Introduction to the RCC for Scholars in the Social Sciences and the Humanities

October 13, 2020

* Slides available at https://github.com/luetgert/IntroRCC

RCC Resources

What is High-Performance Computing (HPC)?

"High-Performance Computing (HPC) is the application of particularly powerful computational infrastructure for computational problems that are either too large for standard computers or would take too long.

A desktop computer generally has a single processing chip, commonly called a CPU. A HPC system, on the other hand, is essentially a network of nodes, each of which contains one or more processing chips, as well as its own memory and large disk arrays."

(Adapted from https://www.nics.tennessee.edu/computing-resources/what-is-hpc)

RCC Resources

How is HPC useful for SSD/HD researchers?

- HPC is designed for researchers who want to engage in computational tasks too big or unwieldy for a laptop or desktop to run (often due to resources, or length of time), or that require special hardware configurations (such as high-end GPUs or secure data storage).

  You can think of the HPC cluster as a second (very large) computer with an environment designed for running large jobs, writing code and/or collaborating with others (sharing data, code and methods).

HPC for Social Scientists: Examples

Mapping the 1919 Chicago Riots (John Clegg)

- https://1919map.rcc.uchicago.edu/

Million Neighborhoods: Mapping Rapid Urbanization in Asia and Africa (Mansueto Institute)

- https://millionneighborhoods.org/#2/8.84/17.54

Others:

- **Addressing the Social Cost of Carbon (Lars Hansen)**
- **Business of Debt (Destin Jenkins)**

HPC for the Humanities: Examples

1. OCR, Image Analysis and Web Scraping pipelines

2. NLP, Computational linguistics, Computational phonology (detecting patterns in language at scale)

3. Textual Optics' Text-Pair Viewer (in development) (networks of thousands of textual parallels/text reuse)

- https://users.rcc.uchicago.edu/~jcarlsen/TPV_test/TPV_small_panelinfo/

4. NNBlake: Poetry generated by a Neural Network trained on the works of William Blake (GPU-based GPT-2)

- https://github.com/rcc-uchicago/BERT-GPT2_tutorial_Summer2020/blob/master/sample%20output/blake_gpt2_generated_poems_no_params.txt

RCC Resources

Are there RCC staff available to help me use HPC?

- Yes! ☺️
- Dr. Brooke Luetgert is the University of Chicago "Computational Scientist for the Social Sciences" luetgert@uchicago.edu
- Dr. Jeffrey Tharsen is the University of Chicago "Computational Scientist for the Humanities" tharsen@uchicago.edu

We also provide general assistance over email or zoom: help@rcc.uchicago.edu

RCC Resources

Ok, so I want to use HPC – what is "Midway"?

- Midway is the name of the university's "High-Performance Computing" (HPC) cluster. **Midway is a shared platform and customizable environment for computational work, including high-end computing and high-performance storage, open to all users who register.**

- The Midway ecosystem is currently comprised of one large main cluster (roughly 30,000 CPUs, many terabytes of RAM and about 15 petabytes of disk space at present) and several other clusters:
  MidwayR for secure data;
  DaLI for large data processing and storage;
  GPU-enabled workspaces;
  plus a number of smaller partitions for specific research groups.

Midway Ecosystem

Not restricted

Protection Level

Midway

Open data

Sensitive

MidwayR2

Moderate

MidwayR3

VM

High

Getting an RCC account

2 Types of accounts:

- Principal Investigator (PI) account
- General user account

Eligibility

- PI = Any UChicago faculty or staff member with PI status
- General User = Any researcher working with or under a UChicago PI

Getting an RCC account (Cont'd)

- To get a PI account on Midway, a PI should submit a request for an RCC **PI Account** via

  https://rcc.uchicago.edu/accounts-allocations/pi-account-request

  The PI will receive a notification email upon approval of the request.

