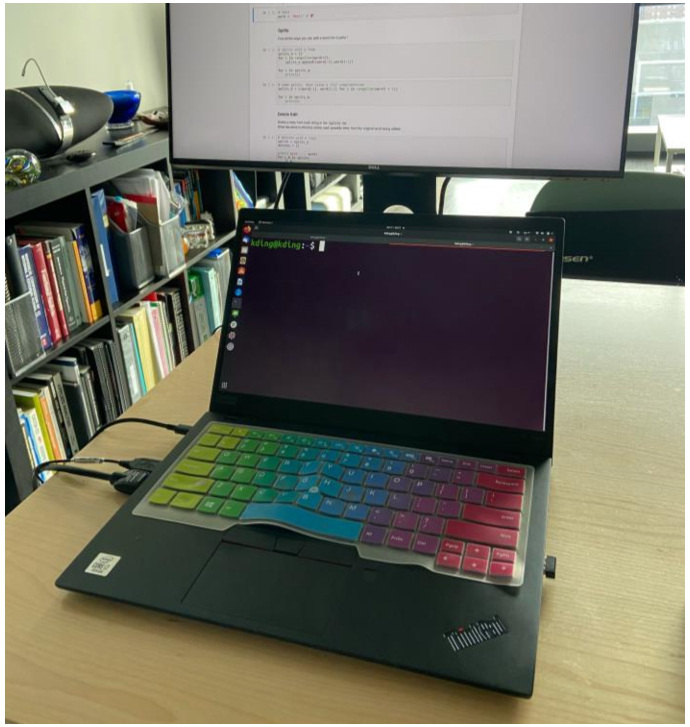
Slurm Workshop

Introduction to High Performance Computing (HPC)

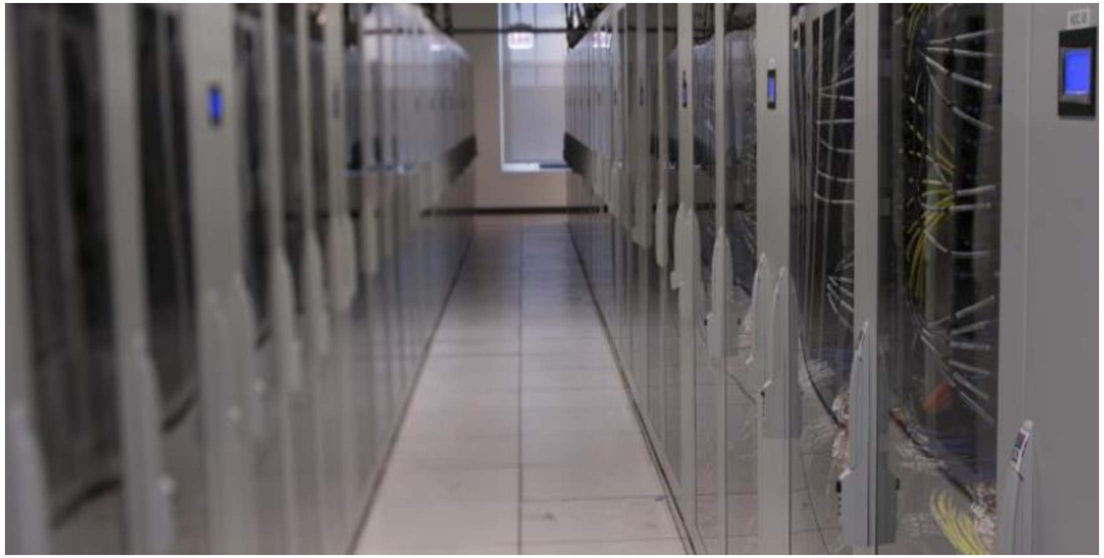
Kaihua Ding, Ph.D.

Slurm skill self-assessment

PC and HPC



```
kding@kding:~$
```



Need large amount of data

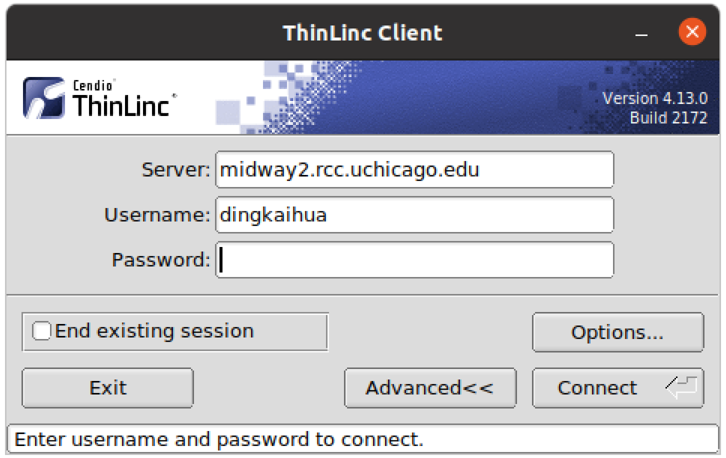
Solve my problems faster

Why HPC?

