

Programming Massively Parallel Multiprocessors and Heterogeneous Systems (Understanding and programming the devices powering AI)

Lecture 0: Class overview and Introductions

Jonathan Appavoo email: jappavoo @ b u . e d u office: CDS 706

~~Programming Massively Parallel Multiprocessors and Heterogeneous Systems (Understanding and programming the devices powering AI)~~

CUDA Programming on GPU's

Lecture 0: Class overview and Introductions

Jonathan Appavoo email: jappavoo @ bu . edu office: CDS 706

Course Synopsis
(See Syllabus on Piazza for details)

Course Prerequisites and Structure

- **Prerequisites**
 - Basic understanding of the C programming language (especially its use of pointers)
 - Undergrad level knowledge of computer architecture
- **Structure**
 - Lectures twice a week for ~7-8 weeks
 - Seminars and Guest Lectures remaining weeks
 - Discussions once a week as needed
 - Office hours

Course Components

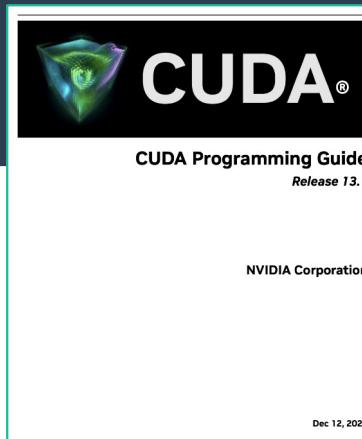
- Lectures 12 -16
- Programming Assignments: 5
- Seminars: student led paper discussion
- Guest Lectures: 5-6 : NVIDIA, IBM, Red Hat, Meta, BU
- Midterm: Lecture Material
- Final: Predominately based on assignments and programming
- *Attendance and Participation in lectures and seminars*

Course Resources

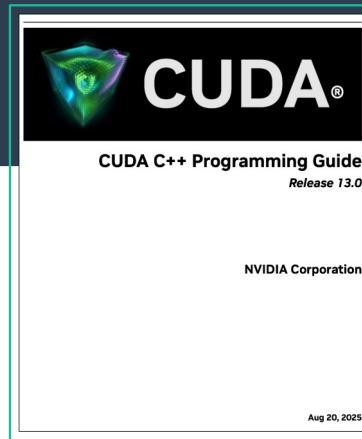
- Primary source of information (besides slides): NVIDIA Inc, "CUDA C++ Programming Guide" (and friends: best practices, API references, white papers, etc)
 - 13.1 vs 13.0 vs 12.6
- Textbook: none
 - but if you need one Kirk and Hwu: "Programming Massively Parallel Processors: A Hands-on Approach (4th edition)" is reasonable.
- Piazza
- Github Classroom
- Gradescope
- MOC/NERC/Red Hat Open AI
 - GPUS: V100, A100 and H100
 - CUDA: 12.6.3
- DGX Sparks
 - GB10's
 - CUDA: 13.1



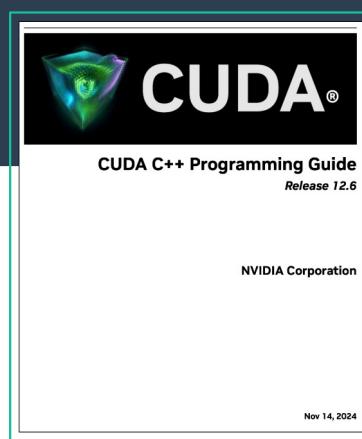
https://bu.bnccollege.com/c/Programming-Massively-Parallel-Processors/p/MBS_6919078_used?currentCampus=8529&rental=true



