



Data Management in the HPC

February 23-24, 2022

Derrick Zwickl

HPC Consultant, Research Computing –
zwickl@arizona.edu

Chris Reidy

Research Facilitation Manager, Research Computing –
chrisreidy@arizona.edu



What you will learn


What: understand how to transfer, organize, and use data on the HPC.

Why: Nice data makes things easier: collaboration, reproducibility => better research

Assume

- Basic understanding of how to use the Unix shell

How

- Overview of storage options on the HPC
 - Methods of transferring data in and out
 - Basic HPC data management /good enough practices
 - Moving data around
- 

Basic Pipeline

- Data input, processing, collaboration, sharing



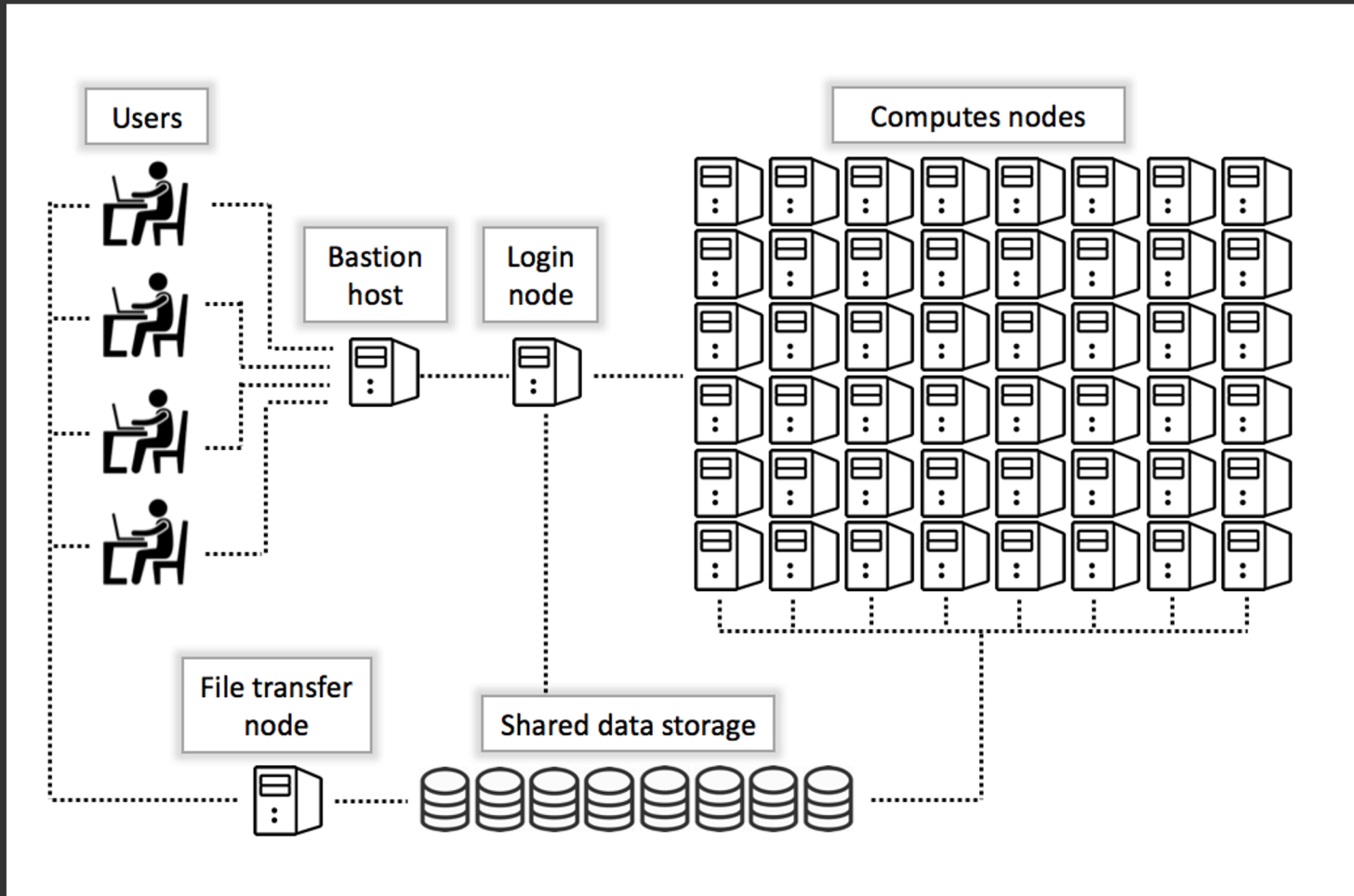


Logging in



THE UNIVERSITY
OF ARIZONA

The diagram of the UA HPC cluster



Login node is like an elevator

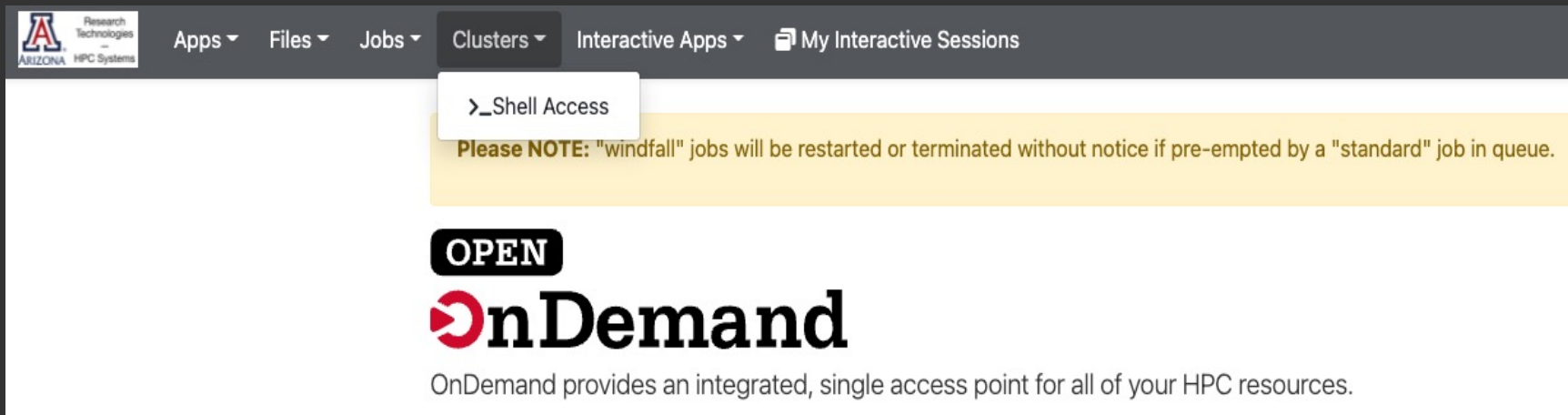


Login node is like an elevator



Cluster Login

- Open ood.hpc.arizona.edu in your web browser and login with your NetID and password.
