Command Line Workshop

2023

Overview

Introduce instructors

About the workshop

Why command line?

UTIA Computational Resources

Overview of workshop and goals

Instructors

Meg Staton

Ryan Kuster

Trinity Hamm

Trevor Freeman

Beant Kapoor

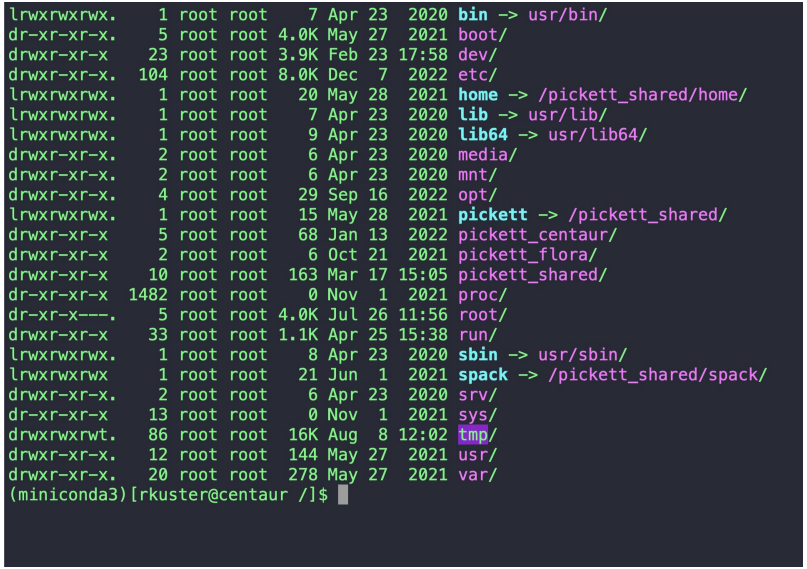
About the workshop

- Everyone is coming here with different levels of experience
- Ask questions!
- This wiki page will contain all the lab materials for the workshop:

  github.com/statonlab/CLI_workshop/wiki

Why command line?

- Used across all scientific fields
- Most tools are free and open-source
- Analyze data in a way that scales
- Perform and document repeatable science
- Work more closely with hardware

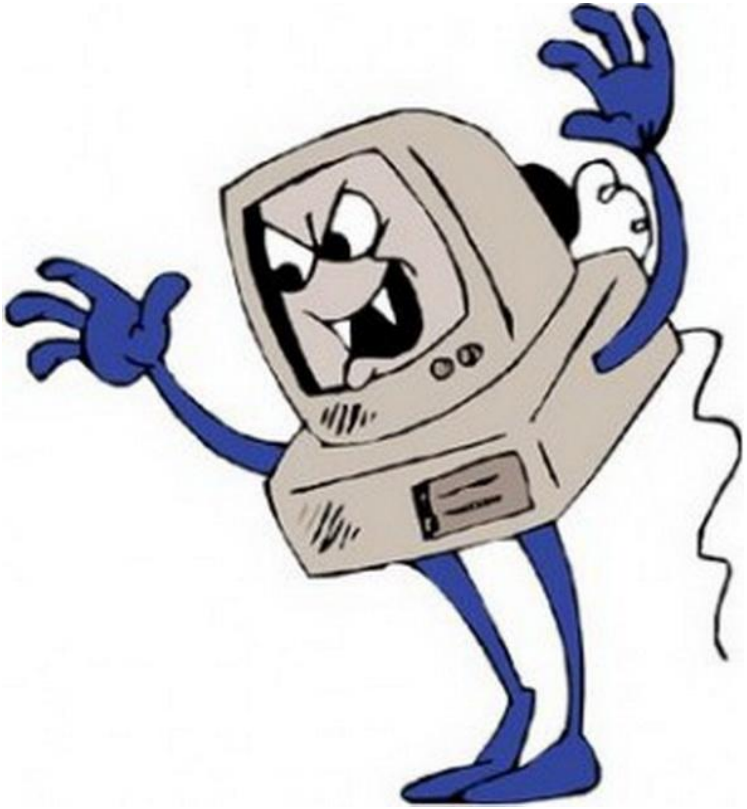


Why is command line so difficult?

- No mouse!
- Everything is text-based (few visuals)
- We're using a very specific language to interact with our machine
- Why else?

Why is command line so difficult?