To use GPU

PC cannot run my program

Connecting to HPC – Tutorial 0



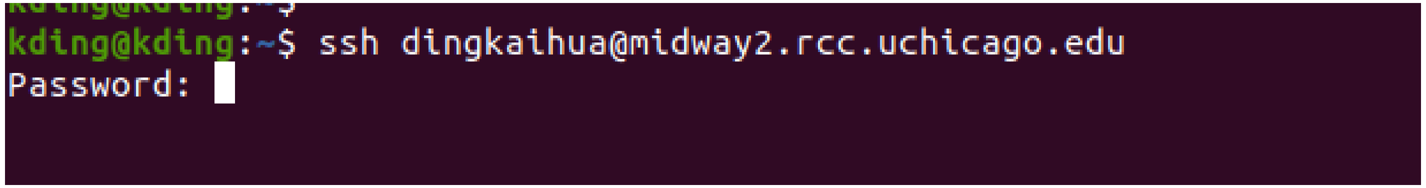
Server: midway2.rcc.uchicago.edu

Username: dingkaihua

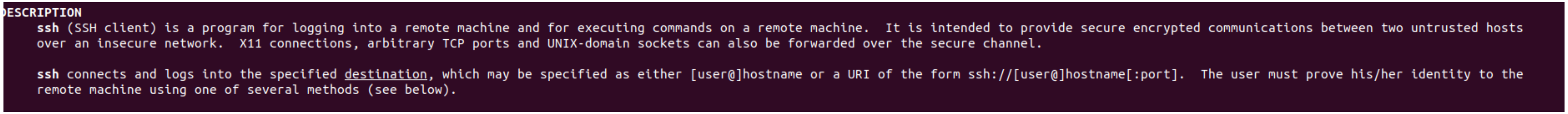
Password:

☐ End existing session

Enter username and password to connect.



```
kding@kding:~$
kding@kding:~$ ssh dingkaihua@midway2.rcc.uchicago.edu
Password: 
```



```
DESCRIPTION
     ssh (SSH client) is a program for logging into a remote machine and for executing commands on a remote machine.  It is intended to provide secure encrypted communications between two untrusted hosts
     over an insecure network.  X11 connections, arbitrary TCP ports and UNIX-domain sockets can also be forwarded over the secure channel.

     ssh connects and logs into the specified destination, which may be specified as either [user@]hostname or a URI of the form ssh://[user@]hostname[:port].  The user must prove his/her identity to the
     remote machine using one of several methods (see below).
```

HPC can only be used through the internet!

HPC storage and data transfer?

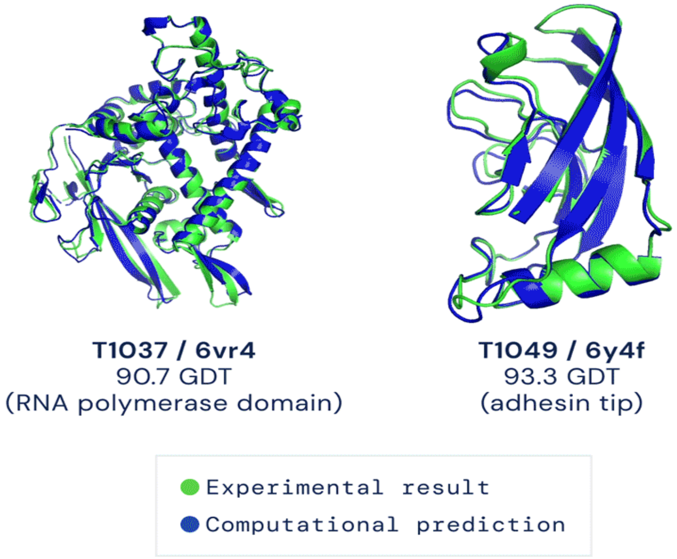
- File system -- since we are sharing the cluster, we each need our separate copies of storage!
- Data mapping & viewing:
  - http – hypertext transfer protocol
  - SAMBA -- re-implementation of the SMB networking protocol
- Data transfer
  - scp -- Secure copy protocol
  - Globus – GridFTP, file transfer protocol

Using internet protocols to manage data transfer and data

https://rcc.uchicago.edu/docs/data-transfer/index.html#http-web-access

Software on HPC

231GB - Call of Duty: Modern Warfare



2.2TB - AlphaFold

Huge software needs to be managed differently.

⇩

Environment modules (http://modules.sourceforge.net/)

Benefit of HPC

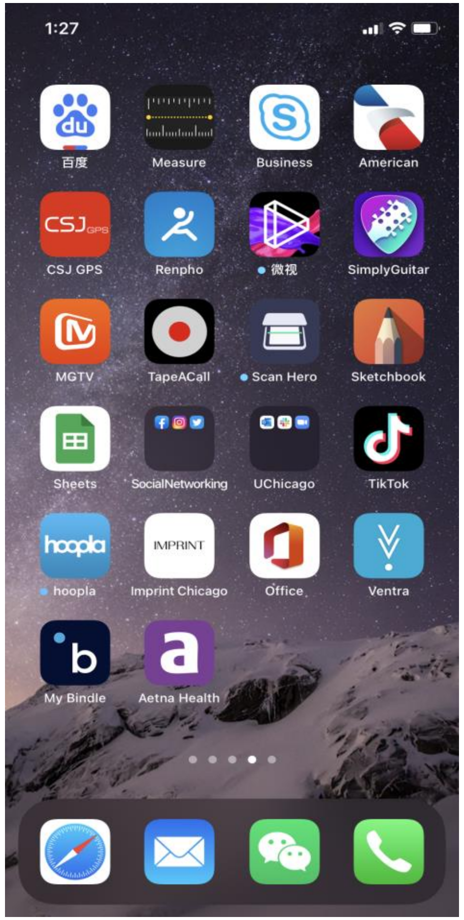
Benefit 1: software that require large memory and large storage will be able to run on HPC.

