CSE 40947/60947 Cross-Layer Design for CMOS and Beyond-CMOS Technologies

Class: Zoom Class Room (https://notredame.zoom.us/j/96857669879)
Monday & Wednesday 2:20pm - 3:35pm

Instructor: Dr. X. Sharon Hu
Zoom Office Room (https://notredame.zoom.us/j/5746316015)
574-631-6015, shu@nd.edu
Office Hours: Monday 1:00pm–2:00pm & Wed. 3:45pm–4:30pm or by appointment

Course Overview and Objectives

As Moore's Law based device scaling and accompanying performance scaling trends are slowing down, it is increasingly challenging for von Neumann architectures and traditional digital processing to meet the performance and energy targets for many demanding applications such as training large neural networks and extract information from huge amount of multi-modality data (image, video and audio, etc.). Emerging CMOS and beyond CMOS technologies are being actively investigated by both industry and academia. To take full advantage of these technologies, cross-layer design spanning from devices to circuits to architectures to algorithms is of critical importance.

This course reviews the state-of-the-art advances in cross-layer design techniques for employing CMOS and beyond-CMOS technologies in compute/data-intensive applications especially from the machine learning and big data areas. It focuses on identifying/modeling unique features of algorithms, architectures, circuits and devices and exploiting these features to improve overall application-level performance. Special attention will be given to addressing challenges including accuracy, energy, memory requirement, process variation, limited precision, and reliability. Some example technologies considered include deep sub-micron CMOS, resistive RAM, spintronics and ferroelectrics.

At the end of this course, you are expected to be able to do the following:

- Identify bottlenecks and special challenges of executing a given algorithm on a given hardware.
- Evaluate and benchmark different hardware in terms of application-level accuracy, latency, energy, reliability.
- Apply cross-layer design principles to model and design hardware for specific applications
- Use knowledge of the underlying hardware to develop more efficient algorithms.
- Conduct algorithm and hardware co-design
- Summarize and explain research results from journal/conference papers in the relevant areas.

Pre-requisits:

1. Undergraduate-level electronic circuit design
2. Undergraduate-level algorithm and computer architecture
3. Familiarity with basic machine learning and neural network algorithms

Reading Material:

1. There is no required textbook.
2. Relevant papers from leading conferences (e.g., International Symposium on Computer Architecture, Design Automation Conference, Conference on Neural Information Processing Systems, AAAI Conference on Artificial Intelligence) and journals (e.g., IEEE Transactions on Parallel and Distributed Computing, IEEE Transactions on Computers, IEEE Transactions on Knowledge and Data Engineering).

Course Format and Grading:

The lectures will be all online via zoom at the regular class time. Lectures will be recorded. However, students are strongly encouraged to attend lectures synchronously as most lectures will be conducted in a group meeting format with live discussions. Specifically, the lectures consist of a mixture of the following

- Instructor-led in-class discussions of the assigned reading material.
- Student presentations on the assigned conference/journal papers.
- Student presentations of the class projects.

The course grades will be determined by the following three components:

Class presentation and participation	25%
Homework assignment	15%
Final project (including the proposal, the progress and final report)	60%

Topics to be Covered

- Representative technologies (e.g., FinFETs, tunnel FETs, FeFETs, resistive RAM and other non-volatile devices)
- Representative compute/data-intensive applications (e.g., supervised and unsupervised learning, nearest neighbor, clustering)
- Traditional accelerator architectures and near- and in-memory computing architectures
- Device/circuit properties (e.g., process variation, noise, and endurance) that impact application-level properties (e.g., accuracy, latency and energy)
- Algorithm and hardware codesign techniques (e.g., data representation, resource allocation and mapping, online resource management, energy/latency/accuracy tradeoff)
- Cross-layer modeling and application-level benchmarking

Schedule

- Introducton: 2
- Technologies: 4
- DNN accelerators (inference and training): 4
- Alternative neural networks: 4

- Traditional accelerator architectures and near- and in-memory computing architectures
- Device/circuit properties (e.g., process variation, noise, and endurance) that impact application-level properties (e.g., accuracy, latency and energy)
- Algorithm and hardware codesign techniques (e.g., data representation, resource allocation and mapping, online resource management, energy/latency/accuracy tradeoff)
- Cross-layer modeling and application-level benchmarking

Course Policies:

- We may use video cameras during class and office hours, so please make sure to dress appropriately. For attending lectures, please use headphones or headsets to minimize disruption, and use a quiet place with a strong internet connection.
- Lecture notes are mainly for guiding in-class discussions and should not be used in place of the reading assignments.
- Reading assignments will be posted at least one week ahead (except the first lecture). Students are expected to do the reading assignment **before the lecture**.
- Homework should be submitted online at `Sakai` (unless specified otherwise) prior to the start of the class on the due date. Homework will be accepted up to two days after the due date. Late homework will receive a deduction of 20% of the total points received for each additional day. However, if a student abuses this privilege by routinely handing in homework late, the privilege will be withdrawn.
- Students will work in teams for the class project. Instructions on how to form a team will be given with the assignment. An evaluation form must be filled for each team-oriented assignment. More details will be given later.
- Inquiries about graded assignments will be accepted only if made **within 3 days** after they are handed back. Such inquiries should be made in writing, which clearly explains the complaints. Only after reviewing the written complaints, can the instructor make any grade adjustments.

Academic Integrity:

According to the University of Notre Dame Undergraduate Academic Code of Honor, "members of the University community are expected to embrace and adhere to the following pledge:

> *As a member of the Notre Dame community, I acknowledge that it is my responsibility to learn and abide by principles of intellectual honesty and academic integrity, and therefore I will not participate in or tolerate academic dishonesty.*"

No academic dishonesty in any form is tolerated. The University's Honor Code (http://honorcode.nd.edu/) reminds our community of our shared purpose both within the institute of academia and as members of a broader humanity; the statement also outlines policy violation procedures. Any questions regarding academic integrity, particularly regarding assignments in this course, should be directed to the instructor.

Privacy Practices in This Course:

This course is a community built on trust; in order to create the most effective learning experience, our interactions, discussions, and course activities must remain private and free from external intrusion. As members of this course community, we have obligations to each other to preserve privacy and cultivate fearless inquiry. We are also obliged to respect the individual dignity of all and to refrain from actions that diminish others's ability to learn. Please note the following course principles:

- **Using learning materials:** Course materials (videos, assignments, problem sets, etc) are for use in this course only. You may not upload them to external sites, share with students outside of this course, or post them for public commentary without my written permission.
- **Using live class recordings:** We are recording class meetings to provide everyone in the class with useful study aids. These recordings will be available for review through Sakai. The University strictly prohibits anyone from duplicating, downloading, or sharing live class recordings with anyone outside of this course, for any reason.
- **Sharing student information:** Our materials and activities may provoke argument or spirited discussion; some of us may volunteer sensitive personal information. Do not share others's personal information on sensitive topics outside of our course community. Student work, discussion posts, and all other forms of student information related to this course are private.
- **Sharing course information with others** Sharing private information about our course community (including discussions, activities, presentations, student work, etc) with others for the purpose of inviting external attention, intrusion, ridicule, or harassment is an egregious breach of trust.

Violating these principles will be handled according to the applicable academic honor code.

Diversity & Inclusion:

The University of Notre Dame is committed to social justice and diversity. I share that commitment and strive to maintain a positive learning environment based on open communication, mutual respect, and non-discrimination. In this class we will not discriminate on the basis of race, sex, age, economic class, disability, veteran status, religion, sexual orientation, color or national origin. Any suggestions as to how to further such a positive and open environment will be appreciated and given serious consideration.