CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

CSDS 451: Designing High Performant Systems for AI

Instructor: Sanmukh Kuppannagari
TuTh: 4:00 PM – 5:15 PM | White 326
Fall 2023

Faculty Contact Information:
Office: Olin 506
Email: sanmukh.kuppannagari@case.edu
Office Hours: TBD
Best way to contact: Simply drop by during the office hours or send me an email to book an appointment outside office hours. Please avoid asking questions on email as it is not a scalable solution. Please see course guidelines for more information.
Slack Workspace: TBD

Teaching Assistants
Nathaniel Tomczak, Email: nkt8@case.edu

Course Overview and Objectives: The objective of the course is to give a broad overview of the challenges and opportunities that exist in designing high performance AI systems. The course is designed to cater to two types of audiences:

- Students working on data science projects who want to understand how to perform faster training or inference of their AI/ML models.
- Students working on parallel algorithms, or hardware acceleration, who want to understand modern techniques for accelerating data science applications.

On the theory side, the course will cover basics and some recent advances in improving the performance of state-of-the-art AI/ML techniques including Convolutional Neural Networks (CNN), and Transformer based Large Language Models (LLM).

On the practical side, the course will enable students to implement the optimizations to better grasp the concepts. Additionally, the course will discuss state-of-the-art programming languages and frameworks for accelerating AI such as Sycl/DPC++ for targeting CPU+Accelerator architectures and pytorch lightning for distributed AI/ML model training.

The focus will be primarily on algorithmic optimizations as opposed to device specific optimizations.

Course Pre-requisites: CSDS 310 or CSDS 410 or graduate standing

Recommended Background: Experience with Programming: Preferably C/C++ and/or Python. Mathematical maturity.

Reading Materials: As the course will discuss leading cutting-edge techniques in AI acceleration, it will derive content from a number of recently published papers. The course slides will explain topics in sufficient detail for the class, however, the list of the papers will be provided before class and students will have the opportunity to review them if they desire.

Grading Policy: Students will be graded on a scale of A (>90%), B (>80%), C (>70%), D (>60%), and F. The grades will be composed of the following components:

Written Assignments (30%): 3 assignments on the theoretical topics covered in the class. The assignments will test the understanding of the algorithms covered in the class.
Programming Assignments (30%): 3 programming assignments on the programming frameworks covered in the class.
Mid-term Exam (10%): An exam during the middle of the semester to test the theoretical topics covered in the class.
Final Exam (10%): A final exam to test the theoretical topics covered in the class.
Project (20%): Students will need to do a project using one (or more) of the programming frameworks discussed in the class to accelerate an AI/ML model discussed in the class using one (or more) optimization strategies discussed in the class.

Late Assignments: Late submission of assignments is permissible, but with a grade reduction of 10% per day up to at most three days. To ensure fairness in such a big class, I am afraid I will have to follow this policy strictly. The only exceptions will be for medical reasons, which will require a written validation from a doctor.

General Course Guidelines:

- Please **start** the assignments, especially the programming assignments, **early**.
- Please **take advantage** of the **office hours**.
- **Asking questions in the class is encouraged.** Discussions in the class are one of the best ways of collaborative learning.
- **Please check the course canvas regularly.** I will post regular announcements, assignments, and grades on canvas.

Course Schedule (Tentative)

Week	Lecture	Topics	Tasks
1	Theory 0 (8/29)	Course Introduction, Overview of Topics	PA 0 out
1	Theory 1 (8/31)	Introduction to heterogeneous computing, Introduction to Key Kernels in AI/ML	
2	Practice 1 (9/5)	Using CWRU HPC. Training AI/ML model using Pytorch	
2	Theory 2 (9/7)	Processor-Memory Architecture – Modeling, Analysis, Challenges, Opportunities for Optimization	**PA 0 due**
3	Theory 3 (9/12)	Systolic Array Architecture - Modeling, Analysis, Challenges, Opportunities for Optimization	PA 1 out
3	Theory 4 (9/14)	Cluster of Accelerators – Modeling, Analysis, Challenges, Opportunities for Optimization	

4	Theory 5 (9/19)	Analyzing Parallelization Techniques	
4	Theory 6 (9/21)	Strategies for Parallelization	**PA 1 due**; WA 1 out
5	Theory 7 (9/26)	Strategies for Parallelization	
5	Practice 2 (9/28)	Heterogeneous Programming using OneAPI	
6	Practice 3 (10/3)	Heterogeneous Programming using OneAPI	
6	Theory 8 (10/5)	Accelerating CNN: Basics, Convolution Algorithm	**WA 1 due** PA 2 out
7	Theory 9 (10/10)	Accelerating CNNs: Matrix Multiplication based convolution algorithm - im2col, kn2row	
7	Theory 10 (10/12)	Accelerating CNNs: Matrix Multiplication based convolution algorithm – scalar MM, Winograd	
8	**Exam (10/17)**	**Midterm Exam**	
8	Theory 11 (10/19)	Accelerating CNNs: Introducing Sparsity – Grouped Convolutions	**PA 2 due** WA 2 out
9	Break (10/24)	Fall Break	
9	Theory 12 (10/26)	Accelerating CNNs: Introducing Sparsity – ShuffleNets, CondenseNets	
10	Theory 13 (10/31)	Accelerating CNNs: Other Sparsity Techniques	
10	Practice 4 (11/2)	Pytorch Lightning	**WA 2 due** PA 3 out
11	Theory 14 (11/7)	Accelerating LLMs: Basics	
11	Theory 13 (11/9)	Accelerating LLMs: Optimizations for Training	
12	Theory 14 (11/14)	Accelerating LLMs: Optimizations for Fine-tuning	
12	Practice 5 (11/16)	Project Discussion	**PA 3 due** WA 3 out; Project Out
13	Theory 15 (11/21)	Accelerating on Clusters: Parallelism Paradigms	
13	Break (11/23)	Thanksgiving Break	
14	Theory 16 (11/28)	Accelerating on Clusters: Megatron	
14	**Exam**	**Final Exam**	

	(11/30)		
15	Theory 17 (12/5)	Accelerating on Clusters: DeepSpeed	
15	Theory 18 (12/7)	Conclusion	**WA 3 due**
Final		No Final	**Project Due**

Accessibility Policy: In accordance with federal law, if you have a documented disability, you may be eligible to request accommodations from Disability Resources. In order to be considered for accommodations, you must first register with the Disability Resources office. Please contact their office to register at 216.368.5230 or get more information on how to begin the process. Please keep in mind that accommodations are not retroactive.

Academic Integrity: Students at Case Western Reserve University are expected to uphold the highest ethical standards of academic conduct. Academic integrity addresses all forms of academic dishonesty, including cheating, plagiarism, misrepresentation, obstruction, and submitting without permission work to one course that was completed for another course. Please review the complete academic integrity policy. Any violation of the policy will be reported to the Dean of Graduate Studies.

Policy on using AI Tools: Use of AI tools such as ChatGPT is encouraged. You can use these tools to understand more about a concept or conduct further research around it. However, be aware that these tools are notorious at providing incorrect or false information. So, it is a good idea to examine the original source. For assignments, it is ok to search the background concepts related to the questions using the AI tools. However, please refrain from directly searching the questions or submitting AI generated text as your own. Same goes for the programming assignments. You can search the algorithms using the AI tools, or ask them to help you in debugging your code, but, to maximize your learning experience, please refrain from directly asking these tools to generate the code for you.

Additional Resources:
Kevin Smith Library - https://researchguides.case.edu/cds
Contact: Daniela Soloman if you are looking for a reading material that is unavailable.