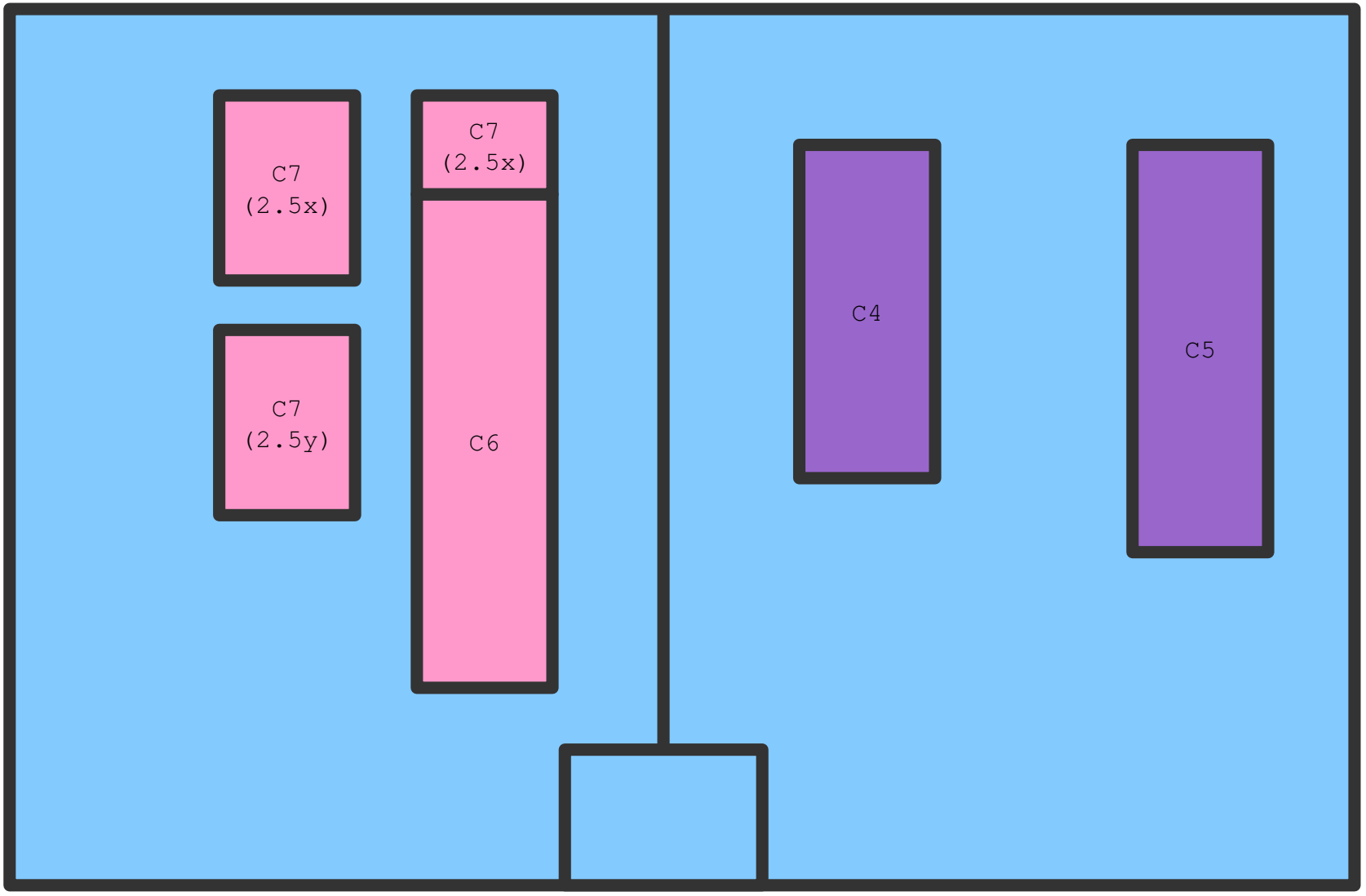
| | COSMA5 | COSMA6 | COSMA7 |
|---|---|---|---|
| Nodes | 300 | 575 | 400 (in 2019) |
| Cores/node | 16 | 16 | 28 |
| Cores | 4800 | 9200 | 11200 |
| Memory/node (GB) | 128 | 128 | 768 (may become 512 in 2019) |
| Memory/core (GB) | 8 | 8 | 28 |
| Total memory | 38TB | 74TB | 230TB |
| Login nodes | 2 (currently 3) | 1 (currently 0) | 2 |
| Operating system | CentOS 7.4 | CentOS 7.4 | CentOS 7.4 |
| | | | |
| | | | |