- From the “Clusters” drop-down menu choose “shell access”



The screenshot shows the top navigation bar of the OOD web interface. The 'Clusters' menu is open, showing a dropdown with the option '>_Shell Access'. Below the menu, there is a yellow warning box with the text: 'Please NOTE: "windfall" jobs will be restarted or terminated without notice if pre-empted by a "standard" job in queue.' Below the warning box, there is a large 'OPEN' button and the 'OnDemand' logo. The text below the logo reads: 'OnDemand provides an integrated, single access point for all of your HPC resources.'

or use ssh from
command line

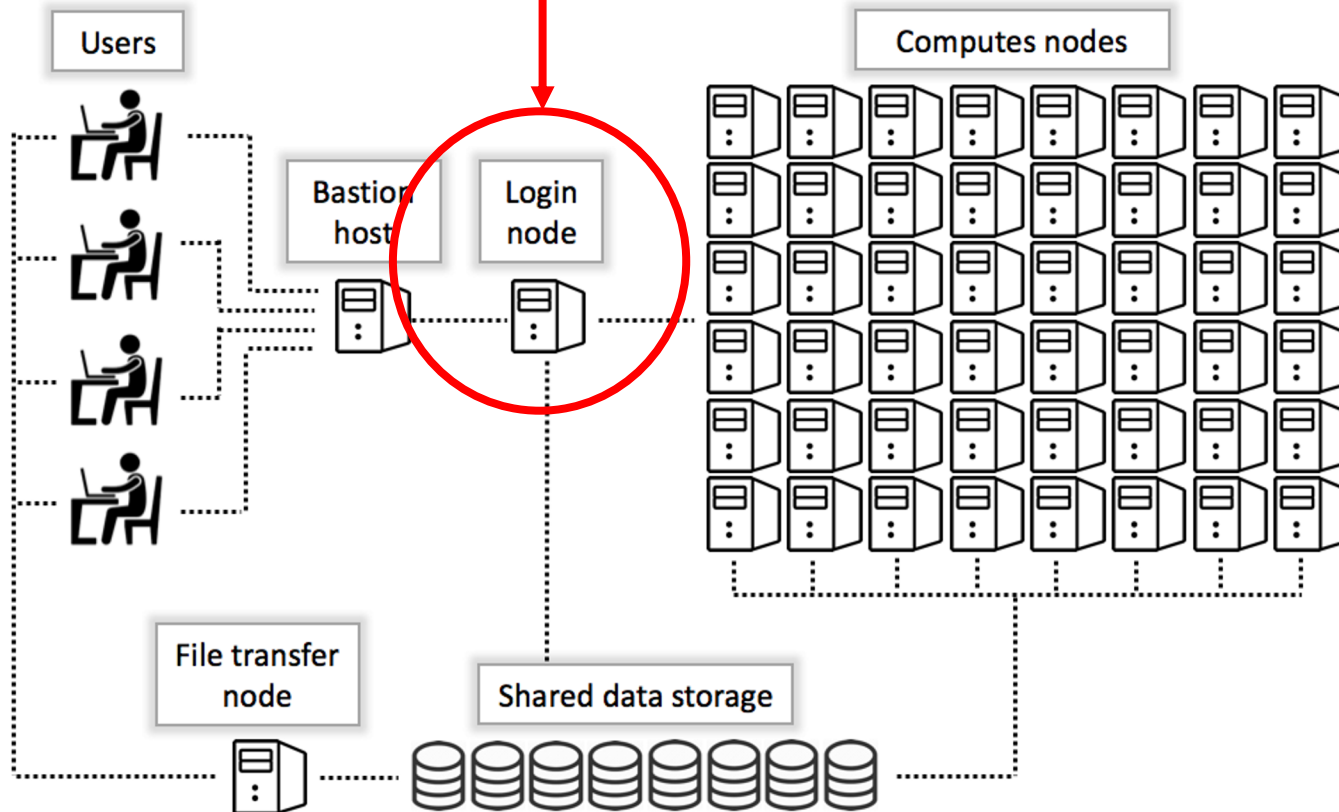
```
~/ $ ssh zwickl@hpc.arizona.edu
Last login: Tue Feb 22 15:50:18 2022 from on-campus-10-138-68-75.vpn.arizona.edu
This is a bastion host used to access the rest of the RT/HPC environment.

Type "shell" to access the job submission hosts for all environments
-----

[zwickl@gatekeeper ~]$ shell
```