- Once the PI account is created, students, researchers and collaborators can apply for an RCC **General User Account** under the PI. https://rcc.uchicago.edu/accounts-allocations/request-account

Accessing Midway

Once I have an account, how do I access Midway? Is it free?

- Once you have an account, you can use **ThinLinc** or **ssh** to log in to the cluster, where you can run small jobs on the login nodes and request resources from the larger pool of comput nodes via the "**sinteractive**" command or the **SLURM** job scheduler.

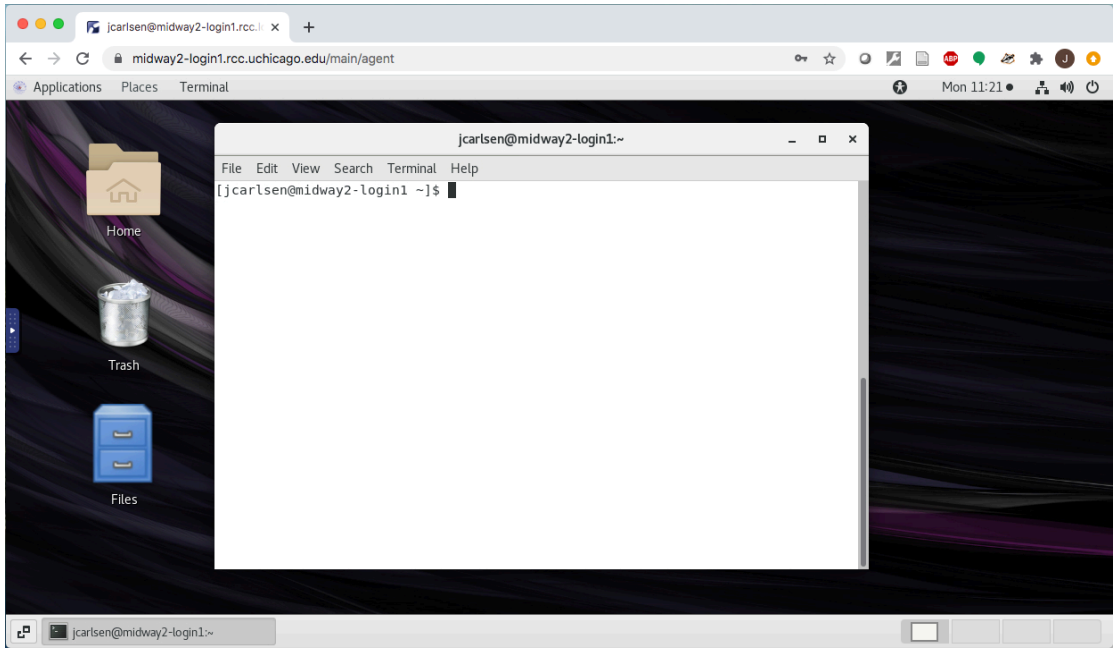
  ThinLinc: https://midway2.rcc.uchicago.edu/

- **Use of the cluster is free of charge**, though we do need to charge PIs for large amounts of disk space, specifically for each 1 TB above the initial free allocation of:
  30GB for each General User in /home and
  500GB of shared group space in /project2 for each PI.

Accessing Midway

What OS runs on Midway? How do I start and use applications on the cluster?

