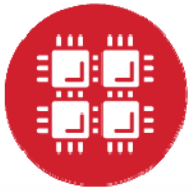


Ohio Supercomputer Center

An OH·TECH Consortium Member

How to Choose the Right File System for Your Work

Summer Wang
1224 Kinnear Road, Columbus, OH 43212

April 6, 2017

The OH-TECH Consortium

Ohio Supercomputer Center provides high performance computing, software, storage and support services for Ohio's scientists, faculty, students, businesses and their research partners.

OARnet connects Ohio's universities, colleges, K-12, health care and state and local governments to its high-speed fiber optic network backbone. OARnet services include co-location, support desk, federated identity and virtualization.

OhioLINK serves nearly 600,000 higher education students and faculty by providing a statewide system for sharing 50 million books and library materials, while aggregating costs among its 90 member institutions.

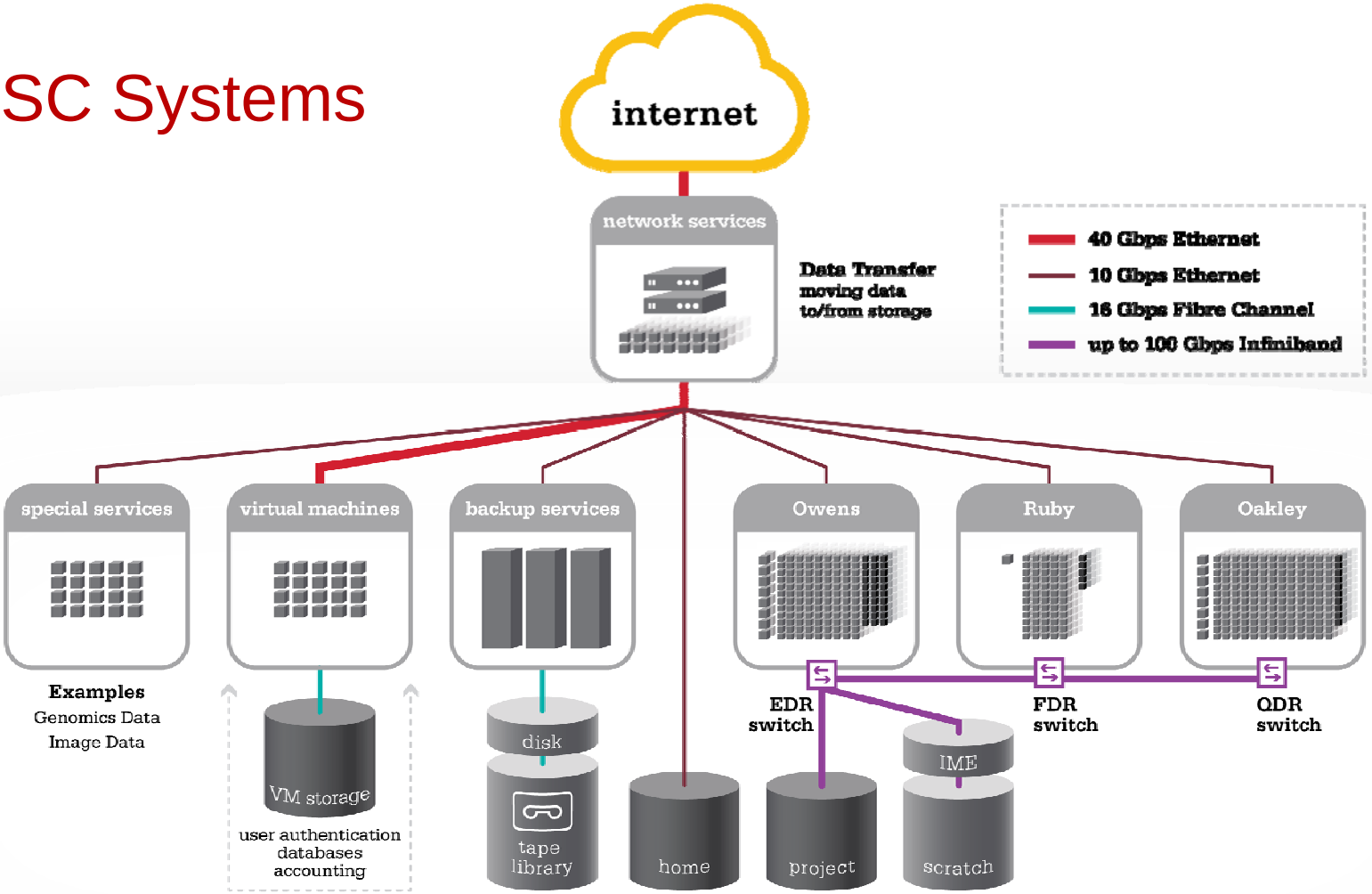
eStudent Services provides students increased access to higher education through e-learning and technology-enhanced educational opportunities, including virtual tutoring.