Benefit 2: **paralleled** software will run much faster on HPC than on PC

Supercomputer → More than one computer

Serial software will not magically run faster on supercomputer. The programmer need to parallel them.

HPC computation



Just click and run

- What if 1000 users are clicking on the same app?
- What if 1000 AlphaFold computations are requested at the same time?
- Whose job should start first?

HPC job scheduler – a simplified toy problem

Setup:

- One shared resource (e.g., a processor).
- Many "jobs" to do (e.g., processes).

Question: In what order should we sequence the jobs?

Assume: Each job has a:

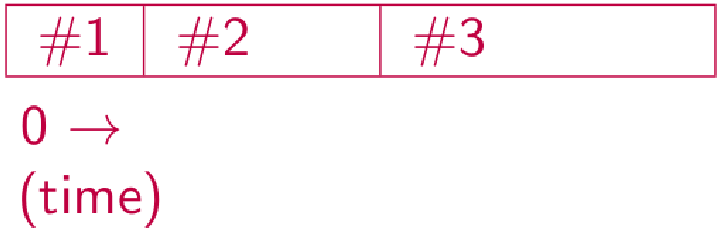
- weight w_j ("priority")
- length l_J

HPC job scheduler – a simplified toy problem

Definition: The completion time C_j of job j = Sum of job lengths up to and including j.

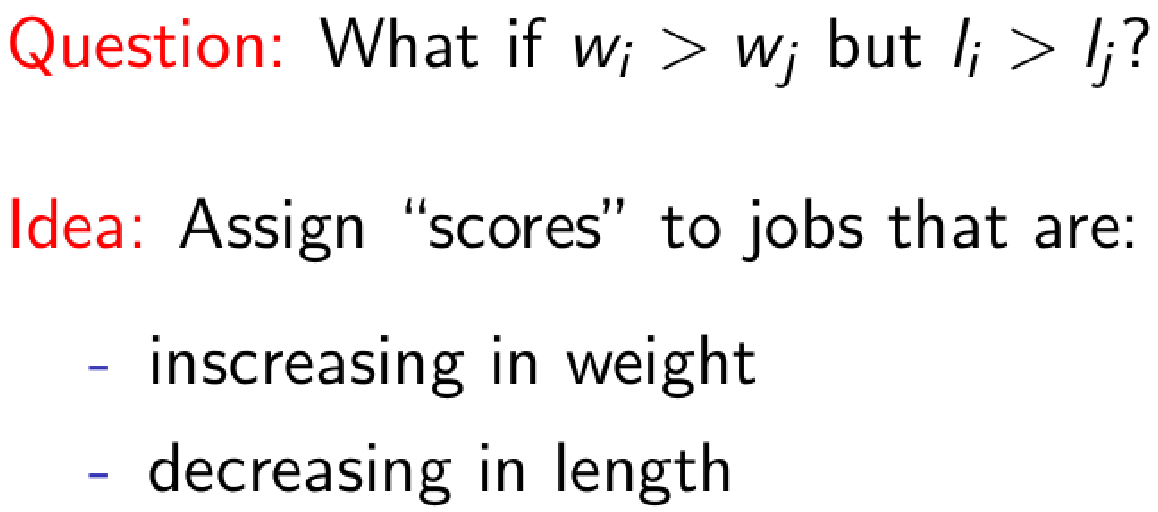
Example: 3 jobs, $l_1 = 1, l_2 = 2, l_3 = 3$.

Schedule:



Question: What is C_1, C_2, C_3? 1, 3, 6

HPC job scheduler – a simplified toy problem



Question: What if $w_i > w_j$ but $l_i > l_j$?

Idea: Assign "scores" to jobs that are:

- inscreasing in weight
- decreasing in length

HPC batch job scheduling is NP-complete

- https://www.mendeley.com/catalogue/9ed825e9-8bae-303d-b7bc-84e0f4051411/
- Clay Mathematics Institute
- https://www.claymath.org/millennium-problems

HPC job scheduler -- SLURM

- Slurm stands for "simple linux utility for resource management"
- Slurm fair tree algorithm
  - https://slurm.schedmd.com/SLUG19/Priority_and_Fair_Trees.pdf
- Slurm documentation
  - https://slurm.schedmd.com/documentation.html