- So much lingo!
  - Operating systems
  - Terminal
  - Shell
  - CPUs/cores/threads
  - RAM vs. Storage
- Intersection of many topics:
  - Computer hardware
  - Computer science
  - Programming



Linux vs Unix

- UNIX
  - Operating system developed in the 1970s at Bell Labs
  - Copyrighted name
  - Usually costs a lot of money (can be free for certain types of development)
  - Common for large corporate computers (servers and mainframes)

Linux vs Unix

Linux

- Linus Torvalds was frustrated that UNIX required a license
- So he wrote his own operating system from scratch that mimics UNIX
- released in 1991
- Free to use, open source
- One of the most prominent and important free, open source software projects

Linux Distributions

Because it is open, it has been ported and cloned, yielding many “flavors” or “distros”, all slightly different

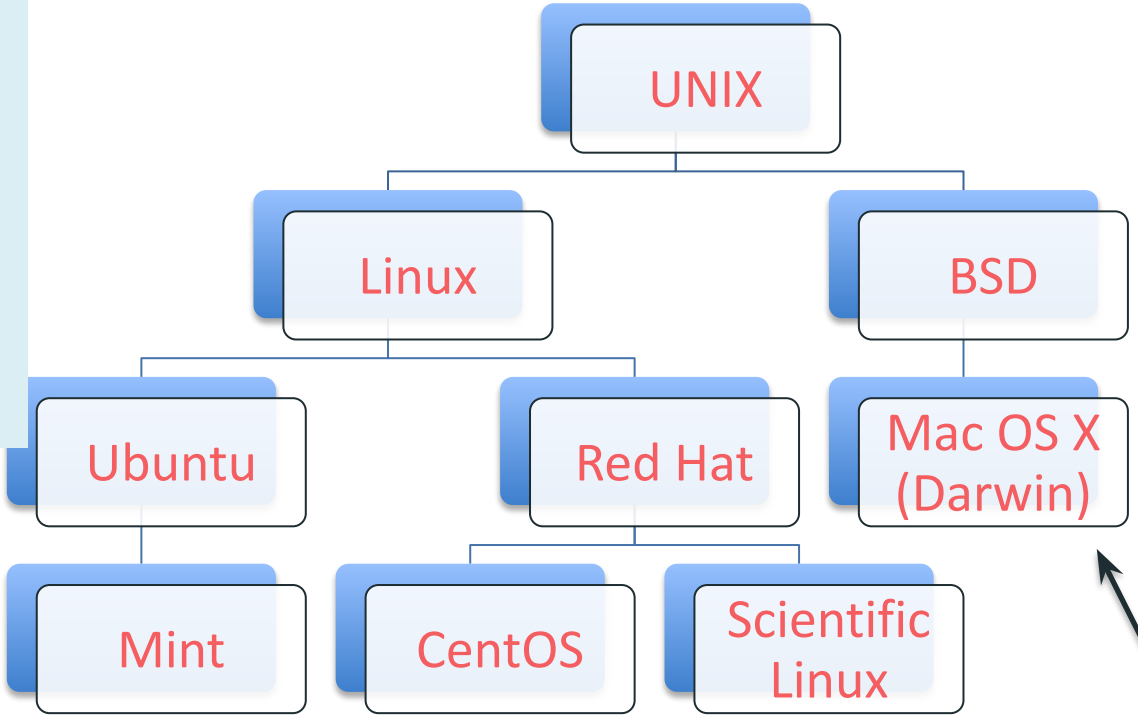
debian

ubuntu

FreeBSD

All in the same family.

Commands and software on one will (usually) work on the others.



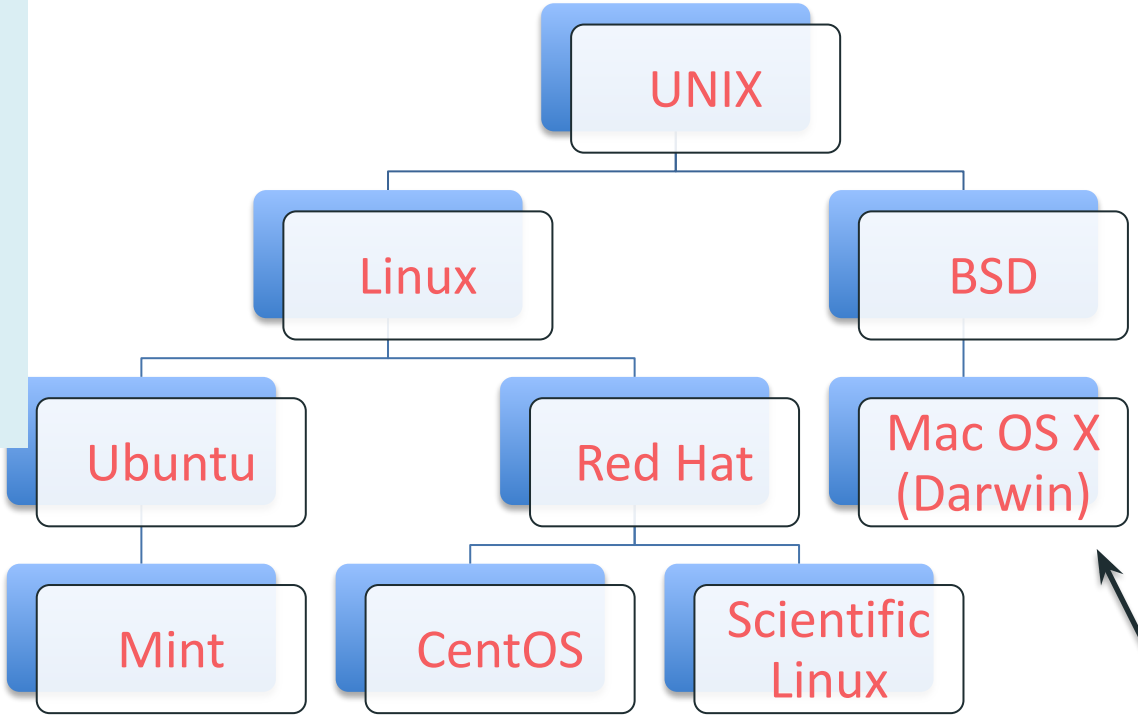
Most Linux personal or desktop computers

