Supercomputing User Training

Module 4: Moving Data In and Out of Pawsey

Pawsey Training Series

Supercomputing User Training

1. Supercomputing Introduction
2. Logging In
3. Filesystems Overview
4. Moving Data In and Out
5. Using Software Modules
6. Using Software Containers
7. Accounting Model Overview
8. Job Scheduling Overview
9. Running Jobs
10. Testing Job Runs
11. Managing Project Data

Outcomes for this Module

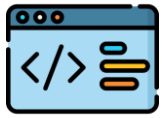
- Describe what is the purpose of the data mover nodes
- Describe what data staging and archiving mean at Pawsey
- Remotely move data in and out of the scratch filesystem

✓ Prerequisite knowledge:

- ✓ **Bash shell basics**
- ✓ **User Training 02: Logging In**

Watch for These Signs!

Definition of new concepts



Hands-on coding (demo)

Best practices

Exercises and solutions

Warnings (bad practices)

Links to user documentation

Icons made by Freepik from www.flaticon.com

Pawsey

Australian Government

NCRIS National Research Infrastructure for Australia

CSIRO

Curtin University

Murdoch University

Government of Western Australia

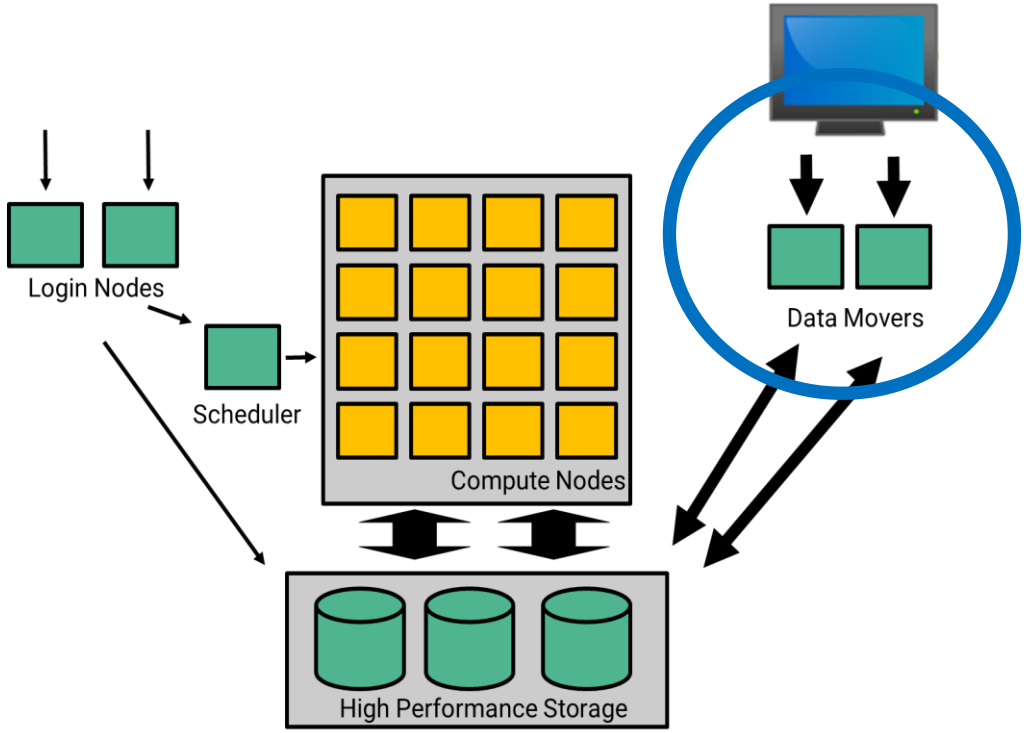
ECU Edith Cowan University

The University of Western Australia

Moving Data In and Out of Pawsey

Data Mover Nodes

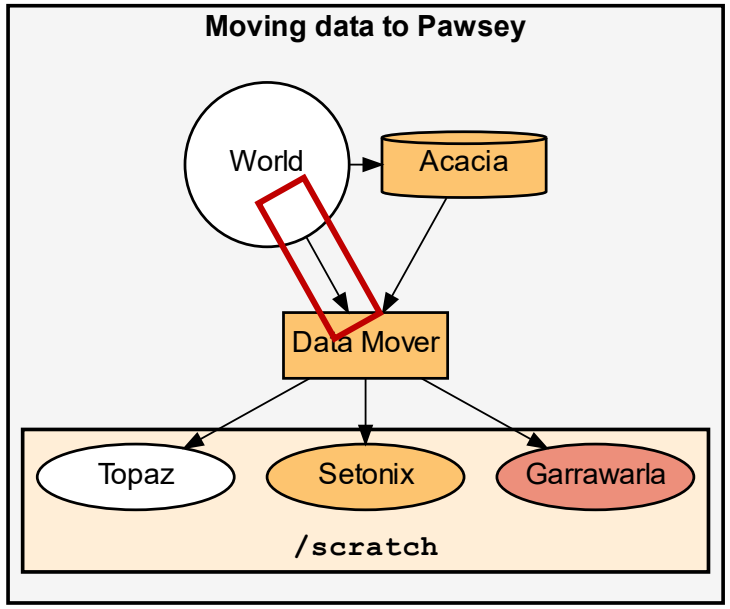
- Externally connected servers dedicated to data transfers between high performance and other storage (in-situ and external)
- Connected to all Pawsey global storage locations
- Enable data transfers in and out Pawsey systems
- Useful to transfer large amounts of data <u>without overloading the login nodes</u> of the supercomputer



