

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 @ b u . e d u 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**

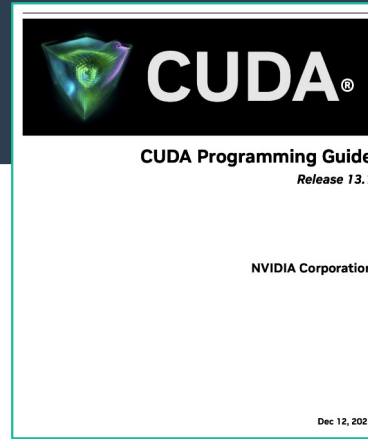
- Lectures twice a week for ~7-8 weeks
- Seminars and Guest Lectures remaining weeks
- Discussions once a week as needed
- Office hours

Course Components

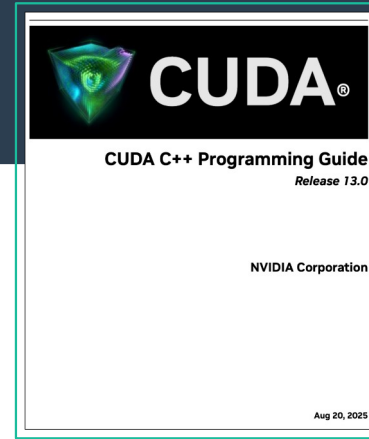
- Lectures 12 -16
- Programming Assignments: 5
- Seminars: student led paper discussion
- Guest Lectures: 5-6 : NVIDIA, IBM, Red Hat, Meta, BU
- Midterm: Lecture Material
- Final: Predominately based on assignments and programming
- Attendance and Participation in lectures and seminars

Course Resources

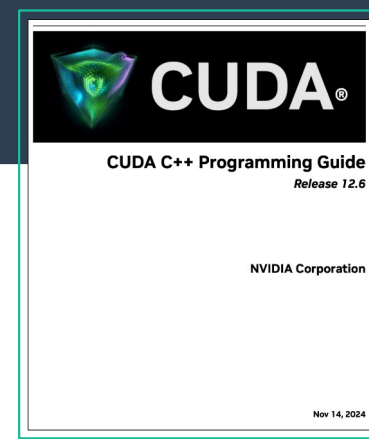
- **Primary source of information (besides slides): NVIDIA Inc, "CUDA C++ Programming Guide" (and friends: best practices, API references, white papers, etc)**
 - 13.1 vs 13.0 vs 12.6
- **Textbook: none**
 - but if you need one Kirk and Hwu: "Programming Massively Parallel Processors: A Hands-on Approach (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



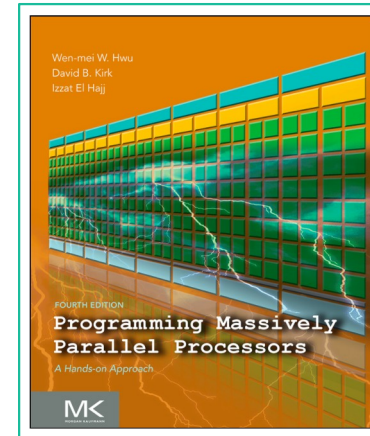
13.1



13.0



12.6



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

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.educ.nerc.mghpcc.org/

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

Lectures: Tuesday and Thursday 11:00F

Staff: Professor Appavoo BU email: jap@bu.edu

Office hours: **Location:** CDS 706 T

Midterms: 75 minute in-class midterm

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

Today, most computer systems are heter Central Processing Units (CPUs) and G gabytes of high-bandwidth memory. Para 100x speedup over similar single-core CP tions, signal processing, financial modeli widespread access and availability of heter AI revolution.

While the CPU's familiar von Neumma general-purpose computing, adding the I GPU has proven critical in providing the tion. Understanding the heterogeneous co is the focus of this class.

The course covers "general purpose" — in a heterogeneous system. The course in including its programming model and syn computing on GPUs, parallel algorithms, C

The focus of the course will be on per programs, GPUs are often used as accele than they would on a multi-core CPU. Sp will require you to evaluate the various n course is not about getting applications but rather about gaining a fundamental u concerning performance.

Acknowledgments: This class borrows Toronto offering of ECE1782H "Program 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)		

