## Course Information

- **CS 59200-MLS** Machine Learning Systems
- **Instructional Modality:** In-person lectures (the first two weeks of the lectures will be on zoom).
- **3 credit hours**
- **Prerequisites:** Grad standing or permission of the instructor. It would be helpful to take a class on introduction of machine learning.

## Instructor(s) Contact Information


- **Instructor:** Xupeng Miao
- **Office Location:** TBD
- **Homepage:** https://hsword.github.io
- **Email Address:** xupeng@purdue.edu
- **Office hours:** TBD


## Course Description

Artificial intelligence (AI) techniques, especially recent advances in generative large language models (LLMs), have surpassed human predictive performance in a variety of real-world tasks. This success is enabled by the recent development of Machine learning (ML) systems (MLSys) that provide high-level programming interfaces for people to easily prototype different ML models on modern hardware platforms.

In this course, we will explore the design of modern ML systems by learning how an ML model written in high-level languages is decomposed into low-level kernels and executed across hardware accelerators (e.g., GPUs) in a distributed fashion. Topics covered in this course include: neural networks and backpropagation, programming models for expressing ML models, automatic differentiation, deep learning accelerators, distributed training techniques, computation graph optimizations, automated kernel generation, memory optimizations, etc. The main goal of this course is to provide a comprehensive view on how existing ML systems work. Throughout this course, we will also learn the design principles behind these systems and discuss the challenges and opportunities for building future ML systems for next-generation ML applications and hardware platforms.

## Learning Outcomes

By the end of the semester, you should be able to:

1. Learn how to read MLSys papers and conduct MLSys research.
2. Understand the state-of-the-art ML systems, especially for their design and research contributions.
3. Explore how to modify the implementation of existing ML systems and optimize their performance.

## Course Logistics

The instruction is mostly lecture-based. The class schedule will be posted and updated in Brightspace or the course website (TBD).

## Assignments

Grades will be based on class participation and the final project.

Reading assignments: 20%
Paper presentation: 20%
Course project: 50%
Class participation (Discussion): 10%

**Reading Assignments**

(Adapted from David Held's 16-881 Spring 2021 and Zhihao Jia's 15-849 Spring 2022)

Starting from the third week, we will be reading and discussing two to three papers during each class (the paper list will be posted with the class schedule). Every paper review deadline will be the time when our class for the corresponding paper discussion begins. Your paper reviews should consist of at least the following three paragraphs:

- 1 short paragraph summarizing the first paper, in your own words (do not copy sentences from the paper)
- 1 short paragraph summarizing the second paper, in your own words (do not copy sentences from the paper)
- 1 short paragraph on any connections you see between the papers, such as:
    - Compare and contrast
    - How one could apply ideas from one paper to solve the problem in the other paper
    - A new idea that would incorporate results from both papers etc

**Paper Presentation**

You will present one paper in the paper discussion class once a semester. You should submit your presentation slides (in PDF) before our class for your paper presentation begins.

**Course Project**

The course project will be completed by groups of 1-3 students. If necessary, we will provide some potential candidate project ideas in the area of machine learning systems. Still, you are also more than welcome to bring your own ideas that are related to your research. The Course project will have the following three components:

One-page proposal (5%)
Final course project presentation (15%)
Final course project report (30%)