# Managing, Sharing and Moving Big Data

Tracy Teal and Greg Mason
Institute for Cyber Enabled Research

- Data storage options
- Storing and accessing data on the HPCC
- Transferring data to and from the HPCC
- Sharing smaller data files with collaborators
- Transfering and sharing big data through Globus
- Recommended data management practices

# What is Big Data?

Big data is any data that is larger than you're used to handling.

It's still just data though, and good data management practices are important.

# What is data?

Data is what you measured or facts you collected.

- Files with data
- Information about the data

# Data

Sequence files

Measurement files from machines

Excel files with measurements

# Metadata

Information about the data – sample, location, experimental conditions, etc.

# Managing data

The key thing about data is that you probably want to keep it. You spent a lot of time and/or money collecting it, and unless you have it in a secure place and know what it is, it's useless.

# Things to think about when managing and storing data

- Backups
- How frequent are backups
- Location – on or off site
- Accessibility – how often do you need to access it
- Sharing – does it need to be shared with collaborators or other lab members
- Integrity – can the data be corrupted
- Raw files – do you have certain datasets that

# Options for storing data

- Computer
- External hard drive (on or off site)
- Backup service – Mozy, Dropbox, others
- Lab or departmental servers
- Amazon
- MSU HPCC

# What is the HPCC?

High Performance Computing Cluster

Founded in 2004

Run by iCER

- Provide a level of performance beyond what you could get and reasonably maintain as a small group
- Provide a variety of technology, hardware and software, that would allow for innovation not easily found

# iCER

Institute for Cyber-Enabled Research

Established to coordinate and support multidisciplinary resource for computation and computational sciences. The Center's goal is to enhance MSU's national and international presence and competitive edge in disciplines and research thrusts that rely on advanced computing.

# HPC Systems

FREE*

- Large Memory Nodes (up to 2TB!)
- GPU Accelerated cluster (K20, M1060)
- PHI Accelerated cluster (5110p)
- Over 540 nodes, 10000 computing cores
- Access to high throughput condor cluster
- 363TB high speed parallel scratch file space
- 50GB  replicated file spaces
- Access to large open-source software stack and specialized bioinformatics VMs
- User and research space – up to 1TB for

# Managing data on the HPCC

- Each User gets 50gb of home directory
- Users can request up to 1tb for free (paid for by the Vice President of research)
- Users can purchase additional disk space at $175/TB-year
- Shared research space is available
- Snapshots of disk are taken every hour
- Full offsite backups are taken every day

