The University of Queensland

Research Computing Centre

HPC Storage User Guide

Document Status

This update: July 14 2020 by David.Green @ uq.edu.au

The Storage Subsystem

Storage Technologies on HPC

Local Disk on Awoonga/FlashLite/Tinaroo

These three HPC clusters have access to the same network storage options (see below) but also have local disks that are best accessed using the TMPDIR environment variable within batch jobs. The disk capacities are details in the section below about Quotas and Limits.

High Performance Network Scratch Space on Wiener

Instead of local disk on each Wiener node, space is available on a high performance network filestore mounted at /scratch.
You will find a folder for your Faculty/School/Institute under /scratch and you will be able to write into that subfolder.

Cluster Local GPFS Storage on Awoonga/FlashLite/Tinaroo

These three clusters share storage that is based on a high performance IBM GPFS storage system.
That storage was installed at the end of June 2021 to replace the DDN GridScaler device that was part of the procurement of FlashLite in 2015.

- Each node in Awoonga, FlashLite and Tinaroo clusters is a native GPFS client for that storage system. This enhances the data performance.
- The locations hosted on this are /sw /home and /scratch (/scratch/user and /scratch/project)
- There are strict quotas being enforced on /home and /scratch.
- There are purging policies being applied to the /scratch to maintain a usable capacity.

~~Cluster Local BeeGFS Storage on FlashLite~~ No Longer Supported

The nodes in the FlashLite cluster are equipped with 4.8TB of solid state NVME drives.
A prototype BeeGFS storage cluster had been deployed to aggregate space from multiple FlashLite nodes into a single distributed storage device.

However, the performance and capacity of our new /scratch filesystem has made it unnecessary to keep the BeeGFS storage, so it has been disassembled.

QRIScloud Collection Storage

The physical proximity to the QRIScloud infrastructure, means that HPC clusters can also directly mount QRIScloud collection storage via the NFS protocol. Historically these have been HSM based collections. More recently they have been handled as MeDiCI/RDM style collections.

UQ RDM / Medici Data Fabric Storage

The UQ RDM collection storage is presented to the HPCs like a QRIScloud collection via the Medici Data Fabric.
Currently that is done using the same NFS based mechanism used for other QRIScloud data collections.
We are evaluating improvements to the attachment of the Medici Data Fabric to the HPC to enhance the performance.
The Medici data fabric is also available on the Weiner HPC system.

The RDM collection storage is a multi-layer, multi-site data caching and replication system.
You see your collection in one of these layers at one of the sites.
When you make changes to your data at a site, those changes propagate laterally, to other sites, and vertically through layers at the primary site (which for HPC connected "Q collections" is our data centre at Springfield)
Your data may have been pushed offline due to inactivity and your collection reaching its low water mark (quota).
It will still show up in the file list, but the size may be smaller than expected.
This is a "file stub" and the file must be retrieved from the HSM layer before you will be able to work with its contents.

How is my collection data stored and accessed?

The RDM/MeDiCI Data Fabric stores vast amounts of data.
However, it does not do so solely by powering a lot of disk drives all of the time.
Instead, RDM storage (specifically the Q collections attached to the HPCs) are housed in a multi-level cache infrastructure (stack) built upon a "hierarchical storage management" system.

Layer	Storage Technology
GPFS Cache	Disk
HSM Cache	Disk
HSM Spin Down	Zero Watt Storage
HSM Archive	Robotic Tape Storage

This full storage stack resides at only one site ("home").
The other sites you may have need to connect to (e.g. for scientific instruments or your R:\ drive or the cloud.rdm.uq.edu.au) just have the GPFS cache that you interact with.
That GPFS cache layer stays in sync with "home".

Some newer collections and older ones with specific characteristics are held on a device called UQ03.
That device combines the GPFS cache and the HSM cache in a single layer.
This combination provides a more seamless experience for big collections and collections that are subject to a high rate of "churn".

On the HPCs, you always access your RDM Q collection data from the GPFS Cache Layer.
On Awoonga/FlashLite/Tinaroo, this cache layer gets mounted onto the HPC at a location such as `/QRISdata/Q1234`.
On Wiener, it is the QBI replica of your data that you interact with. That is mounted at `/afm01/Q1/Q1234`
The "HSM Spin Down Layer" powers down after a period of inactivity to conserve energy and to reduce our carbon footprint.