Login node

```
[dshyshlov@login2 ~]$
```



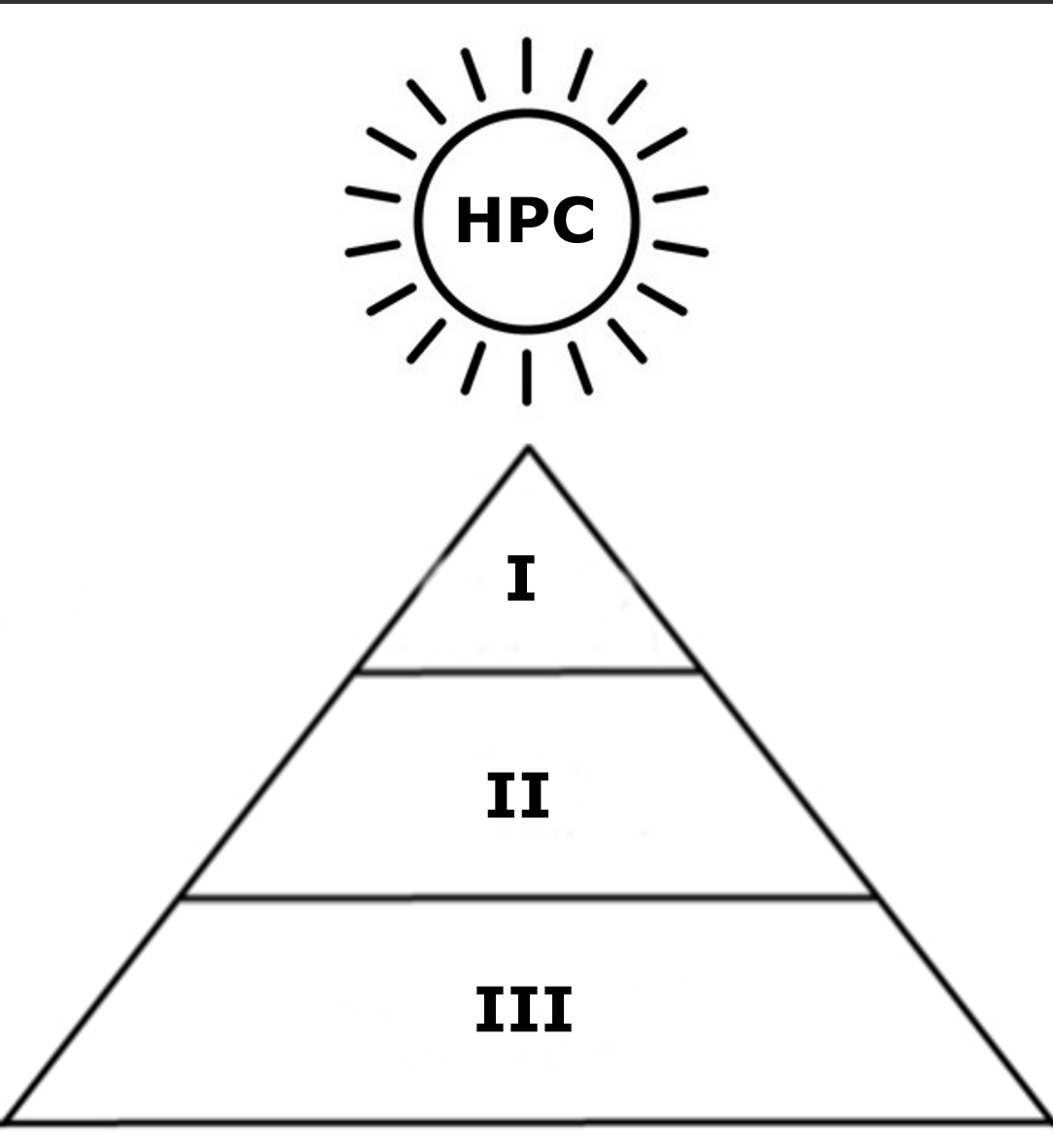


Data Storage Options



THE UNIVERSITY
OF ARIZONA

HPC Storage



- Tiered storage
- HPC storage:
Active data
- Google drive:
Non-active data
- Cloud storage:
Archived data