# COSMA networks

- Mostly, you don’t need to know this:
  - COSMA uses Infiniband for inter-node communication (MPI) and file system access
    - FDR10 for COSMA5/6 (40Gb/s)
    - EDR for COSMA7 (100Gb/s)
    - 2:1 blocking ratio
  - All nodes also connected by 1G Ethernet for communication/control/homespace

# 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...

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

As part of its normal functionality, 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.

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup

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

DiRAC

# SAFE for DIRAC

*Service Administration from EPCC*

safe

**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.

Any SSH key you register here will be included 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 |
| **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 | |

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup

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

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

| | |
|---|---|
| **Email Address** ★ | ali@framscouts.org.uk |
| **Nationality** ★ | United Kingdom |
| Title (Mr,Mrs,Dr etc.) | Mr |
| **First Name** ★ | Alastair |
| **Last Name** ★ | Basden |
| **Institution for reporting** ★ | University of Durham |
| **Department** ★ | Physics |
| Phone number (include International code e.g. +44 for UK) | + followed by numbers and spaces |
| **Opt out of user Emails** ★ | |
| **Address Line 1** ★ | Department ofPhysics |
| Address Line 2 | South Road |
| Address Line 3 | |
| Address Line 4 | |
| **Town/City** ★ | Durham |
| **Postcode** ★ | DH13LE |
| **Country** ★ | United Kingdom |
| SSH Public key | Browse... id_rsa.pub |
| **Gender** ★ | Male |
| **Career stage** ★ | Postgraduate Researcher: Ph.D. or other PG research degree |

Register

DIRAC SAFE guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC Account Signup

https://safe.epcc.ed.ac.uk/dirac/SignupServlet

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**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

DIRAC SAFE guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC Change SAFE Password

https://safe.epcc.ed.ac.uk/dirac/PasswordChangeRequestServlet/112

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**Please set a password for use with the SAFE**

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

DIRAC SAFE guide

SAFE is an EPCC product

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

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC Welcome

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

Your details | Service information | Login accounts | Help and Support

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**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 guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC SAFE

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

Your details | Service information | Login accounts | Help and Support

DiRAC

# SAFE for DIRAC

*Service Administration from EPCC*

safe

**SAFE for DIRAC**

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

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.

If a menu is coloured orange it indicates a pending request that needs your attention.

DIRAC SAFE guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC Login account Request

https://safe.epcc.ed.ac.uk/dirac/TransitionServlet/User//-/Transition=Cho

Your details | Service information | Login accounts | Help and Support

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**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 - Durham

Next

DIRAC SAFE guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC Login account Request

https://safe.epcc.ed.ac.uk/dirac/TransitionServlet/User/

Your details | Service information | Login accounts | Help and Support

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**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 | Machine | Type | Description | Policies |
|---|---|---|---|---|
| ◉ | cosma | | The Durham COSMA machine | • Accounts named after dirac-id ✔<br>• Only one account per person is allowed ✔<br>• Users can register a public key to access the machine ✔<br>• Password authentication is not allowed |

Next

DIRAC SAFE guide

SAFE is an EPCC product

File Edit View History Bookmarks Tools Help

DIRAC SAFE Signup Successful | DIRAC SAFE Login | DIRAC Login account Request

https://safe.epcc.ed.ac.uk/dirac/TransitionServlet/User/

Your details | Service information | Login accounts | Help and Support

DiRAC

SAFE for DIRAC

*Service Administration from EPCC*

safe

**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

| | |
|---|---|
| **Requested username ★** | dc-basd2 |
| **IP range to connect from ★** | 0.0.0.0/24 Comma seperated list of ip ranges in the form x.x.x.x/y |

Request

DIRAC SAFE guide

SAFE is an EPCC product

# 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
  - A shared facility
    - Usually with extra RAM
  - Via ssh:
    - ssh USERNAME@login.cosma.dur.ac.uk
    - ssh USERNAME@login7.cosma.dur.ac.uk


Credit: fotolia.com

# ssh on COSMA

- Authentication requires an SSH key:
  - A key has 2 parts
    - Private part (keep this very safe!)
    - 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

# Generating an ssh key

- ssh-keygen -t rsa -b 4096
  - This will ask for a passphrase
    - Please use one – this protects your private key
  - This will create:
    - id_rsa  (private key – keep this safe)
    - id_rsa.pub (public key – upload this to SAFE)
  - We will then append your id_rsa.pub key to the .ssh/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.