Research & Innovation Center will operate, when opened, as the proving grounds for next-generation technology infrastructure innovations and a catalyst for cutting-edge research and collaboration.

Agenda

- Different File Systems at OSC
- File Management
- Data Sharing
- File Transfer Mechanisms

OSC Systems


internet
network services
Data Transfer
moving data
to/from storage
40 Gbps Ethernet
10 Gbps Ethernet
16 Gbps Fibre Channel
up to 100 Gbps Infiniband
special services
virtual machines
backup services
Owens
Ruby
Oakley
Examples
Genomics Data
Image Data
VM storage
user authentication
databases
accounting
disk
tape
library
home
project
IME
scratch
EDR
switch
FDR
switch
QDR
switch

Overview of OSC File Systems

File System	Total Size	Bandwidth	Space Purpose	Type
Home	900 TB	10 GB/s	Permanent	WAFL in an NFS appliance
Project	3,400 TB	30-50 GB/s	Long-term	GPFS
Global Scratch	1,000 TB	30-50 GB/s (*100 GB/s through IME)	Temporary	GPFS
Local Scratch	Varies	Varies	Temporary	Varies
Tape	5,500 TB	3.5 GB/s	Archive	LTO tape

Quota Display

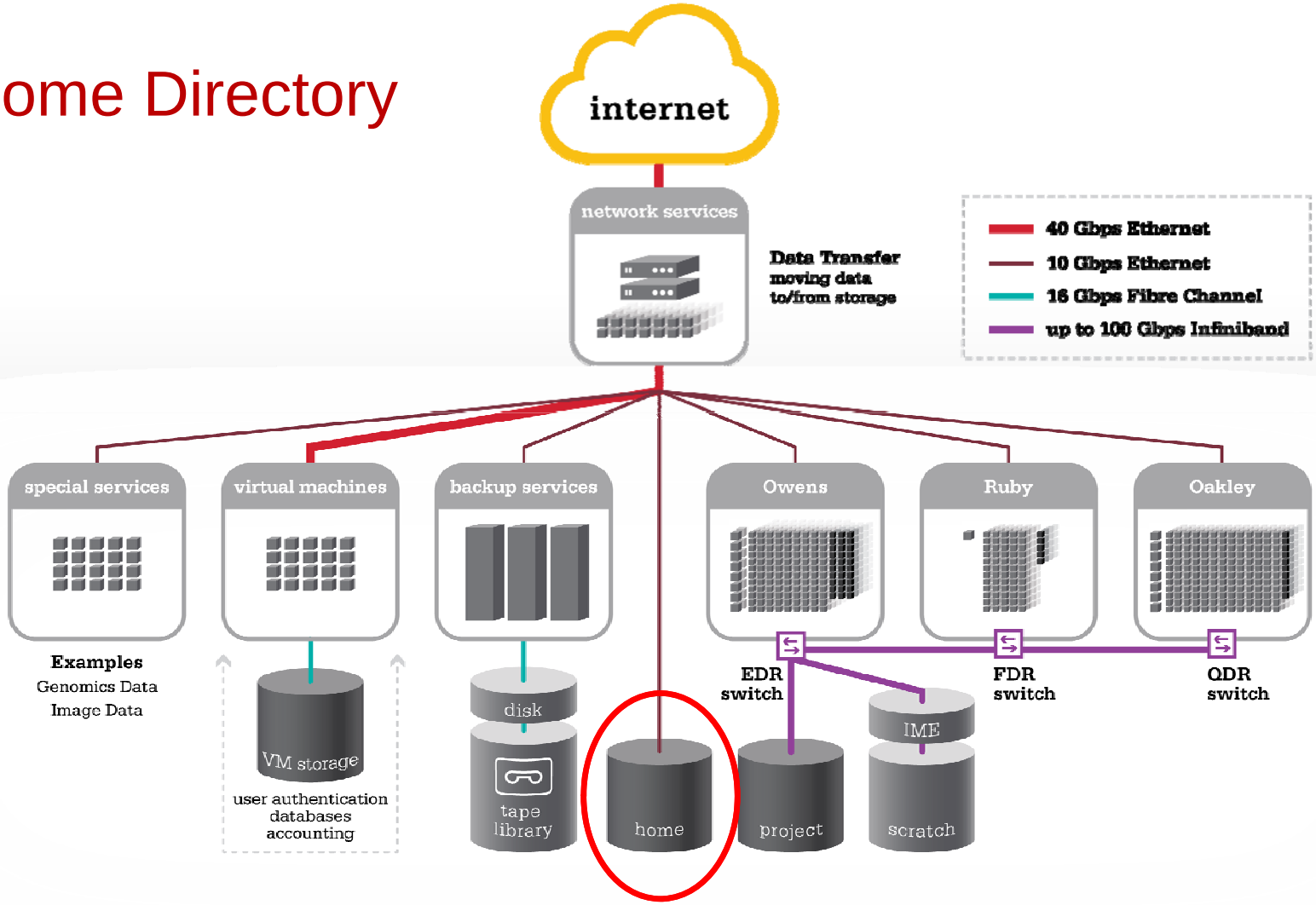
- Quota including your Home, Project, and Global Scratch is displayed at login
  - This information is collected twice a day

```
As of 2017-03-28T16:02:07.000000 userid xwang on /fs/scratch used 1288GB of quot
a 0GB and 2181550 files of quota 0 files
As of 2017-03-28T16:02:33.482001 userid xwang on /users/PES0691 used 824.2MB of
quota 500GB and 1 files of quota 1000000 files
As of 2017-03-28T16:02:47.657370 userid xwang on /users/oscgen used 423.0GB of q
uota 500GB and 852764 files of quota 1000000 files
```

