Introduction to ROAR Collab: Interactive Jupyter and R Studio Sessions

Nitheesha Nakka

C-SoDA Brownbag Spring 2024

PennState

Goals

- Introduce HPC
- Introduce ROAR Collab
- Jupyter Notebook and RStudio Interactive Session
  - Overview key terms and concepts
  - Load packages

Background

High Performance Computing (HPC) System

- Supercomputer
- Commonly used
  - eg. Businesses, banks, healthcare, other universities
- Physical hardware stored in high security facility on campus

ROAR Collab

- ROAR = Penn State's Supercomputer
  - High performance research cluster
  - RISE
- Uses
  - 1) Interactive Session- Coding "live"
  - 2) Batch Job- Submitting prewritten code and data

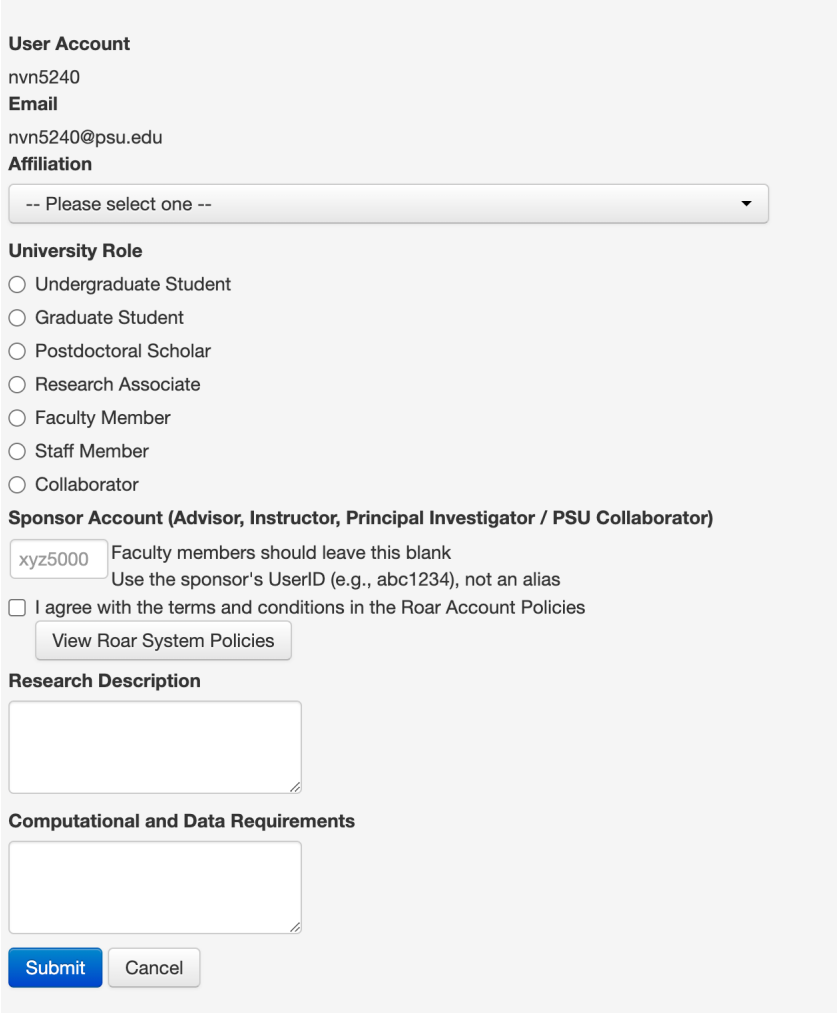
Your computer

ROAR

Getting Started: Account Setup

Account Setup

- Request an Account:
  - https://www.icds.psu.edu/account-setup/
- Submits a ticket to I-ASK
- All Non-faculty need a PI for a ROAR account
  - Advisor or Course Instructor