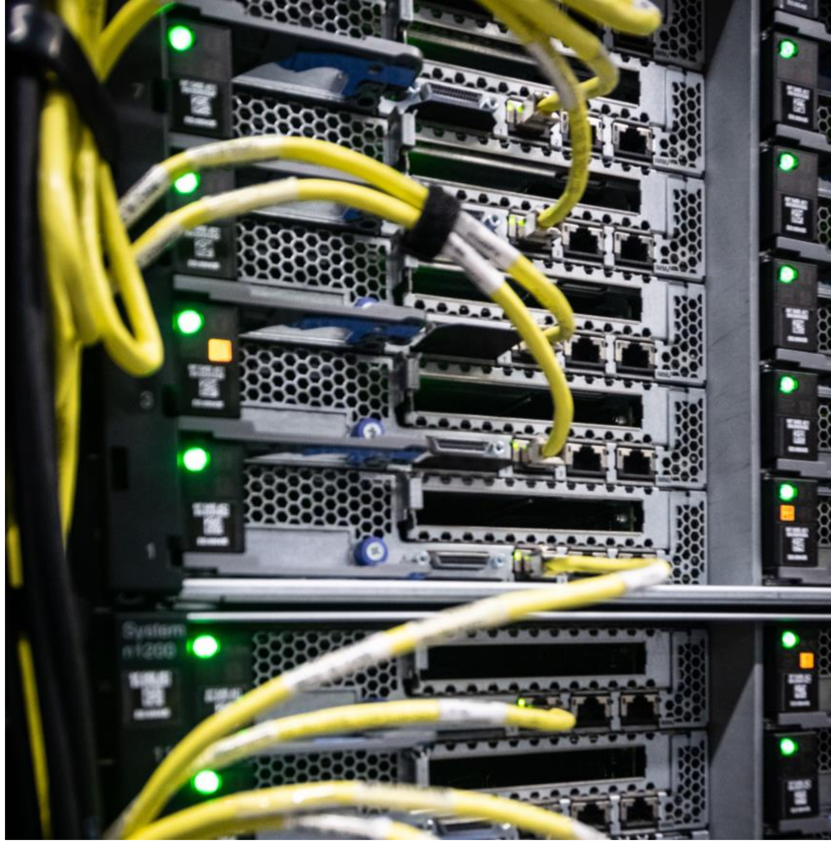
HPC job scheduler -- Slurm

- Slurm is just a software performing scheduling!
- Over 60% of TOP500 user Slurm
  - https://en.wikipedia.org/wiki/Slurm_Workload_Manager
- Job scheduler (Slurm) will be here to stay and running jobs on clusters cannot be fully automated
  - Halting problems is undecidable
  - https://en.wikipedia.org/wiki/Halting_problem
  - NP-hard, even harder than NP-complete

First poll: what is Slurm?

Q&A

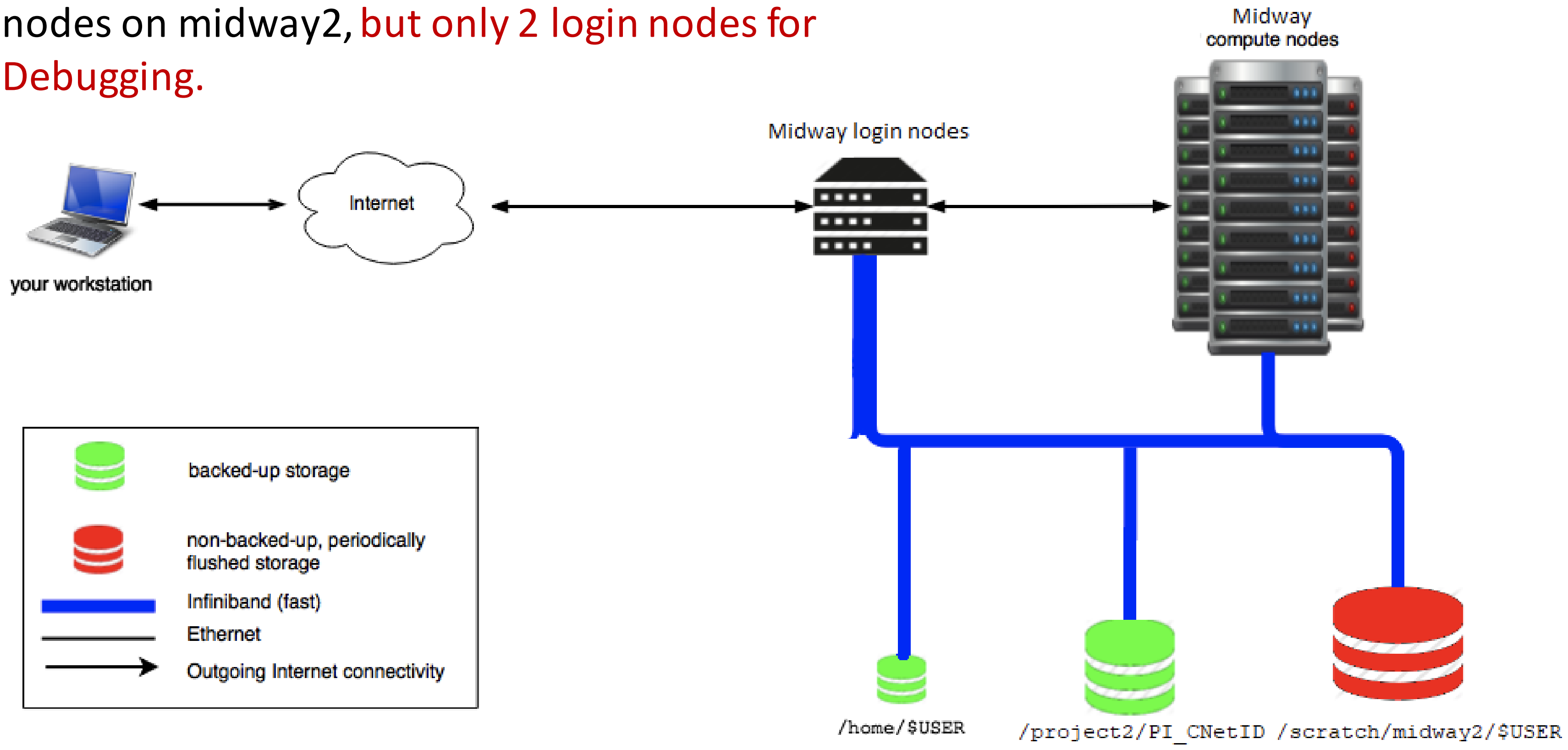
5-minute break



- Midway is a constellation a of many compute systems and storage with various architectures coupled together in one system.
- SLURM is the software used to manage the workload on Midway
- https://top500.org/

Schematic of the Midway Cluster

There are about more than 1300+ nodes compute nodes on midway2, but only 2 login nodes for Debugging.