- Use command **`quota`** to get the most updated quota information of your home directory

```
[xwang@owens-login01 ~]$ quota -u xwang
Disk quotas for user xwang (uid 16059):
     Filesystem  blocks   quota   limit   grace   files   quota   limit   grace
10.11.200.26:/oscgen
                443570508  524288000 524288000          852821  1000000 1000000
```

Home Directory



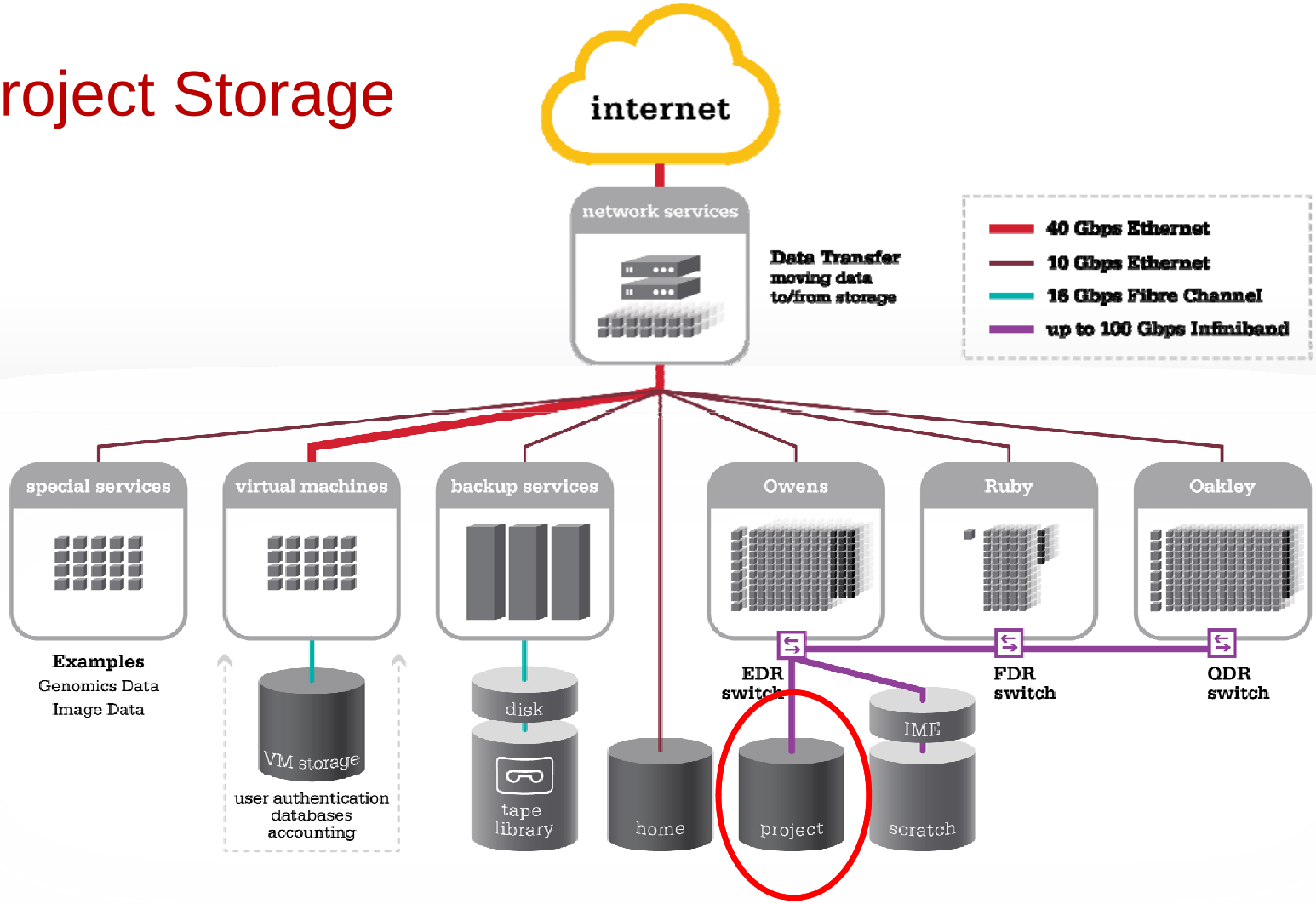
Home Directory

- This storage is provided to each OSC userID
  - Path: **`/users/projectID/userID`**
  - Environment variable: **`$HOME`** or **`~userID`**
  - Quota (fixed): 500 GB of storage and 1,000,000 files
- Visible from all login nodes and all compute nodes
- Backup: daily with two copies of files in tape
- Purge: no
- Lower performance; not recommended for heavy I/O job

