CS 854 Advanced Topics in Computer Systems: Model Serving Systems for GenAI

Term: Fall 2024
Instructor: Hong Zhang
Email: honzhang@uwaterloo.ca
Class Meetings: TBA
Class Room: TBA
Office Hours: By appointment

Overview

We are witnessing an explosion of Generative AI (GenAI) applications. The latest GenAI models such as GPT-4 have achieved unprecedented performance in various tasks such as code generation, text classification, and problem reasoning. However, serving GenAI applications, i.e., deploying trained GenAI models on a compute cluster and conducting model inference for incoming user requests, presents challenges in system design.

This seminar-based course will introduce you to the key concepts and the state-of-the-art in model serving systems for emerging Generative AI (GenAI) and encourage you to think about either building new tools or how to apply an existing one in your own research.

The course will start with an overview of datacenters and cloud computing, which form the backbone of model serving systems. We will then introduce model serving systems for traditional DNN before the GenAI era. Finally, we will focus on various important topics for serving systems for GenAI, including efficient batching, memory and cache management, request scheduling and load balancing, and compound AI systems such as Retrieval-Augmented Generation. Note that this course is NOT focused on AI methods. Instead, we will focus on how one can build efficient serving systems for existing AI methods.

Paper review and presentation

Every student is expected to present one paper in this course, each presentation will be 30 min + Q&A/discussion. You can select one paper from the paper list [TBA], and you may also choose to present your preferred paper that covers other topics in the area of model serving system for GenAI. In the presentation, you should:

- Provide necessary background and motivate the problem.
- Present the high-level idea, approach, and/or insight (using examples, whenever appropriate) in the required reading as well as the additional reading.
- Discuss technical details so that one can understand key details without carefully reading.
- Explain the differences between related works.

- Identify strengths and weaknesses of the required reading and propose directions for future research.

Post-Presentation Panel Discussion

To foster a deeper understanding of the papers and encourage critical thinking, each lecture will be followed by a panel discussion. This discussion will involve two distinct roles, simulating an interactive and dynamic scholarly exchange.

Roles and Responsibilities

1. **The Authors**

- The group that presents the paper and the group that writes the summary will play the role of the paper's authors.
- Responsibility: As authors, you are expected to defend your paper against critiques, answer questions, and discuss how you might improve or extend your research in the future, akin to writing a rebuttal during the peer-review process.

2. **The rest of class**

- Critically assess the paper, posing challenging questions and highlighting potential weaknesses or areas for further investigation. Your goal is to engage in a constructive critique of the paper, simulating a peer review scenario.

Course Project

You will have to complete a term-long research project. The topic should be closely related to model severing systems for GenAI. It should be approved by the instructor and have original contribution. Surveys are not permitted as projects; instead, each project must contain a survey of background and related work.

You must meet the following milestones (unless otherwise specified in future announcements) to ensure a high-quality project at the end of the semester:

This course will have a term-long research project. The research project is expected to be novel and related to computer architecture and systems. Both individual and group projects are allowed but the expectations for a group project will be higher.

1. Project proposal:
   Students are expected to form groups of 1-2 members and finalize project groups by the time of the proposal submission. Each group needs to submit a proposal document to the instructor and present the proposed idea in class.

2. Mid-term checkpoint:
   This checkpoint aims to track the progress of each research project. Each group needs to present the main ideas of their project and demonstrate some initial results that support the ideas. Students are expected to provide feedback to other groups after each presentation.
3. Final report and presentation:
   Each group will submit their final report and present the project to the whole class at the end of the term. The final report will be in the form of a research paper. The score of the course project will be primarily based on the final report and presentation.

Grading

Grades for this seminar class will be calculated as follows:

10%	In-class discussion
10%	Paper reviews
30%	Paper presentations
50%	Course project

Course Schedule

The course schedule is available in (TBA). The schedule will be updated as the class moves forward.

Academic Integrity

This course strictly follows the academic integrity policy of the Faculty of Math. Any academic misconduct will be reported to the Office of Academic Integrity.

Specific to this course:

- Students should complete paper reviews on their own without referring to other students' ideas. Sharing paper reviews with other students before assignment submission is prohibited.
- Students may reuse or refer to the presentation from the authors but needs to explicitly mention the origin of the slides.
- Course projects should be original. Any ideas learned from other students or the literature needs to be properly cited.
- Collaboration and discussion are encouraged but every student or group (for project) needs to complete the assignment or project on their own.