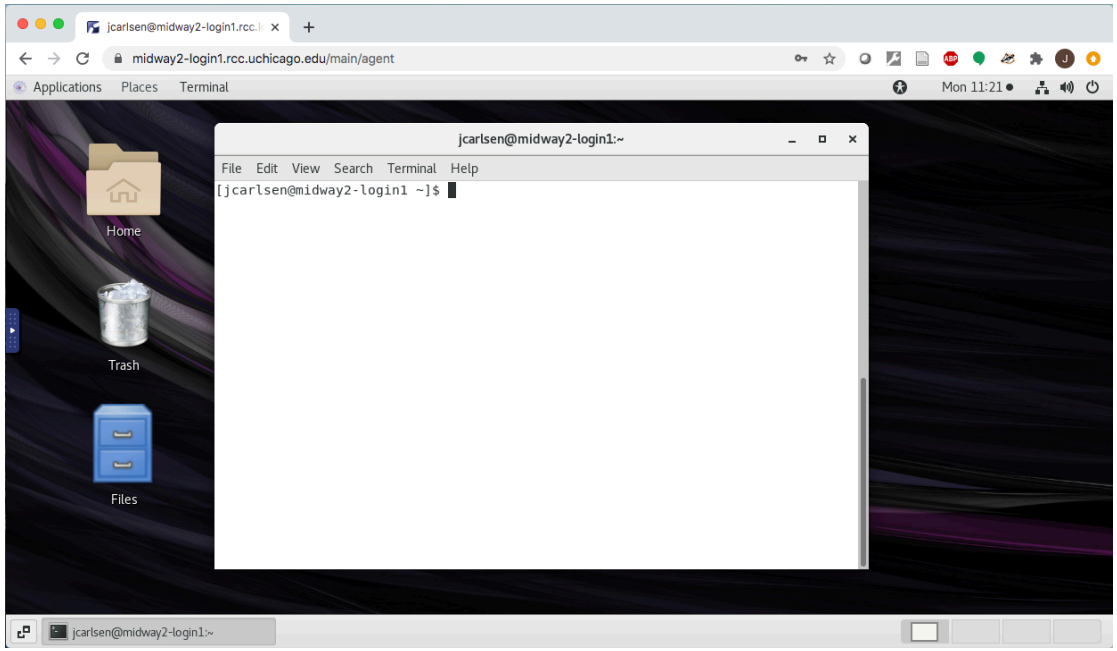
- Midway runs Scientific Linux 7.2, and ThinLinc provides a bare-bones "virtual desktop" graphic interface:



Accessing Midway

What OS runs on Midway? How do I start and use applications on the cluster?

- The **Terminal** (Linux command line) is the main way most users interact with the system.



Finding and using software on Midway

To see the available list of modules you can load

$ module avail [name]

To see your list of currently loaded modules

$ module list

To load and run the **RStudio** module

$ module load rstudio (loads the default module: R 3.6.1)

$ rstudio

To load and run the **Jupyter Notebook (Python)** module

$ module load python (loads the default Anaconda and Python 3.7.6)

$ jupyter notebook

(will open in a Firefox browser)

Computing on Midway

Login nodes

Should only be used for compiling your code, submitting jobs, transferring files, and short test runs

Interactive sessions

To be used for using software with a graphic interface

Batch jobs (the SLURM scheduler)

The primary way we recommend users use Midway

You need to create a job submission script

Interactive sessions

To request an interactive session with 4 cores on one node
(each node on the main partition has 28 cores) and 8GB of total memory
for 12 hours:

```
$ sinteractive --time=12:00:00 --nodes=1 --ntasks-per-node=4 --mem=8G
```

To request 12 hours on a GPU node with 1 GPU and 24G memory:

```
$ sinteractive --time=12:00:00 --partition=gpu2 --gres=gpu:1 –mem-per-cpu=24G
```

The **RCC User Guide** contains a great deal of helpful information for connecting to Midway, moving data, and setting up and running jobs:
https://rcc.uchicago.edu/docs/

Best Practices and Useful Commands

- How do I check my disk storage allocations?
  $ quota

  How do I check my PI's available service units?
  $ accounts balance

- How do I upload data & download my research results?
  SFTP (e.g. FileZilla), or the "scp" command

- How to monitor your usage and your account
  $ myq (while a job or sinteractive session is pending or running)
  $ sacct (when a job has finished, to see the exit status and stats)

- How to schedule jobs using SLURM (SBATCH syntax):
  https://rcc.uchicago.edu/docs/using-midway/index.html#batch-jobs