Most large computational servers including ISAAC

Apple Computers (UNIX-like core)

All in the same family.

Commands and software on one will (usually) work on the others.



Most Linux personal or desktop computers

Most large computational servers including ISAAC

Apple Computers (UNIX-like core)

Shell

- A **shell** is any user interface allowing access to the functions of an operating system
  - Command line (CLI)
  - Graphical (GUI)
- Most often “shell” or “terminal” refers to a CLI

Shell Variety – sh and bash

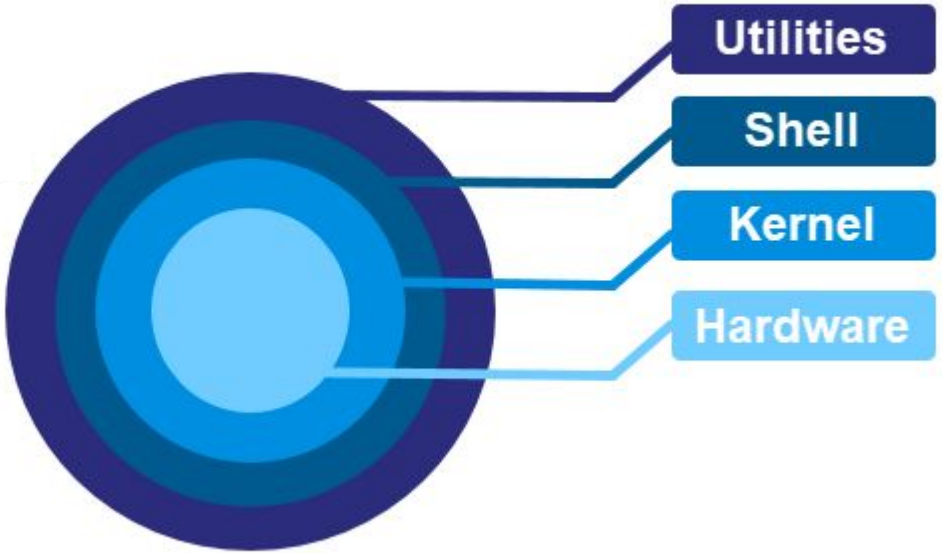
- sh or Bourne Shell
  - the original shell still used on UNIX systems and in UNIX-related environments
  - No longer standard shell, but still available on every Linux system for compatibility reasons.
- bash or Bourne Again SHell
  - the standard shell
  - a set of add-ons and plug-ins to the bourne shell
  - Good for beginners, loved by many pros
  - This is our shell
- Others: zsh or Z shell (Mac), csh or C shell, tcsh or turboC shell

Text editors

- Programming must be done with a text editor
  - Plain text, not Word
  - Nano is okay but not very feature-rich
- Recommended (and free):
  - vim (for the brave)
  - emacs (for the brave who want to argue with vim'mers)
  - Nano – for beginners

How does it all fit together?