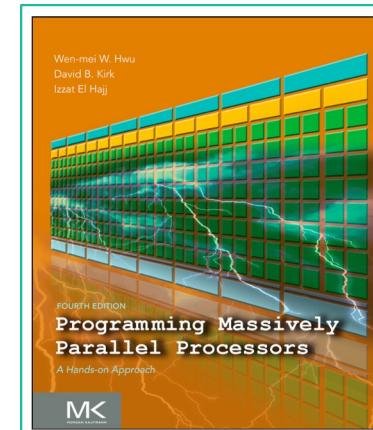
13.1



13.0



12.6



Grade Breakdown

- **Participation, Seminars & Attendance**
 - 30%
- **Programming Assignments:**
 - 40%
- **Midterm:**
 - March 19th: 15%
- **Final:**
 - 15%
-

Syllabus

The details,
including a
tentative weekly
calendar

Please read and
ask any questions
you have on
Piazza

syllabus

CAS CS 391 - Spring 2026: Programming Massively Parallel Multiprocessors and Heterogeneous Systems (Understanding and programming the devices powering AI)

Piazza: <https://piazza.com/bu/spring2026/cs391/home>

Programming/GPU environment: apps.edu.nerc.mghpc.org/

Gradescope: <https://www.gradescope.com/>

Lectures: Tuesday and Thursday 11:00AM

Staff: Professor Appavoo BU email: jai@csail.mit.edu

Office hours: Location: CDS 706 TBA

Midterms: 75 minute in-class midterm

Course Description: After decades, we MultiProcessors (MPP) has become common (GPUs), access to MPPs is no longer restricted Computing on Super-Computers.

Today, most computer systems are heterogeneous. Central Processing Units (CPUs) and Graphics Processing Units (GPUs) have terabytes of high-bandwidth memory. Parallel computing offers up to 100x speedup over similar single-core CPUs. Applications include signal processing, financial modeling, and machine learning. The widespread access and availability of heterogeneous computing is driving the AI revolution.

While the CPU's familiar von Neumann architecture is well suited for general-purpose computing, adding the GPU has proven critical in providing the performance needed. Understanding the heterogeneous computation model is the focus of this class.

The course covers "general purpose" — parallel computation in a heterogeneous system. The course includes parallel computation, including its programming model and synchronization primitives, parallel computing on GPUs, parallel algorithms, and parallel data structures.

The focus of the course will be on performance. In general, programs, GPUs are often used as accelerators rather than they would on a multi-core CPU. Specifically, the course will require you to evaluate the various models of computation. The course is not about getting applications to run faster, but rather about gaining a fundamental understanding concerning performance.

Acknowledgments: This class borrows heavily from the University of Toronto offering of ECE1782H "Programming Massively Parallel Multiprocessors and Heterogeneous Systems". Thank you, Dr Stumm.

8 Detailed Syllabus Calendar

The following is the tentative calendar changes and updates will be posted on Piazza.

Date	Activity/Topics	Readings	Assignments
Tue 01/20/26	Class Overview and Introductions		
Thu 01/22/26	Motivation and Challenges (lec1:1-32)		