SDE- Secure Data Enclave

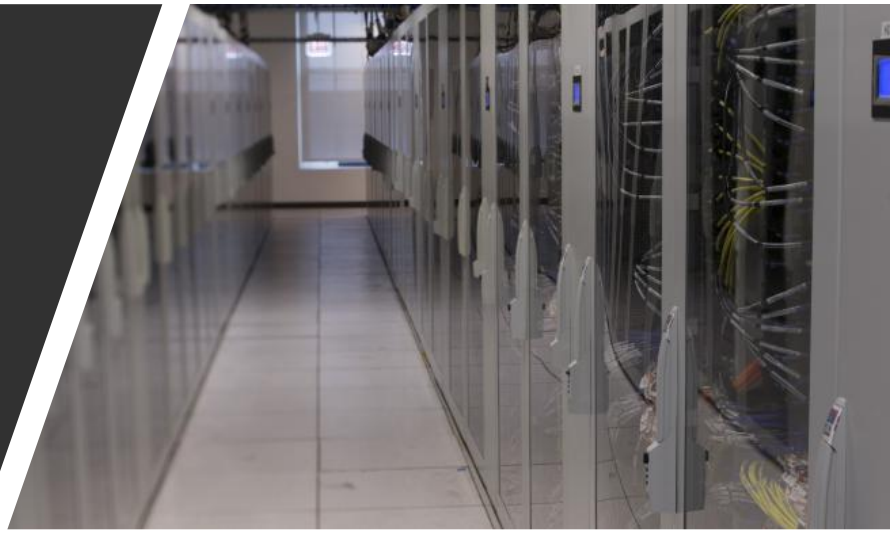
Your data will dictate the HPC compute environment that you require. When you apply to University Research Administration (URA) for a Data Use Agreement (DUA) with a third-party provider or to the Internal Research Board (IRB) for data that you are generating, data that contain personal identifying information (names, addresses, financial data, medical information, etc.) may require restricted access, encryption or even offline only storage. MidwayR is a small cluster designed to meet these additional requirements and provide you with the flexibility and exceptional performance comparable to Midway.

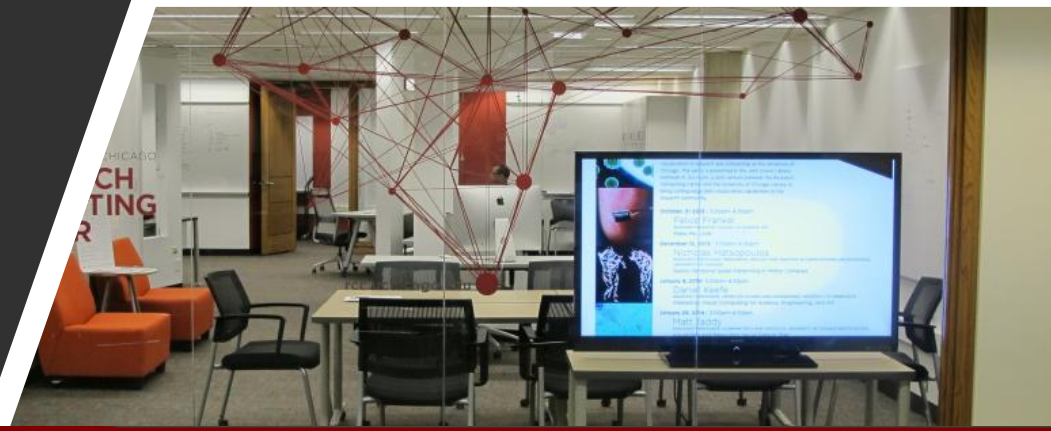
Where are we located?

The RCC is a unit in the Office of Research and National Laboratories

Regenstein Room 216:
Walk-in lab and Help Desk
help@rcc.uchicago.edu

Central office:5607 S. Drexel Ave
Data center: 6045 S. Kenwood Ave





Resources

Slides and brief notes from today are available at https://github.com/luetgert/IntroRCC