Getting Started: File Transfer Tools

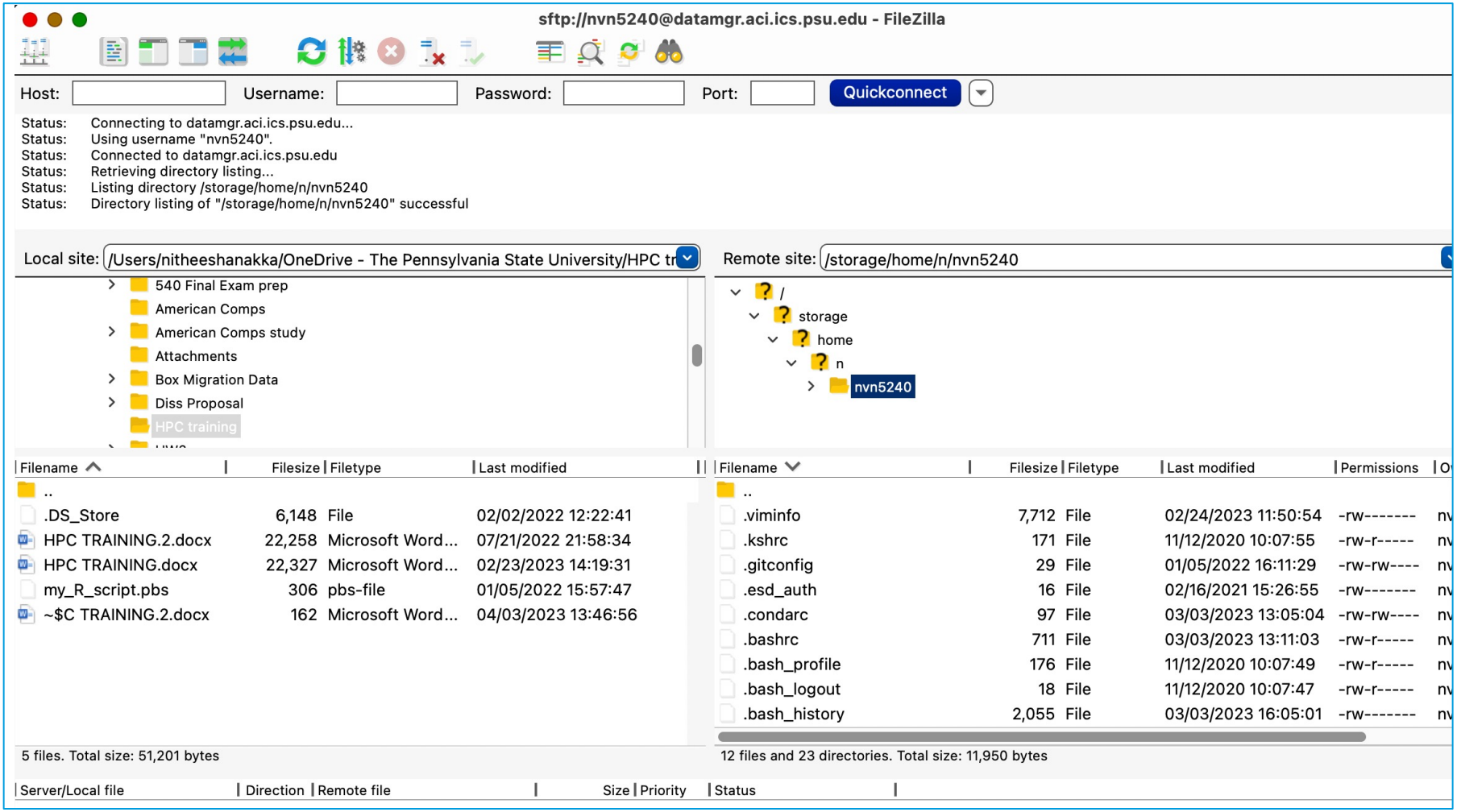
FileZilla *Client*-

Submit jobs and shift files *through* FileZilla

- Operates like a gateway
- Download here: https://filezilla-project.org/

- Other: command line, WinSCP Secure Protocol Client
  - Download here: https://winscp.net/eng/download.php
- Set-Up
  - username= nvn5240
  - password =
  - Port = 22
  - Host = submit.hpc.psu.edu
  - After doing the above 1 time you can click “quick connect” to go directly to home directory
  - 2-Factor authentication every login
- Drag & Drop
  - Local --> Remote
  - Can also delete files/ folders in filezilla

