Beginners' Guide to the Science Cluster for Students at the Institute of Astronomy

Introduction

These notes refer only to the setup of the Science Cluster at the Institute of Astronomy. If you work with one of the research groups that runs its own computing facilities (e.g. Gaia, IMAXT), or in other departments, at least some of these notes will not apply.

Staying safe on-line

Who is responsible for the security of computers?

We all are!

- IT team works hard to make sure our systems are safe and secure
- You need to play your part too

What can you do:

- Apply the latest software updates for OS and applications on your devices
- Run antivirus software on your laptop
  - Make sure it is up to date
  - Windows defender is good - make sure it's not disabled
  - Mac users can download McAfee Endpoint Protection: https://help.uis.cam.ac.uk/service/security/antivirus (also available for Windows)
- Always keep your passwords private
  - Use a different password for each account
  - Consider using a password manager app (e.g. lastpass)
  - Enable 2- / multi-factor authentication for critical services, if available
- Be aware of spear-phishing: targeted phishes, sometimes using social engineering
  - Watch out for links that take you to fake web sites, particularly in unexpected emails - hover to get a preview of where it will take you
  - Check for the padlock in URL bar when you have to type e.g. passwords
  - Check the certificate in the padlock for critical sites
  - Be very careful before opening attachments in unexpected emails - who knows what the file will do
- If you think you have been caught by a scam, don't be embarrassed
  - Change any passwords that might have been compromised, immediately
  - Please tell us so that we can check things out

- UIS Stay Safe Online training: https://help.uis.cam.ac.uk/service/security/cyber-security-awareness

Passwords

UIS provides many of the services that you use, e.g. email (ExOL), High Performance Computing. You use your UIS (Raven) password to access these services. For some of these services, your username will be CRSid@cam.ac.uk, and for others it will be CRSid.

You will also need to use your UIS (Raven) password to access the IoA local web-pages (Intranet: https://local.ast.cam.ac.uk), and you will need to be on the VPN.

The IoA Science Cluster uses CRSid as the username but uses different password servers to those used by UIS.

- CRSid@cam.ac.uk/UIS password - to access many centrally managed University services e.g. email
- CRSid/UIS password - to access the IoA Intranet
- CRSid/Science Cluster password - to log in to Science Cluster (Linux) computers.

How to change your IoA Science Cluster password

You can change your password on any Science Cluster system that you can login to, and the change will take effect immediately.

The command to change your password is `passwd`. Please choose a secure password: at least 8 characters long, at least one number, at least one upper case letter, at least one lower case letter, and at least one non-alphanumeric character. Do not use whole words that can be found in a dictionary, and do not use passwords that you use elsewhere.

What is the IoA Science Cluster?

The Science Cluster is a group of networked servers and desktops that provide a consistent environment for your scientific computing at the IoA.

- The main operating system used is RedHat Enterprise Linux (RHEL) and currently the systems, apart from a few servers, are on RHEL 7.
- Systems are managed by the IoA IT team
- Users do not have root or sudo access

- All systems see the same networked filesystems - your data is available on every system
- Various access rules are designed to protect the integrity of the system
- More details at `https://local.ast.cam.ac.uk/computing`. Hint: try the search box.

Window Manager

There are two main desktop environments, GNOME and KDE. Gnome is the default. Which you choose is a matter of personal preference.

- If you are an IoA PhD or MPhil student you will have a desktop system in your office that you can log into directly at the console using your Science Cluster username and password.
- Part III and MASt students will need to log in from another system such as their laptop either by ssh or through the NoMachine servers calx011 and calx026.

System reboots (every Monday morning)

Please note that systems at IoA reboot weekly at 00:35 on Monday mornings. This is to ensure that all security patches are activated regularly.

If you need to run long jobs spanning a weekend there are ways to have them save their state across reboots. Check the Intranet for references to checkpointing software `dmtcp`.

You may also request an exemption from the reboot of your desktop workstation by emailing helpdesk. Exemptions are not possible for public server systems.

Logging in from outside:

The IoA computer systems are subject to a security model which only allows you to log into public servers and your own desktop in specific ways.

muon gateways: https://local.ast.cam.ac.uk/computing/remote-access-security/gateway-accessible-machines

(There used to be a cat called Muon at the IoA http://www.sverre.com/images/muon_sv_640_framed.jpg)

NX server (shared and your IoA workstation). If your NoMachine session is running on `calx011.ast.cam.ac.uk` or `calx026.ast.cam.ac.uk` and you need to run cpu intensive jobs, please ssh to `calx079.ast.cam.ac.uk` or `calx092.ast.cam.ac.uk`.