# ssh key example

```
cosma7:~$ ssh-keygen -t rsa -b 4096
Generating public/private rsa key pair.
Enter file in which to save the key (/home/user/.ssh/id_rsa):
Created directory '/home/user/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/user/.ssh/id_rsa.
Your public key has been saved in /home/user/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:T4Ey6lfKrcFB/fjELAuV7THm/MAAis1xdY5IKLnWhFE user@computer
The key's randomart image is:
+---[RSA 4096]----+
|   .=E.+.. .      |
|   ++o= + B       |
|   .=+ = B B      |
|   o .o + % +     |
| .   . o S %      |
|    . o * B o     |
|     . * 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

# X-forwarding

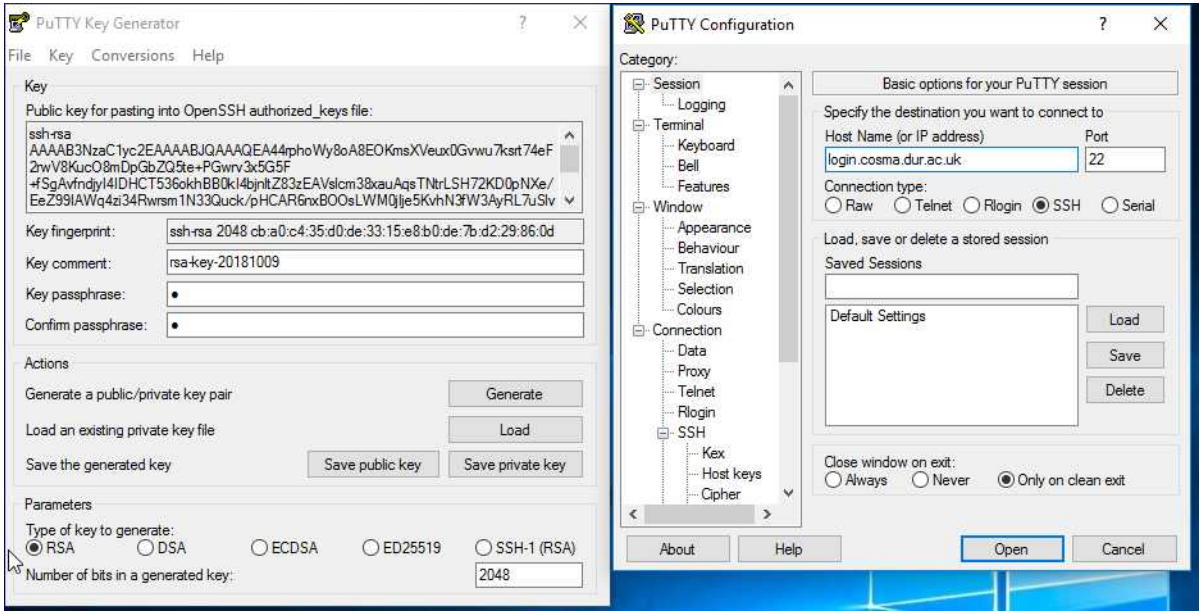
- If you want graphical access:
  - ssh -X USER@login.cosma.dur.ac.uk
  - (or -Y if that doesn’t work well – less secure)
- You can then start remote programs which will display windows locally, e.g.
  - text editors
  - plots of your data
  - movies etc are not recommended unless you have an excellent connection
- Note: X-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

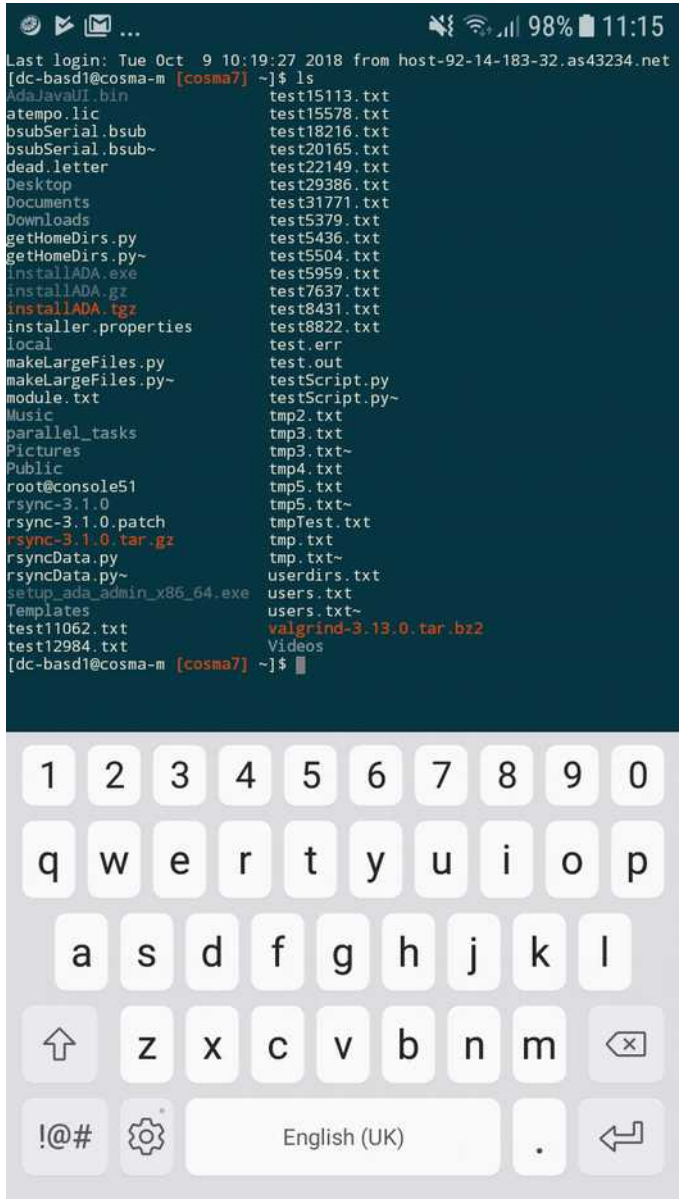
# COSMA passwords

- Are only used to authenticate with the cosma website
  - (to access usage statistics, etc)

# ssh from other platforms

- On Windows, use e.g. putty
  - You will need to run puttygen first to create the ssh key pair
- On Android, use e.g. JuiceSSH
  - You will need to generate an ssh key pair (using the app)



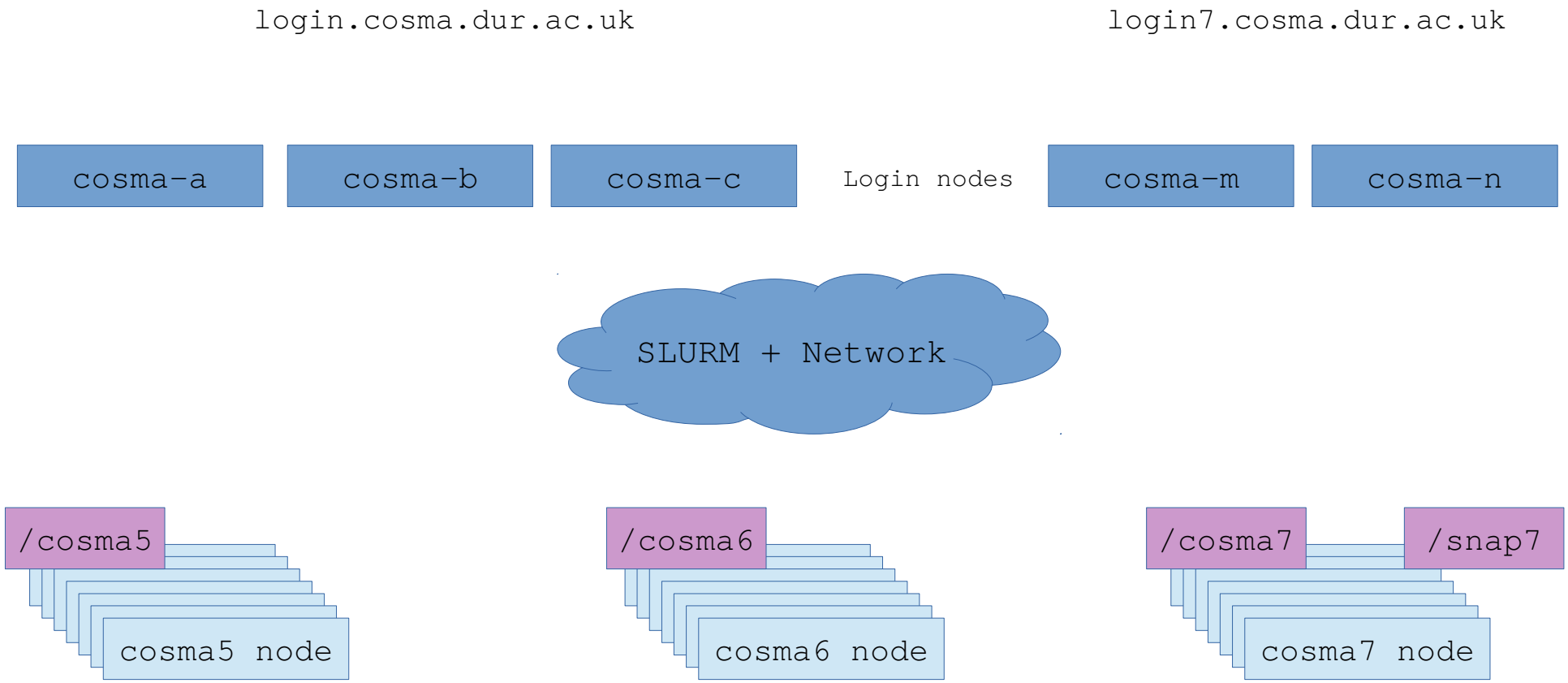
# Typical workflow

- 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:
    - 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, either start firefox over X (slow), or forward the ssh connection from your desktop:
  - ssh -D 1234 USER@cosma-m.cosma.dur.ac.uk
  - chromium-browser –proxy-server="socks5://localhost:1234" https://webpage.on.cosma

# COSMA login nodes

- Currently, COSMA5 and COSMA7 have login nodes
  - COSMA5 has 3 nodes: cosma-a, cosma-b and cosma-c
  - COSMA7 has 2 nodes: cosma-m and cosma-n
- Shortly, one COSMA5 node will be moved to COSMA6
- All login nodes offer access to all facilities
  - File space, job queues, libraries, tools, compilers, etc
- In general, please use COSMA6 or 7 for DiRAC projects, and use COSMA5 for internal Durham projects
  - Ask (us or supervisor) if not sure
- login.cosma.dur.ac.uk  (round-robin allocation to -a, -b, -c)
- login7.cosma.dur.ac.uk (round-robin allocation to -m, -n)

# Graphical COSMA



# COSMA file system

- Your home folder will be in
  - /cosma/home/PROJECT/USERNAME
  - PROJECT is probably durham
  - 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)
  - /cosma6/data/PROJECT/USERNAME (10TB/2.5PB)
  - /cosma7/data/PROJECT/USERNAME (10TB/2.1PB)
  - /snap7/scratch/PROJECT/USERNAME (Unlimited/346TB)
- Data space is optimally connected to the corresponding COSMA
  - Reading /cosma6/ from COSMA6 will be faster than from COSMA7
  - /snap7 space is temporary storage to use within a single run
    - e.g. for restart points
    - Fast SSDs
    - At one point was the fastest storage in Europe
- /cosma/local/ is where tools and libraries are located

# File system quotas

- quota  (for home space):