- Note: Whatever changes you make either in local or remote DOES NOT automatically make those changes on the “opposite” computer


sftp://nvn5240@datamgr.aci.ics.psu.edu - FileZilla
Host:
Username:
Password:
Port:
Quickconnect
Status: Connecting to datamgr.aci.ics.psu.edu...
Status: Using username "nvn5240".
Status: Connected to datamgr.aci.ics.psu.edu
Status: Retrieving directory listing...
Status: Listing directory /storage/home/n/nvn5240
Status: Directory listing of "/storage/home/n/nvn5240" successful
Local site: /Users/nitheeshanakka/OneDrive - The Pennsylvania State University/HPC tr
Remote site: /storage/home/n/nvn5240
540 Final Exam prep
American Comps
American Comps study
Attachments
Box Migration Data
Diss Proposal
HPC training
/
storage
home
n
nvn5240
Filename
Filesize
Filetype
Last modified
.DS_Store 6,148 File 02/02/2022 12:22:41
HPC TRAINING.2.docx 22,258 Microsoft Word... 07/21/2022 21:58:34
HPC TRAINING.docx 22,327 Microsoft Word... 02/23/2023 14:19:31
my_R_script.pbs 306 pbs-file 01/05/2022 15:57:47
~$C TRAINING.2.docx 162 Microsoft Word... 04/03/2023 13:46:56
5 files. Total size: 51,201 bytes
Permissions
.viminfo 7,712 File 02/24/2023 11:50:54 -rw-------
.kshrc 171 File 11/12/2020 10:07:55 -rw-r-----
.gitconfig 29 File 01/05/2022 16:11:29 -rw-rw----
.esd_auth 16 File 02/16/2021 15:26:55 -rw-------
.condarc 97 File 03/03/2023 13:05:04 -rw-rw----
.bashrc 711 File 03/03/2023 13:11:03 -rw-r-----
.bash_profile 176 File 11/12/2020 10:07:49 -rw-r-----
.bash_logout 18 File 11/12/2020 10:07:47 -rw-r-----
.bash_history 2,055 File 03/03/2023 16:05:01 -rw-------
12 files and 23 directories. Total size: 11,950 bytes
Server/Local file
Direction
Remote file
Size
Priority
Status

Accessing ROAR Collab

- Online Login

  https://www.icds.psu.edu/access-roar-and-roar-collab-online/

  → https://rcportal.hpc.psu.edu

  → HPC Dashboard

Interactive Jupyter Notebook

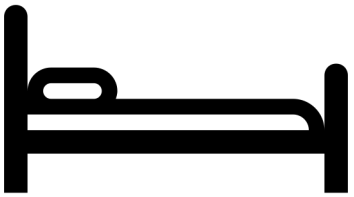
PennState

Resource Request

- Jupyter Interface:
  - Jupyterlab
  - Jupyter Notebook
- Conda Environment Type
- Conda Environement from RISE
- Account
- Partition
- Node Type
- Number of Cores
- Memory per core (in GB)
- Number of Hours

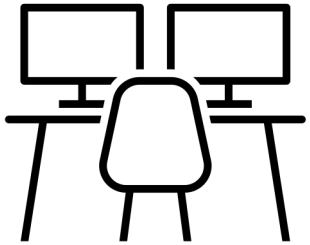
Resource Request: Environments

- Jupyter Interface
- Conda Environment Type
  - custom
  - Predefined environments
- Conda Environment from RISE
  - pytorch
  - Tensorflow
  - Data-science collection
  - Deep-learning



What Are Environments?

- Folder with all the required:
  - packages,
  - software,
  - dependencies,
  - version control