RCC Slurm

- Standard HPC resources
- Completely free, only for UChicago affiliates
- Require faculty approval
- Managed according to SU (service units), 1 SU = 1 core x 1 hour (https://rcc.uchicago.edu/accounts-allocations/calculations-service-units)

Q&A

Summary

- The RCC system is pretty common among Top500
- Management according to SU
- Free for all UChicago affiliates
- State-of-the art hardware

Slurm syntax

```
#!/bin/bash

# Here is a comment
#SBATCH --time=1:00:00              => Time your job is allowed to run

#SBATCH –nodes=1                    => Number of nodes to run on
#SBATCH –ntasks-per-node=1          => Number of cores on each node to use
#SBATCH --mem-per-cpu=2000          => Memory per cpu => 2000Mb or 2Gb
#SBATCH –job-name=MyJob             => Name of the job.
#SBATCH –output= MyJob-%j.ou        => Job output file behaves as stdout for the code.
#SBATCH –error=MyJob-%j.err         => Error file. behaves as stderr for the code.

module load <module name>           => Load any modules you need for your application
#Run your code                      => run the code you want
```

Second poll: Slurm vs parallel computing

Tutorial 1 interactive job

…/Slurm_10142021/Turorial0_Interactive

Tutorial 2: how to submit OpenMP jobs?

```
#!/bin/bash
#Here is a comment
#SBATCH --job-name=MyJob
#SBATCH --output=MyJob-%j.out
#SBATCH --error=MyJob-%j.err
#SBATCH --time=1:00:00
#SBATCH --partition=broadwl
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=10
#SBATCH --mem=16000  #Per Node
##SBATCH --mem-per-cpu=2000  #Per CPU
module load gcc/9.2.0
make -f Makefile
export OMP_NUM_THREADS=8
#Run your code
./norm_prog
```

Slurm has some variables you can use. %j is the job number. When the job runs %j will be expanded to the job number.

Specify number of cores > 1.

OMP_NUM_THREADS is an environment variable.

…/Slurm_10142021/Tutorial2_openmp

Tutorial 3: how to submit GPU jobs?

```
#!/bin/bash
#SBATCH --time=1:00:00
#SBATCH --time=00:10:00
#SBATCH --partition=gpu2
#SBATCH --gres=gpu:1
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=4
#SBATCH --mem=2000
#SBATCH --job-name=MyJob
#SBATCH --output=MyJob_%j.out
#SBATCH --error=MyJob_%j.err


module load cuda/10.1
make -f Makefile
./deviceQ
```

