pawsey

Supercomputing Filesystems

Setonix Phase 1 Release

4 July 2022. Version 1.02

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

Focus for this training

Learning outcomes:

- Compare file management approaches on Setonix and other systems
- Monitor filesystem use and size
- Familiarise yourself with file migration pathways and use cases
- Review recommended changes to your data workflow
- Familiarise yourself with changes to reference datasets
- Review Pawsey's policies and best practices for data

Core Migration Training Modules:

1. Getting Started with Setonix
2. Supercomputing Filesystems
3. Using Modules and Containers
4. Installing and Maintaining Your Software
5. Submitting and Monitoring Your Job
6. Using Data Throughout the Project Lifecycle

pawsey

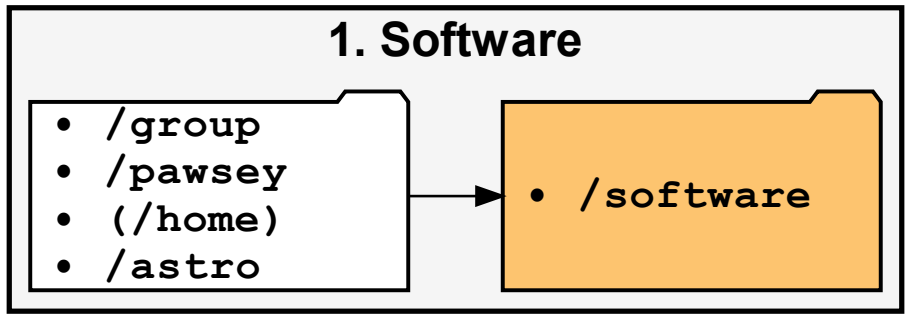
Section 1

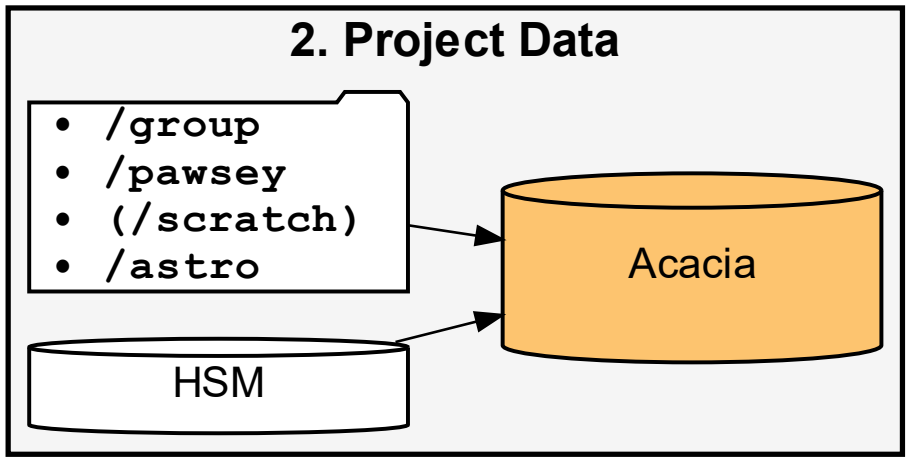
Filesystems at a Glance

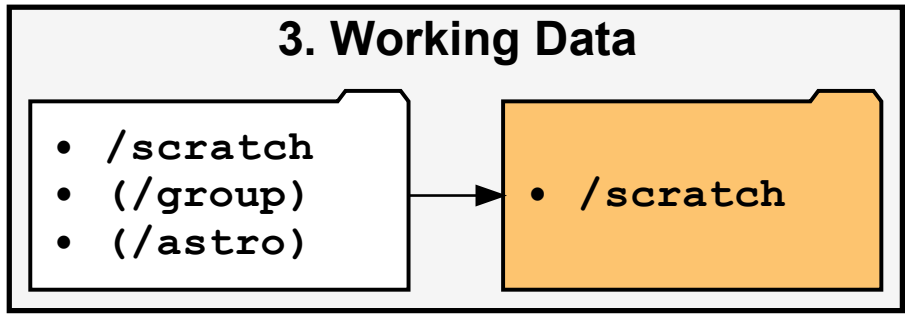
What are the current high performance filesystems used for?

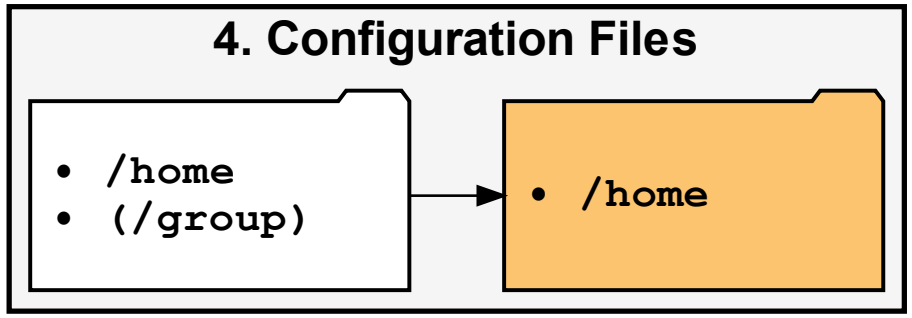
Name	Purpose
`/home`	• Store user **Configuration Files**
`/group`	• Store project-maintained **Software** (and **Slurm Batch Scripts**) • Store **Project Data** (reference datasets, data produced for project-length timescales)
`/scratch`	• Store **Working Data** (data required to run a job/workflow)
`/pawsey`	• Store Pawsey-maintained **Software** • Store Radioastronomy **Project Data** (reference datasets, data produced for project-length timescales)
`/astro, /askapbuf, /askapextend`	• Store Radioastronomy maintained **Software** • Store Radioastronomy **Project Data** (reference datasets, data produced for project-length timescales)

What are the changes to filesystems for Setonix?