Your active data can and should reside in the GPFS Disk Cache layer.
When you don't access your data for a period of time, it can be evicted from the GPFS layer and pushed offline in the HSM.
Refer to the section below entitled "Pre-fetching files from HSM" about how to efficiently retrieve archived data from the HSM.

Ensuring that your RDM Collection is easily accessible from HPCs

When you apply for the RDM collection storage "record" that you want to access from UQ HPC systems (Awoonga/FlashLite /Tinaroo/Weiner), you must make sure

- that you check the box near the bottom that says **The project data needs to be mounted on UQ HPC facilities.**
- that you check any of the boxes that pertain to your data. Some types of data may not be permitted to be stored on the

Medici fabric.
- that you read and understand the various terms and conditions.

This will allocate you an initial 1TB of storage almost immediately that you will be able to access from the HPCs.
The quota can easily be increased as needed.
Your RDM collection is the logical place to retain your results and to be able to share data with your supervisor and other researchers as necessary.

- Once you have a valid RDM MeDiCI collection
  - It will automatically become accessible (via the autofs mechanism) onto Awoonga/FlashLite/Tinaroo HPC systems.
  - It needs to be mounted onto Wiener by the systems team. You request that by submitting an email to helpdesk@qbi.uq.edu.au

What could adversely impact performance of RDM Collection Storage?

You should avoid running computational workloads against your RDM collection- especially if there is a lot of Input/Output.

Running heavy computational workloads in RDM collections can have diabolical impacts on other users of the HPC, QRIScloud storage and the research cloud.
Please use scratch disk space provided in all the HPCs and

- copy in your inputs,
- perform your computations in the scratch spaces ($TMPDIR, /scratch/user/, /scratch/project/ or /nvmescratch (Weiner), then
- copy back your outputs back to your RDM collection as "lumps" (i.e. tar or zip files)

Why does my collection sometimes appear to be owned by the root user?

When you look at the collections from their parent directory /QRISdata, any collection directory that is not mounted will appear to be owned by root.
Once a collection is mounted the ownership changes to that in the mounted fileset.

So to properly look at these, you would need to look inside the collection directory (and thus forcing the mount request). The `ls -l` command would help you do that.

Why do my collection files sometimes appear to be owned by the nobody user?

This can happen under some circumstances.
It is often caused because of the way that the files were added to the collection.

It is important for the usability of your collection that file permissions are correctly set.
RDM collection files and folders should have the following features

- owned by the Q system account for the collection, or, owned by the user who created the file
- belong to the collection RW group (eg. Q1234RW)
- be group readable and writeable but not accessible by others
- folders should have the group writable sticky bit set ("s")
- have other switches (including the "+") set for the RDM mechanism to work flawlessly.

An example is shown here for a directory with some files in it:

```
tinaroo1:/local/home/davidg # ls -salh /QRISdata/Q0286/UQ-RCC/tests
total 9.8G
 32K drwxrws---+ 2 Q0286 Q0286RW  32K Jun 24 10:24 .
 32K drwxrws---+ 6 Q0286 Q0286RW  32K Mar 18  2019 ..
   0 -rw-rwx---+ 1 Q0286 Q0286RW 1.5G Jun 11  2017 tensorflow-1.1.0.sapp.old.20200601
9.8G -rw-rwx---+ 1 Q0286 Q0286RW 9.8G Mar 17  2018 tensorflow-1.6.0-cuda-9.0.sapp
   0 -rw-rwx---+ 1 Q0286 Q0286RW 2.4G Jan 29  2018 tensorflow-gpu-1.4.1.sapp.old.20200601
   0 -rw-rwx---+ 1 Q0286 Q0286RW 2.1G Apr 24  2017 tensorflow.sapp.old.20200601
   0 -rw-rwx---+ 1 Q0286 Q0286RW   51 May 24  2018 textFile
```

If you find files and folders within your collection that show up on the HPC with incorrect ownership, or you are prevented from updating a file, then please submit a support request to rcc-support@uq.edu.au.

Why do my collection files sometimes look empty in output of the ls -salh command?

In the example directory shown in the previous section, you can see that only one file has a size on disk (the left column of `ls -salh` output) of 9.8G(B) that matches the expected file size (in column 6 of `ls -salh` output).
That means that currently the file is in GPFS cache disk and available to use immediately.

The other files in that folder have a size on disk (left column) of 0 instead of their file size.
This means that the file has been put into nearline HSM storage and will need to be retrieved from HSM before being available to use.