- We're communicating with hardware with text via the shell
- The shell is our interface to a critical part of the OS called the "kernel"
- The kernel manages all of the hardware resources



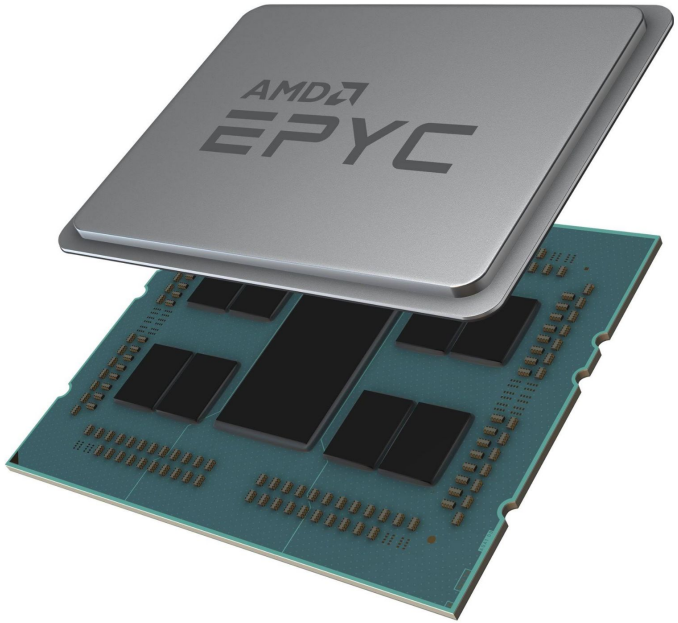
Central Processing Unit (CPU)

- The “brain”
- Executes instructions with electronic circuits
- “Clock Speed” is measured in gigahertz (GHz)
  - Billions of operations per second
  - Benchmarks are a better way of comparing processors
- Manufacturers include Intel and AMD

CPUs are 64-bit now (used to be 32-bit, this refers to the amount of data that can be operated on at one time)

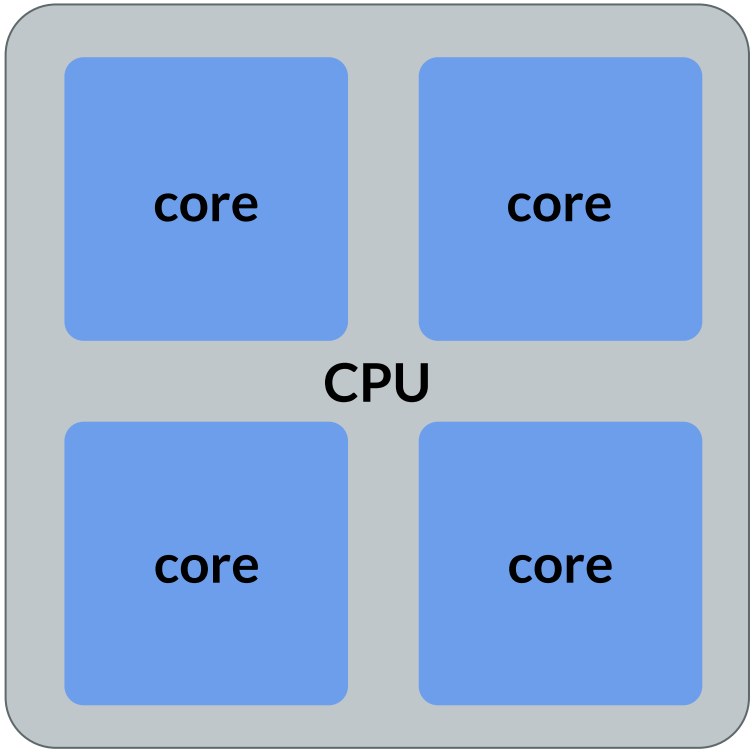
MultiCore Processors

- A processor used to contain only one set of circuits - now processors can contain many “cores”
- Cores are hardware, like the processor they are attached to, and in general can handle a single task
- Some cores are capable of working on parts of multiple tasks at the same time, which are called threads (a virtual division of labor)