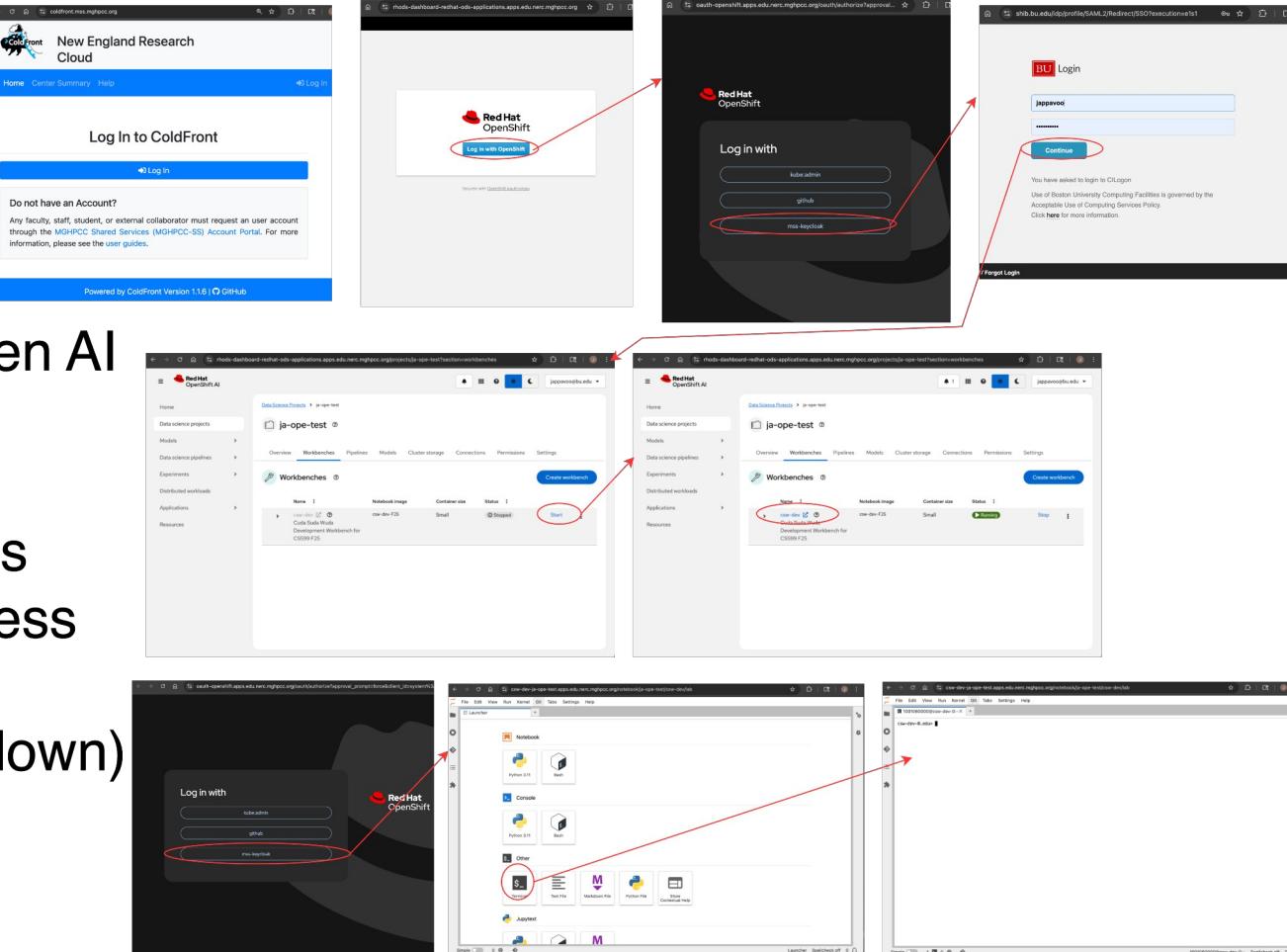
7

Fri 01/23/26	Login & Pthread create	https://hpc-tutorials.llnl.gov posix/ (Sections 1-5 inclusive)	Assignment 1 out
Tue 01/27/26	Multiprocessors		
Thu 01/29/26	GPU architecture		
Fri 01/30/26	[Optional] C and pthread help session		
Tue 02/03/26	Introduction to CUDA programming I		
Thu 02/05/26	Introduction to CUDA programming II		
Fri 02/06/26	batchtools, CUDA Hello World and Error Handling		
Tue 02/10/26	Introduction to CUDA programming III		DUE: Assignment 1 Assignment 2 out