HPC Storage – Tier I

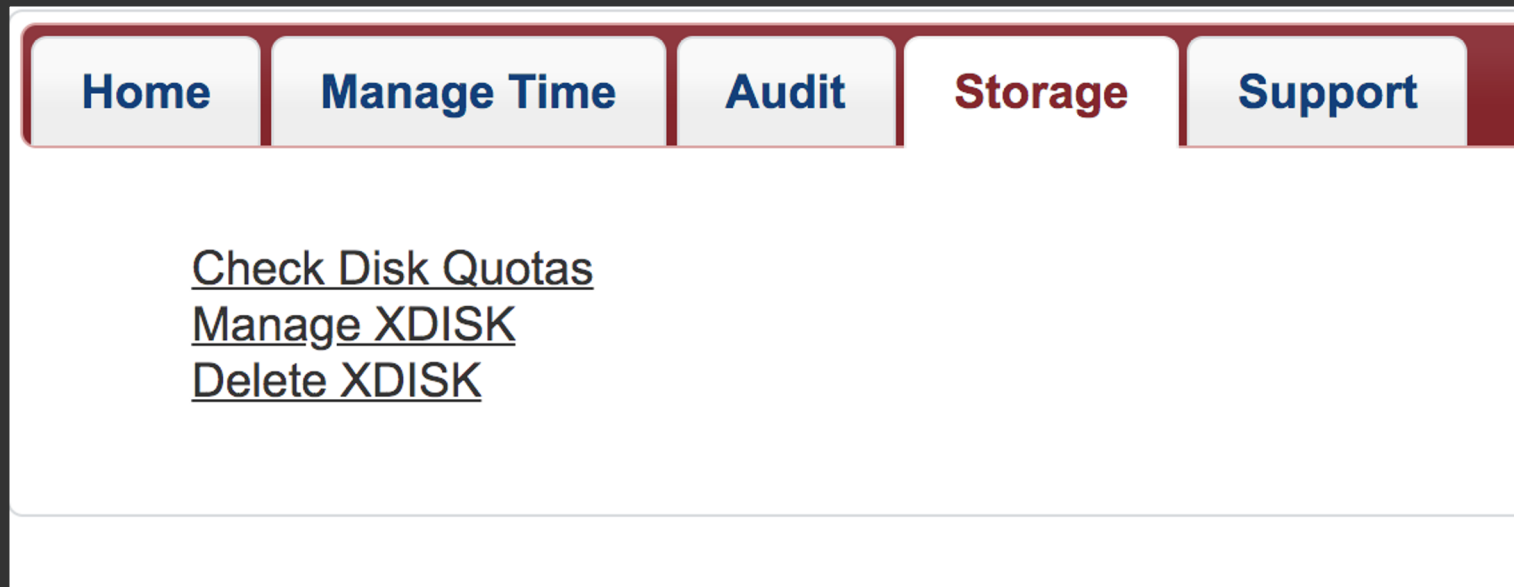
- Three storage options (all FREE):
- User level
 - /home
 - 50 GB
 - aka home directory or ~/
- Group (PI) level
 - /groups
 - 500 GB
 - /xdisk
 - temporary “scratch” storage
 - up to 20 TB for up to 150 days (plus another 150 day extension)
 - Requested and managed by PIs
- No rented storage

HPC Storage – Tier I



Monitoring storage

- portal.hpc.arizona.edu



The screenshot shows a navigation bar with five buttons: Home, Manage Time, Audit, Storage, and Support. The 'Storage' button is highlighted in red. Below the navigation bar, there are three underlined links: [Check Disk Quotas](#), [Manage XDISK](#), and [Delete XDISK](#).

Monitoring storage

- command line - uquota

```
[dshyshlov@login3 ~]$ uquota
```

	used	soft limit	hard limit
/groups/dshyshlov	209.6G	500.0G	500.0G
/home	13.9G	50.0G	50.0G
/xdisk/dshyshlov	3.9G	9.8T	9.8T

Monitoring storage – xdisk

- Xdisk is temporary storage
- Max duration – 150 days
- One extension – another 150 days
- PI's can monitor and manage xdisk on the portal