Specify partition gpu2

Specify number of gpus, like *gpu*:1

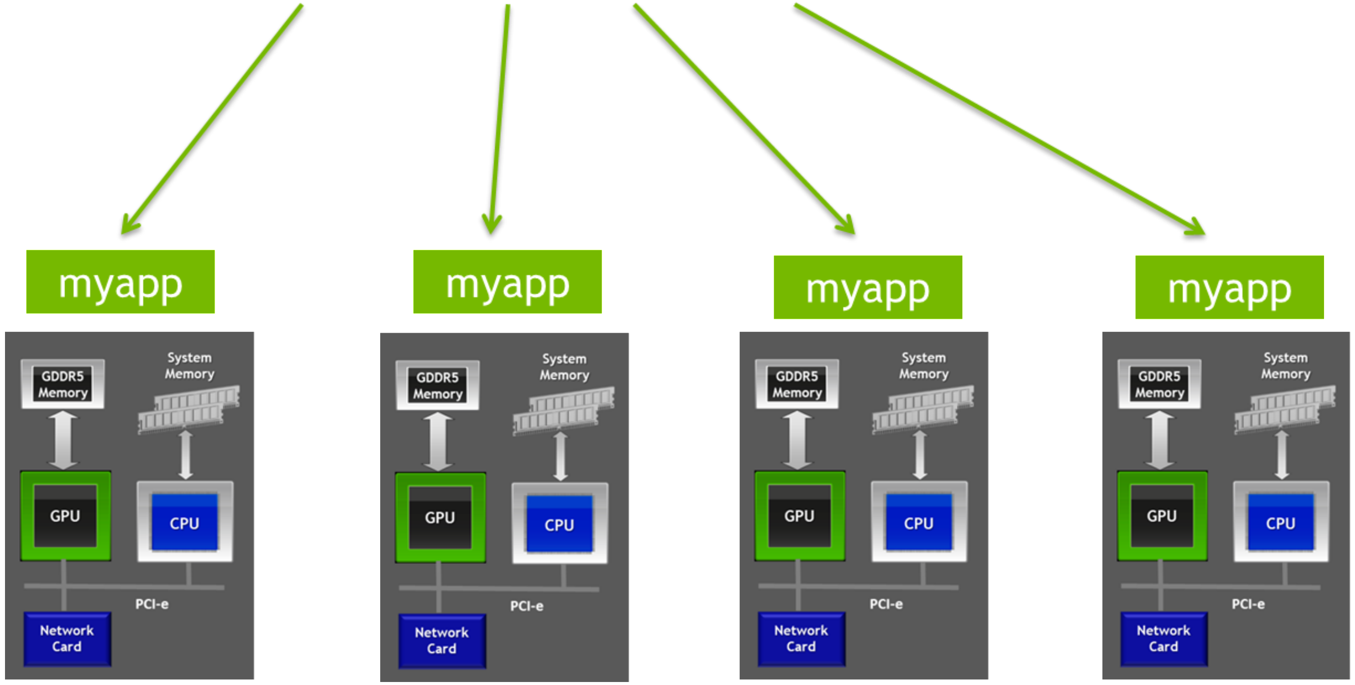
…/Slurm_10142021/Tutorial3_GPU

Third poll: GPU

Tutorial 4: How to submit MPI + GPU jobs?

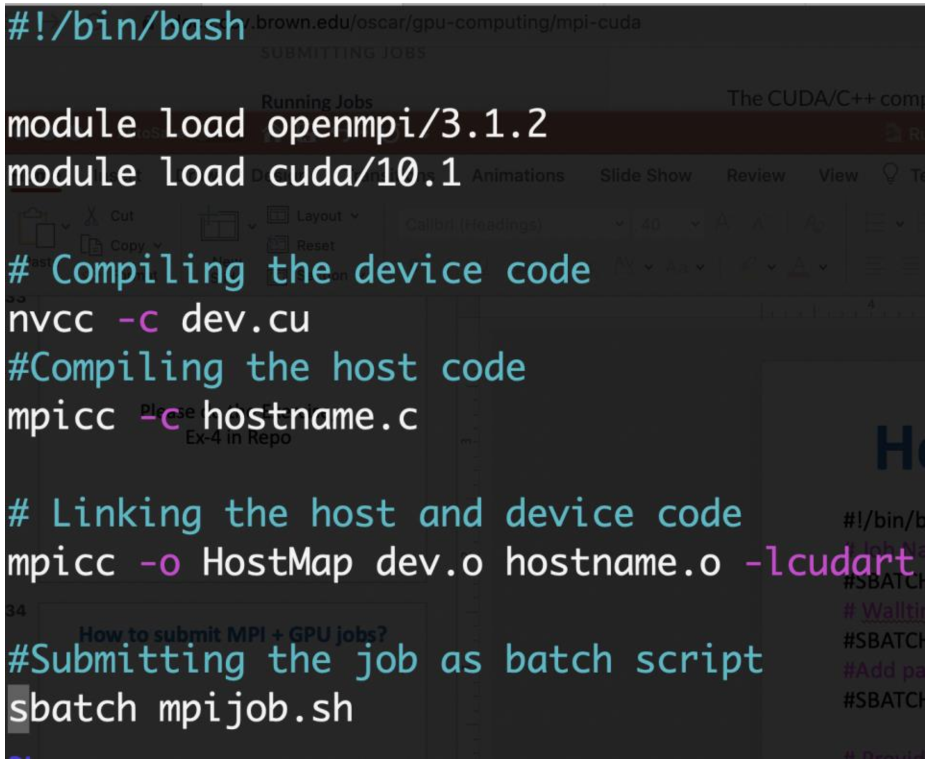
CUDA Aware MPI

```
mpirun -np 4 ./myapp <args>
```



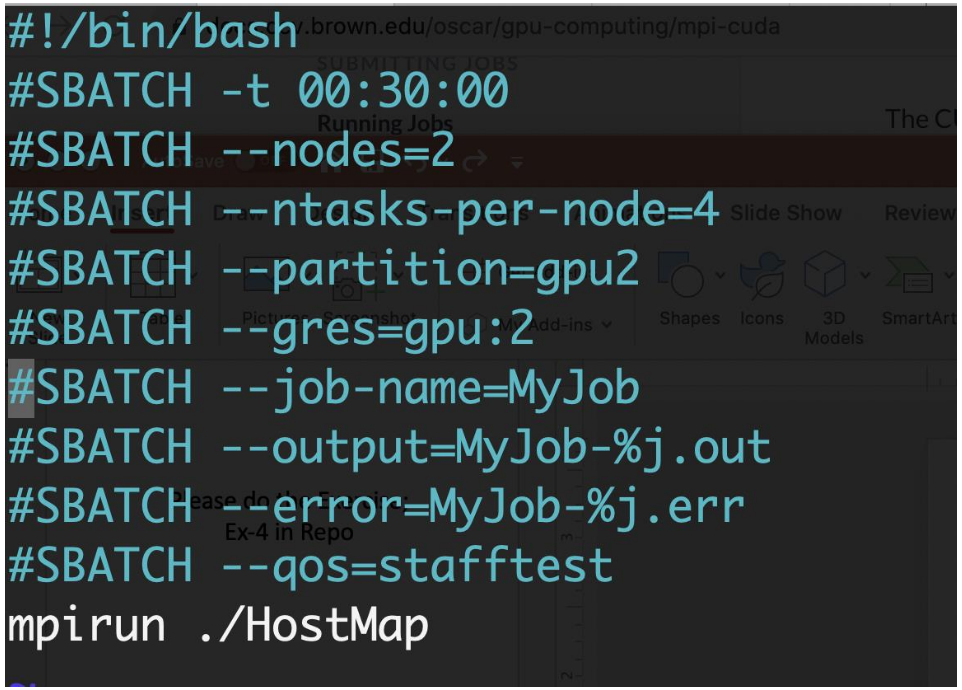
Tutorial 4: How to submit MPI + GPU jobs?