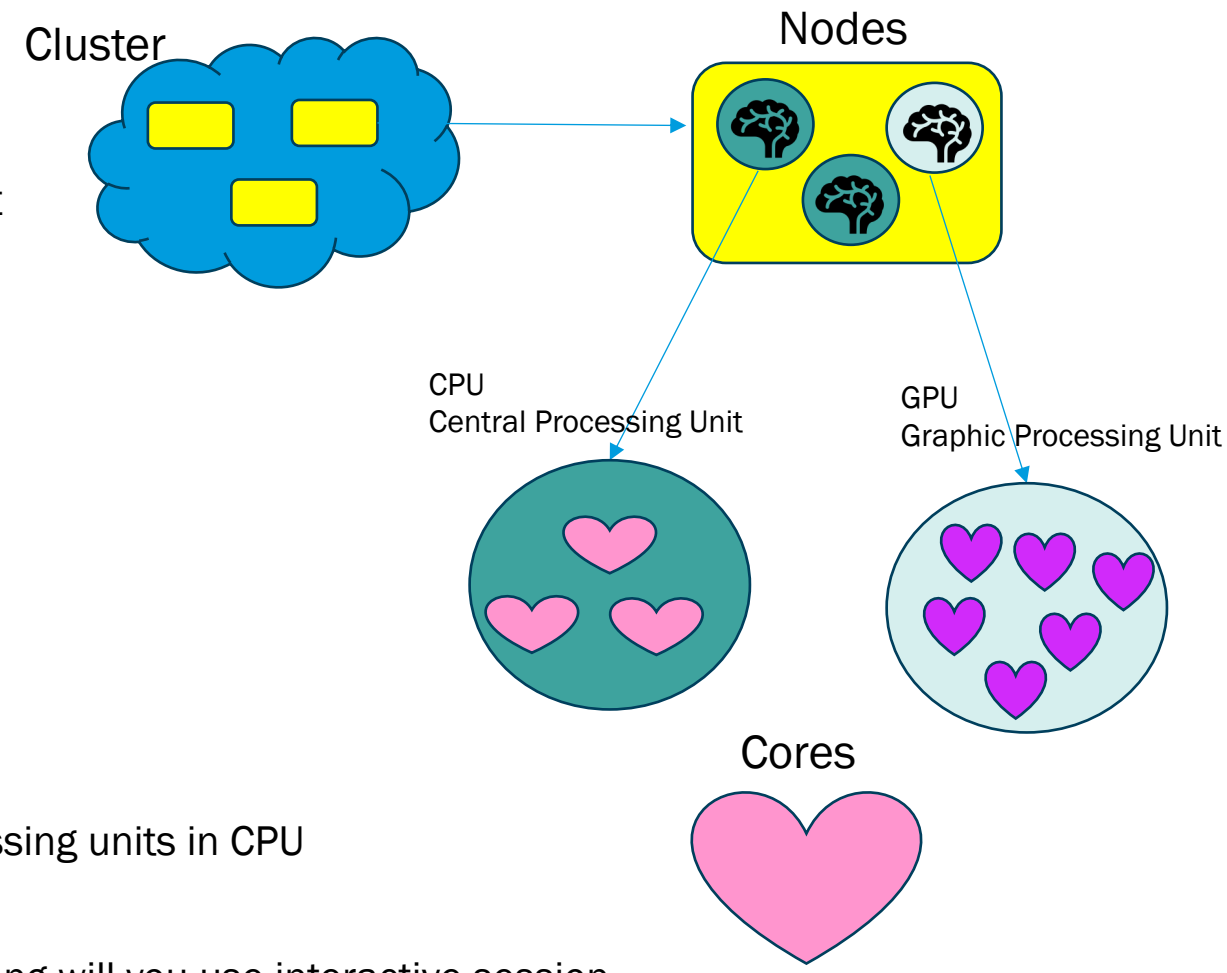


Resource Request: Accounts and Partition

- Jupyter Interface
- Conda Environment Type
- Conda Environment from RISE
- Account: What allocation you will be using
  - Open
    - 100 cores across 20 nodes
    - 48 hours job time
  - Paid Allocation
- Partition: How to allocate resources
  - Open
    - No guaranteed start time
  - Open-suspend
  - Interactive
  - SLA-Prio
    - 1 hour guaranteed start time
  - Burst mode
  - Burst Suspend
    - No guaranteed start time

Resource Request: Nodes and Cores

- Jupyter Interface
- Conda Environment Type
- Conda Environement from RISE
- Account
  - Open
  - paid
- Partition
  - Open
  - Sla-prio
- **Node Type:** machine
  - Standard Cores
  - Basic Cores
  - GPU Cores
  - Interactive Cores
  - Highmem Cores
- **Number of Cores:** processing units in CPU
- **Memory per core (GB)**
- **Number of Hours:** how long will you use interactive session



Resource Request: Launch!

- Jupyter Interface: Notebook
- Conda Environment Type: predefined
- Conda Environement from RISE: pytorch
- Account: open
- Partition: open
- Node Type: standard
- Number of Cores: 1
- Memory per core (in GB): 4
- Number of Hours: 1

LAUNCH!

  - More resources = longer wait
  - Job ID

R Studio Dashboard

PennState

Resource Request:

- Environment Type:
  - Predefined
  - Custom: load the version of R that you want
    - To see all versions type “spider r” in Command line (terminal)
- Select R Version from RISE: v4.2.1 (default)
- Account: open
- Partition: open
- Node Type: standard
- Number of Cores: 1
- Memory per core (in GB): 4
- Number of Hours: 1

LAUNCH!