GPU Lab Infrastructure

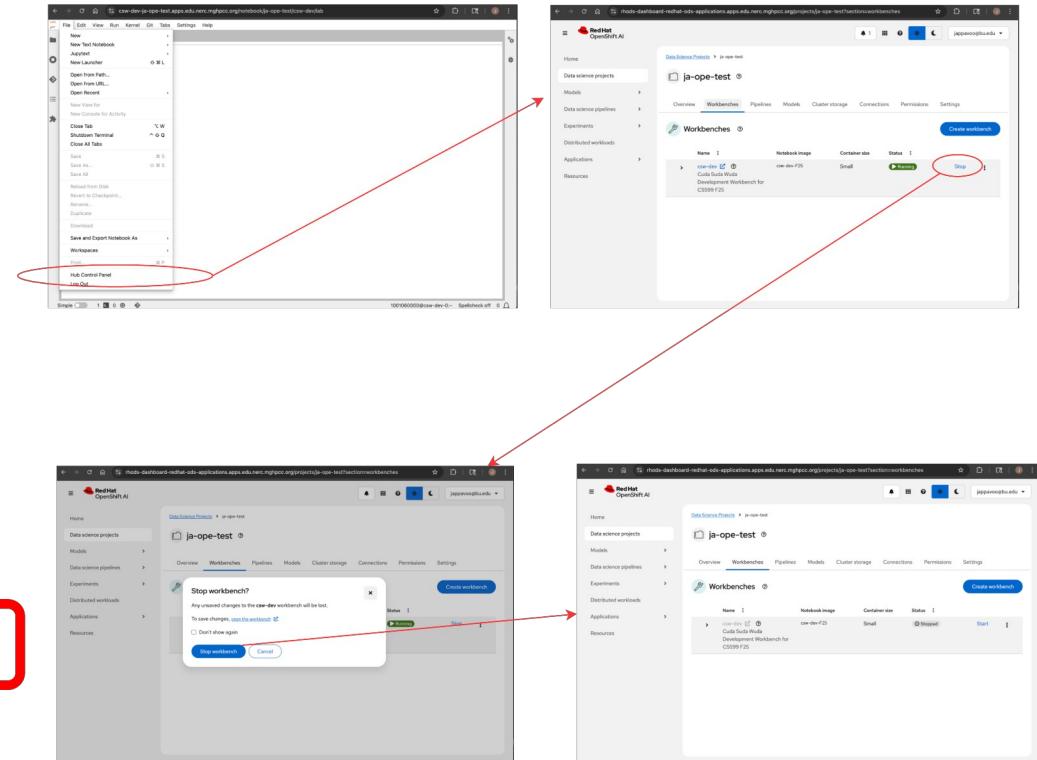
Containerized CUDA + Batch GPUs

- Getting Started
 - ColdFront
 - OpenShift and Red Hat Open AI
- Dev Container
 - RHEL 9 + CUDA 12.6.3
 - Browser Jupyter Lab Access
 - Terminal Remote Shell Access
 - Persistent home directories
 - Idle culler (but please shutdown)
- Batch Queues
 - bqstat, bjobs, brun,



Containerized CUDA + Batch GPUs

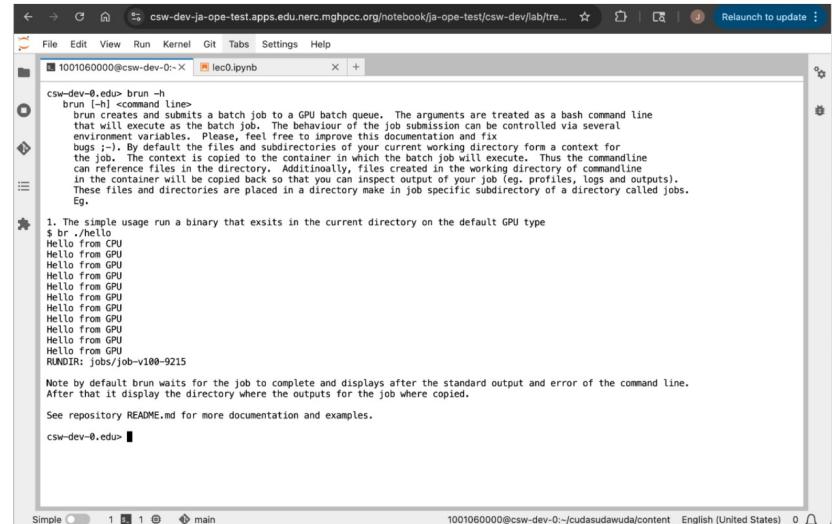
- Getting Started
 - ColdFront
 - OpenShift and Red Hat Open AI
- Dev Container
 - RHEL 9 + CUDA 12.6.3
 - Browser Jupyter Lab Access
 - Terminal Remote Shell Access
 - Persistent home directories
 - Idle culler (but please shutdown)
- Batch Queues
 - bqstat, bjobs, brun,