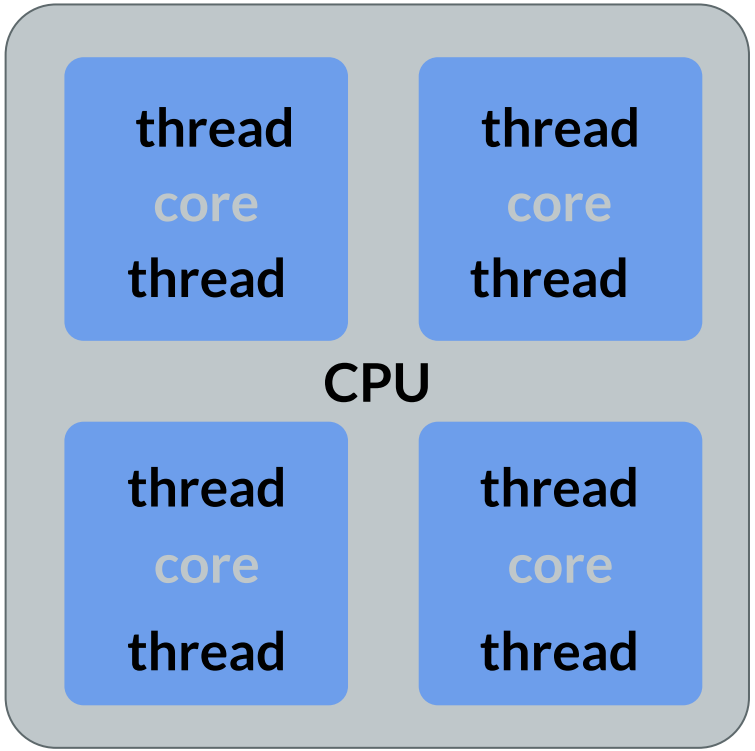
MultiCore Processors

- A processor used to contain only one set of circuits - now processors can contain many “cores”
- Cores are hardware, like the processor they are attached to, and in general can handle a single task
- Some cores are capable of working on parts of multiple tasks at the same time, which are called threads (a virtual division of labor)



MultiCore Processors

- A processor used to contain only one set of circuits - now processors can contain many “cores”
- Cores are hardware, like the processor they are attached to, and in general can handle a single task
- Some cores are capable of working on parts of multiple tasks at the same time, which are called threads (a virtual division of labor)



Memory: RAM vs. Storage

Imagine, instead of analyzing large data files, we're interested in building puzzles
When we're not using them, they sit, stored in memory. They take up space on disk, so think of this as physical storage (how many GB or TB of data am I storing for future analysis?)

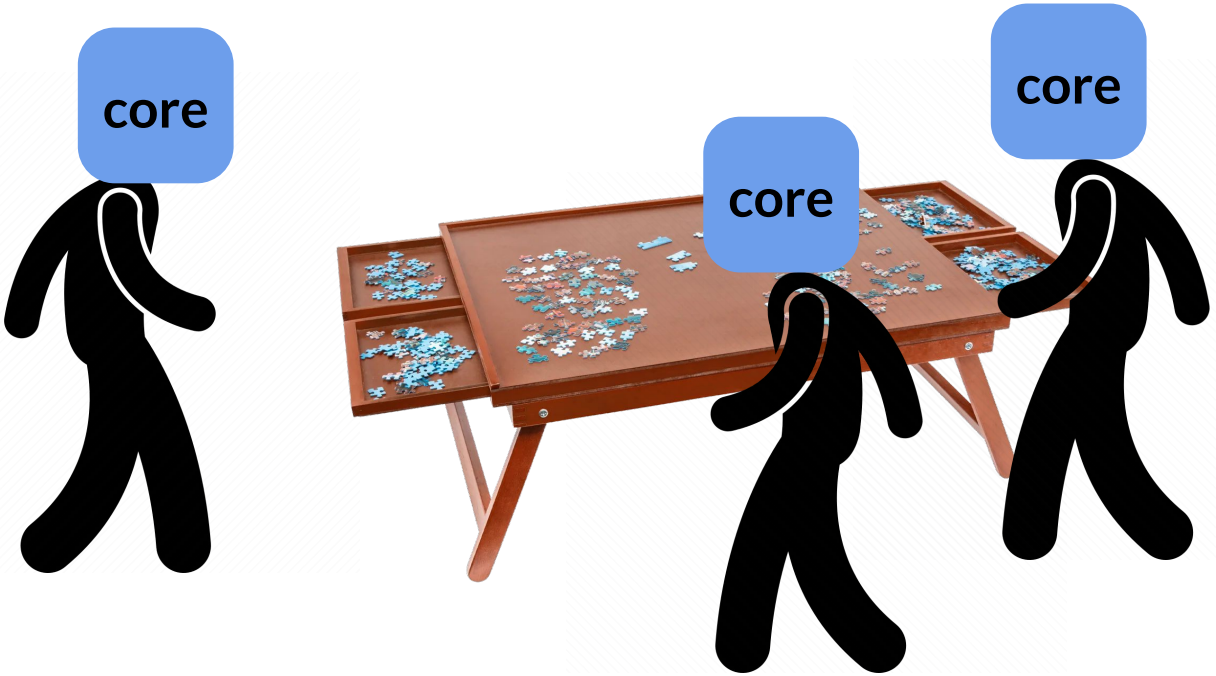
Memory: RAM vs. Storage

When we want to build a puzzle, we have to open the box and actually work on it somewhere. Analogous to the short-term, Random Access Memory (RAM), the size of the table is how much room we have to hold things in working memory.

Memory: RAM vs. Storage

We can dedicate a single core to start working on the puzzle…

Memory: RAM vs. Storage



And if we allow for parallel processing, multiple other cores/threads can assist with the job.

Memory: RAM vs. Storage

At some point, we might not have enough RAM (table space) to process large amounts of data (puzzle).