Error: No space left on device

- Error “no space left on device" when writing data to home directory: the disk is full
  - Check your home directory quota. If you reach quota, consider reducing space usage (to be discussed later)
  - If your usage is less than 500GB, your disk may be filled up with 'snapshot' files (invisible to users and used to track fine-grained changes to your files for recovering lost/deleted files). To avoid this situation in future, consider running jobs with heavy I/O in the scratch file system (examples will be shown later)

Project Storage



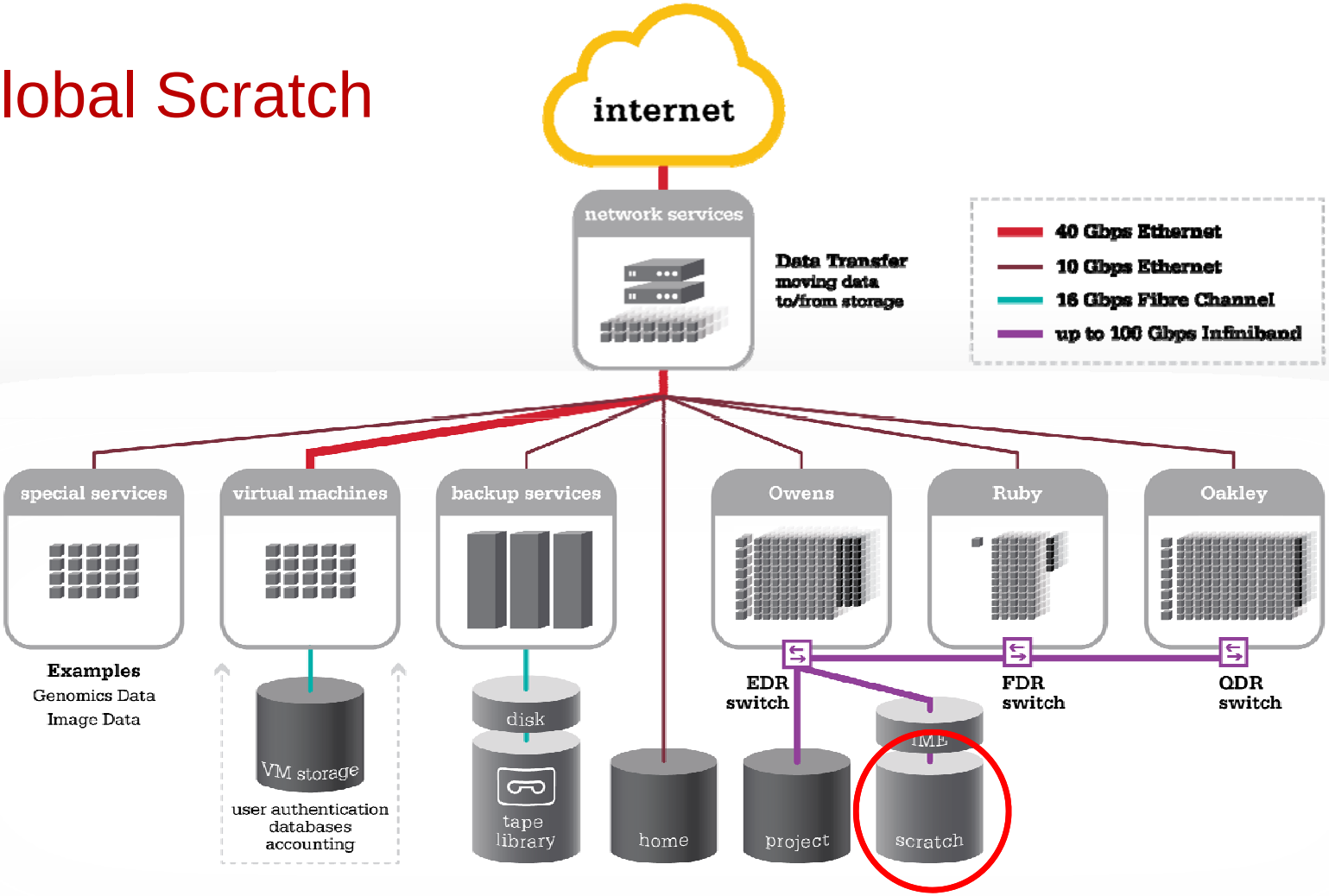
Project Storage

- This storage is provided via request from PI
  - Path: **`/fs/project/projectID`**
  - Quota: by request
  - Once it is allocated, it is shared by all members of the project for a specific period of time
- Visible from all login nodes and all compute nodes
- Backup: daily with single copy of files in tape
- Purge: no
- High performance