Terminal session login

In this section are instructions for starting a simple terminal session from a remote system. First start a terminal on your local system:

Terminal programs

- From Linux use any terminal program
- From Mac OSX start `Terminal` or `xterm`. You will need to make sure that X11 is installed. This is provided by the XQuartz package from https://www.xquartz.org/ first. Starting X11 starts an `xterm` automatically.
- From Windows you will need to install an X11 capable ssh terminal client program, such as MobaXterm Free Edition from https://mobaxterm.mobatek.net/.

From the Astronomy managed VPN

If you can connect to the Astronomy managed VPN, then you will be able to ssh directly to one of the public servers or your desktop (PhDs and MPhils). For example, to log into the server calx024:

```
ssh -Y username@calx024.ast.cam.ac.uk
```

From any other network

If you cannot connect to the Astronomy managed VPN, you need to login through one of our gateway systems: `muon1.ast.cam.ac.uk`, `muon4.ast.cam.ac.uk`, `muon5.ast.cam.ac.uk`, or `muon6.ast.cam.ac.uk`.

Note that gateway systems are low specification systems intended for granting access to the compute servers. **Do not run compute intensive jobs on gateway systems.** Once you are logged on to a `muon` gateway system you can run non-compute intensive commands there or ssh on to one of the compute servers.

In summary, you need to use the following commands to login via the gateway systems:

```
ssh -Y username@muonN.ast.cam.ac.uk
```
(where N=1,4,5,6) and substitute your real username for the string ``username''.

Then, eg:
```
ssh -Y calx024
```
to log in to the compute server.

In these examples `-Y` says to forward the X11 (graphics display) connection (so that you can start graphical applications).

Note that in the case of the onward login you don't need the `username@` or `.ast.cam.ac.uk` because they are assumed to be the same as for the current login on `muonN` if not specified.

Filesystems and disk space

When you log into one of the Science Cluster computers, you will be able to access various files:

- software local to the computer
- software installed on a network disk
- data on data disks
- your personal files in your home directory

What do we mean by 'home directory'?

This is your working directory is when you log in.

Log into muon1:

```
ssh -l username muon1.ast.cam.ac.uk
```

`pwd`
Print Working Directory: This tells you what directory you are currently sitting in.

`ls -lart`
List all files (including dot files) in reverse time order, showing the permissions of owner, group and other users.

The 'dot files' (files beginning with a '.') are usually configuration files for a particular piece of software. They are important for configuring your environment, so don't delete them.

To see other options for the `ls` command, you can look at the man page:
`man ls`

`df -h .`
Display Free disk space in human readable form..
The '.' means the current directory ('..' is the directory above).
This command not only tells you how much free space you have in the current directory, but also tells you where the filesystem is hosted.

`df -h /tmp`
/tmp is a local filesystem (exists only on that computer) that can be written into by anyone, but the files are temporary and will vanish. May be useful if writing out a parameter that only needs to exist for a short period of time on that computer.

Types of disk space

You have a 'home' directory, and can be allocated space on one or more 'data disks'. The difference between them is the type of storage, how resilient it is to failure, if it has snapshots, and how often it is backed up.

home directory: the most valuable/expensive storage, as it has snapshots, is replicated in another building, and is backed up weekly to a third building.

/data/vault: one nightly snapshot, is replicated in another building, and no backup by the IT team

data directory: larger space available, but no snapshots, no replication and no backup by the IT team.

All public file systems are protected from individual drive failure as we use redundant arrays of disks. This should not be considered a backup system since there are many other ways in which you can lose data, including:

- Deleting or overwriting a file by accident
- File system corruption - rare
- Multiple concurrent disk failures - not as rare as you might hope
- Hackers
- Fire
- Theft...

/home

Your initial allocation (under `/home/username`) is 5GB. Although a small allocation, this space is protected by multiple levels of backup so is good for high value files such as your own code and documents you write.

You will find point in time snapshots of your home space taken over the previous 2 months under the `~/.snapshot` directory. Further system level copies of your home directory are also made automatically.