# Accessing the HPCC

- SFTP
- Command line
- Mounting drives on your computer

# SFTP

https://wiki.hpcc.msu.edu/display/hpccdocs/Transferring+Files+to+the+HPCC

Mac

- Filezilla

PC

- MobaXTerm

# Logging in to the HPCC

Mac

If you're on a Mac, you can use the terminal

Applications -> Utilities -> Terminal

Then type

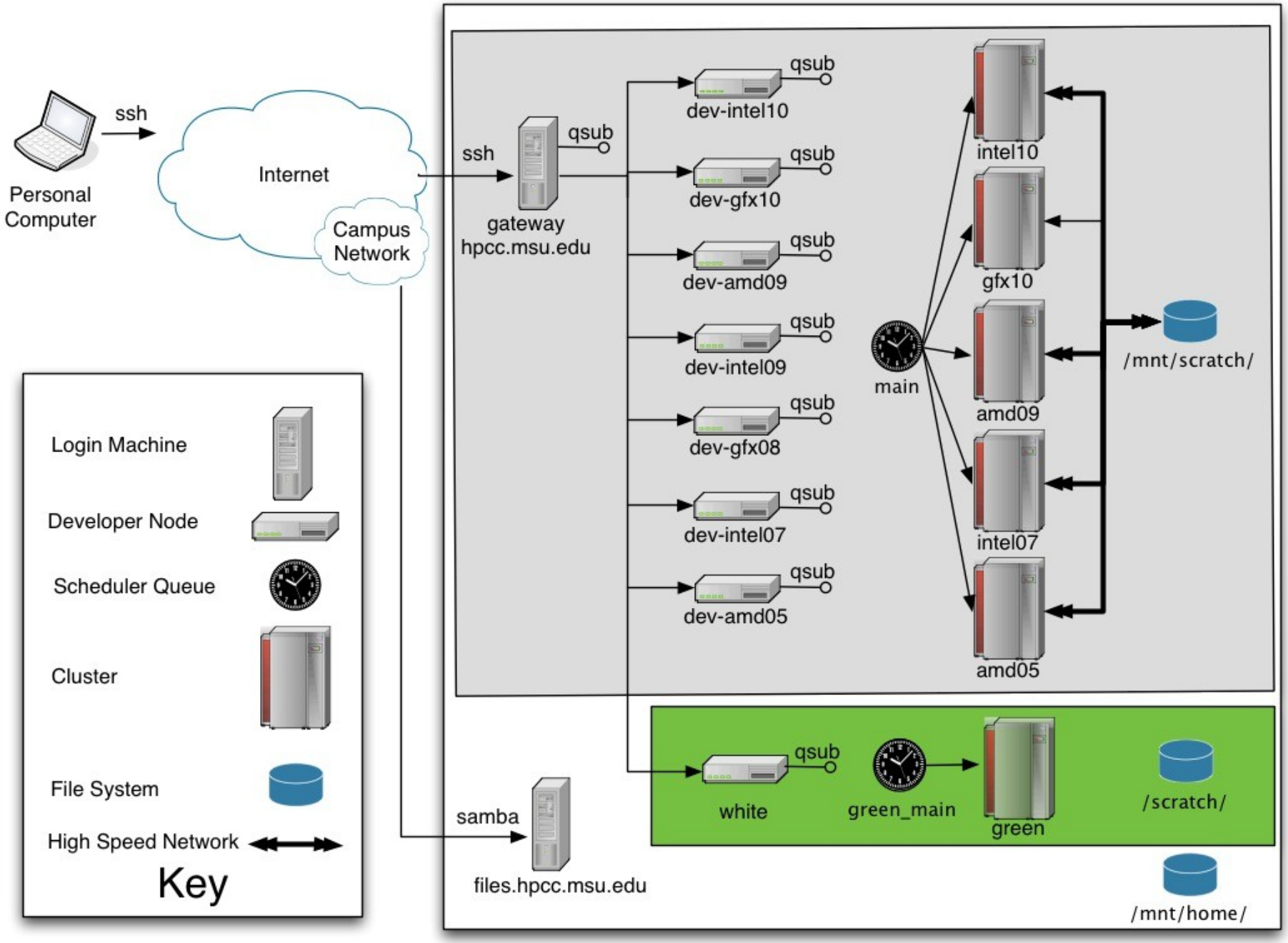
```
ssh -X  msu_username@hpcc.msu.ed
```

You'll be prompted for your MSU password and then you'll be logged in

PC

You can use MobaXTerm

http://mobaxterm.mobatek.net


ssh
Personal Computer
Internet
Campus Network
ssh
qsub
gateway
hpcc.msu.edu
dev-intel10
dev-gfx10
dev-amd09
dev-intel09
dev-gfx08
dev-intel07
dev-amd05
main
intel10
gfx10
amd09
intel07
amd05
/mnt/scratch/
Login Machine
Developer Node
Scheduler Queue
Cluster
File System
High Speed Network
Key
samba
files.hpcc.msu.edu
white
green_main
green
/scratch/
/mnt/home/

# Getting around at the command line

Shell cheat sheet

https://github.com/swcarpentry/boot-camps/blob/master/shell/shell_cheatsheet.md

ls
cd
mkdir
quota

# Mapping your home or research space to your home computer

https://wiki.hpcc.msu.edu/display/hpccdocs/Mapping+HPC+drives+to+a+campus+computer

# Using VPN if you're off campus

http://network.msu.edu/public/ssl-vpn.html

# Research space

User space versus research space file permissions

Requesting research space

https://contact.icer.msu.edu/contact

# Sharing with a collaborator

- Request an MSU ID for your collaborator and request access for them to your shared research space
- MSU FileDepot
- Globus

# Requesting access for collaborators

Link on New Account Request form

https://contact.cl.msu.edu/request.php?service=netidpurchase

# FileDepot

https://filedepot.msu.edu

# Globus

Globus transfer is easy, fast, secure, and reliable.

http://globus.org

https://wiki.hpcc.msu.edu/display/hpccdocs/Transferring+data+with+Globus

Sharing data through Globus

MSU has a plus account and can give out a limited number of endpoint IDs

# So what do I do about data file management?

(ideas my own and not necessarily those of iCER or MMG)

- Create a database that will keep track of your data files – this can be Access, Filemaker, SQL, whatever is best for your group to work with
- In the database keep track of the metadata and the filename of the data associated with that metadata

  e.g. 2014-05-07 fly leg_region Morgan 2014-05-07_fl_1.ols
- Figure out a file naming scheme that will create unique file names. Using the date helps.
- Store those files in shared research space on the HPCC