Extra

PennState

Installing Packages

A) Install in Jupyter Notebook or Rstudio Session

B) Install in session in terminal/ command line

A) salloc -N 1 -n 4 --mem-per-cpu=1024 -t 3:00:00
- Opens compute note

B) Launch R or Python

C) install package

C) Conda Installer

1) salloc -N 1 -n 4 --mem-per-cpu=1024 -t 3:00:00
   - cat .bashrc : checks if there is anything in bash RC
2) module load anaconda
3) conda create -n name_of_environment
4) conda activate name_of_environment
5) conda install -c conda-forge r-tidyverse
   - https://anaconda.org/conda-forge/r-tidyverse
6) conda deactivate

Installing Packages

A) Install in Jupyter Notebook or Rstudio Session

B) Install in session in terminal/ command line

A) salloc -N 1 -n 4 --mem-per-cpu=1024 -t 3:00:00
   - Opens compute note

B) module load R or Python

C) install package

C) Conda Installer

1) salloc -N 1 -n 4 --mem-per-cpu=1024 -t 3:00:00
   - cat .bashrc : checks if there is anything in bash RC
2) module load anaconda
3) ~~conda create -n name_of_environment~~
4) conda activate name_of_environment
5) conda install -c conda-forge r-tidyverse
   - https://anaconda.org/conda-forge/r-tidyverse
6) conda deactivate

When to Use ROAR Collab

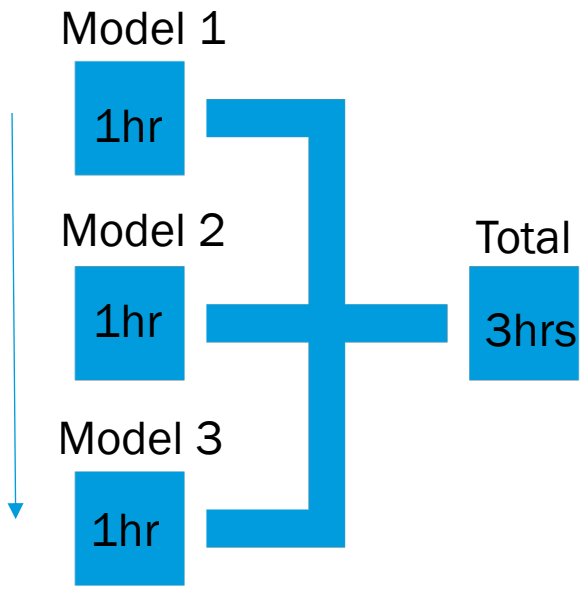
Large data

Computing power

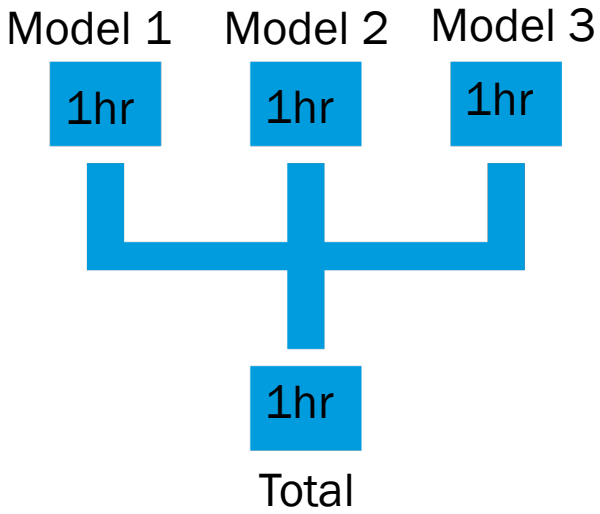
Parallelization

When to Use ROAR Collab: Parallelization

Serial Computing: 1 core



Parallelization: 3 cores



Resources

ROAR Collab Intro:
https://www.icds.psu.edu/roar-collab/
ROAR User Guide
https://www.icds.psu.edu/roar-collab-user-guide/
ICDS Guidelines
https://www.icds.psu.edu/wp-content/uploads/2020/04/ICDS-ACI-P030-Access-Control-Formatted_v3.1.pdf
I-Ask Help Desk Virtual Office hours

Parallelization:
doParallel R package:
https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstartedParallel.pdf
foreach():
https://cran.r-project.org/web/packages/foreach/vignettes/foreach.html
Ipythonparallel:
https://www.anyscale.com/blog/parallelizing-python-code

Myself
Email: nvn5240@psu.edu
Github: https://github.com/NitheeshaN

Thank you!

PennState