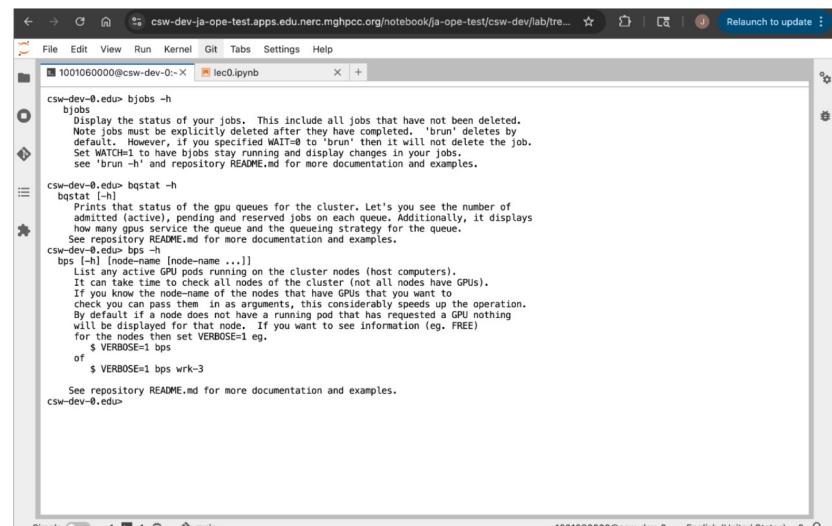
Batch Queues

<https://github.com/jappavoo/batchtools>

- **brun**: run a binary on a GPU "node"
- **./jobs**: outputs from run/job
- **bjobs**: current jobs
- **bqstat**: status of queues
- **bps**: list activity on GPU nodes (broken)
- misc: **blog**, **bods**, **bwait**



The screenshot shows a Jupyter Notebook cell with the command `brun -h`. The output provides detailed information about the `brun` command, including its purpose (submitting a batch job to a GPU batch queue), how it handles environment variables, and how it copies files from the current working directory into the job's context. It also shows examples of running a binary and copying its output back to the host.