```
Disk quotas for user dc-basd1 (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-basd1
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-basd1
Filesystem     usage      quota      limit      files      quota      limit
--------------------------------------------------------------------------
/gpfs          0GB        0GB        0GB
/cosma5        33.4439GB  5120GB     5200GB     23         0          0
```

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

```
Quota for dc-basd1
Filesystem     usage      quota      limit      files      quota      limit
--------------------------------------------------------------------------
/cosma7        0.00390625MB 0MB        0MB        1          0          0
/cosma6        31.9992GB  0MB        0MB        17         0          0
```

- Eventually, 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

# 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

# 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...

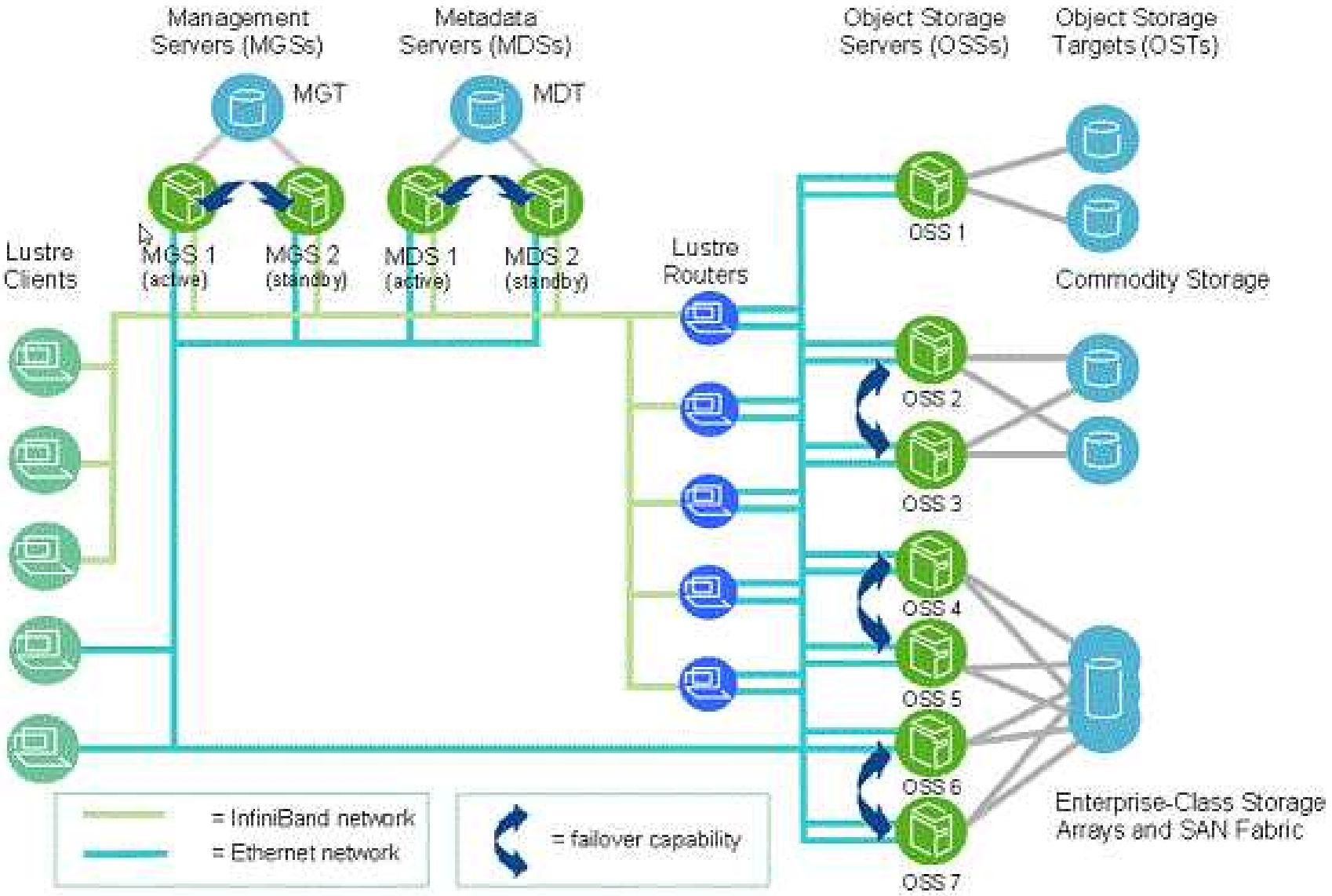
- Use /cosma/home/ for scripts, code, self-compiled libraries, python modules etc.
  - This is backed up
  - No data files from runs
  - No log files from runs
- Use /cosma[567]/data/ for input and output data produced by your runs
  - This is archived to tape
- Use /snap7/scratch/ if running on COSMA7 for staging posts, restart files etc.
  - This is not backed up, and may be removed if older than a few days.

# Parallel file systems

- Storage of data across multiple servers
  - Data is distributed across these
    - Striped