Legend
- Gold = New infrastructure
- White = Existing infrastructure
- (Brackets) = Usage is not recommended

- Note that the old and new `/scratch` and `/home` have the same names but are different filesystems.

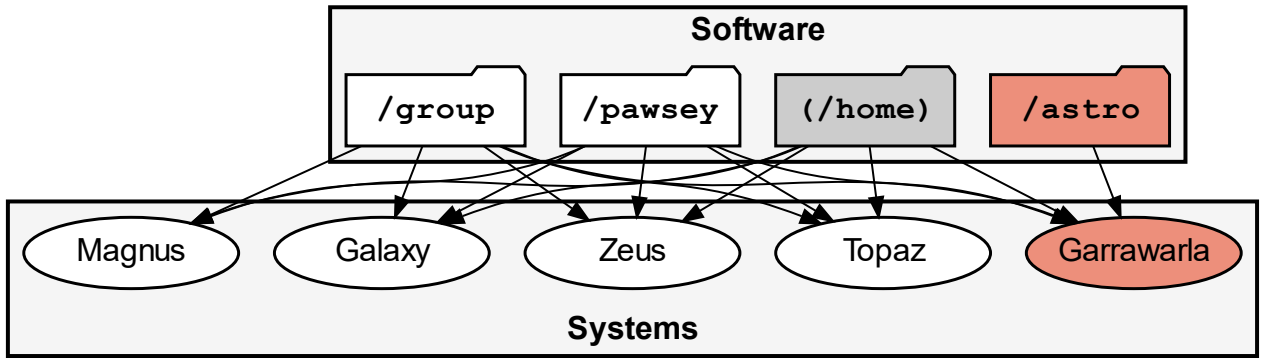
pawsey

Section 2

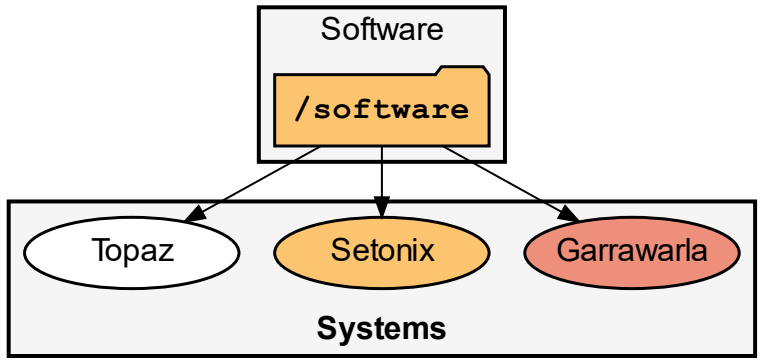
Filesystem Details

Where will software reside?

Pre-Setonix



Setonix



- Software that was previously located on many filesystems will be moved to a single filesystem, simplifying the setup.
- The new `/software` filesystem will contain:
  - Programming environments, libraries and tools
  - Domain software maintained by Pawsey
  - User-managed software stacks
  - Files related to user-managed workflows, such as Slurm batch scripts

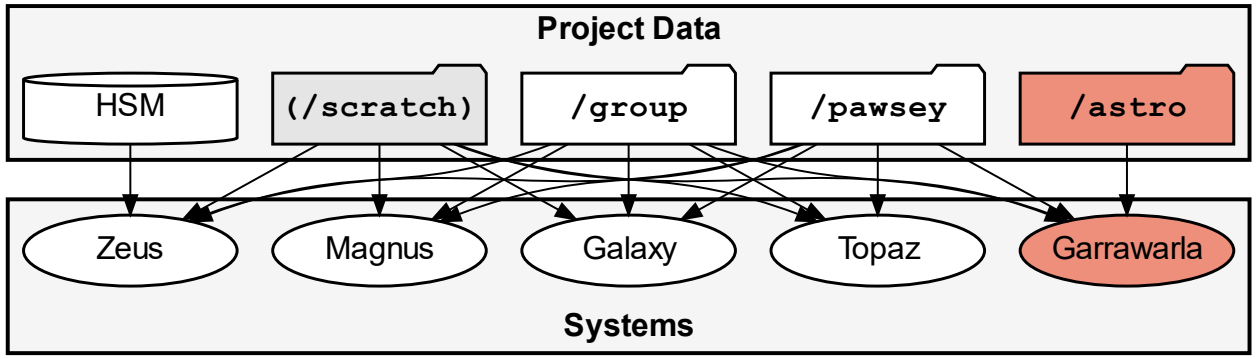
Legend
- Gold = New infrastructure
- White = Existing infrastructure
- Red = Astronomy infrastructure
- (Brackets) = Usage is not recommended

Properties of /software

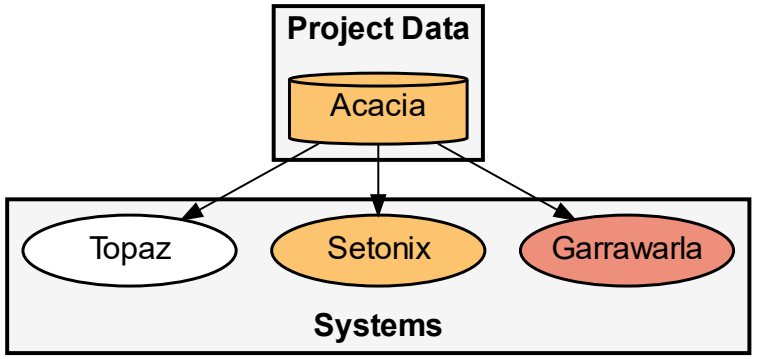
- High-performance parallel LustreFS filesystem.
- 256 GB quota per project; 100,000 file limit per project.
- Data is retained for the lifetime of the project.
- Organised as:
  - Project-wide level: `/software/projects/project-id/`
  - User level: `/software/projects/project-id/username/`
- The `$MYSOFTWARE` environment variable provides a shortcut to the user software location.
- More details @
  - Software Stack
  - Migration Training Module 3: Using Modules and Containers (Materials, Recordings)
  - Migration Training Module 4: Installing and Maintaining your Software (Materials, Recordings)

Where will project data reside?