Snapshots are a space efficient way of creating a backup by keeping a baseline copy and only saving (parts of) files that changed since the baseline. Don't churn the contents of your home space repeatedly because there is not space to keep nearly 90 complete copies of everybody's home directory. Ask for data space instead.

To see how much space you have remaining, use the `df` (disk free) command, e.g.

```
df -h /home/username
```

Note that exhausting your disk space in /home can lead to strange errors, so keep an eye on it.

/data/vault

You may request an allocation up to 20GB on `/data/vault/username` from helpdesk. The level of protection here is lower than for /home. These directories have a single daily snapshot (`/data/vault/username/.snapshot/nightly.0`) which is also mirrored to a separate filesystem. However, *this is not a backup system* because the snapshot and mirror copy is overwritten every day at midnight.

/data

If you need it for your research you can get access to large amounts of disk space, on the `/data/username` filesystem, by sending an email to helpdesk@ast.cam.ac.uk. **In general /data is *not* backed up at all.**

You can see how much space you have available using e.g. `df -h /data/username`.

Network File System (NFS)

The home directory is hosted on a server in the server room, and is shared to the various computers of the Science Cluster,

The NFS clients run software that automatically mounts the home directory on a mount point (/home) when it is required, and umounts it when it is no longer being used. 'Data disks' are also NFS filesystems and are automounted on /data when required.

You can imagine /home and /data mount-points as hat-stands, which are empty until someone uses them. If you look at /data:

```
ls /data
```

you will see only those data disks that are being used on that computer at that time. To ask the automounter to mount your own data directory, you need to refer to the next level down (e.g. /data/username). The automounter will unmount a data directory or home directory that has not been used for ten minutes.

Backing-up your work

If you have data on a `/data` disk *it is vital* that you back this up if you care about it. People sometimes learn this fact the hard way. *You may lose several years of research if you do not back up, and you won't get any allowances for this from the university*.

It is usually not possible, and almost never cost-effective (in the academic environment), to try to recover deleted files, particularly on a shared disk.

Do not try to clean up your directory before backing it up: you could accidentally delete what you intend to backup.

Making your own backup

Alternative ways of backing up your data are either:

- Copy the data to different disks on different computers (perhaps including your laptop) in other locations.
- Copy the data to removable media on Blue-Ray DVD
- Copy the data to cloud storage e.g. Google Drive through Gsuite@Cambridge https://help.uis.cam.ac.uk/service/storage/google-drive or Microsoft OneDrive https://help.uis.cam.ac.uk/service/collaboration/365/onedrive

Do try restoring a backup occasionally to check it works! Be careful not to overwrite your current version of files with the old ones in the backup.

Simple tar backups

The simplest way to backup your work is to make a compressed `tar` archive (similar in concept to a zip file) like this:

```
tar czf superdata2021oct04.tar.gz superdata/
```

which will write a file `superdata2021oct04.tar.gz` containing the contents of directory `superdata/` (and everything under it) to the current directory. Copy it manually to **somewhere else** (different filesystem, different computer system, cloud storage, laptop etc). Other common tar commands:

- List contents of tar file: `tar tzf superdata2021oct04.tar.gz`
- Extract tar file to current directory: `tar xzf superdata2021oct04.tar.gz`

rsync

`rsync` is the command of choice to copy files to a different directory since it only copies files that have changed since you last did this copy.

You need a command like this to copy files to a different directory:

```
rsync -av /home/foo/indir/ /data/foo/outdir/
```

This will replicate any files in indir to outdir. **Note:** the trailing slashes (/) on the directories are *very important*.

For a secure backup make sure to copy files to a different physical device.

rsync will not delete any files in outdir that no longer appear in indir. You need to add the `--delete` option (after `-av`) in order to get this.

If you want to copy files to another computer you can do that but the remote system must be running an ssh server. For example:

```
rsync -av /home/foo/indir/
username@somecomputer.cam.ac.uk:/home/foo/outdir/
```

The `username@` part is optional if you have the same username on the destination computer.

DVD/Blueray backups

The system `calx037.ast.cam.ac.uk` (Hoyle Photocopier room H30) has a Blueray writer. Standard disks can hold 25GB.

To make a backup on DVD or Blueray you need to make a virtual filesystem with your files in it and then burn that filesystem to the disk.

Although you can do this by hand from the command line its easier to use a graphical tool like `brasero` or `k3b` which will hide all the details from you behind a GUI.