- High performance access
  - Simultaneous reads/writes
- COSMA uses:
  - Lustre for /cosma6, /cosma7, /snap7
  - GPFS for /cosma5
  - NFS (not parallel) for /cosma/home and /cosma/local

# 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



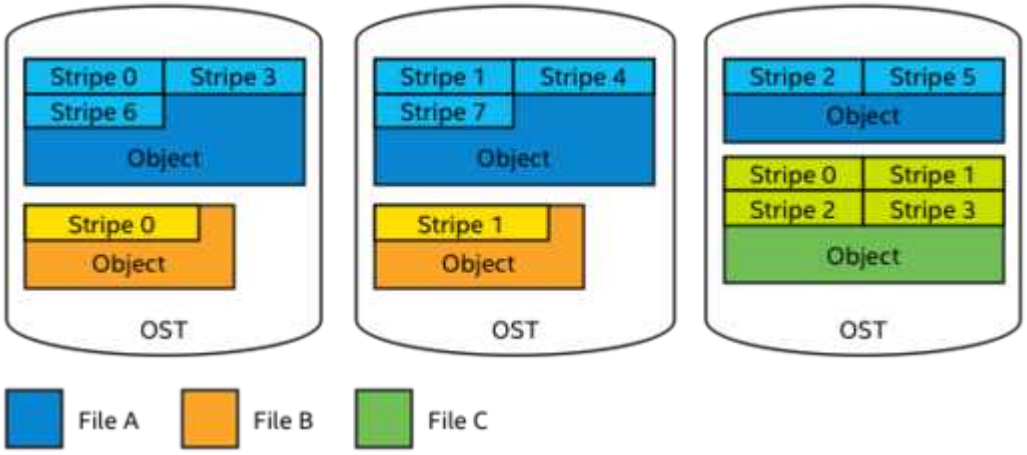

Credit: lustre.org

# Lustre 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)
- Striping is inherited from the directory that a file is in
- Default striping for COSMA is 1 (i.e. not striped)




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

Size of the individual data blocks

Number of OSTs to stripe across

- 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 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”
  - File size is only stored on the OSSs
  - Use “ls” to see if a file exists
  - 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

# 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
- intel_comp/2017
- intel_comp/2018
- openmpi/3.0.1
- intel_mpi/2017
- intel_mpi/2018
- 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

# COSMA5 partitions

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

# COSMA6 partitions

- cosma6
  - Standard cosma6 queue
- cosma6-pauper
  - A poor-mans queue for projects with no resources left
- cosma6-prince
  - When a large number of nodes are required

# COSMA7 partitions

- cosma7
  - Standard queue
- cosma7-pauper
  - For users with no allocation left
- cosma7-prince
  - When a large number of nodes are required

# sinfo

- sinfo
  - Shows a list of the nodes allocated to the different queues