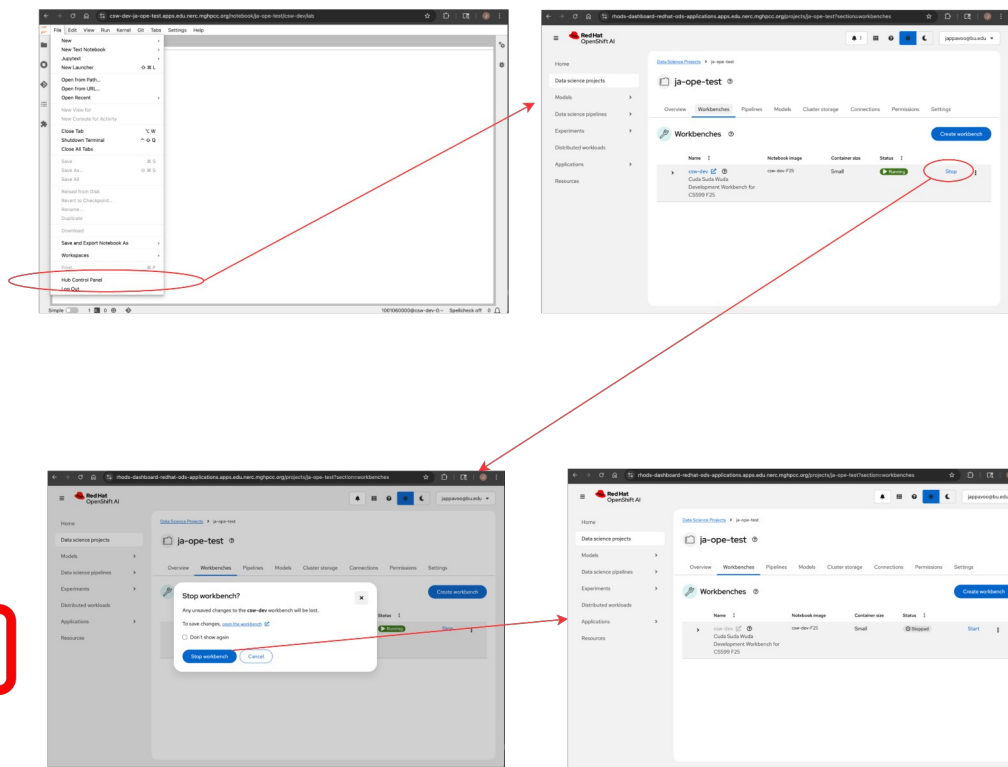
7

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,



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

- # Please help me fix bugs

The screenshot shows a terminal window with the following content:

```

csw-dev-0.edu~ bjobs -h
bjobs
Display the status of your jobs. This include all jobs that have not been deleted.
Note jobs must be explicitly deleted after they have completed, 'brun' deletes by
default. However, if you specified WAIT=0 to 'brun' then it will not delete the job.
Set WATCH=1 to have bjobs stay running and display changes in your jobs.
see 'brun -h' and repository README.md for more documentation and examples.

csw-dev-0.edu~ bqstat -h
bqstat -h]
Prints that status of the gpu queues for the cluster. Let's you see the number of
admitted (active), pending and reserved jobs on each queue. Additionally, it displays
how many gpus service the queue and the queuing strategy for the queue.
See repository README.md for more documentation and examples.

csw-dev-0.edu~ bps -h
bps [-h] [node-name [node-name ...]]
List any active GPU pods running on the cluster nodes (host computers).
It can take time to check all nodes of the cluster (not all nodes have GPUs).
If you know the node-name of the nodes that have GPUs that you want to
check you can pass them in as arguments, this considerably speeds up the operation.
By default, if a node does not have a running pod that has requested a GPU nothing
will be displayed for that node. If you want to see information (eg. FREE)
for the nodes then set VERBOSE=1 eg.
$ VERBOSE=1 bps
or
$ VERBOSE=1 bps wrk-3

See repository README.md for more documentation and examples.
csw-dev-0.edu~

```

Batch Queues

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



Batch Queues

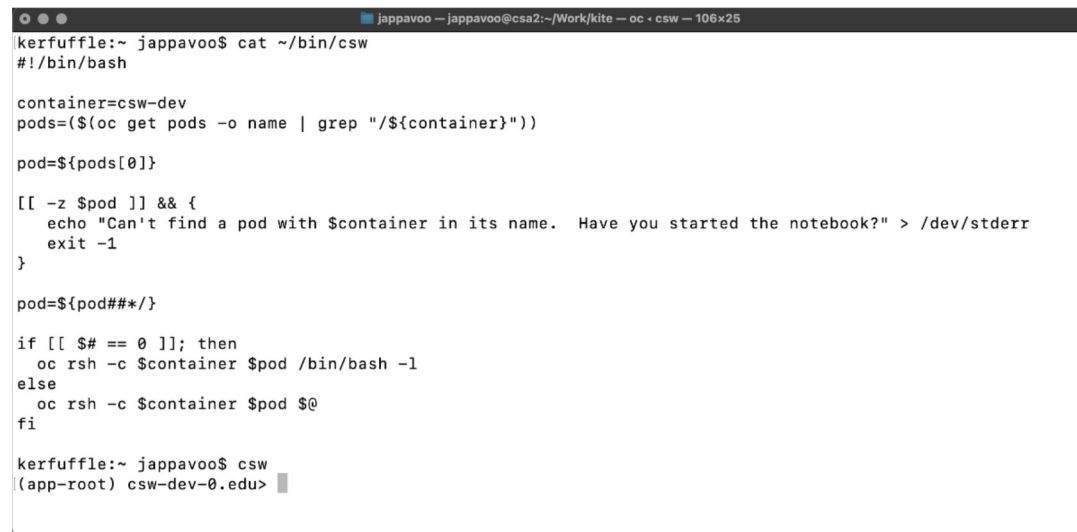
<https://github.com/jappavoo/batchtools>



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@csa2:~/Work/kite — oc - csw — 106x25
kerfuffle:~ 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 $@
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?