Scaling up inside one computer

Sphinx

- 64 logical cores (32 actual cores)
- threading creates multiple virtual cores out of each physical core (usually two threads per core).
- = 64 jobs at once
- 512 Gb of RAM
- Lives in the UTK server room with other computers

This is by not a particularly large server by research standards.

High performance computing (HPC)

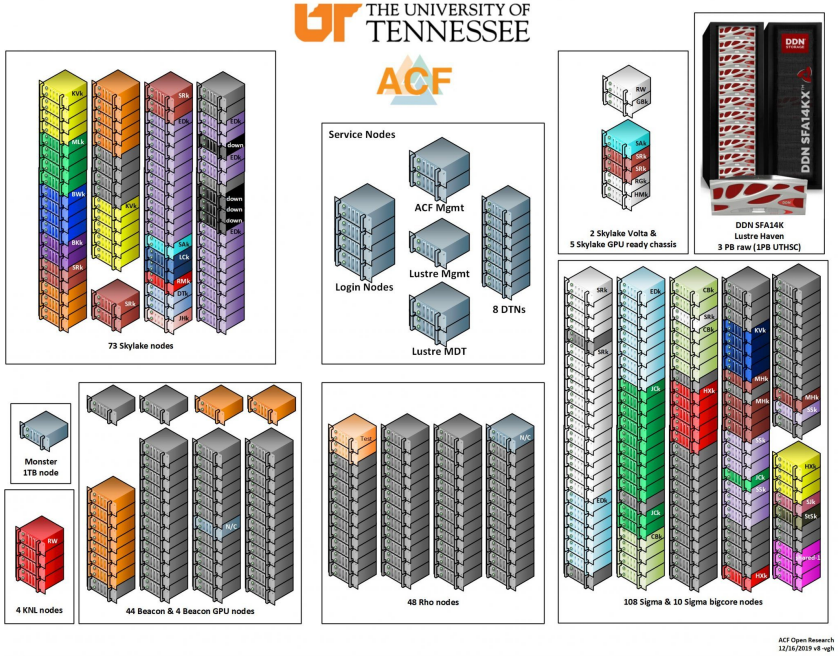
- Large data is often too much for your laptop to handle
- Move to a large computer or clusters of computers – ie “remote” computers
- called **servers**
- HPCs are also called supercomputers

Type the commands

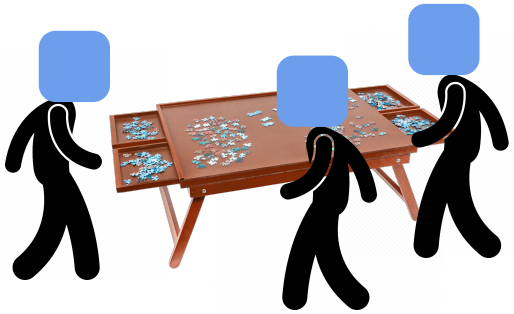
Actual work being done here

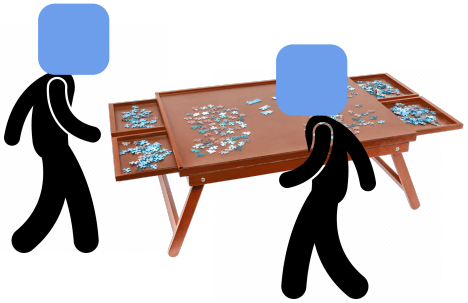
ISAAC-NG

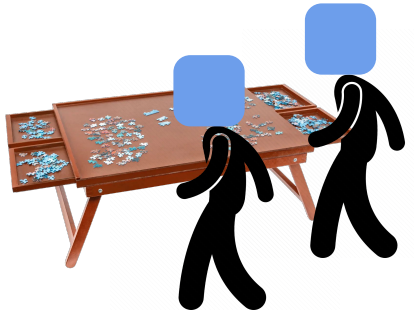
- https://oit.utk.edu/hpsc/isaac-open-enclave-new-kpb/
- Scientific Linux
- Any UTK affiliated researcher (student, faculty or staff) can get access
- Higher priority access for those who buy in
- 5,632 nodes across 115 nodes