Logging into other shared systems on site

There are a number of shared systems that you can log into to run jobs.

The command to log into another machine is `ssh -Y hostname`. The `-Y` option allows graphical output to be routed back to your local system. For example:

```
ssh -Y calx024  [or ssh -Y calx024.ast.cam.ac.uk]
[type your commands here in this new shell]
[type exit to return to your computer]
```

https://local.ast.cam.ac.uk/computing/resources/shared-machines

Running science analysis software

We are now using Environment Modules to start all new science software packages. A module sets up your environment to run a particular piece of software, so that it finds the correct libraries.

Environment modules make software packages available for you to run and allow you to change versions easily. See https://local.ast.cam.ac.uk/computing-user-guide/programming/module-environment

You can use the `module avail` commmand to see what is available. Here is truncated output from that command:

```
> module avail
Aladin/7.015                                fftw/3.3.4_double
openmpi/gcc/4.5.0/1.4.2
Aladin/7.5beta                              fftw/3.3.4_float
openmpi/gcc/4.7.2/1.6.5
Aladin/8.040                                fftw/3.3.6_double
openmpi/gcc/6.4.0/2.1.1
anaconda2/20161102                          fftw/3.3.6_float
openmpi/gcc/7.2.0/2.1.1
anaconda2/20161208                          fposs/P98
openmpi/intel/11.1/1.4.2
.
.
.
```

If you want to know what one of these module files does, use e.g. `module help anaconda2`. To load a module you can use e.g. `module load anaconda2`.

Python

There are many versions of python on the Science Cluster.

New student accounts are set up to activate the anaconda python 3 install.

There are two commonly used branches of python, 2.7 and 3.x. Python 2.7 became end-of-life at the end of 2019 so it is strongly recommended that new projects use python 3. Note however, that much locally written software may only work under python version 2.7. You will probably need to switch back to python 2.7 to run this:

```
module unload anaconda3
module load anaconda2
```

There are some major incompatibilities between python 2 and python 3 so code changes will be needed when changing major version.

To change the default version of python when you log in, edit your `~/.profile` shell initialization file so that the lines:

```
#module load anaconda2 # Anaconda python 2.7
module load anaconda3 # Anaconda python 3
```

become:

```
 module load anaconda2 # Anaconda python 2.7
#module load anaconda3 # Anaconda python 3
```

Log out and log back in to get the changed default.

nice'ing your job

If you are planning to run a long job (longer than a few minutes) on a **public** system, you should *nice* your job. Niced jobs run at a lower priority level than interactive jobs and so impact interactive performance less. Nice levels run between -19 (high priority) and 19 (low priority). Programs normally run at level 0 and you can only reduce this priority by setting higher nice levels. You would use:

```
nice -n 10 ./myjob
```

to run the job `myjob` at nice level 10 (the required nice level for long running jobs).

The easiest way to see how loaded a computer is (how many things are running on it), is to use the `top` command which shows the `topmost', or most CPU intensive, jobs running in real time. Press `q` to exit it (press `?` to get other possible commands).

Disconnecting and reconnecting from remote sessions

Normally, when you log out of a session it closes all your running programs (except under some circumstances - see Linux tutorial for details). It is however possible to disconnect sessions and leave (particularly) interactive programs running in a way that you can reconnected to them later.

Simple terminal sessions

You can make terminal sessions that you can disconnect from and reconnect to later using the `tmux` command, which also allows multiple terminal sessions in the same connection. After logging in to the remote system using `ssh` type `tmux` to start a tmux session; then run your programs.
If you want to disconnect, type `ctrl+b` followed by `d` (most `tmux` commands use `ctrl+b` followed by something - try using `ctrl+b` then `?`).
You can then reconnect later by logging into the same computer and typing `tmux attach`.
To close your tmux session type `exit`.

NoMachine remote desktop sessions

Running NoMachine sessions can be disconnected by closing the display window on your client system. When you make a connection back to the server later you will be offered the options to either resume the running session or to start a fresh session.

Batch job submission

The *HTCondor* system allows you to send jobs to a much larger pool of systems, including other desktop systems. You set the requirements for your job in a submit file and add it to the queue to be run. Jobs are matched to systems with sufficient resources to run them and run fairly (nobody can hog the queue) from the queue among the available computers.
Documentation is in the Users' Guide on the Intranet.