We recommend that you use the **recall_medici** command on the HPC to trigger a recall of a file you need.
You do not have to use the recall_medici command as long as the operation you are performing to trigger the recall from HSM is able to wait without failing for the file contents to be retrieved from the HSM layer.

When and how to use the recall_medici script

If the processing you are doing is likely to fail because the file is not in the GPFS cache layer (and the processing does not handle the wait properly), then you must `recall_medici` the files before you attempt the processing that might fail.

You can do that recall_medici step in one of three modes:

1. at the command line shell (login node or qsub -I session) - use a wildcard
2. as part of a preprocessing batch job (which subsequent jobs need to wait for completion) to retrieve all the files needed next
3. just-in-time - as a preceding step within each processing batch job

The third option could cause your processing batch jobs to need extra walltime because each processing job will be waiting for its file(s) to be recalled.

Why do my collection files sometimes fail to read or appear to have invalid format?

Very rarely we see situations where a file in GPFS cache is incompletely read back from HSM.
The output of an **@ls -salh** command shows a file size on disk (column 1) that is non-zero but less than the expected size (column 6).
In this situation, you should contact rcc-support@uq.edu.au

What can be done if I have deleted or damaged a collection file or folder accidentally?

The good news is that when a file is removed or modified in the GPFS layer it does not mean it is gone for good.
In most circumstances, we should be able to retrieve the contents of the file from the HSM and reinstate it.
The caveats on that is that the file has to have been successfully written to HSM some time before being deleted.
Depending on the time between creation and accidental deletion/modification we often can recover from a HSM copy which is not purged immediately.

In this situation, you should contact rcc-support@uq.edu.au

Sometimes my data appears to be missing

The RDM collection storage is a multi-layer, multi-site data caching and replication system.
Occasionally, the different sites can get "out of sync".
For example, files created on, or uploaded to, the HPC, may not appear immediately in your mapped network drive on your PC.
There have also been issues with the synchronisation to the cloud.rdm.uq.edu.au service endpoint.
Usually, just waiting a few hours is sufficient to allow for the data to be replicated between sites.

If the data at one site is out of sync for a prolonged period, you should submit a support request to

- ITS HelpDesk if the data is available at HPC but not available at R: drive or cloud.rdm.uq.edu.au
- Conversely, rcc-support@uq.edu.au if the data is available in the R drive but has not made it to the HPC.

Storage Quotas and Limits

See also the section on HPC Enhanced Access Mechanism

Current Settings

Cluster	File System	Access	GB Limit	Files Limit	Time Limit	Comment
A, F & T	**/home**	GPFS	50	1,048,576	none	Unlimited retention but no centralised backups. ~~Filesystem snapshots provide some scope for recovery from accidental deletion. (see storage related tricks section below)~~
A, F & T	**/scratch/user**	GPFS	150	100,000	Time based deletion policy.	Deletion is commencing soon.
A, F & T	**/scratch /project**	GPFS				Quota settings depend on project. Deletion policies are being implemented.
A, F & T	**/RDS/Q1234**	NFS				Quotas and timelimits as per QRIScloud collection allocation.
A, F & T	**/RDS/Q9876**	NFS/GPFS				Quotas and timelimits as per RDM/Medici collection allocation.
Awoonga	**/state /partition1**	local disk	200		job duration	Use **$TMPDIR** in running jobs on nodes. Purged on job completion.

FlashLite	**/state /partition1**	local disk	386		job duration	Available but not often used. You will need to purge manually.
FlashLite	**/nvme**	local disk	4400		job duration	Use **$TMPDIR** in running jobs on nodes. Purged on job completion.
Tinaroo	**/state /partition1**	local disk	852		job duration	Use **$TMPDIR** in running jobs on nodes. Purged on job completion.

All users should be using the variable $TMPDIR for their work and NOT hard coding the paths /state/partition1, or /nvme explicitly.

The rquota command

The `/usr/local/bin/rquota` command (usually, in your path) will inform you of your current usage, and quotas, across quota-controlled filesystems.
A quota limit value of 0 implies it is unlimited!

```
uqdgree5@tinaroo1:~/Tests/Tinaroo/PBS> rquota
FileSet        Used(GB) Limit    Files      Limit
30days_group   0        0        0          0
/home          10       20       49932      204800
/30days        395      1000     23219      3145728
/90days        74       400      71503      1048576

Group quotas
FileSet        Used(GB) Limit    Files      Limit
/groups         3071     2048     0          928383
```

What is $TMPDIR?