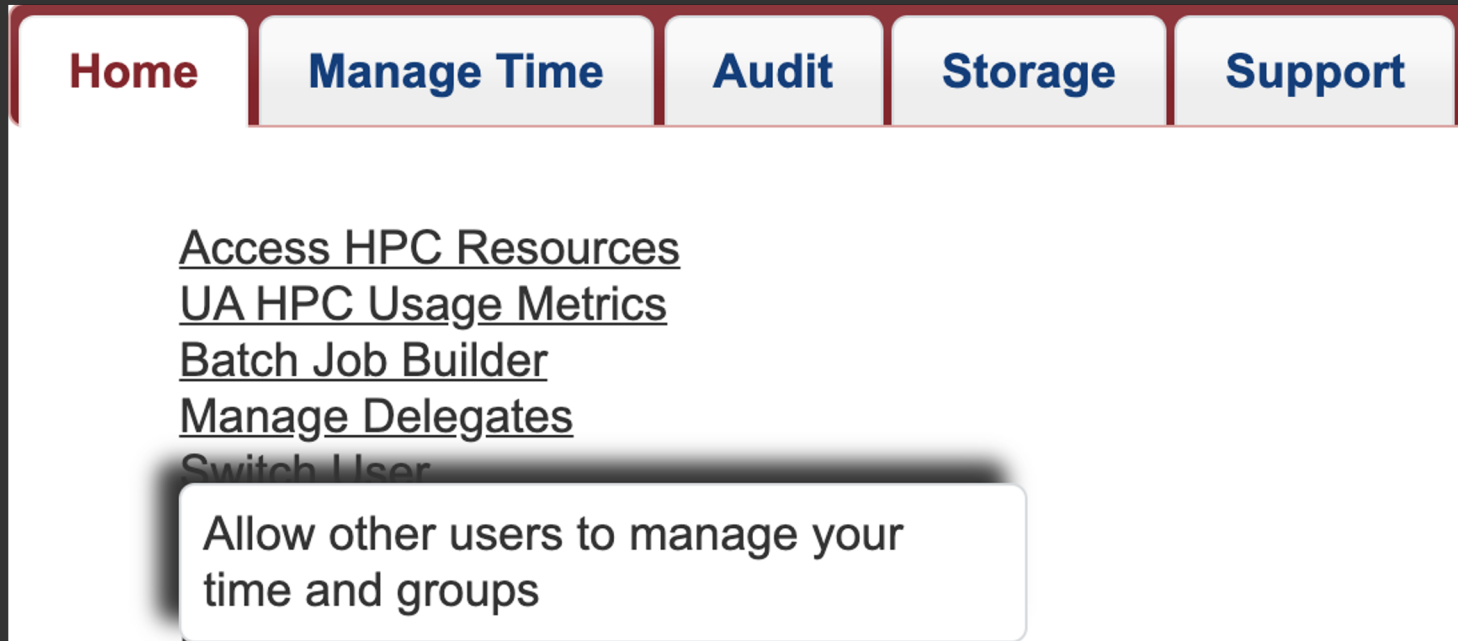
The screenshot shows a web portal with a navigation bar containing 'Home', 'Manage Time', 'Audit', 'Storage', and 'Support'. The 'Storage' tab is active. On the left, there are links for 'Check Disk Quotas', 'Manage XDISK', and 'Delete XDISK'. A modal window titled 'XDISK' is open, displaying the following information:

Current Allocation	/xdisk/dshyshlov
Remaining Days	129
Group	dshyshlov
Size (GB)	10000
Duration (days)	299

An 'Ok' button is located at the bottom right of the modal window.

Delegating xdisk permissions

- PI's can delegate permissions to manage xdisk.