Pre-Setonix



Setonix



- Project data is data that persists for the lifecycle of the project.
  - Includes project-specific reference datasets; calibration datasets.
  - Was previously located on many different filesystems or in the HSM.
  - Will now be stored on the Acacia object storage system.
- The old /group filesystem will be temporarily available as /oldgroup on the Setonix copy partition.

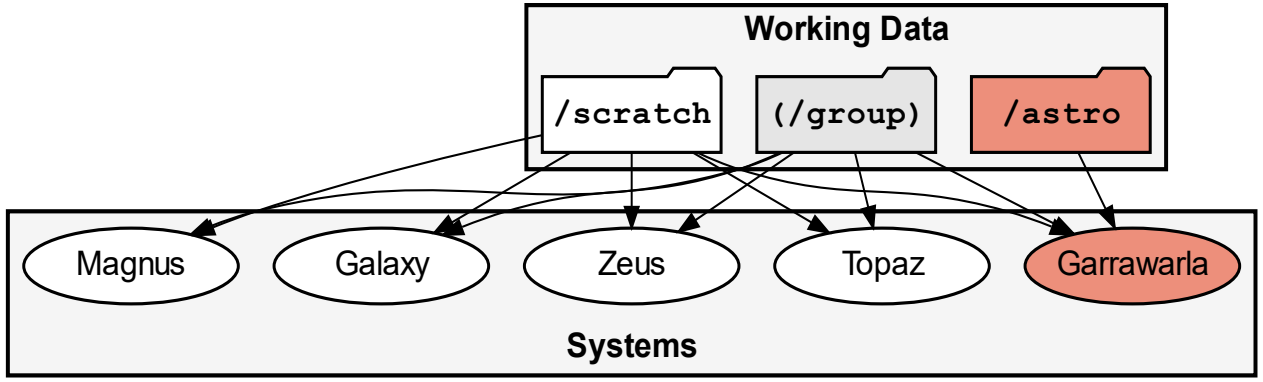
Legend
- Gold = New infrastructure
- White = Existing infrastructure
- Red = Astronomy infrastructure
- (Brackets) = Usage is not recommended

Acacia Object Storage

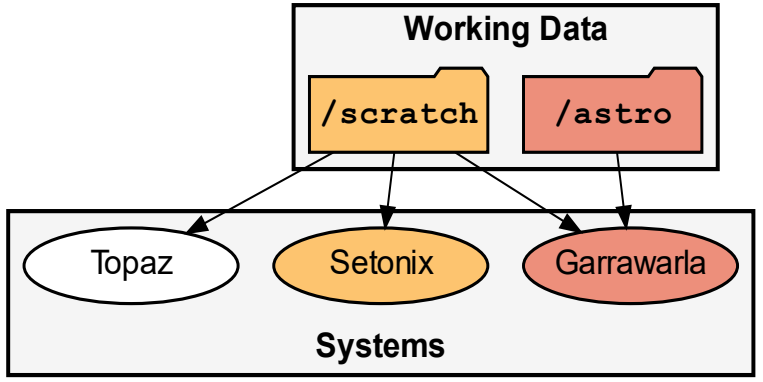
- The Acacia object storage system:
  - Is a high-performance, large-volume Ceph object storage using the S3 interface
  - Is used for storage of project data for the length of the project
  - Provides different functionality and operates differently to traditional filesystems such as /home, /scratch, and /group
- Merit allocation supercomputing projects will have 1 TB of project-wide allocation on Acacia by default.
- Up to 10 TB can be requested via the Service Desk, and larger allocations must apply separately.
- Staging data to and from Acacia are covered in Migration Training Module 6: Using Data Throughout the Project Lifecycle. (Materials, Recordings)

Where will working data reside?

Pre-Setonix



Setonix



- Use the new `/scratch` filesystem for working data on Setonix.
  Working data should be:
  - Actively read or written to by jobs currently queued, running or recently completed.
  - Staged onto the `/scratch` filesystem within days of when it will be used by jobs on the system.
  - Moved off the `/scratch` filesystem within days of the jobs completing on the system.
- The `/scratch` filesystem of Setonix will be accessible on Zeus data mover nodes as `/newscratch`.
- `/astro` will continue to be available for Garrawarla.

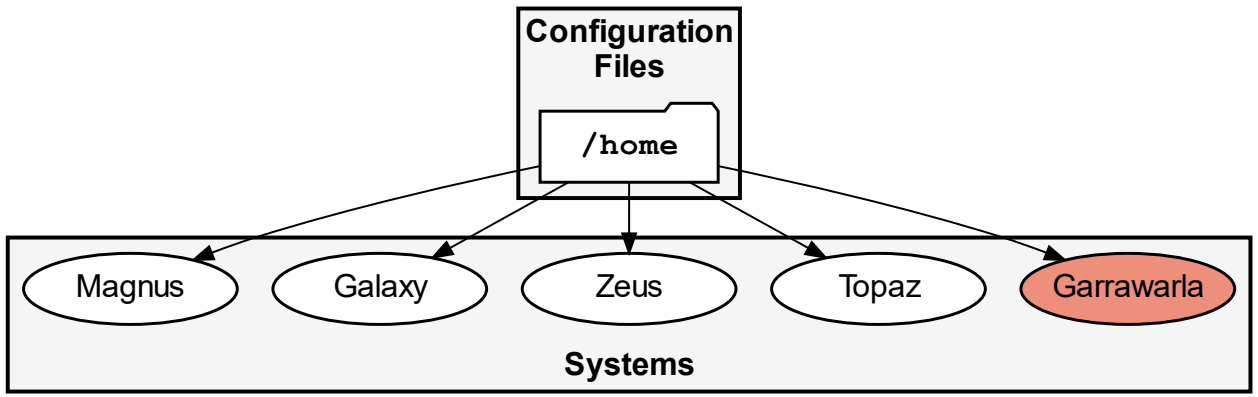
Legend
- Gold = New infrastructure
- White = Existing infrastructure
- Red = Astronomy infrastructure
- (Brackets) = Usage is not recommended