Compilation



```
#!/bin/bash

module load openmpi/3.1.2
module load cuda/10.1

# Compiling the device code
nvcc -c dev.cu
#Compiling the host code
mpicc -c hostname.c

# Linking the host and device code
mpicc -o HostMap dev.o hostname.o -lcudart

#Submitting the job as batch script
sbatch mpijob.sh
```

Job Submission



```
#!/bin/bash
#SBATCH -t 00:30:00
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=4
#SBATCH --partition=gpu2
#SBATCH --gres=gpu:2
#SBATCH --job-name=MyJob
#SBATCH --output=MyJob-%j.out
#SBATCH --error=MyJob-%j.err
#SBATCH --qos=stafftest
mpirun ./HostMap
```

…/Slurm_10142021/Tutorial4_GPU

Tutorial 5: How to submit array based jobs?

```
#!/bin/bash
# Job Name
#SBATCH --job-name=arrayjob
# Walltime requested
#SBATCH --time=0:10:00
#Add partition
#SBATCH --partition=broadwl-lc

# Provide index values (TASK IDs)
#SBATCH --array=1-16

# Use '%A' for array-job ID, '%J' for job ID and '%a' for task ID
#SBATCH --error=maths%A-%a.err
#SBATCH --output=maths%A-%a.out

# single core
#SBATCH --ntasks-per-node=1
#SBATCH --mem-per-cpu=2000

# Use the $SLURM_ARRAY_TASK_ID  variable to provide different inputs for each job
input=$((SLURM_ARRAY_TASK_ID*1000+2))
echo "Running job array number: "$SLURM_ARRAY_TASK_ID  "input " $input
```

…/Slurm_10142021/Tutorial5_GPU

Q&A

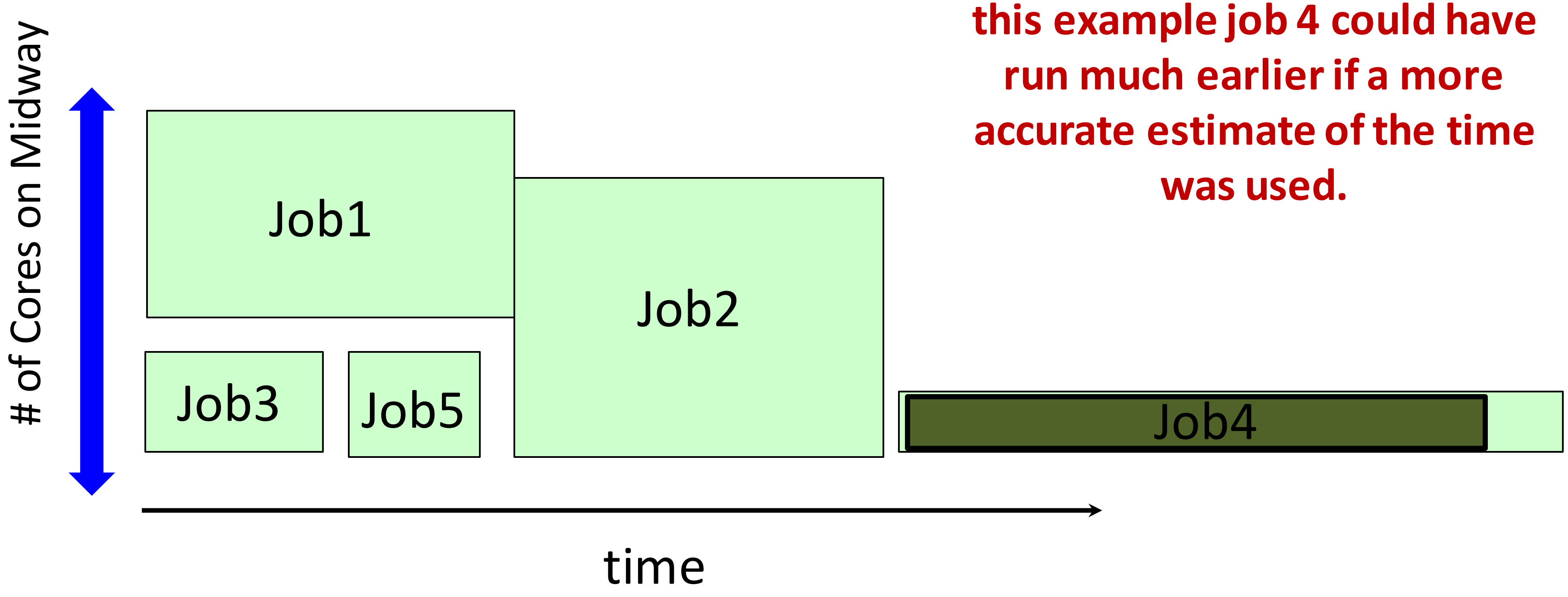
5-minute break

What if I need an entire node, large memory or specific features?