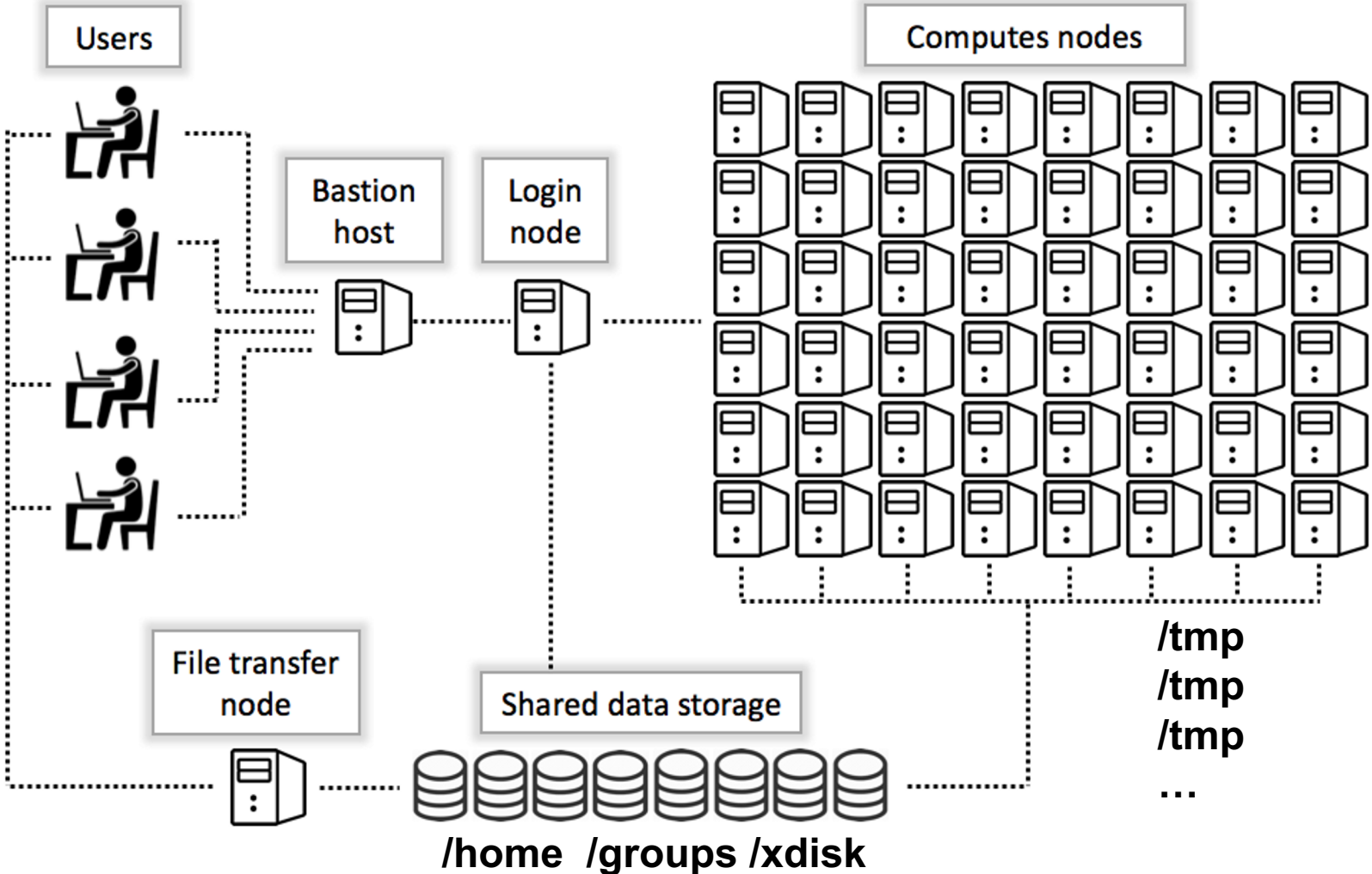


The image shows a screenshot of a web application interface. At the top, there is a navigation bar with five tabs: "Home", "Manage Time", "Audit", "Storage", and "Support". Below the navigation bar, there is a list of links: "Access HPC Resources", "UA HPC Usage Metrics", "Batch Job Builder", "Manage Delegates", and "Switch User". A dropdown menu is open under "Manage Delegates", displaying the text "Allow other users to manage your time and groups".

HPC Storage

- Compute node disk
 - /tmp
 - local disk on compute nodes
 - not connected to the main storage
 - the best performance for calculations
 - Ocelote – 800 GB is available on each node
 - Puma – 1.4 TB is available on each node
 - data is removed once the job is finished, so need to script copying input/output data to the main storage