How to Apply for Project Storage

- Information from PI Sent to OSC Help:
  - How much space your project will need
  - How long you need this space
  - Why you need the space
  - What measures you are taking and will continue to take to optimize the space usage (deleting unneeded data, compressing data, moving infrequently used data to archive)

Global Scratch



Global Scratch Storage

- This storage is shared by all OSC users
  - High-performance, high-capacity, temporary space
  - Path: **`/fs/scratch`**
  - Quota: no
- Visible from all login nodes and all compute nodes
- Backup: no
- Purge: yes. Data that have not been accessed for more than or equal to 180 days will be removed from the system every Wednesday
- High performance (similar to Project without through IME)

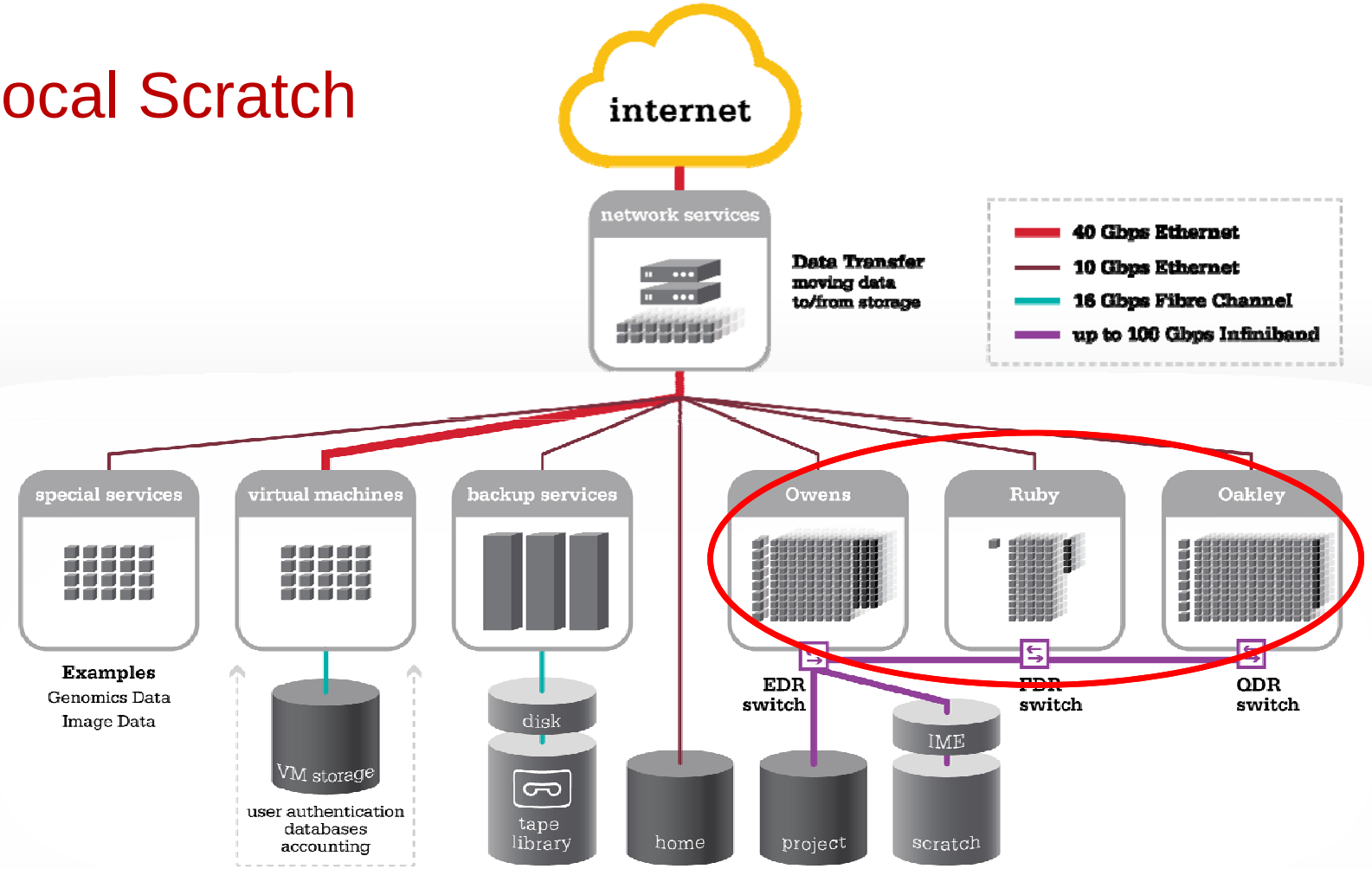
Exemption to Scratch Purge Policy

- Information Sent to OSC Help in a Timely Manner:
  - OSC HPC username
  - Path of directories/files that need exemption to file deletion
  - Duration: till MM/DD/YY (The max exemption duration is 180 days)
  - Detailed justification