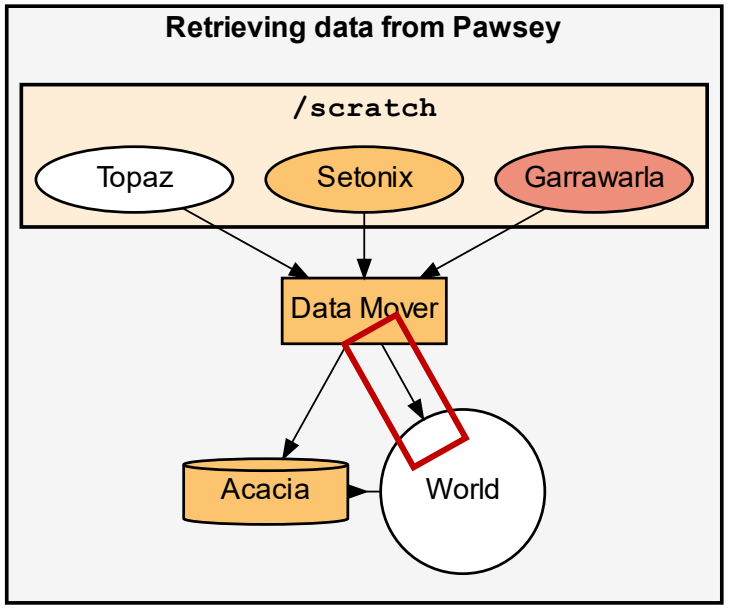
Staging and Archiving Data at Pawsey

Data Staging (at Pawsey)
Copying data from the world to Pawsey servers (pushing).

Data Archiving (at Pawsey)
Copying data from Pawsey servers to the world (pulling).



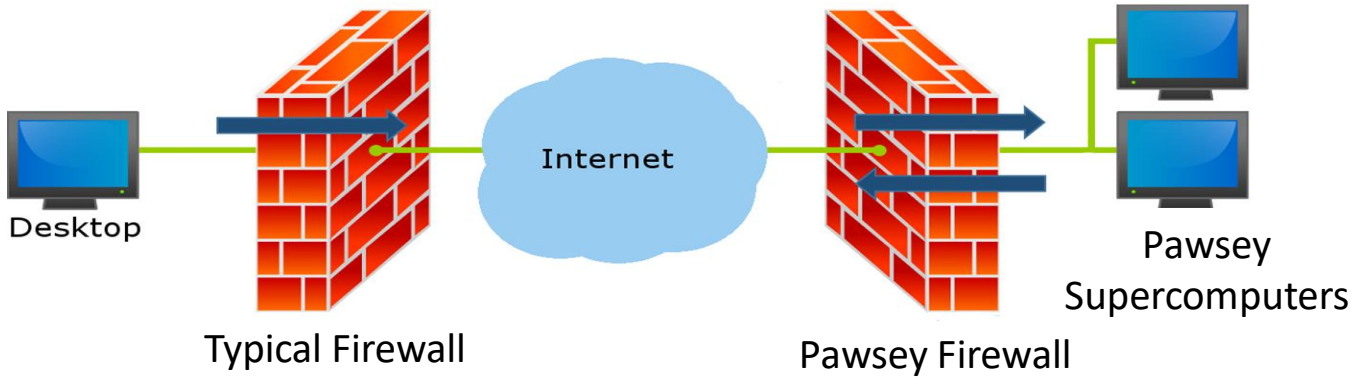
Use the data mover nodes to stage and archive data at Pawsey.
Be mindful. This method is important for large datasets, not to overload the login nodes. To build good habits, also apply this to small datasets.



** More information on Acacia in User Training 11: Managing Project Data*

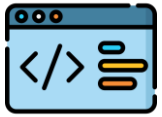
Access to Data Mover Nodes

- Remote (external) access via SCP/SFTP using SSH port
  - Pawsey firewalls allow both incoming/outgoing SSH connections
  - Most organisations & home routers block incoming SSH connections
  - Server name: `data-mover.pawsey.org.au`



- Also, internal Pawsey access via `copy` partition for processing via job scheduler
  - More on this later

Remote Data Transfers: How To



DEMO: Transfer data back and forth using `scp`

The recommended command line tools are `scp` or `rsync`

- Ensure username/password are correct

Graphical User Interface (GUI) clients can also be used

- Common ones: MobaXTerm, FileZilla, WinSCP, CyberDuck
- Ensure username/password are correct
- Multiple failed tries can trigger IP blacklisting
- If you can, disable login auto-retry

Pack together large number of files (>100) to transfer into a single tarball archive.
This method will result in more efficient, faster data transfers.