- Exclusive node --- #SBATCH –exclusive
- Useful commands
  - Jobinfo –j jobid
  - rcchelp qos
- Specific hardware -- #SBATCH –constraint=v100
  - *nodestatus* command

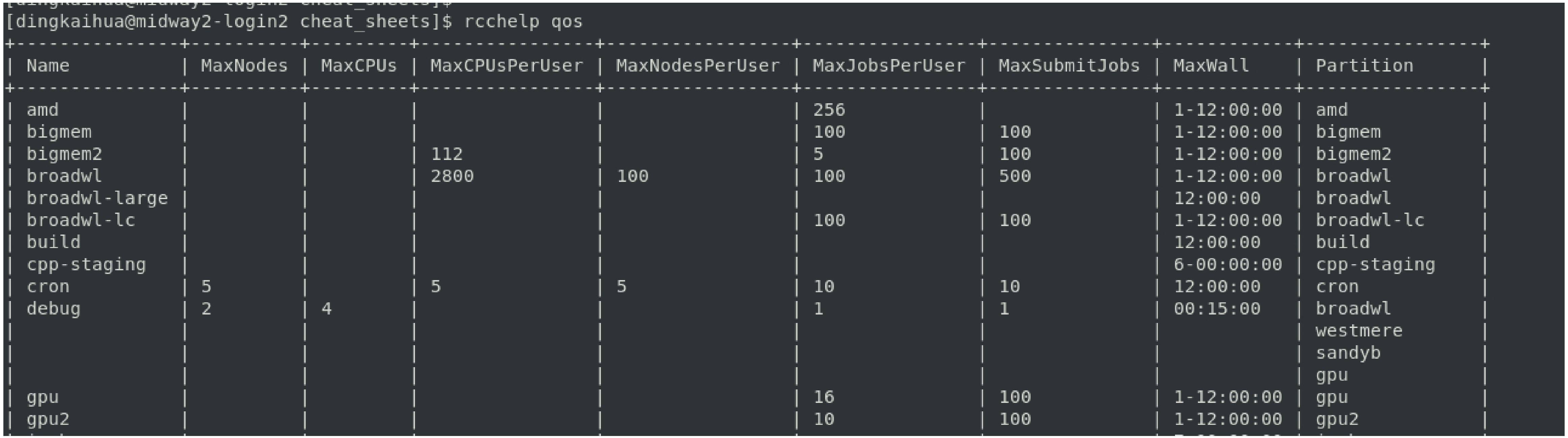
Job Priority

Why has someone else's job started before mine?



You can take advantage of backfill by specifying wall time as accurately as you can. In this example job 4 could have run much earlier if a more accurate estimate of the time was used.

Make an informed decision Slurm setups



```
[dingkaihua@midway2-login2 cheat_sheets]$ rcchelp qos
+---------------+----------+---------+----------------+-----------------+----------------+---------------+------------+---------------+
| Name          | MaxNodes | MaxCPUs | MaxCPUsPerUser | MaxNodesPerUser | MaxJobsPerUser | MaxSubmitJobs | MaxWall    | Partition     |
+---------------+----------+---------+----------------+-----------------+----------------+---------------+------------+---------------+
| amd           |          |         |                |                 | 256            |               | 1-12:00:00 | amd           |
| bigmem        |          |         |                |                 | 100            | 100           | 1-12:00:00 | bigmem        |
| bigmem2       |          |         | 112            |                 | 5              | 100           | 1-12:00:00 | bigmem2       |
| broadwl       |          |         | 2800           | 100             | 100            | 500           | 1-12:00:00 | broadwl       |
| broadwl-large |          |         |                |                 |                |               | 12:00:00   | broadwl       |
| broadwl-lc    |          |         |                |                 | 100            | 100           | 1-12:00:00 | broadwl-lc    |
| build         |          |         |                |                 |                |               | 12:00:00   | build         |
| cpp-staging   |          |         |                |                 |                |               | 6-00:00:00 | cpp-staging   |
| cron          | 5        |         | 5              | 5               | 10             | 10            | 12:00:00   | cron          |
| debug         | 2        | 4       |                |                 | 1              | 1             | 00:15:00   | broadwl       |
|               |          |         |                |                 |                |               |            | westmere      |
|               |          |         |                |                 |                |               |            | sandyb        |
|               |          |         |                |                 |                |               |            | gpu           |
| gpu           |          |         |                |                 | 16             | 100           | 1-12:00:00 | gpu           |
| gpu2          |          |         |                |                 | 10             | 100           | 1-12:00:00 | gpu2          |
```

Qos table shows the association and limitation of each partition.

There is an optimal number of nodes / cores to request to achieve the best speedup.

- Amdahl's law
- https://en.wikipedia.org/wiki/Amdahl%27s_law
- More nodes does not lead to faster computation eventually

Job submission and monitoring

Slurm Commands

Command	Description
`sbatch script.sbatch`	Submits `script.sbatch` job script
`squeue –u $USER or myq`	Reports the status of your jobs
`sacct –u $USER`	Displays accounting data for your job(s)
`scancel jobid`	Cancels a running job or removes it from the queue
`scontrol show job jobid or jobinfo`	Displays details of a running job

Recommended online resources

- User guide on running jobs on Midway
  - https://rcc.uchicago.edu/docs/running-jobs/index.html

- Details Slurm documentation
  - https://slurm.schedmd.com/sbatch.html

- SLURM Cheat Sheet
  - https://slurm.schedmd.com/pdfs/summary.pdf

Q&A

Slurm skill self-assessment

Thank you!

Please feel free to reach out for more questions!
help@rcc.uchicago,
dingkaihua@uchicago.edu