Inside a Batch Job

The TMPDIR is an environment variable (with value $TMPDIR) that is defined within every running batch job.
Each running batch job has a temporary location created that is

- is unique to the job and
- is owned and accessible to only the job owner.

No one else can access the contents of your $TMPDIR directory while your job is running.
Upon job completion or termination, the location specified by $TMPDIR is removed.

Outside of a running PBS job

The TMPDIR is also a defined environment variable in your unix command shell.
It has the value $TMPDIR that corresponds to your personal directory in the /30days filesystem.
This feature allows for ease of job script testing and for handling situations where software expects a $TMPDIR to exist.
The contents of the $TMPDIR (when not inside a running batch job) are left in their /30days location.

See also Storage Best Practices below about why and how to best use $TMPDIR in batch jobs.

Storage Best Practices

- Use $TMPDIR when you need local disk for your batch jobs.
  The $TMPDIR directory is created in /state/partition1 automatically as part of your batch job and is removed for you automatically at the end of the job.
- Saving user data randomly into local disk on a node (outside of $TMPDIR mechanism) can adversely impact other users. Please DON'T do that.
- Compute Nodes are periodically rebuilt and the local disk space is reformatted, so do not rely on using local disk on compute nodes unless via $TMPDIR.
- Don't forget that $TMPDIR is unique for each job and job-array sub-job.
  Although the path may be the same, the $TMPDIR directory will probably contain different files on different nodes and for different jobs.
- If you need to work with many small files, please keep them bound together in a single archive file (ZIP or tar).
  Copy the archive file to local disk (i.e. $TMPDIR) before unpacking the files to work on them in local disk space.

To figure out which directories have the most files, you can use the **`du`** command and the options **`--inodes`** and **`--max-depth`** combined with the `sort` and `head` commands to get a summary of your most populous folders.

```
uqdgree5@tinaroo1:~> du --inodes --max-depth 1 /home/uqdgree5/Tests | sort -nr | head -5
900     /home/uqdgree5/Tests
270     /home/uqdgree5/Tests/FlashLite
260     /home/uqdgree5/Tests/Tinaroo
```

```
115     /home/uqdgree5/Tests/Nimrod
84      /home/uqdgree5/Tests/PBSPro
```

Suggested Data Workflow

We suggest that you consider the following as a model for your data arrangements:

- use /home for small nonvolatile items such as documentation, source codes, PBS files and the like
- use /30days as scratch space for job outputs before copying them back to HSM or /90days or /home
- use /90days for reference or input data files (keep them tar-ed or zip-eds or like fastqScreenDB (if that is a reference input) or packed inputs.
- use the UQ research data HSM or a QRIScloud collection to hold (compressed) tar archives of your input data.

- unpack the compressed input data onto local disk on the compute nodes ($TMPDIR) if there is space
  Awoonga nodes have 200GB of local disk. Tinaroo and FlashLite have more local disk.

Storage Related Tricks and Tips

Stale File Handles

From time to time you may experience the dreaded "Stale file handle" message.

This means that the machine you are working on (usually a login node) has lost its connection to some part of the storage infrastructure backend.
It can affect /home and the other filesystem served by the HPC cluster GPFS device, as well as /QRISdata collection storage.
The stale file handle situation can be due to a failure of the backend storage sub-system, but can also be due to a mounting problem for just the node you are on.

In most situations, simply waiting will allow time for the connection to be re-established automatically.
However if you do not wish to wait then you can login to a different node (see the Connecting to HPC User Guide for how to connect to a specific login node, instead of the alias name)
If you waited, and/or tried another entry point and you still get stale file handles, then please check the RCC Active Incidents page.
Submit a support request (mailing rcc-support@uq.edu.au) if there is no mention of filesystem related problems on the Incidents page.

Addendum: The stale file handle issue is most prevalent on Tinaroo1.
If you logged into tinaroo.rcc.uq.edu.au or tinaroo1.rcc.uq.edu.au, then try logging into tinaroo2.rcc.uq.edu.au directly (see note above about Connecting to HPC User Guide)

Recovering from an accidental deletion in /home

There is a "backup", of sorts, that is done on the /home file system on the Awoonga/FlashLite/Tinaroo HPCs.

The GPFS storage subsystem has a feature called snapshots.
Approximately daily, it will note any differences (to the previous snapshot version) of all the files in your home directory.
It quietly works away each morning before dawn.