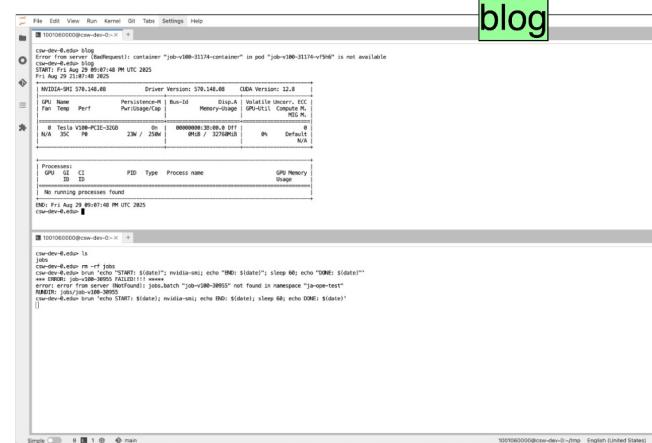
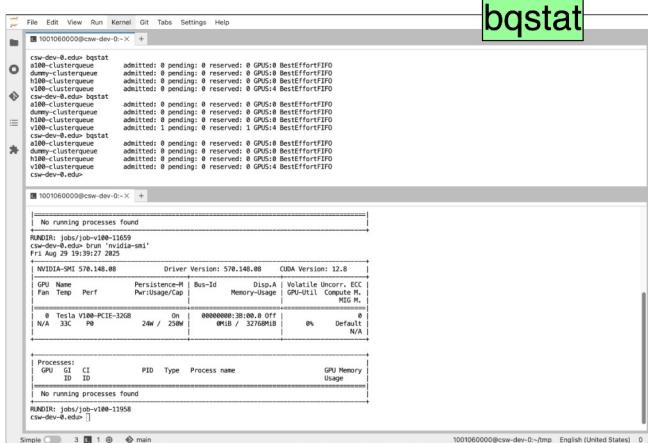
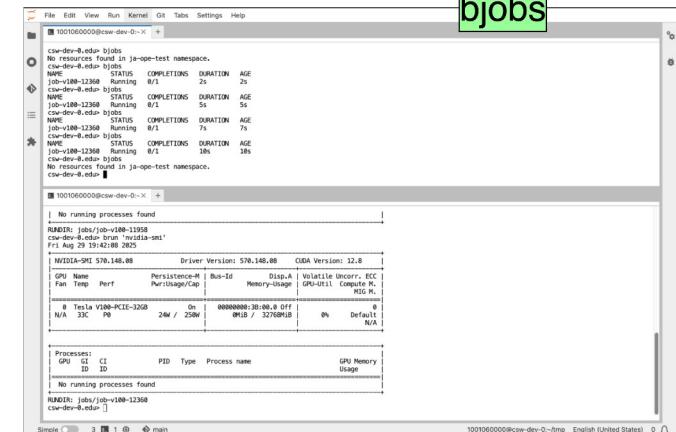
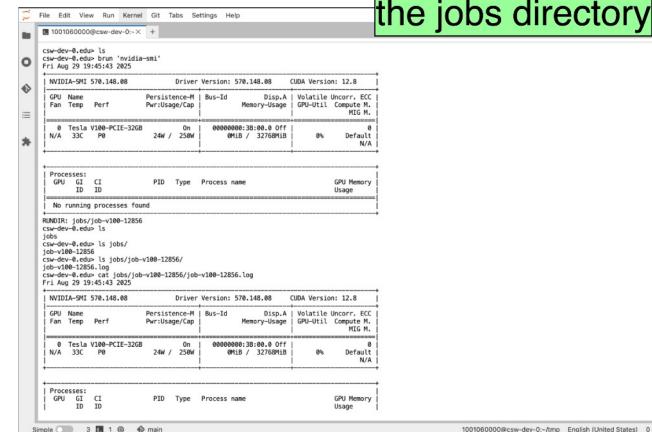
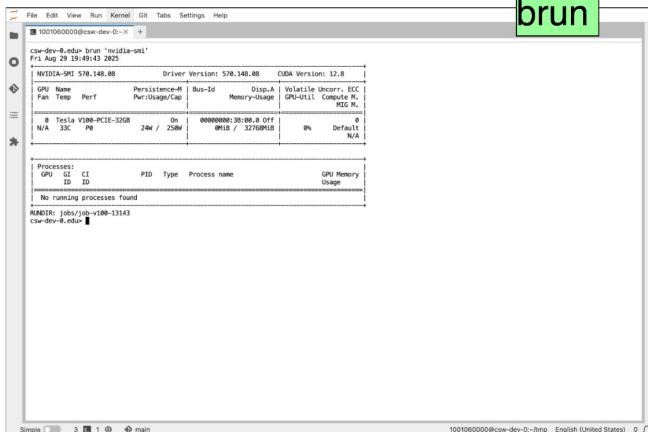
The screenshot shows a Jupyter Notebook cell with the command `bjobs -h`. The output explains that `bjobs` displays the status of all jobs, including completed ones. It notes that jobs must be explicitly deleted after completion. The command also includes options for monitoring changes and handling specific job IDs.

