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 | Olin 305
Fall 2024

Faculty Contact Information:
Office: Olin 506
Email: sanmukh.kuppannagari@case.edu
Office Hours: Tuesdays, 3-4 pm and Fridays 2-3 pm (in-person only)
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.

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:

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

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%): 2 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	Lecture 1 (8/27)	Course Introduction, Introduction to heterogeneous computing, Introduction to Key Kernels in AI/ML	PA 0 out
1	Lecture 2 (8/29)	Processor-Memory Architecture – Modeling, Analysis, Challenges, Opportunities for Optimization	
2	Lecture 3 (9/3)	Processor-Memory Architecture – Data layout, GPU Programming Model	
2	Lecture 4 (9/5)	Using CWRU HPC. Training AI/ML model using Pytorch	**PA 0 due** WA 1 out
3	Lecture 5 (9/10)	Data Parallel Programming using GPU Architectures	
3	Lecture 6 (9/12))	Data Parallel Programming and Task Parallelism	
4	Lecture 7 (9/17)	Designing GPU Algorithms and Parallel Program Analysis	

4	Lecture 8 (9/19)	Heterogeneous Programming using OneAPI	**WA 1 due** PA 1 out
5	Lecture 9 (9/24)	Heterogeneous Programming using OneAPI	
5	Lecture 10 (9/26)	Pytorch Lightning	
6	Lecture 11 (10/1)	Systolic Array Architecture – Modeling, Matrix Multiplication	
6	Lecture 12 (10/3)	Midterm Review and Accelerating CNN: Basics, Convolution Algorithm	PA 1 due
7	Lecture 13 (10/8)	Accelerating CNNs: Matrix Multiplication based convolution algorithm - im2col, kn2row, Scalar MM	
7	**Exam (10/10)**	**Midterm Exam**	WA 2 out
8	Lecture 14 (10/15)	Accelerating CNNs: Introducing Sparsity – Grouped Convolutions, Depthwise convolution, sparsity and pruning techniques	
8	Lecture 15 (10/17)	Accelerating LLMs: Basics	**Project Teams Due**
9	No Lecture (10/22)	Fall Break	
9	Lecture 16 (10/24)	Accelerating LLMs: Sparse Attention	**WA 2 due** PA 2 out
10	Lecture 17 (10/29)	Accelerating LLMs: Attention as Graph Processing I	
10	Lecture 18 (10/31)	Accelerating LLMs: Flash Attention	
11	Lecture 19 (11/5)	Accelerating LLMs: Kernel methods and other optimizations	
11	Lecture 20 (11/7)	Cluster of Accelerators – Modeling, Analysis, Challenges, Opportunities for Optimization	**PA 2 due** WA 3 out **Project Topics Finalized**
12	Lecture 21 (11/12)	Communication Primitives on Cluster	
12	Lecture 22 (11/14)	Distributed Data Parallel Algorithms on Cluster	
13	Lecture 23 (11/19)	Distributed Training on Cluster I	
13	Lecture 24 (11/21)	Distributed Training on Cluster II	**WA 3 due**
14	**Exam (11/26)**	**Final Exam**	
14	No Lecture (11/28)	Thanksgiving	
15	Lecture 25 (12/3)	Survey on AI Accelerators	

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

15	Lecture 26 (12/5)	Conclusion, Future Research Directions	
Final	(12/11)	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.