For each user, there are multiple "shadow" copies of your home directory and all the files therein.
For any file in your home directory, you should find a timestamped file system entry.
For example, `/home/.snapshots/20190408_0311/uqdgree5/` is a snapshot taken at 03:11AM on April 8 2019.
If I deleted a file from my home directory some time after that point in time, I would be able to "roll back" to that April 8 version.
You own all of your snapshot files so you can copy back over onto the proper home directory the one that you removed or messed up.

The snapshots only go back a month or so, so don't wait too long!

Pre-fetching files from HSM

Due to the finite amount of disk space available for collections, larger files, and files that have not been used for some time, may be evicted from the main disk layer of the UQ RDM and pushed offline to low energy disk or tape media.
The act of accessing any offline file will trigger a recall of the file.
Unfortunately, not all research software seems capable of waiting for the file to be retrieved.
Users sometimes experience file read or input/output errors.

To avoid this sort of disruption to your work, we recommend that you use the `/usr/local/bin/recall_medici` command.
It accepts a single filepath, or a glob (a wildcard).
That command will wait until the offline copy is read back into disk, before proceeding.

Here is an example of it being used.

> Notice that the `ls -salh` output has two columns with filesize information.
> The filesize figure in Column 1 is the size of the data actually stored on (MeDiCI/UQ-RDM) disk layer.

The filesize figure in Column 6 is the size of the file as expected when it is on the disk.
If the Column 1 size is zero it means that the file is "offline".
If the Column 1 size is matches the Column 6 size, it means that the file is "online".
Sometimes the Column 1 size will be slightly larger than Column 6. This can happen for small files due to rounding to whole disk blocks.
Whenever a file is offline and required it will be recalled from the hierarchical storage (HSM) subsystem. This step can take some time to complete.

```
uqdgree5@tinaroo1:~> ls -salh /QRISdata/Q0837/demo.bam
32G -rwxr-x--- 1 uqdgree5 qris-uq 32G Nov 14  2018 /QRISdata/Q0837/demo.bam

Now the file has been evicted from the GPFS cache layer by a systems administrator. Check it again now ...

uqdgree5@tinaroo1:~> ls -salh /QRISdata/Q0837/demo.bam
0 -rwxr-x--- 1 uqdgree5 qris-uq 32G Nov 14  2018 /QRISdata/Q0837/demo.bam

So now the file exists in the HSM subsystem but not on MeDiCI disk.

Then use the recall_medici command to trigger the recall of the file from HSM
I am also curious to see how long it takes... About 15 minutes.

uqdgree5@tinaroo1:~> date; /usr/local/bin/recall_medici /QRISdata/Q0837/demo.bam ; date
Thu Nov 28 15:15:09 AEST 2019
Thu Nov 28 15:30:40 AEST 2019

uqdgree5@tinaroo1:~> ls -salh /QRISdata/Q0837/demo.bam
32G -rwxr-x--- 1 uqdgree5 qris-uq 32G Nov 14  2018 /QRISdata/Q0837/demo.bam

It does not hurt if you run recall_medici on a file that is already on disk.
In fact it should finish very quickly.

uqdgree5@tinaroo1:~> date; /usr/local/bin/recall_medici /QRISdata/Q0837/demo.bam ; date
Thu Nov 28 15:38:02 AEST 2019
Thu Nov 28 15:38:45 AEST 2019

uqdgree5@tinaroo1:~> ls -salh /QRISdata/Q0837/demo.bam
32G -rwxr-x--- 1 uqdgree5 qris-uq 32G Nov 14  2018 /QRISdata/Q0837/demo.bam
uqdgree5@tinaroo1:~>
```

Downloading Data from External Sources

If you have an option for where you pull data from, endeavour to use a source that is **"on-net"**.
The term **"on-net"** means that it does not involve commercial charging rates for the download.
Generally speaking, sites hosted within Australia, Universities and research institutions are on-net.
Because of peering arrangements that AARNet has, it is not necessarily the case that commercial providers are not on-net.
You need to check.

The on-net/off-net status of a download source can be ascertained using a tool provided by AARNet.
It is linked off their Network Operations page

HOW TO CHECK IF AN IP ADDRESS IS ON-NET
Enter an IP Address in the Network Address Query Tool

You just need to enter the server name into the tool as data-cbr.csiro.au (without the ftp:// or http:// and file path info).
For example the data server at CSIRO is On-Net

Checking data-cbr.csiro.au
150.229.21.196 is domestic (On-Net) network 150.229.0.0/19 originated by AS6262 via path 6262

Cambridge University is on-net but outside of Australia, obviously:
Checking www.cambridge.ac.uk
131.111.150.22 is international research (On-Net) network 131.111.0.0/16 originated by AS786 via path 20965 786

The Toshiba corporation has two entries, one is on-net the other is off-net.
It would always be preferable to use the **on-net** location where ever possible.