Properties of `/scratch`

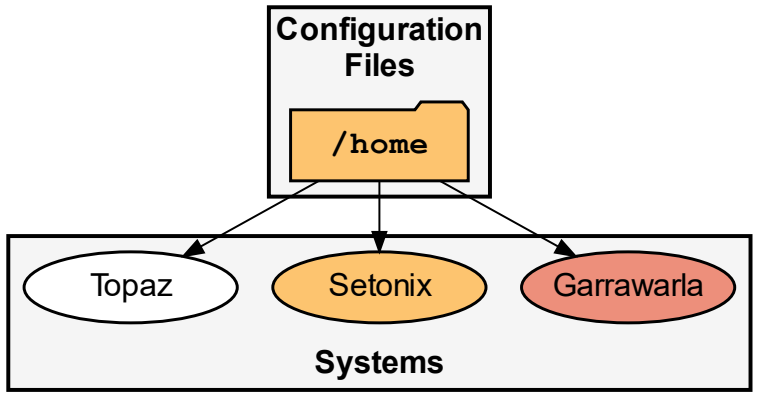
- `/scratch` is a:
  - High-performance parallel LustreFS filesystem with:
    - 1PB capacity limit per project
    - 1 million file limit per user
    - 30-day purge policy
  - Organised as:
    - Project-wide level: `/scratch/`*`project-id`*`/`
    - User level: `/scratch/`*`project-id/username`*`/`
- The $MYSCRATCH environment variable provides a shortcut to the user-level location.
- More details @ Pawsey Filesystems and their Usage

Where will configuration files reside?

Pre-Setonix



Setonix



Legend
- Gold = New infrastructure
- White = Existing infrastructure
- Red = Astronomy infrastructure
- (Brackets) = Usage is not recommended

- `/home` should contain software configuration files:

  Examples: `~/.bashrc` for the Bourne again shell; and `~/.vimrc` for the Vim editor.

- Other types of data might need to be redirected away from `/home`.
  Examples: software installations to `/software`; temporary data to `/scratch`.
- Workflow-specific configurations should be placed in `/software`.

Properties of /home

- The /home available on pre-Setonix will be temporarily available as /oldhome on the Setonix data mover nodes.
- /home is a:
  - NFS filesystem (not designed for large-scale parallel access) with:
    - 1 GB quota per user
    - 10,000 file limit per user
  - Organised as /home/*username*/
- The ~ symbol or $HOME environment variable provides a shortcut.
- Do not set up job environments using login scripts in /home.
- More details @ Pawsey Filesystems and their Usage

```
magnus-1$ pwd
/home/username

magnus-1$ ls
pre-setonix files
```

```
setonix-01$ pwd
/home/username

setonix-01$ ls
setonix files
```

Monitoring Filesystem Usage

- Filesystem usage can be checked when logged into Setonix

/home

```
$ quota -s
Disk quotas for user username (uid userid):
     Filesystem   space   quota   limit   grace   files   quota   limit   grace
172.18.0.100:/home
                   114M   1024M   1024M             374   10000   10000
```

/scratch and /software

```
$ pawseyAccountBalance -s
Compute Information
-------------------
          Project ID     Allocation         Usage     % used
          ----------     ----------         -----     ------
          pawsey0001          25000         18019       72.1

Storage Information
-------------------
/scratch usage for pawsey0001, used = 15.14 TiB, limit = 2048.00 TiB
/software usage for pawsey0001, used = 776.73 GiB, limit = 256.00 GiB
```

Other System Impacts

- Garrawarla and Topaz are currently using some of the existing filesystems.
- Migration of `/home` and `/group` will impact Garrawarla and possibly Topaz:
  - Garrawarla filesystem migration will occur following the completion of the initial Setonix migration, in consultation with the MWA team.
  - Topaz migration is contingent on the Setonix Phase 2 deployment.
- These migrations will involve transfer of files to the new filesystems similarly to the migration described for Setonix in the following slides.

Summary of Key Changes with Impacts

Existing filesystem	New filesystem	Impacts / Considerations
/group (for software)	/software	• Is a filesystem to store project-specific software • Provides performant access to software and the central location simplifies setup • May require fresh installations of domain software by project groups
/group (for project data)	Acacia	• Is an object storage system to store project data • Provides large-volume, project-scale object storage accessible via S3 interface • Requires explicit management of data using /scratch to stage • May require migration of files from /oldgroup using the Setonix copy partition
/scratch	/scratch	• Purpose unchanged; filesystem to provide fast, large-volume work storage • New filesystem, larger and faster. • Migration of files from pre-Setonix /scratch can be done on Zeus copyq partition where Setonix /scratch is mounted as /newscratch
/home	/home	• Purpose unchanged; filesystem to store configuration files • New filesystem; may require migration of files from /oldhome
/pawsey	/software	• Software managed by Pawsey will be migrated by Pawsey staff
/astro (for software)	/software	• Migration of project-managed radio-astronomy software
/astro (for data)	Acacia	• Migration radio-astronomy project data

pawsey

Section 3

Migrating Data to Setonix

Planning Your Data Migration

- For each type of file already discussed, consider what files exist on the old filesystem, whether they need updating, and where they should go:
  - Configuration files — From old `/home` → Setonix `/home`
  - Working data — From old `/scratch` or `/group` → Setonix new `/scratch`
  - Project data — From old `/group` → Acacia
  - Slurm batch scripts — From old `/home`, `/scratch` or `/group` → Setonix `/software`
- Reinstall software (with newer versions where appropriate) instead of moving it.
- Move files that are no longer actively needed to institutional storage or delete them.
- More details @ Migrating data to Setonix in the Setonix Migration Guide.