```
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[7001-7036,7038-7063,7065-7142,7144-7147]
```

- sinfo -p cosma7
  - Show only cosma7 nodes
- See the man pages for further information:
  - man sinfo

# squeue

- Shows the current state of the queues
- e.g. squeue -p cosma
  - Shows only cosma
  - 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     cosma   LWHalo dc-rega4  R      14:56     36 m[5124-5136,5146-5156,5159-5
86282     cosma  L400_de dc-pfei1  R    2:37:07     32 m[5184,5219-5242,5244-5250]
86281     cosma  L400_de dc-pfei1  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

```
#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=userid  # 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 part 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
```

# 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[567]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, if compiling for a COSMA6 queue on the COSMA7 login nodes, do not optimise too far
  - C6 nodes are older and illegal instructions may result
  - (C5 node CPUs are identical to C6)
- A COSMA6 login node will be made available to alleviate this problem shortly

# Jupyter hub

- A Jupyter hub will shortly be made available on the COSMA7 login nodes
  - But, this is not a good HPC paradigm and should really only be used for analysis of results

# Future updates

- Singularity
  - A modularised container facility for HPC
- COSMA7 expansion
  - Up to 300 nodes by the end of October
  - Up to 400 nodes by March 2019
- COSMA8
  - Expected sometime in 2020-2021 as part of DiRAC3
- Hierarchical storage management
  - Automatic archiving of unused files
    - Debate about automatic retrieval
  - Releasing filesystem space for other users