More details @

- Transferring Files in/out Pawsey Filesystems
- Data Storage and Management Policy (governing data stored at Pawsey)

OUTPUTS: Remote Data Transfers: How To

```
$ echo "Hello from local" >my-file
$ scp my-file username@data-mover.pawsey.org.au:/scratch/project-id/username/
..
my-file                                       100%   17     2.4KB/s   00:00


$ # login onto Setonix
$ ssh username@setonix.pawsey.org.au
$ ls /scratch/project-id/username/
my-file
$ exit


$ scp username@data-mover.pawsey.org.au:/scratch/project-id/username/my-file ./file-is-back
..
my-file                                       100%   17     2.4KB/s   00:00
$ ls
file-is-back my-file
```

EXERCISE: Staging Data

1. Create some (e.g. 3) dummy files on the shell terminal in your local machine
   - Commands:

```
$ echo "File number 1" >file1
$ echo "File number 2" >file2
$ echo "File number 3" >file3
```

2. Create a tarball archive using the dummy files
   - Command:

```
$ tar czf archive1.tgz file?
```

3. Push the tarball archive to your scratch space in Setonix using the Data Mover nodes
   - Template command: `scp <filename> <username>@<datamover>:/scratch/project-id/username/`

4. Login into Setonix and verify the tarball archive is there

OUTPUTS: Staging Data

```
$ scp archive1.tgz username@data-mover.pawsey.org.au:/scratch/project-id/username/
..
archive1.tgz                                                  100%  175    23.6KB/s   00:00

$ ssh username@setonix.pawsey.org.au

$ cd $MYSCRATCH
$ ls
archive1.tgz
```

- The server name for Data Mover nodes is `data-mover.pawsey.org.au`
- Password is requested for both `scp` and `ssh`
- To verify the transfer
  - Use `ssh` to login into `setonix.pawsey.org.au`
  - You can use `$MYSCRATCH` as a shortcut to your scratch directory

EXERCISE: Archiving Data

1. Login into Setonix and go to your scratch directory (tip: use the variable `$MYSCRATCH`)

2. Create a dummy file there
   - Command:

```
$ echo "Hello from Setonix" >setonix-file
```

3. Note the full path of the file (tip: `pwd` can help)
4. Logout from Setonix using the `exit` command
5. From your local machine, pull the dummy file from your Setonix scratch directory
   - Template command: `scp <username>@<datamover>:<filepath>/<filename> .`

OUTPUTS: Archiving Data

```
$ ssh username@setonix.pawsey.org.au

$ cd $MYSCRATCH
$ echo "Hello from Setonix" >setonix-file
$ pwd
/scratch/project-id/username

$ exit

$ scp username@data-mover.pawsey.org.au:/scratch/project-id/username/setonix-file .
..
setonix-file                                                  100%   22     3.5KB/s   00:00
```

Data Transfers using the `copy` partition

- Enables initiating data transfers from the Data Mover nodes, rather than remotely from your computer

- Suitable for
  - Internal Pawsey transfers (e.g. between `/scratch` and Acacia)
  - Transfers between centres or institutions rather than laptops
    - Remote access to the external centre/institution is needed

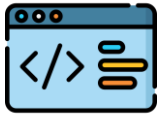
- Use the Pawsey job scheduler (Slurm)
  - Interactive session for small (quick) data transfers
  - Batch script for large data transfer
  - Scheduler option: `--partition=copy`
  - Common tools including `scp` and `rsync` available

Pack together large number of files (>100) to transfer into a single tarball archive.
This method will result in more efficient, faster data transfers.

**** More information on Acacia in User Training 11: Managing Project Data***
**** More information on scheduling jobs in User Training 9: Running Jobs***

Summary

- Terms we learnt
  - Data Staging at Pawsey
  - Data Archiving at Pawsey



- Tasks we learnt
  - Move data in and out scratch: `scp`

- Use the data mover nodes to stage and archive data at Pawsey
- Pack together large number of files (>100) to transfer into a single tarball archive
- Use multiple tarballs (e.g., one per directory) for extremely large transfers

Pawsey

Australian Government

NCRIS National Research Infrastructure for Australia

CSIRO

Curtin University

Murdoch University

Government of Western Australia

ECU Edith Cowan University

The University of Western Australia

Getting Help

Getting Help

https://support.pawsey.org.au

Pawsey has extensive User Support Documentation.

Areas covered include:

- System user guides
- Knowledge Base
- Pawsey-supported software list
- Maintenance logs
- Policies and terms of use

For further assistance, contact the help desk, via User Support Portal.

Help us to help you by providing details, such as:

- Which resource
- Error messages
- Location of files
- SLURM job id
- Your username if having login issues
- Never tell us (or anyone) your password!

Become a Pawsey Friend and receive our Newsletter:
https://pawsey.org.au/pawsey-friends/

Q & A Session