Memory: RAM vs. Storage

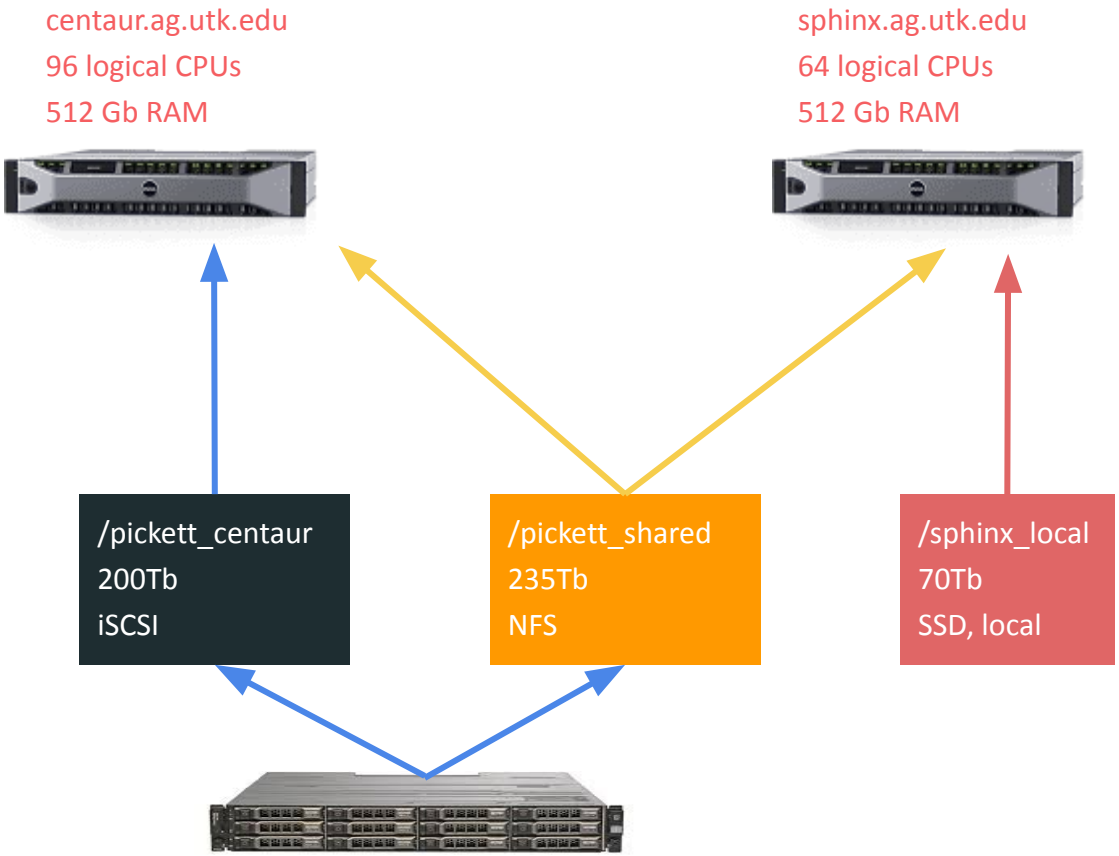






Distributed computing gets tricky.

Introduction - UTIA Computational Resources



Sphinx is faster than Centaur but has lower memory. Be judicious in using it and move all files out of Sphinx once you are done with them.

We reserve the right to remove unattended files from Centaur if they are taking up too much memory.

SSH

- Secure shell
- Network protocol – secure channel of communication between a client and a server
- This is the most common way to connect to a remote computer
- Its encrypted and its available on all UNIX/Linux systems
- To log into UTIA-CR, you should ssh as follows:

ssh yourusername@sphinx.ag.utk.edu

If off campus...

