Introduction to Data Storage at Research Computing

Gerardo Hidalgo-Cuellar

Email: ghidalgo93@gmail.com
Slides: https://github.com/ghidalgo93/RCTutorial

Key Takeaways

Why High Performance Computing (HPC) has specific storage needs

What storage systems users have access to at CU Research Computing (RC)

What each storage system should or shouldn't be used for

HPC Storage | RC Storage | Wrap up & Questions
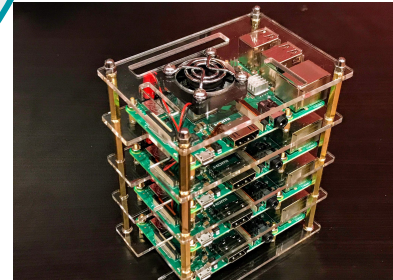
High Performance Computing

Personal Computers

- Personal Data
- Photos
- Games
- Applications...

HPC

- Quantum mechanics
- Weather forecasting
- Molecular sciences
- Early universe...

High Performance Computing

Submit Job

Get Results

Management service

Computer Hardware
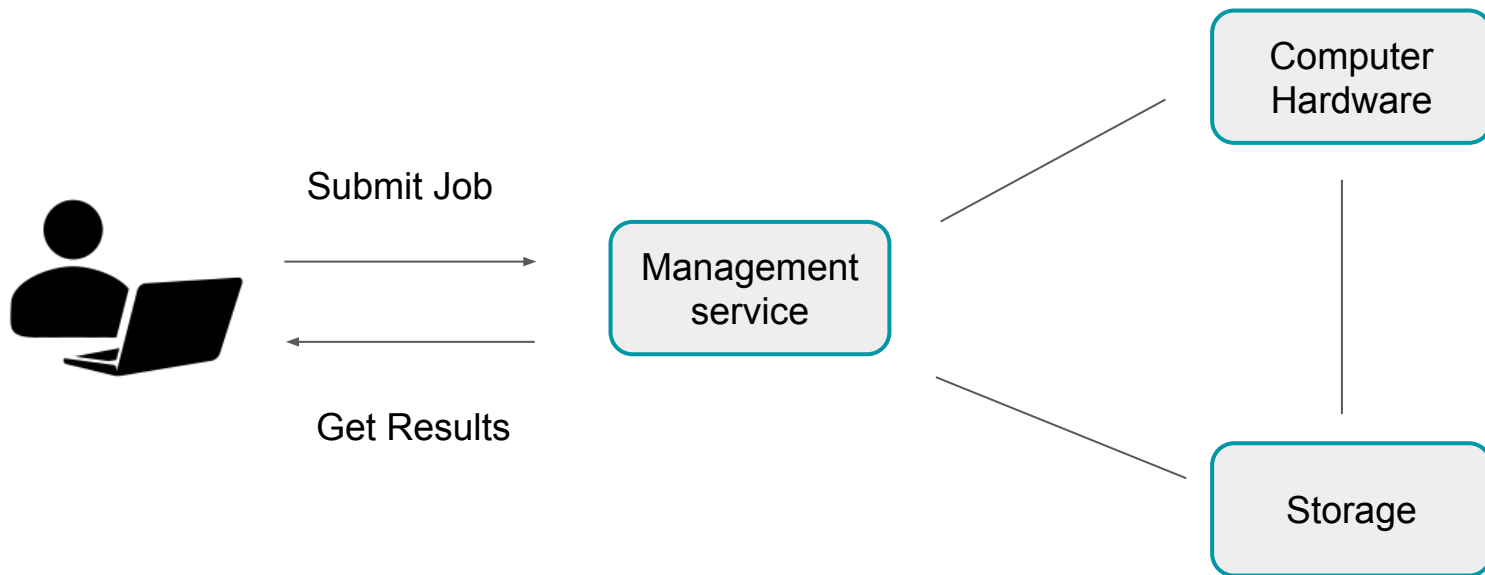
Storage

Research Computing Storage

Storage

- /home - user space
- /projects - shared space
- Local storage - e.g. your computer
- PetaLibrary - storage for a fee
- Scratch Space - temporary

Example Case:

- Melisa wants to use RC resources to model early universe conditions

- We're going to follow her journey via the RC storage system

- Has an RC account and has successfully accessed the network via the login node

```
Login              /home/<user>
Node
                                    /projects/<user>

                                              Scratch
                                              space
```
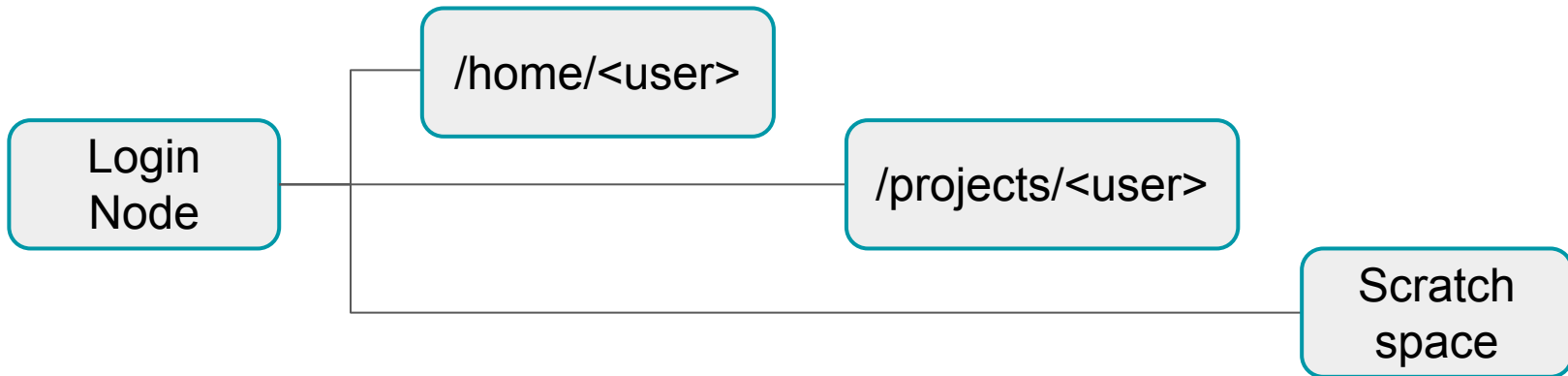
/home

/home/<user>

Login Node
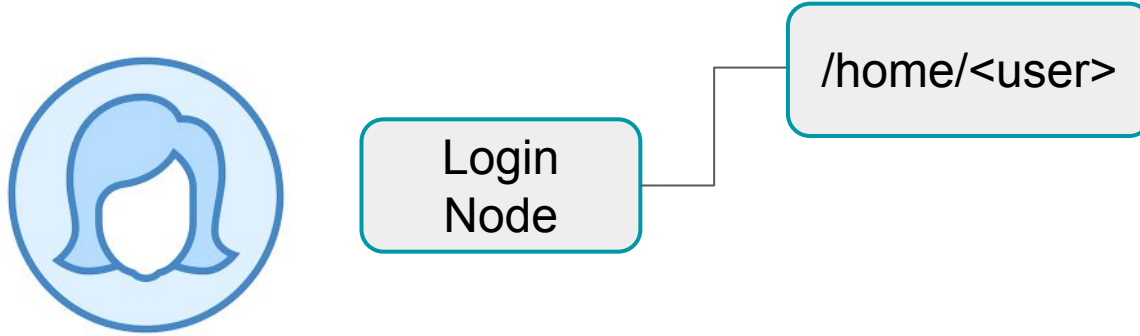
Examples:

- Slurm job script she's been working on
- Editor configurations

/home

/home/<user>
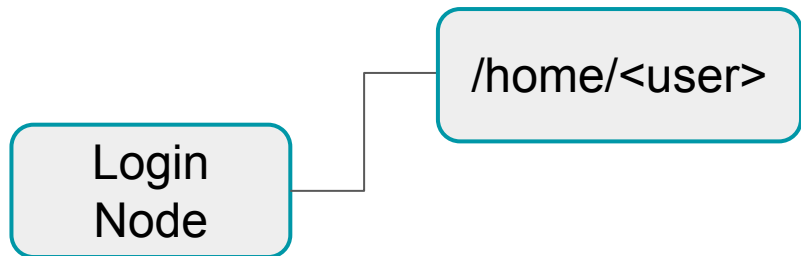
Login
Node

- User Space
- Low storage - 2gb
- Visible from all nodes
- Regularly backed up

For	**Not for**
- Source code - Small compiled programs - Job scripts	- Intensive data read/write - Sharing contents

/projects

Login
Node ———— /projects/<user>

Examples:

- Small dataset of test positions
- Shared job scripts to be run

/projects

```
┌──────────────┐        ┌──────────────────────┐
│ Login        │────────│ /projects/<user>     │
│ Node         │        │                      │
└──────────────┘        └──────────────────────┘
```
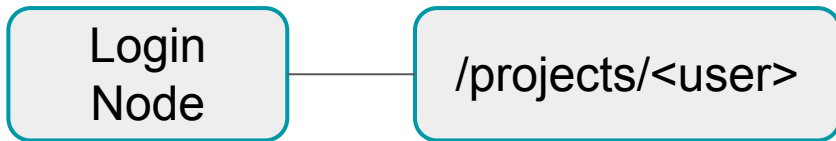
- Shared space
- Medium storage - 250gb
- Visible from all nodes
- Regularly backed up

For

● Store software builds
● Small data sets
● Sharing contents

Not for

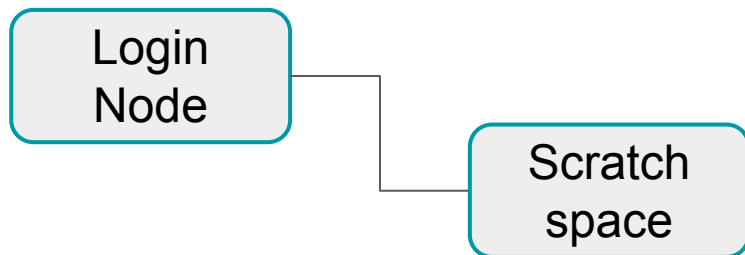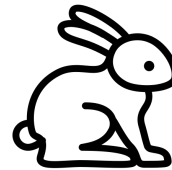| |
|---|
| ● Intensive data read/write |

Scratch Space

Login
Node

Scratch
space

Examples:

- Where her data will be written to throughout the job
- Summit: /scratch/summit/<user>
- Blanca: /rc_scratch/<user>

Scratch Space

Login Node ── Scratch space

- Temporary
- High storage - 10tb
- NOT backed up
- Purged when needed (max 90 days)

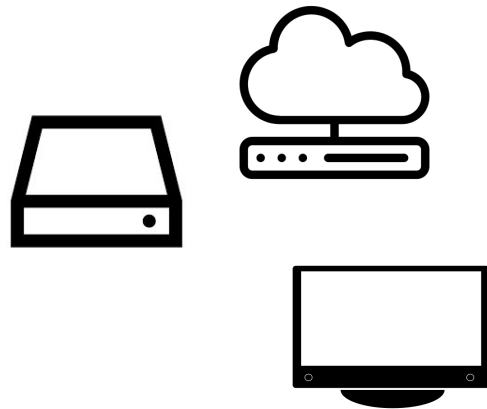| For	Not for
● Intensive data read/write on compute nodes	● Long-term storage

Long Term Storage

???

Data to analyze

Long Term Storage 🐢

Local Storage

- Your computer, hard drive, cloud storage, etc.

PetaLibrary

- High capacity storage for a fee
- NSF subsidized service
- Backed up & Non-Backed up services

Some examples!

- Job scripts

- User configs

- Small datasets

- Source code

- Intensive data writes

- Store data long-term

- /home - user space 🐢

- /projects - shared space 🐢

- Local storage 🐢

- PetaLibrary - storage for a fee 🐢

- Scratch Space - temporary 🐰

Wrap Up

Why HPC has specific storage needs

- Lots of data read/write
- Shared resources
- High performance dependent

What storage systems RC users have access to

- Permanent vs Temporary
- Network vs External

What each storage system should or shouldn't be used for

- Different storage/data transfer needs

Thank you

Questions?

Citations

Images

Content

Images	Content
• Wikipedia commons • vectorstock.com • medium.com • raspberrypi.org	1. https://www.colorado.edu/rc/resources 2. https://curc.readthedocs.io/ 3. https://researchcomputing.princeton.edu/support/knowledge-base/data-storage 4. https://www.weka.io/learn/hpc-storage-explained/ 5. https://searchstorage.techtarget.com/definition/parallel-file-system 6. https://www.usgs.gov/core-science-systems/sas/arc 7. Introduction to HPC: https://www.youtube.com/watch?v=bkLVuNfiCVs

Backup slides

High Performance Computing (HPC)

"...the practice of aggregating computing power in a way that delivers much higher performance than one could get out of a typical...computer" -USGS

HPC file systems (general)

- **Collaborative**
- **Scalable**
- **Data integrity and security**
- **Available**
- **Transfer**

Jargon/Definitions

Used	Definition
I/O	input/output, data transfer
<placeholder>	YOUR input
.file_name	A hidden directory
HPC	High Performance Computing