Checking www.toshiba.com
163.171.217.16 is international transit (Off-Net) network 163.171.217.0/24 originated by AS54994 via path 2914 7473 54994
Checking www.toshiba.com
163.171.197.87 is domestic peering (On-Net) network 163.171.197.0/24 originated by AS54994 via path 4826 23686 23686 23686 23686 23686 23686 54994

un-tar to a different place

You don't always have to unpack your tar file into the same directory as the tar file

```
tar -xf archive.tar -C /target/directory
```

So for example if your compressed inputs were on /90days but you want to unpack them into $TMPDIR in a running job then do this

```
tar -zxf /90days/$USER/inputs.tgz -C $TMPDIR
```

Stacking Bricks

You may find that your data overwhelms you (or at least your quota settings).

To assist you with managing your files and folders, there are a few tools available:

- **fpart** - a partitioning tool for files and folders (`module help fpart`)
- **parallel** - a tasking tool that will manage parallel execution of a set of tasks (`module load parallel`)
- **bricks** - in-house tools for directory-aware partitioning (dpart) and HSM-aware copying (hsync).

If you **`module load bricks`**, you will automatically load the fpart and parallel modules.

The `fpart` and `dpart` utilities will create sets of files each containing a list of files and folders.
Those lists of filenames can be fed to `parallel` for processing.
The `hsync` command understands how to work well with the hierarchical storage (HSM).

Collections on the HSM

Some of the QRIScloud collections are housed on our hierarchical storage management (HSM) infrastructure (called Tier 3).

- HSM is the generic name for the technology that combines
  - a small amount of disk space,
  - a large amount of tape storage and
  - software that applies policies to try to keep active files on disk and inactive data on tape.
    Our HSM is based on the Data Migration Facility (DMF) product that was developed by SGI that is now been incorporated into Hewlett Packard Enterprise.
    DMF is one brand of HSM, like a ... is one brand of car.

Carpe Orbis (Seize the Disk)

If you have a QRIScloud collection that is housed on the hierarchical storage infrastructure (Tier 3) you are able to pre-emptively recall your data from the tapes so that the data is already in the disk cache when you need it.
If you don't use a HSM file for a period of time, it will be migrated to multiple tapes and the disk copy removed.
If the disk cache starts to fill up, then the oldest files will be pushed to tapes to make space.

If your collection is housed on HSM, it will look like this when you run the /bin/df -h command.
(the bare `df` command is actually a wrapper that kindly does not show you all of the mounted collections)

```
tinaroo1:~ # /bin/df -h /RDS/Q0275
Filesystem                          Size  Used Avail Use% Mounted on
10.255.120.226:/tier3a1/Q0275/Q0275  120T  116T  4.4T  97% /QRISdata/Q0275
```

The DMF Commands

The commands that are helpful when you want to be more proactive in your use of the HSM are:

- `dmls` allows you to list HSM file attributes as well as regular file system information
  - REG means regular file (on disk only)
  - DUL means dual state (on disk and on tape)
  - OFL means offline (on tape only)
  - MIG/UNM means transitioning MIG-rating to tape or UNM-igrating from tape
- `dmfind` allows you to search for filenames based on HSM attributes (e.g. -state OFL) .
- `dmget FILENAME` will queue up the retrieval of the file from the tape storage and place it on the disk cache.
- `dmput -r FILENAME` this copy a disk file to tapes and will release the disk space ASAP

It is often useful to combine dmfind with dmput or dmget (or other linux commands). See the dmfind manual page for further information and examples.

Why does it take so long to get a file?

Why the HSM can take a long time to retrieve a file...

There are only so many tape drives in our robotic tape silos.
The tape drives switch between reading data and writing data depending on what needs to be done.

The access to data on tapes via those tape drives is handled by a queue based system.
The queue depth can change by orders of magnitude in short amounts of time.

Lately, there has been an unusually heavy amount of writing to tapes that has needed to be done.
This has meant increased competition for read operations.

If there are a lot requests for reads or writes then it takes longer to recall a specific file.

The best thing you can do to mitigate waiting for HSM data to be recalled is to fetch it ahead of when you need it by using the **`dmget`** command. (See DMF commands section for more information)

Storage Schematic