New Scratch Storage Policy for Public Comments

- Key Modifications:
  - Purge period: change from 180 days to 120 days
  - Exemption is allowed, but renewal of existing exemption request is forbidden
- Public Comment Period: April 3-28, 2017
  - All comments are welcome
  - We will provide response to comments after the public comment period closes
- Link

Local Scratch



Local Scratch Storage

- This storage is on compute node for use during the user's batch job
  - Path: **`/tmp`** after ssh/rsh to the node
  - Environment variable: **`$TMPDIR`**
  - Space: 800GB (Oakley) and 1 TB (Ruby & Owens)
- Visible from the local compute node only
- Backup: no
- Purge: yes. Data is purged at the completion of the PBS job. Be sure to copy output back at the end of job
- Highest performance because data does not have to be sent across the network and handled by a file server

Using Scratch Storage in Job for Temporary Files

- Two Main Advantages:
  - Access is fast
  - Avoid temporary files generated in your home directory
- **`$TMPDIR`** or **`$PFSDIR`**
  - **`$TMPDIR`**: local disk of each node. The files in this directory are NOT visible from all the nodes in a parallel job; each node has its own directory
  - **`$PFSDIR`**: a temporary directory on Global Scratch file system. It is used in a parallel job of which a single directory is shared by all the nodes a job is running on

Example: Use **$TMPDIR** in Job

```
#PBS -N fluentjob
#PBS -l walltime=5:00:00
#PBS -l nodes=1:ppn=1
#PBS -l software=fluent+1
#PBS -j oe
# The following lines set up the FLUENT environment
module load fluent
# Move to your working directory where the job was submitted from
cd $PBS_O_WORKDIR
# Copy files to $TMPDIR and move there to execute the program
cp test_input_file.cas test_input_file.dat run.input $TMPDIR
cd $TMPDIR
# Run fluent
fluent 3d -g < run.input
# Finally, copy files back to your home directory
cp result.cas $PBS_O_WORKDIR
```

Read and write files directly from $TMPDIR; access is fast

Read and write files directly from your home directory; access is slow

Using **$PFSDIR** in Job

- Similar to **$TMPDIR**
- However, the **$PFSDIR** directory won't be created by job prologue. You will have to create the **$PFSDIR** directory by yourself if you use this directory.
- See here for more information: https://www.osc.edu/supercomputing/known-issues/critical_change_about_using_pfsdir_directory_at_osc

File Management

- How to Lower Disk Space Usage:
  - Identify and delete data you do not need
  - Compress large, rarely used files
    - Use **`gzip`** or **`bzip2`** commands
  - Combine large numbers of small files
    - Use **`tar`** command
  - Archive data (the service will be provided by OSC later)
  - Transfer data to local file system (to be discussed later)
  - Web documentation:

  https://www.osc.edu/resources/getting_started/howto/howto_reduce_disk_space_usage

Data Sharing

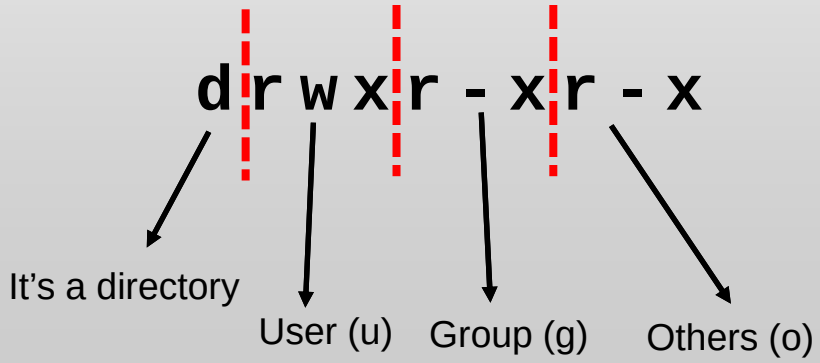
- Data Security and Integrity
  - Sharing data with others must be done carefully
  - The chances for data loss increase when more users have access to the data. Permissions should be set to the minimum necessary to achieve the desired access
- Sharing Your Data with Other OSC Users
  - Change file permissions using **`chmod`**
  - Access Control Lists (ACLs)

Unix File Permissions

- Check file permissions using the command **`ls -l`**
- By default:

	Permission
Home	**`drwxr-xr-x`**
Project	**`drwxrwx---`**
Global Scratch	**`drwxr-xr-x`**
Local Scratch	**`drwx------`**

  - **`r`**: read
  - **`w`**: write
  - **`x`**: execute



chmod ("change mode") Command Syntax

Description	Letter/ Operator	Octal Digit
Read permission	**r**	4
Write permission	**w**	2
Execute (or access for directories) permission	**x**	1
No permission	**-**	0
The permissions that the User who owns the file has for it	**u**	
The permissions that other users in the file's Group have for it	**g**	
Permissions that Other users not in the file's group have for it	**o**	
Permissions that All users have for it	**a**	
Adds the specified modes to the specified classes	**+**	
Removes the specified modes from the specified classes	**-**	
The modes specified are to be made the exact modes for the specified classes	**=**	