HPC Storage



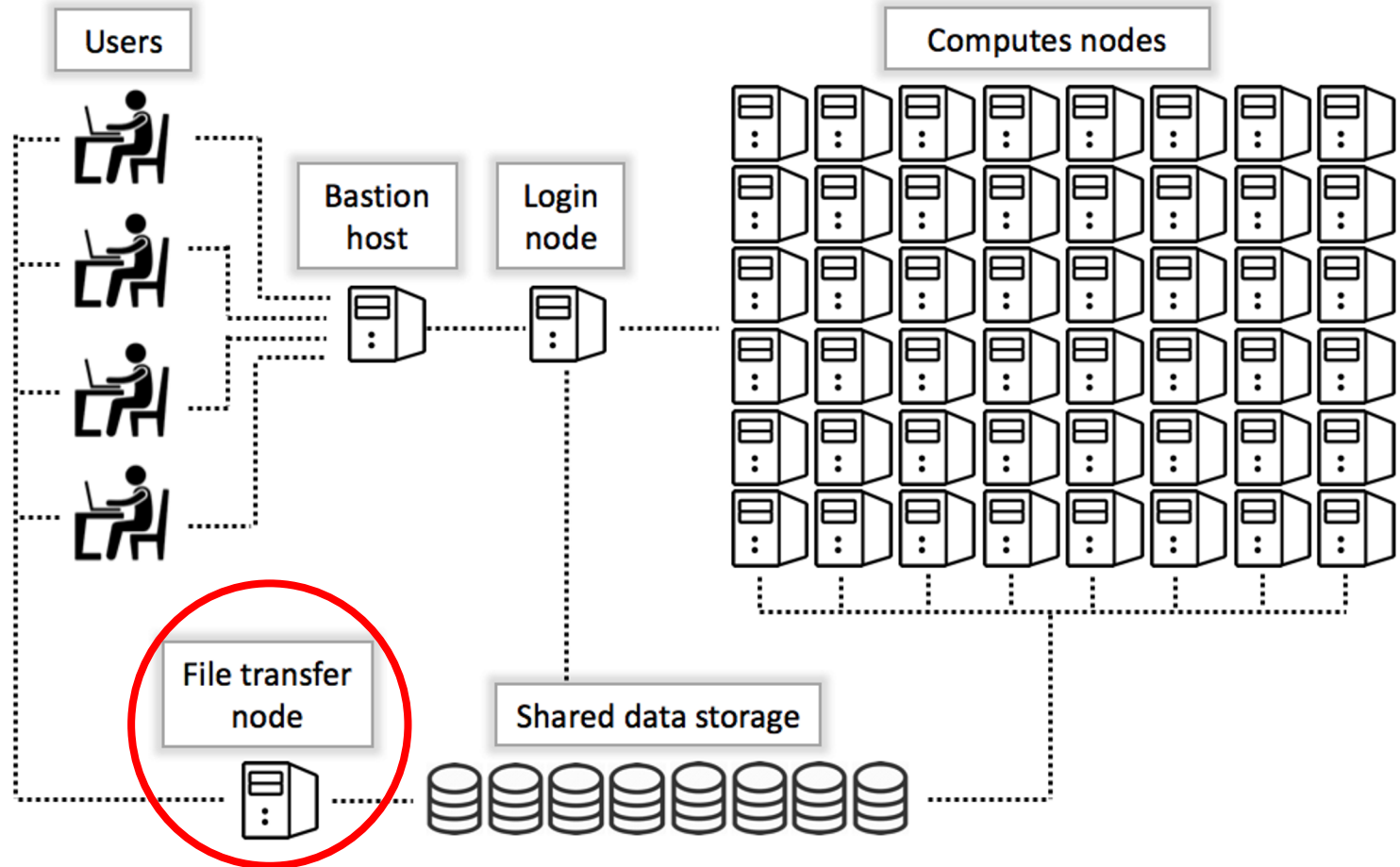


Transferring Files



THE UNIVERSITY
OF ARIZONA

Transferring Files



filexfer.hpc.arizona.edu

File transfer between local computer and HPC

- GUI



- Command line

- `wget`

- `scp`

- `sftp`

File transfer in web browser

- Open On Demand: ood.hpc.arizona.edu

The screenshot displays the OOD web interface. At the top, there is a navigation bar with the Arizona Research Technologies HPC Systems logo and menu items: Apps, Files, Jobs, Clusters, Interactive Apps, and My Interactive Sessions. On the right, it shows 'Help', 'Logged in as zwickl', and 'Log Out'. A yellow banner contains a note: 'Please NOTE: "windfall" jobs are terminated without notice if pre-empted by a "standard" job in queue.' Below the banner is a toolbar with buttons: '>_ Open in Terminal', '+ New File', 'New Directory', 'Upload', 'Download', 'Copy/Move', and 'Delete'. A left sidebar shows a tree view with 'Home Directory', '/groups', and '/xdisk'. The main content area shows the current directory path '/ home / u2 / zwickl /' with a 'Change directory' button and a 'Copy path' button. There are checkboxes for 'Show Owner/Mode' and 'Show Dotfiles', and a 'Filter:' input field. Below this is a table listing files and directories:

	Type	↑↓	Name	↑↓	Size	↑↓	Modified at	↑↓
<input type="checkbox"/>	Folder		bin	⋮	-		2/22/2022 2:18:45 PM	
<input type="checkbox"/>	Folder		containerTest	⋮	-		2/22/2022 9:55:45 AM	
<input type="checkbox"/>	Folder		DLcheckpointing	⋮	-		2/18/2022 4:50:14 PM	

Command line file transfer

- Command line transfer is easy with scp command

copy recursively,
maintaining permissions



- `scp -rp what_to_copy where_to_copy`

remote cluster location
default target is home directory



```
~/Desktop/ $ scp -rp data_dir zwickl@filexfer.hpc.arizona.edu:
```

```
~/Desktop/ $ scp -rp data_dir zwickl@filexfer.hpc.arizona.edu:/groups/chrisreidy/
```

```
~/Desktop/ $ scp -rp zwickl@filexfer.hpc.arizona.edu:data_dir .
```

To download, use the remote
location as the first argument





Managing Files



THE UNIVERSITY
OF ARIZONA

Sharing data between HPC users

- You can share data with another HPC user without moving the data
- Open file permissions
- Create symbolic links



File permissions

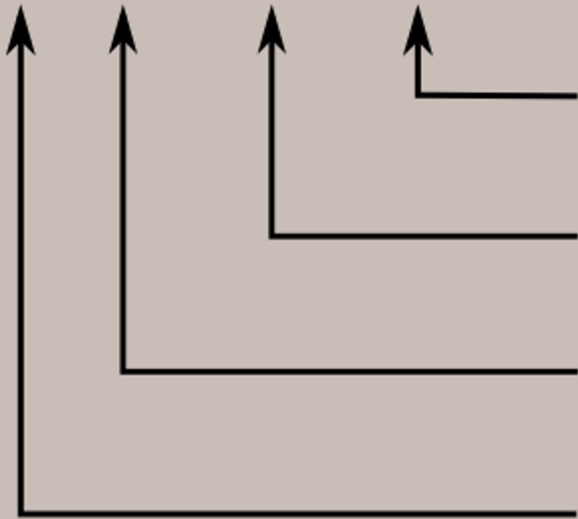
- Check file permissions – `ls -l`

```
-rw-r--r-- 1 dshyshlov rc 639 Oct 10 15:58 script.pbs
```

Owner

Group

- rwx rwx rwx



Read, write, and execute permissions for all other users.