Please help me fix bugs

Please help me fix bugs

Batch Queues

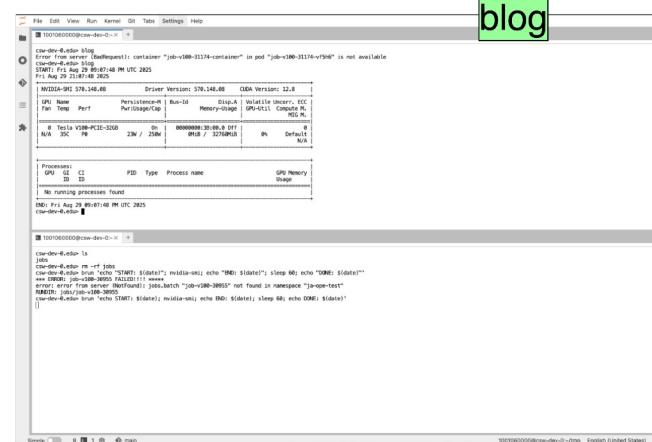
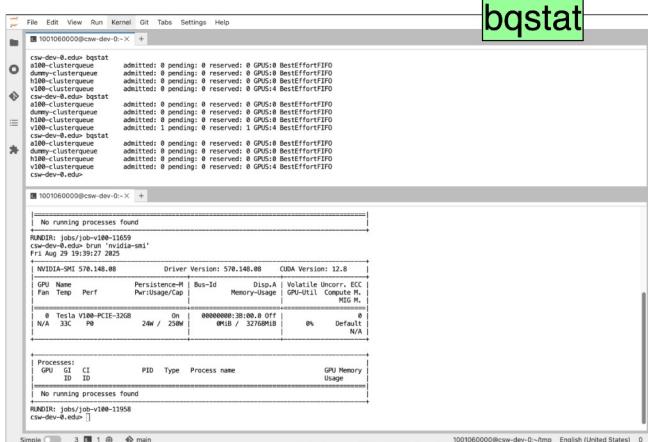
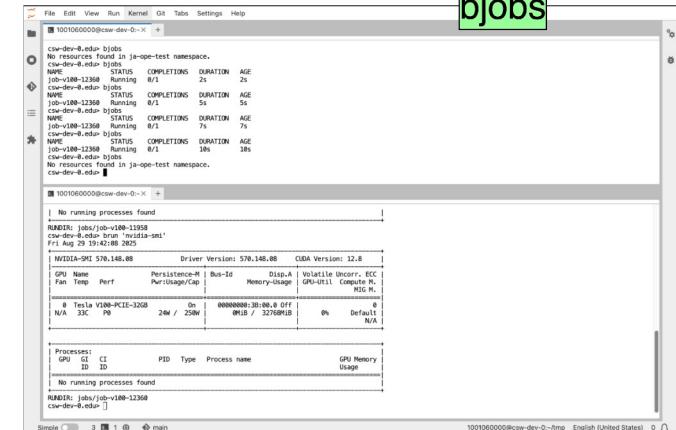
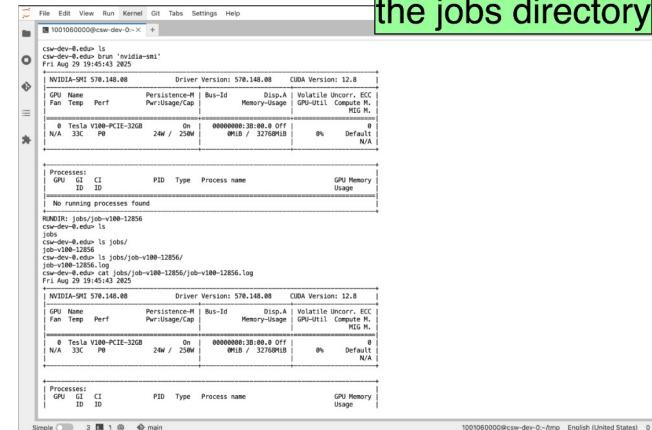
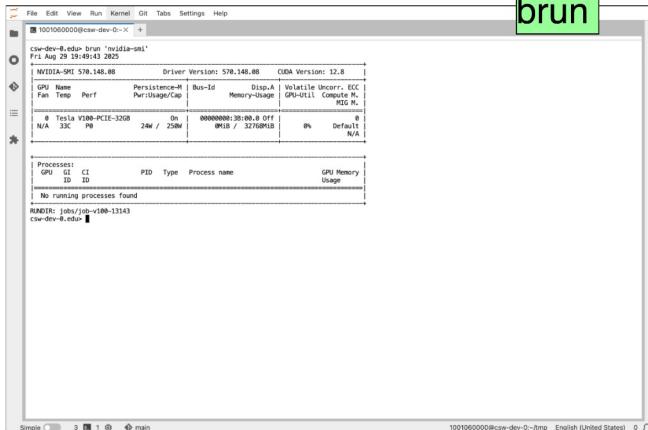
<https://github.com/jappavoo/batchtools>