Example: Change File Permissions Using **chmod**

- Symbolic Mode

  `chmod u=rwx,g=rx, o=r file`

- Octal Mode

  `chmod 754 file`

  – 7 is the combination of permissions 4+2+1 (read, write, and execute) for user

  – 5 is 4+0+1 (read, no write, and execute) for Group

  – 4 is 4+0+0 (read, no write, and no execute) for other

- For Directory: with **`–R`** (Recursive)

  `chmod –R u=rwx,g=rx, o=rx directory`

- For more information: **`man chmod`**

Access Control Lists (ACLs)

- An ACL is a list of permissions associated with a file or directory
  - An extension of the traditional UNIX permission concept
  - Allow more complex and sophisticated control for sharing data, like granting access to files to an individual user
  - Supplant the standard UNIX group permissions with the **`mask`** property. Changing the group permissions will invalidate the ACLs set for a file or directory
- ACLs on our systems:
  - Home directory: NFSv4 ACL
  - Project and Global Scratch storage: POSIX ACL

Home Directory: NFSv4 ACL

- To Set and Modify ACLs:
  `nfs4_setfacl [OPTIONS] COMMAND file`
- To View ACLs: **`nfs4_getfacl file`**
- More info:
  - **`man nfs4_acl`**
  - **`man nfs4_setfacl`**
  - **`man nfs4_getfacl`**

Project & Global Scratch Storage: POSIX ACL

- To Set and Modify ACLs:

  `setfacl [OPTIONS] COMMAND file`

  - To set ACLs: **`setfacl -m u/g:ID:rw- file`**
  - To remove ACLs: **`setfacl –x u/g:ID file`**
- To Display ACLs: **`getfacl file`**
  - A '+' character to the Unix permissions string indicates that ACLs are set for the file or directory
- More info:
  - **`man acl`**
  - **`man setfacl`**
  - **`man getfacl`**

Demonstration

- **chmod**
- **NFSv4 ACL**
- **POSIX ACL**

Transferring Files at OSC

- Using Unix Commands on a Linux/Mac Machine
- Using an SFTP Client on a Windows/Mac Machine
- Using OSC OnDemand
- Using Globus

Transferring Files Using Unix Commands

- For small files, connect to login node
- For large files, connect to **`sftp.osc.edu`** (file transfer only)
- **`sftp:`**
  - Encrypts data before it is sent across the network
  - Resume interrupted transfers, list directory, remove remote file
- **`scp`**
  - Secure Shell (SSH) for data transfer and utilizes the same mechanisms for authentication
- **`rsync`**
  - It synchronizes files and directories between two different locations (or servers)
  - Copies only the differences of files that have actually changed

Transferring Files Using an SFTP Client

- For small files, connect to login node
- For large files, connect to **`sftp.osc.edu`** (file transfer only)
- Software
  - WinSCP
  - Cyberduck
  - FileZilla

Transferring Files Using OSC OnDemand

- OnDemand is our "one stop shop" for access to OSC High Performance Computing resources via a web browser:
  - **Transfer files through a web-based File Explorer**
  - Create, edit, submit, and monitor jobs
  - Run GUI applications
  - Connect via SSH
- File Transfer Limit:
  - Upload: file size <10 GB, up to 6 files simultaneously
  - Download: no size limit, up to 6 files simultaneously

Demonstration

- OnDemand:

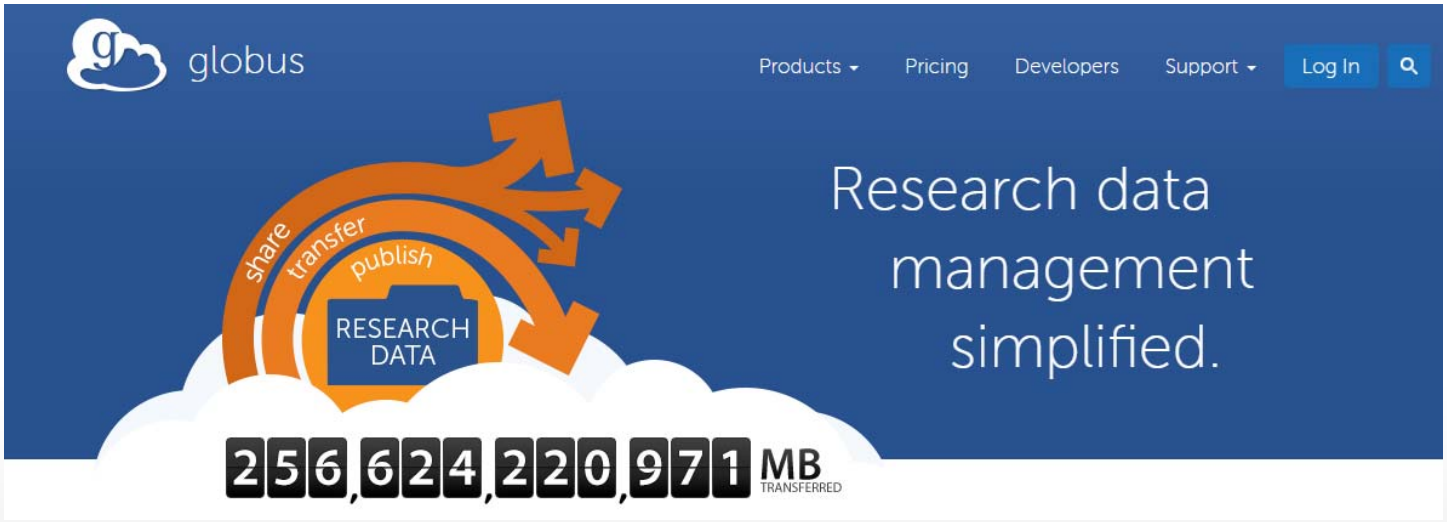
www.ondemand.osc.edu

- Web Documentation:

https://www.osc.edu/resources/online_portals/ondemand/file_transfer_and_management

https://www.osc.edu/resources/online_portals/ondemand

Globus Overview



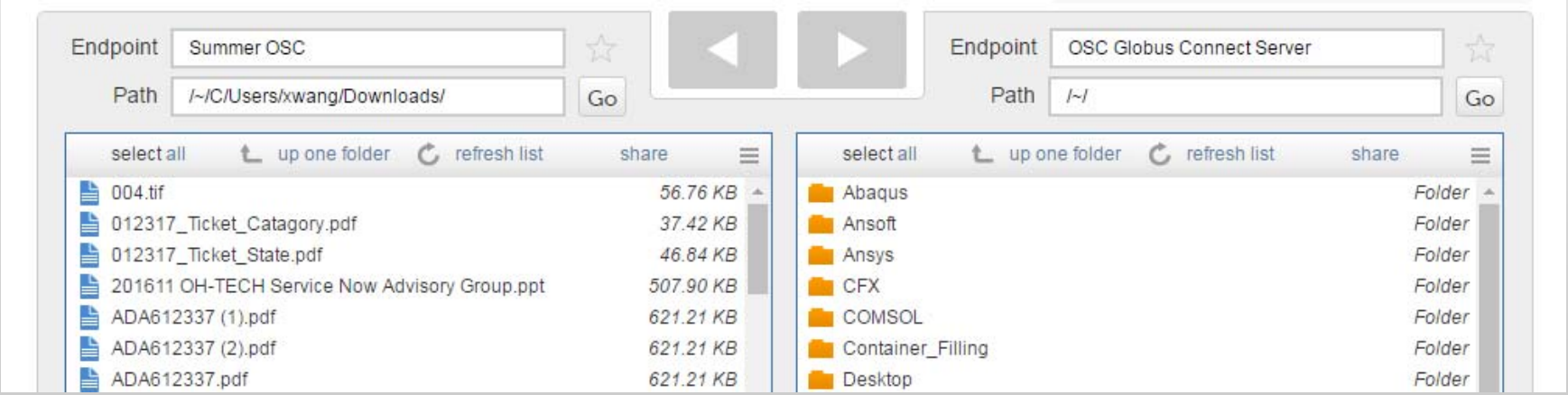
- Developed and maintained at the University of Chicago
- An online service that uses GridFTP for high-speed, reliable, and secure data transfer
- Manage file transfers, monitor performance, retry failures, recover from faults automatically when possible, and report the status of your data transfer
- Good for large files, and large collections of files

Transferring Files Using Globus

- With Globus, files are always transferred between two "endpoints". One endpoint is always a Globus server, such as the one at OSC. The other endpoint can be a second server, or a Globus connect personal endpoint, which is a desktop application
- Transfers are managed by a master server at globusonline, which manages credentials, sets up, and monitors transfers. However, data streams directly between the two endpoints.

Transferring Files Using Globus Between OSC and Your Personal Desktop

- Install Globus Connect Personal
  - Create an account: https://www.globus.org/
  - Download and Install Globus Connect Personal Client
- Login to Globus. Find "Transfer Files" under "Manage Data" (OSC endpoint: osc#gcs)



Demonstration

- Globus
- Web Documentation:

https://www.osc.edu/resources/getting_started/howto/howto_transfer_files_using_globus_connect

Questions

Summer Wang

Client Engineer, Ohio Supercomputer Center

xwang@osc.edu

Like Us on Facebook:
https://www.facebook.com/ohiosupercomputercenter
Follow Us on Twitter: https://twitter.com/osc

1224 Kinnear Road
Columbus, OH 43212
Phone: (614) 292-9248