Read, write, and execute permissions for the group owner of the file.

Read, write, and execute permissions for the file owner.

File type:
- indicates regular file
d indicates directory

File permissions

- View file permissions in OOD

The screenshot shows the Open OnDemand (OOD) web interface. At the top, there is a navigation bar with the Arizona Research Technologies logo and menu items: Apps, Files, Jobs, Clusters, Interactive Apps, and My Interactive Sessions. On the right, it shows 'Help', 'Logged in as zwickl', and 'Log Out'. Below the navigation bar is a yellow warning banner: 'Please NOTE: "windfall" jobs will be restarted or terminated without notice if pre-empted by a "standard" job in queue.' Below the banner is a toolbar with buttons: '>_ Open in Terminal', '+ New File', 'New Directory', 'Upload', 'Download', 'Copy/Move', and 'Delete'. On the left, there is a 'Home Directory' sidebar with links for '/groups' and '/xdisk'. The main content area shows the current directory path as '/ home / u2 / zwickl /' with a 'Change directory' button and a 'Copy path' button. Below the path, there is a text 'Check this box' with a red arrow pointing to the 'Show Owner/Mode' checkbox, which is checked. To the right of this checkbox are 'Show Dotfiles' and a 'Filter:' input field. Below this is a table listing files and directories. The table has columns for Type, Name, Size, Modified at, Owner, and Mode. The files listed are 'bin', 'containerTest', and 'DLcheckpointing', all owned by 'zwickl' with mode '755'. A red arrow at the bottom right points to the 'Mode' column, with the word 'Permissions' written in red below it.

Home Directory

- /groups
- /xdisk

↑ / home / u2 / zwickl / Change directory Copy path

Check this box → Show Owner/Mode Show Dotfiles Filter:

Showing 15 of 47 rows - 0 rows selected

Type	↑↓	Name	↑↓	Size	↑↓	Modified at	↑↓	Owner	↑↓	Mode	↑↓
<input type="checkbox"/>	Folder	bin	<input type="text"/>	-		2/22/2022 2:18:45 PM		zwickl		755	
<input type="checkbox"/>	Folder	containerTest	<input type="text"/>	-		2/22/2022 9:55:45 AM		zwickl		755	
<input type="checkbox"/>	Folder	DLcheckpointing	<input type="text"/>	-		2/18/2022 4:50:14 PM		zwickl		755	

Permissions

Numerical file permissions

Octal Value	File Permissions Set	Permissions Description
0	---	No permissions
1	--x	Execute permission only
2	-w-	Write permission only
3	-wx	Write and execute permissions
4	r--	Read permission only
5	r-x	Read and execute permissions
6	rw-	Read and write permissions
7	rwX	Read, write, and execute permissions

Numerical file permissions

- Permissions can be summarized with the digits 1-7
- One number appears for each of the user, group and all permission categories
- So, $777 = rwxrwxrwx$
 $755 = rwxr-xr-x$

Change file permissions

- Command to change permissions – `chmod`
 - `chmod +x filename` - make file executable for everyone
 - `chmod g+rwx filename` - open all permissions for your group
 - `chmod 777 filename` - open all permissions for everyone
 - `chmod -R 777 filename` - same as above but recursively (for all the subdirectories and files)

Symbolic links

- Create a soft link to a file or directory
 - In `-s path/to/the/destination link_name`
 - perfect for situations when you need to share read only data
 - requires permissions to link directory of another user
- Examples:
 - Create a shortcut to your `/groups` in your `/home`
 - In `-s /groups/PInetID ~/my_groups`
 - Share data on xdisk (permissions!)
 - In `-s /xdisk/PInetID/project ~/project_data`

File & Space Management Tools

- Show disk usage command - du

```
$ du scripts/  
24 scripts/slurm_testing  
8 scripts/puma/test  
380 scripts/puma  
12 scripts/elgato  
3304 scripts/ocelote  
36 scripts/chaining  
7264 scripts/
```

- Useful flags:
 - -h – human readable format
 - -s – display only the total
 - --max-depth=N – limit depth of subdirectories to N

```
$ du -sh scripts/  
7.1M scripts/
```


Operate on Data

- Things that can “break” the system:
 - heavy use of the login node
 - too many jobs
 - too many files
 - heavy I/O jobs
 - copying GB of data

Preview for Part 2

- File and folder tidiness
- Scriptable methods of transferring data into/out of the HPC
- Additional tools for managing files and projects
- Data archiving