High Performance Computing

The University runs a high performance computing centre https://www.hpc.cam.ac.uk/ which you may be able to get access to if your supervisor deems it appropriate for your project. See also our Intranet at: https://local.ast.cam.ac.uk/computing/resources/high-performance-computing

When things don't work

Self help: Check for known problems or system maintenance

If things aren't working as expected, please check this isn't a known problem or due to planned maintenance. This information is posted to:

- Our Intranet: https://local.ast.cam.ac.uk/computing
- Login message of the day: That's all the text that scrolls past when you log in

Self help: Read the documentation on the Intranet

A common problem when software doesn't work as expected is that you haven't set it up as documented.

- https://local.ast.cam.ac.uk/computing/resources/software
- Or use the Intranet search box

Self help: Check your environment modules

- Which modules do I have loaded?
  `linux> module list`
- Which modules are available?
  `linux> module avail`
- Load a module:
  `linux> module load <modulename>`
  e.g. `linux> module load gsl`
  initializes the default (usually the latest installed version) of the gnu scientific library
  e.g. `linux> module load gsl/1.13`
  initializes a specific (older) version of gnu scientific library
- Unload a module:
  `linux> module unload <modulename>`

Self help: Check disk space and whether the problem is specific to one computer

Use the df command to check that you have space free in your home directory and any data directory where you need to write files. Try running the command on a different computer, to find out if the problem is specific to one computer.

Send an email to helpdesk@ast.cam.ac.uk
https://local.ast.cam.ac.uk/computing/getting-support

When you contact helpdesk we will do our best to resolve your problem. It's much easier for us to do this if you give us as much information about the problem as you can. For example,

The important points are to tell us:

- Which computer you are logged into when having problems (use `hostname` to find out)
- Whether you get the same result on a different computer
- What exactly you are trying to achieve
- What you typed
- Any error messages - these can be really helpful
- Any specific time constraints

Cutting and pasting from your session is good.

Printing

Printing from Science Cluster (calx systems)
A list of available printers is at:
https://www.ast.cam.ac.uk/computing/printers.scanners.copiers. The printing dialogue from modern programs on the Science Cluster will allow you to choose from a list of printers.

Alternatively, you can print Postscript, PDF and text files directly to the printers from the command line. The Unix command to print files is `lpr` or `lp`. Do try to view files before printing to make sure they look okay.

Plain text files will render better if printed from a text editor such as `gedit` or `emacs`, rather than sending them directly to the printer.

To use a particular printer you need to determine its queue name. This is the first part of the hostname that is written on the printer (ie the bit before `.ast.private.cam.ac.uk`). So for the printer called `colour5.ast.private.cam.ac.uk` the queue name is `colour5.`

The command to view what is on the *queue* (assumed to be `queue_name` in these examples), i.e. list the jobs waiting to be printed is:

```
lpq -P queue_name
```

To print a file on the printer:
```
lpr -P queue_name filename
```

If you realise you didn't want to print that file, find the number of the job you printed (see the output of the `lpq -Pqueue_name` command, and do:
`lprm -P queue_name job_number`

Setting a default printer on the Science Cluster

If you are nearly always printing to the same printer you should set a default printer. To make `queue_name` your default printer, add a line like the following to the file `.profile` in your home directory:
`export PRINTER=queue_name`

You won't need to use -P option on the lpr command for that printer after logging in again and the correct printer should be selected automatically in print dialogues.

Printing from your laptop

You can print directly to the IoA printers from your laptop after following the setup instructions at
https://www.ast.cam.ac.uk/computing/printers.scanners.copiers/setup.guides

Office packages

Your University of Cambridge Microsoft account gives you Office 365 ProPlus on your personal devices. You have access to Office365 Online (https://portal.office.com from the Science Cluster or to downloadable Office apps for your (Windows or Mac) laptop at https://help.uis.cam.ac.uk/service/user-accounts-security/accounts-passwords/microsoft-accounts/copy of ees-software.

Alternatively, on RHEL you can do do word processing, make presentations, or use spreadsheets, using the `libreoffice` suite. This is a very capable office suite, particularly so when used with its own file formats.

`libreoffice` can read (many) Microsoft Office documents but sometimes files don't render as expected because the file format is proprietary and has to be determined by trial and error by the developers.

Finding information, data and papers

https://people.ast.cam.ac.uk/~rmj/lectures/intro_computing/finding_information/