Migrating Your Data to Setonix

- The old /home and /group filesystems will be mounted temporarily on the Setonix `copy` partition to support migration of data, with a prefix of “old”:
  - `/oldhome`
  - `/oldgroup`
- The `/scratch` filesystem of Setonix will be accessible on Zeus `copyq` partition as `/newscratch`.
- For smaller file transfers (less than 10 GB or 1000 files) use `salloc` and an interactive session in the `copy` partition to manually move files.
- Create Slurm batch scripts to automate larger file transfers using the `copy` partition and use `sbatch` to submit the job.
- Use `tar` to combine and compress large numbers of files.
- Use multiple tarballs (e.g., one per directory) for extremely large transfers.
- Use `rsync` to allow resuming larger file transfers.
- More details @ Migrating data to Setonix in the Setonix Migration Guide.

Migrating Your Data to Setonix: Example Batch Script

Migrating data from old home and old group filesystems is best done using the Setonix data mover nodes, where they are mounted as /oldhome and /oldgroup.

Slurm batch script directives

```
#!/bin/bash --login
#SBATCH --account=project
#SBATCH --partition=copy
#SBATCH --ntasks=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=48:00:00
```

Example 1:
Using `rsync` to copy files

```
OLDPATH=/oldgroup/project/user
NEWPATH=/scratch/project/user
rsync -vhsrl --chmod=Dg+s $OLDPATH/directory $NEWPATH/.
```

Example 2:
Using `tar` to reduce the number of files

```
cd $OLDPATH
tar zcfv directory.tar.gz directory
cp directory.tar.gz $NEWPATH
cd $NEWPATH
tar zxfv directory.tar.gz
```

Migrating Your Data to Setonix: Example Batch Script

Migrating data from old scratch to new scratch is best done using Zeus data mover nodes, where Setonix /scratch is loaded as /newscratch.

Slurm batch script directives

```
#!/bin/bash --login
#SBATCH --account=project
#SBATCH --partition=copyq
#SBATCH --ntasks=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=48:00:00
```

Example 1:
Using `rsync` to copy files

```
OLDPATH=/scratch/project/user
NEWPATH=/newscratch/project/user
rsync -vhsrl --chmod=Dg+s $OLDPATH/directory $NEWPATH/.
```

Example 2:
Using `tar` to reduce the number of files

```
cd $OLDPATH
tar zcfv directory.tar.gz directory
cp directory.tar.gz $NEWPATH/.
cd $NEWPATH
tar zxfv directory.tar.gz
```

Pawsey

Section 3

Transferring Data to Setonix

Best Practices When Transferring Files

- Do not transfer a large number of files (>100). Use the tar command to first pack the files into a single tar file for transfer.
- Use the data mover nodes when transferring large amounts of data.
  - Data mover nodes are accessible through either:
    - The generic hostname `data-mover.pawsey.org.au`, or
    - The `copy` partition for batch processing through the scheduler.
- Review the Data Storage and Management Policy which governs data stored on Pawsey infrastructure.
- More details @ Transferring files in the Supercomputing Documentation

Transferring Data: Remote Access

Using the Command Line

- The Setonix filesystems are available externally via the login and data mover nodes.
- For larger transfers, use data mover nodes via `data-mover.pawsey.org.au`
  This avoids congesting the login nodes.
- Available clients will depend on your own terminal environment. Examples include:
  - scp
  - rsync

Using a GUI Client

- GUI clients can also be used.
- Use the hostname `data-mover.pawsey.org.au`
- Commonly used clients include:
  - MobaXTerm
  - FileZilla
  - WinSCP
  - Cyberduck

More details @ Transferring files in the Supercomputing Documentation.

Transferring Data: Setonix Data Mover Nodes

Using an Interactive Session

- Use only for smaller data transfers.
- Use the `salloc` command to access the copy partition and then manually use commands to copy files.
  - Common programs such as `scp` and `rsync` are available

Using a SLURM Batch Script

- Target the `copy` partition.
- Use `tar` to combine and compress large numbers of files.

- Remote access to the external system will be needed for the transfer.
- Suited for transfers between centres or institutions rather than laptops.
- More details @ Transferring files in the Supercomputing Documentation.