Please help me fix bugs

Batch Queues

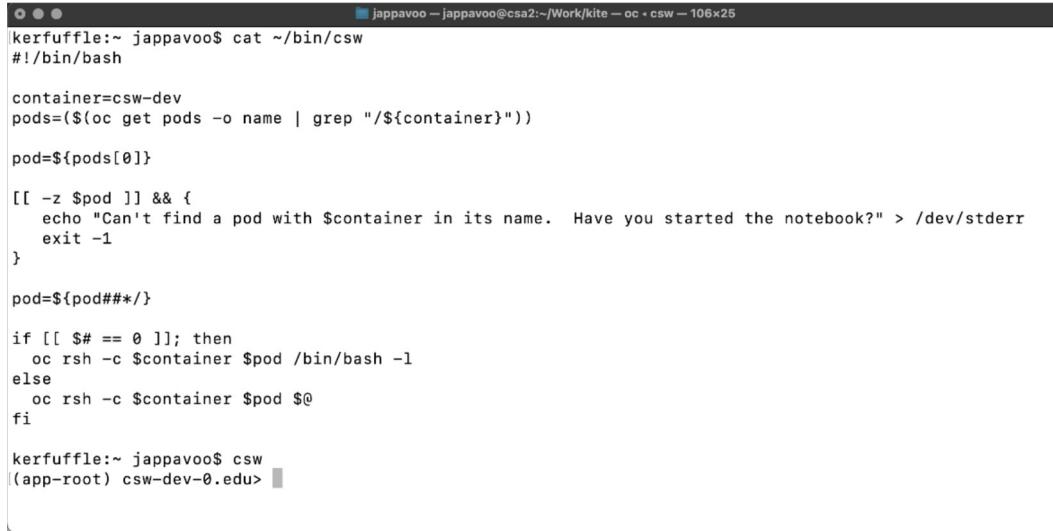
<https://github.com/jappavoo/batchtools>



Terminal Interface

Alternative to using browser

- Install OpenShift CLI client (oc)
 - <https://console.apps.edu.nerc.mghpcc.org/command-line-tools>
- Get and execute login command
 - <https://oauth-openshift.apps.edu.nerc.mghpcc.org/oauth/token/display>
- Start csw-dev workbench via Red Hat OpenShift AI
- Use 'oc rsh' and friends
 - see my csw script for an example of how I start shells sessions



```
jappavoo:~ jappavoo$ cat ~/bin/csw
#!/bin/bash

container=csw-dev
pods=$(oc get pods -o name | grep "/${container}"))

pod=${pods[0]}

[[ -z $pod ]] && {
    echo "Can't find a pod with $container in its name. Have you started the notebook?" > /dev/stderr
    exit -1
}

pod=${pod##*/}

if [[ $# == 0 ]]; then
    oc rsh -c $container $pod /bin/bash -l
else
    oc rsh -c $container $pod $0
fi

kerfuffle:~ jappavoo$ csw
(app-root) csw-dev-0.edu>
```

Concurrency, parallelism, scalability, me and you

Our backgrounds, our understanding and
why we are here

Some Ice Breakers

Well for geeks anyway

- What does "Parallel Computing" make you think/feel?
- How many cores does your laptop have?
- What is the largest number of cores that a system you have worked on has had?
- What is the largest number of threads you have explicitly created in a program?
- Have you written a pthreaded application? How long did you debug it?
- What are your CS interests?
- Why brings you to this class?