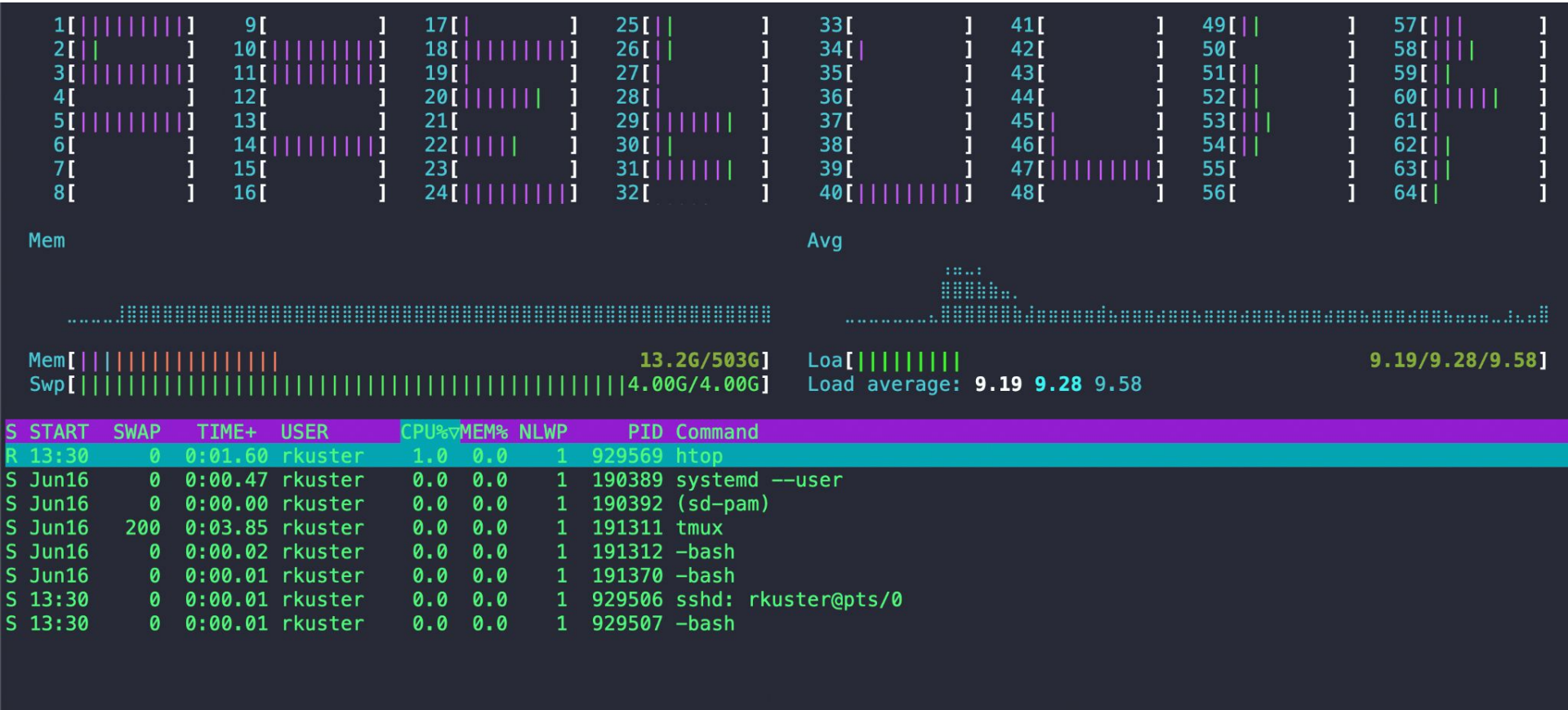
Use the VPN:

Install and activate VPN Software Pulse Secure: https://utk.teamdynamix.com/TDClient/2277/OIT-Portal/KB/ArticleDet?ID=123517

Sharing the Servers

- It can get a bit crowded!
- Commands to track server usage:
  - htop
  - ps -ef | grep <yourusername>
- There should always be at least 1 CPU free so the server doesn't crash and other users can login

Monitoring processes with htop



Storage Management

- Centaur has 200 Tb of storage, while Sphinx as 70 Tb. The shared space, available on both servers, has 235 Tb.
- Due to the nature of data that we work with, do not take these numbers for granted: storage can become sparse quickly.
- The best solution is to be proactive and **establish best practices to reduce excessive memory usage** so that we do not have to purge every file without your permission.

Storage Good Citizenship

- What to keep?
  - Raw data
    - Do not keep if it is publicly available. Instead add to the README about where to find it. (Save the fastq-dump command!)
  - Scripts
  - Final analysis outputs
- COMPRESS
    - .tar.gz is smaller than either alone
    - Sam -> BAM
    - Compress files or small folders, not whole project directories
- What to throw away?
  - Intermediate analysis files
  - Files from abandoned analysis

Example

How do I check the size of a file?

ls -lh

How do I check the size of a folder?

du –skh <folder>

How do I compress a file?

tar -cvzf file.tar.gz file.txt

How do I decompress a file?

tar -xvzf file.tar.gz

The UTIA Computational Resources Calendar

- After this class, by request, you may be added to the UTIA Computational Resources Calendar.
  - This offers an opportunity to book RAM/cores and let others know when you will be using the server for an intensive process.
  - Always update whenever you start a new process!
  - If a job finishes earlier than expected, also be sure to update so other users will know they can run their jobs now.

Why Command Line?

- Speed, Power, Modularity
- Can customize commands and run software with a few keystrokes
- Can send information between programs with pipes
- Can operate on hundreds to millions of files
- Can have many different jobs running simultaneously across many computers
- Easier to write code and release software – which means you can use the best software from open source researchers
- Modular workflows and components
  - We can experiment with different pieces of software at each stage of analysis (or substitute in our own!)
  - Reuse
  - Examine results at each stage of analysis

Day 1 workshop goals

Connect to the computing resources

Navigate a UNIX environment

Apply first commands

Download and investigate datasets

Manipulate/view files and directories

A few final notes

- Please ask questions!
- If you need any help we'll, put a sticky note on your laptop and someone will be right over to assist:
  - Red sticky = something is wrong/confusing
  - Green sticky = everything worked out