pawsey

Section 4

Data Workflows

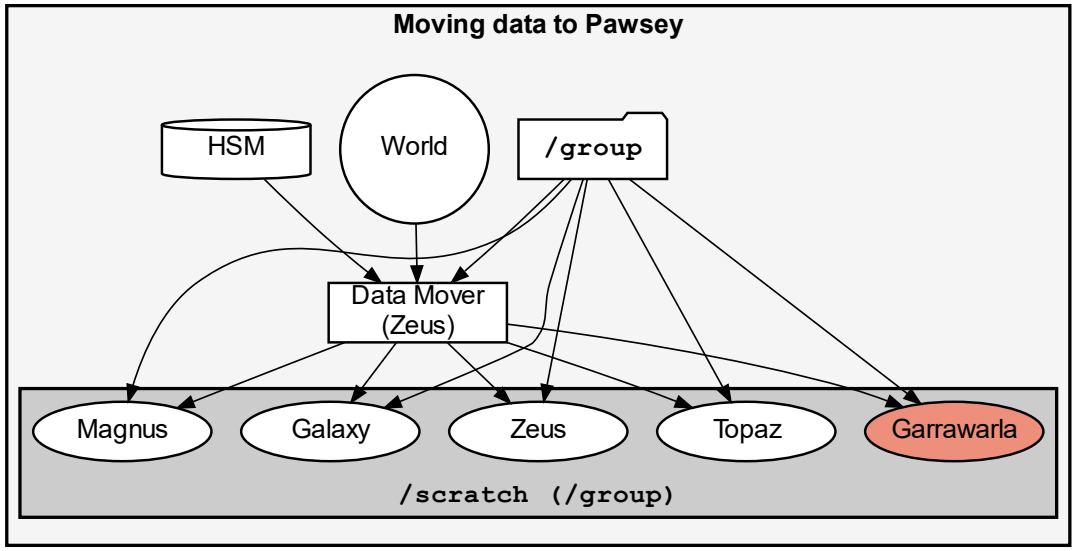
Data Workflows: Pre-Setonix

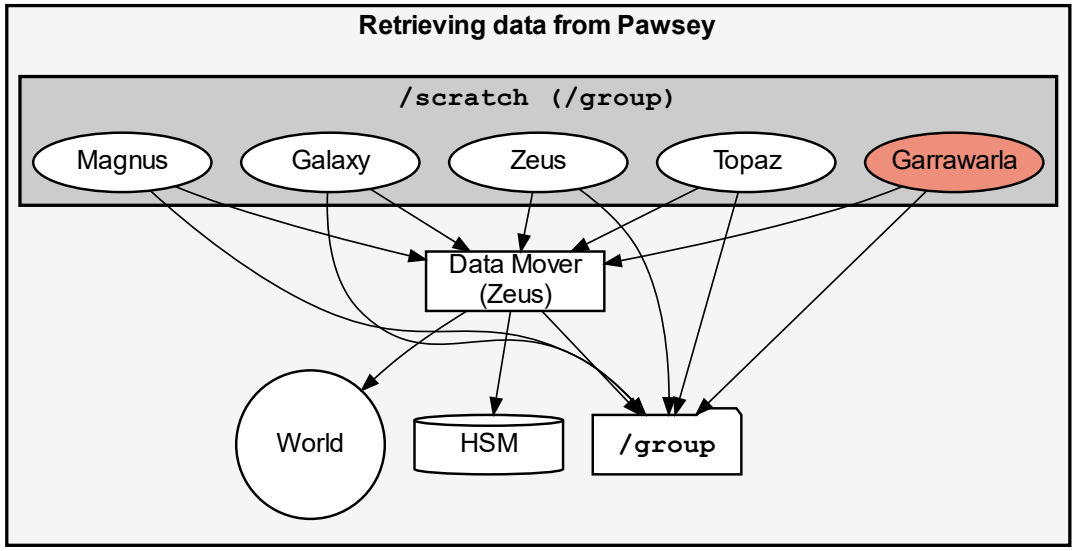
Moving data in

1. Pull data in, potentially from external sources.
2. (Optional) Pre-processing of data.
3. Run job on Magnus (or other systems), producing data on `/scratch` (or `/group`, not recommended).

Moving data out

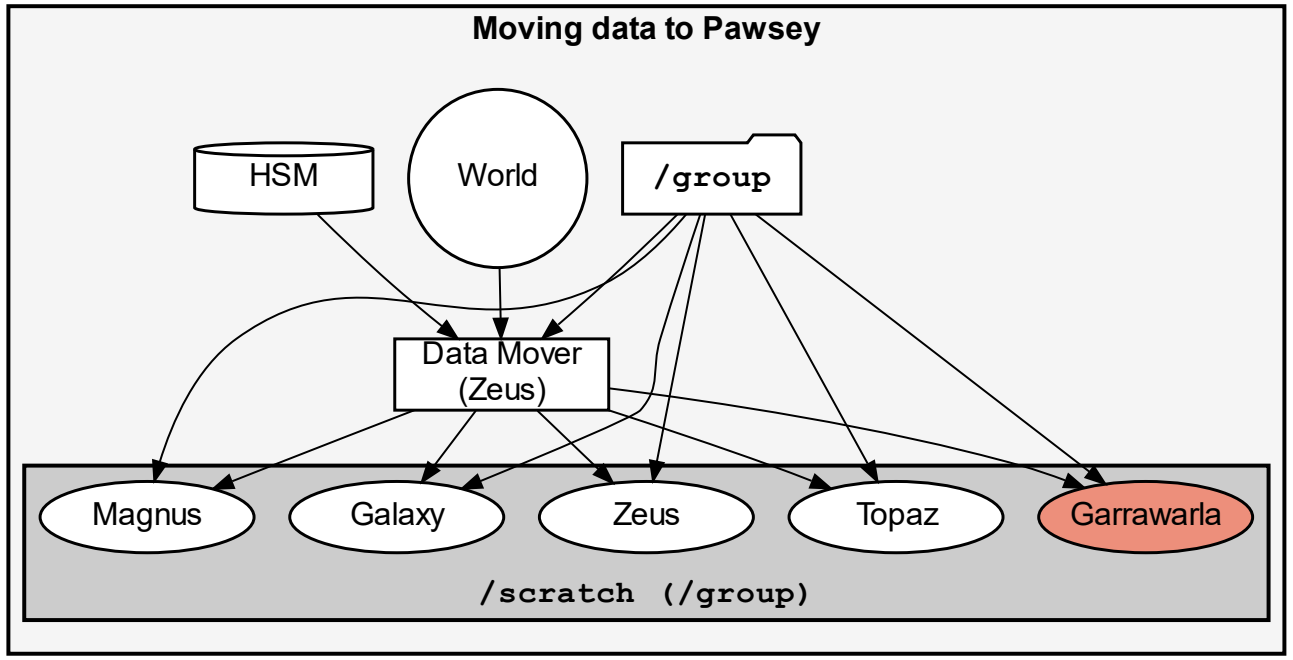
1. Run job on Magnus (or other systems), producing data on `/scratch` (or `/group`, not recommended).
2. Launch `copyq` job on Zeus to copy data to HSM tape storage or outside Pawsey.
3. If data is required later, mv it to `/group` (if small).



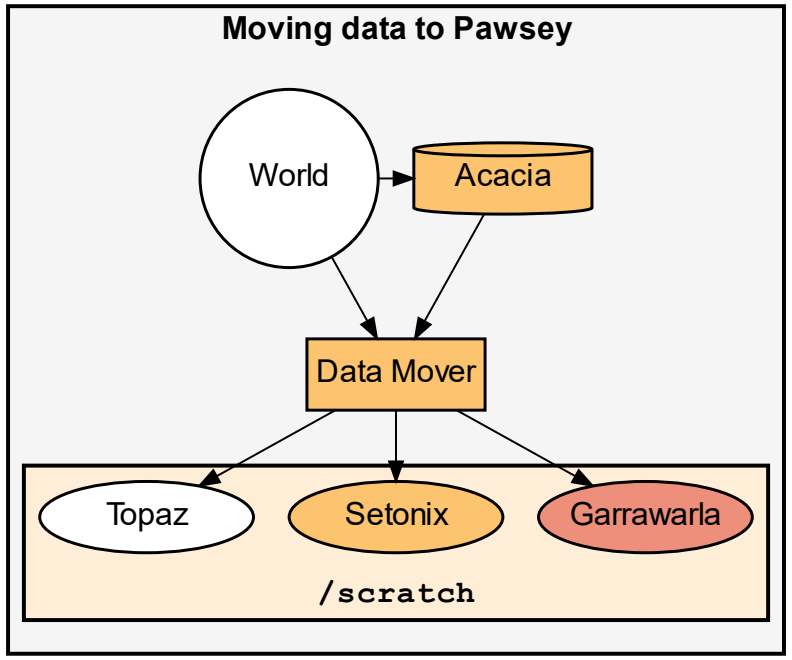


Workflow Changes on Moving Data to Pawsey

Pre-Setonix

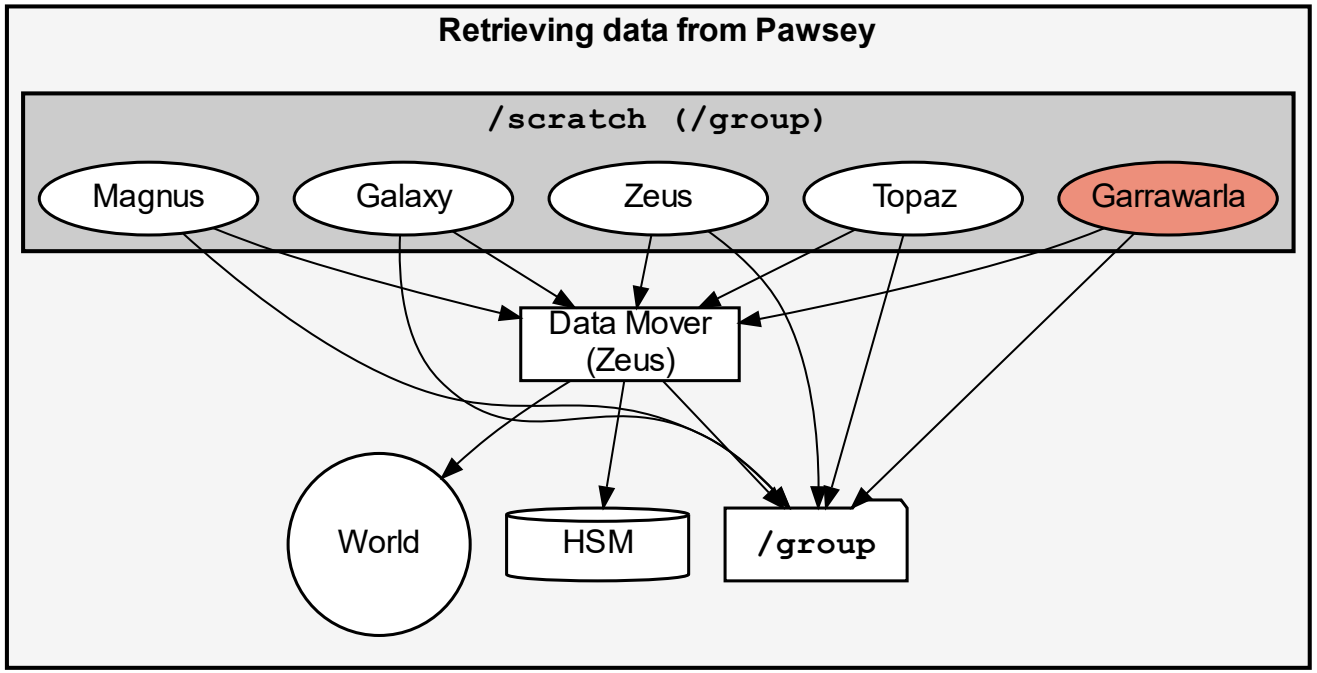


Setonix

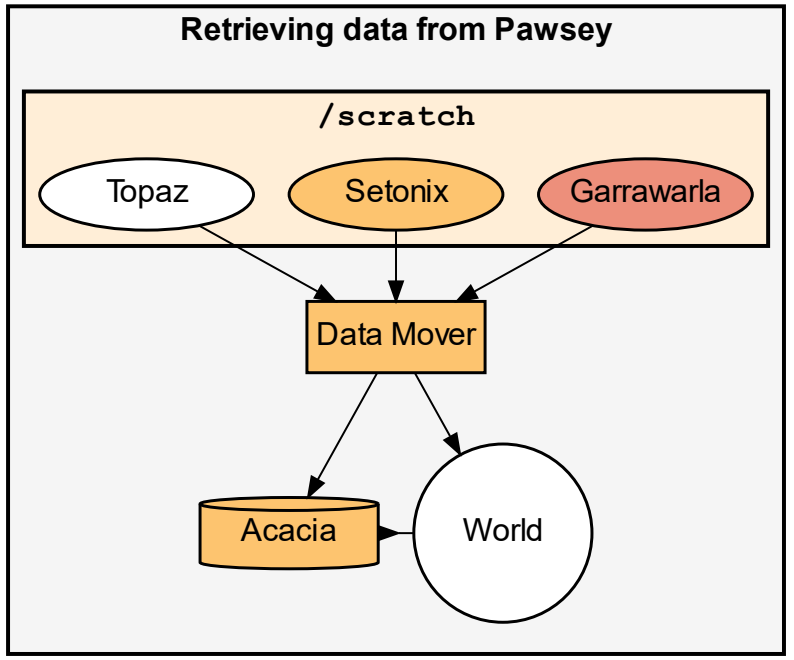


Workflow Changes on Moving Data from Pawsey

Pre-Setonix

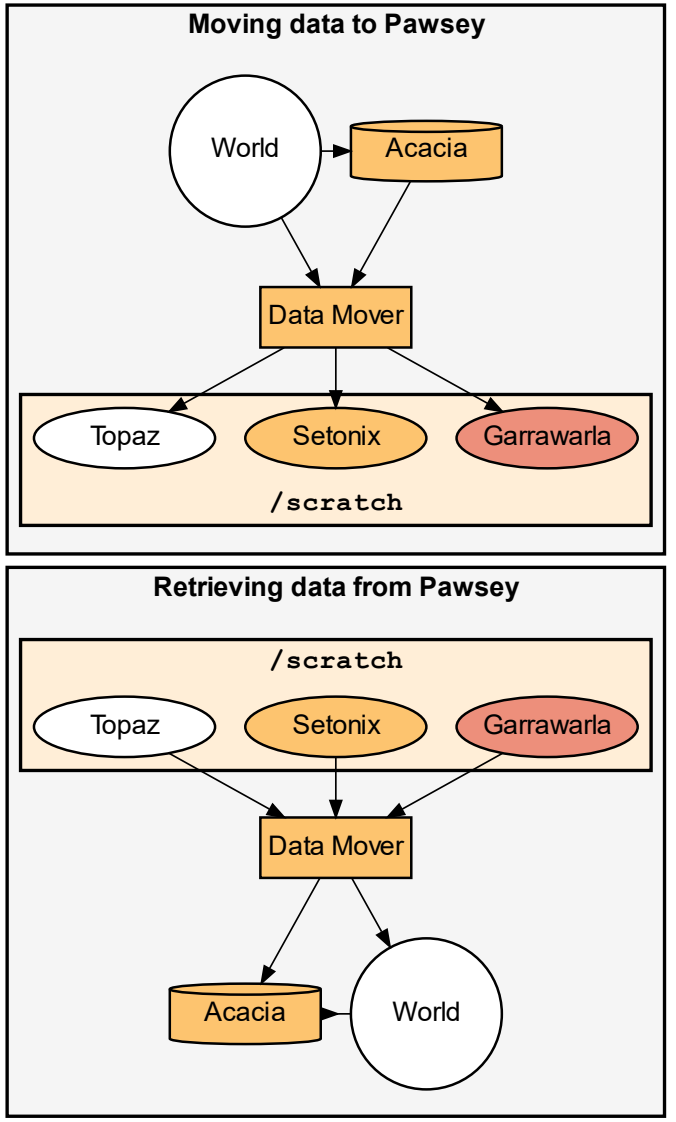


Setonix



Key Changes

- Movement of data in workflows will be simplified since data mover nodes are part of Setonix.
- HSM and `/group` will be replaced by the Acacia storage system, which provides large-volume, project-length storage.
- Because data mover nodes are part of Setonix, you do not need to use a different cluster to move data following the migration.
- Setonix data mover nodes are available via:
  - The `copy` partition in SLURM
  - Externally via the `data-mover.pawsey.org.au` hostname
- More details @ Data workflows



Section 5

Reference Datasets on Setonix

Pawsey

Using Reference Datasets on Setonix

- Reference data sets are a common set of data used widely in a particular research field.
- Pawsey staff make these available in one place to avoid multiple, redundant copies.
- On Setonix, each data set will be contained in a subdirectory:

  `/scratch/references/`*`subdirectory/`*

- Examples:
  - `/scratch/references/askap/` (radio astronomy)
  - `/scratch/references/mwa/` (radio astronomy)
  - `/scratch/references/blast/` (bioinformatics)
- More detail @ Reference Datasets

pawsey

Section 6

Pawsey Data Policies and Practices

Data Migration Best Practices

- Proactively monitor filesystem usage, both in size and number of files.
- Review files on the old filesystems, and only transfer files that are necessary.
- Transfer files to the appropriate filesystem and folders.
- Don’t forget there is a 30-day purge policy on `/scratch`.
- Be aware of the Filesystems Policies.
- For Acacia best practices, refer to the Acacia User’s Guide and the Acacia Videos Playlist on Pawsey’s YouTube channel.
- More details @ Pawsey Filesystems and their Usage.

How do I get help?

Migration Documentation & Migration Guides

- Setonix Migration Guide
- Setonix User Guide
- Supercomputing Documentation

Migration Training Materials & Video Recordings

- Upcoming Migration Training
- Recordings: Pawsey YouTube Setonix Migration Phase 1 Playlist
- Materials: Setonix Migration Training Materials (PDFs)

Help Desk

- Help Desk
- Email: help@pawsey.org.au

pawsey

Thank you for attending!

Please complete this short survey:

https://www.surveymonkey.com/r/Y3YFYHQ