Course Syllabus
ECE 4950/6950 Special Topics in ECE: AI Hardware
Fall 2026, TuThu 2:55pm-4:10pm, Upson 142

---

1. Course Information

Lectures:	Tuesday & Thursday 2:55pm-4:10pm, 142 Upson Hall
Website:	https://www.csl.cornell.edu/courses/ece4950fa26
CMSX:	https://cmsx.cs.cornell.edu
Ed:	https://edstem.org/us/courses/100429

Instructor:	Zhiru Zhang, zhiruz@cornell.edu
Office Hours:	Thursday 5:00-6:00pm, Online

Course Texts:
◦ Lecture slides/notes posted on the course website
◦ Austin et al., How to Scale Your Model, Google DeepMind, online, 2025.

Supplementary Materials:
◦ Required and supplementary readings posted with the course schedule.

2. Course Description and Objectives

Artificial intelligence (AI) is increasingly shaped by the capabilities and limits of the hardware systems used to train, serve, and specialize modern models. As deep neural networks (DNNs) and large language models (LLMs) continue to grow in scale, the design of efficient AI systems requires a working understanding of algorithms, models, arithmetic, memory hierarchy, parallel execution, and accelerator architecture. This course studies the principles and practice of AI hardware across both general-purpose and specialized platforms.

The course offers an introductory exploration of AI hardware design and analysis, connecting first-order performance models with hands-on implementation. Example topics include DNN computation, hardware specialization, roofline performance modeling, systolic arrays, TPUs, GPUs, kernel programming, LLM inference and serving, sparsity, quanti- zation, and specialized LLM accelerators. Through lectures, readings, homework, labs, and a final project, students will develop the skills to reason about AI hardware from workload characteristics through architecture and implementation, and to make informed design choices across performance, efficiency, programmability, and scalability.

2.1. Prerequisites

The required prerequisite for this course is ECE/ENGRD 2300 or equivalent background. Students are expected to have a strong understanding of digital hardware and basic knowledge of computer architecture. Prior knowledge of machine learning, deep learning, LLMs, or GPU programming would be helpful, although not required.

2.2. Target Audience and Learning Outcome

This course is open to graduate students and advanced undergraduates who are interested in understanding how modern AI workloads are mapped to efficient hardware systems and how hardware architecture influences AI model performance, cost, and energy efficiency.

Upon completion of this course, students will be able to: (1) Explain the core computational patterns in modern AI models; (2) Analyze and compare AI hardware architectures in terms of efficiency, scalability, and programmability; (3) Use kernel programming frameworks (e.g., Triton) to implement and optimize AI inference workloads; (4) Understand how model optimization techniques affect the accuracy–efficiency trade-off.

3. Course Organization

This course includes a combination of lectures, tutorials, homework assignments, lab assignments, two exams, and a course project.

3.1. Lectures

The lecture sessions will cover the following topics before the final project period begins. These topics are tentative and may be covered in a different order. Please refer to the course website for the detailed up-to-date lecture schedule.

Background and Modeling
Introduction .... 1 session
Basics of neural networks .... 1 session
Hardware specialization .... 2 sessions
Roofline analysis .... 1 session
AI Accelerator Architectures
Systolic arrays .... 1 session
TPUs .... 1 session
GPUs .... 3 sessions
Other accelerators .... 2 sessions
LLM Systems and Optimizations
LLM inference .... 2 sessions
Sparsity .... 1 session
Quantization .... 1 session
LLM serving .... 1 session
LLM training .... 1 session
Other Topics
Kernel programming .... 1 session
AI for hardware design .... 1 session

3.2. Participation

Students are expected to attend lectures and actively participate in technical discussions during regular lecture, in-class tutorials, and guest lectures.

3.3. Quizzes

There may be short pop quizzes during lectures to cover key topics discussed in the current or previous lecture, including required readings. If quizzes are used, the overall quiz grade will be determined by the average of all quizzes, *excluding the two lowest scores.*

3.4. Readings

There will be required readings associated with most lectures. Students are expected to complete the assigned readings before class and participate in the corresponding in-class discussion. Reading assignments will be posted on the course website.

3.5. Assignments

A total of eight assignments will be released in sequence on CMSX, including four homework assignments and four lab assignments. Homework assignments are designed to help students solidify the important concepts covered in lectures and readings. Lab assignments will provide hands-on experience with AI hardware design, analysis, and implementation. **Students enrolled in ECE 6950 are required to complete the optional component of each lab assignment and project when such an optional component is included.**

3.6. Exams

There will be an in-class prelim on Thursday, October 8. The exam will be open-book and open-notes. A final exam will also be held during the university's scheduled final exam period.

3.7. Course Project

Students will have about three weeks to work in small groups on a final project based on their interests, choosing from a list of suggested topics. The students will meet with the instructor and TAs during regular class hours to track project progress. The students are expected to submit the project report and source code by Friday, December 11.

4. Course Policy

This section outlines various policies regarding enrollment, grading, assignment submission, academic integrity, and accommodations for students with disabilities.

4.1. Enrollment

This course is open to graduate students and advanced undergraduates. Students are expected to take this course for credit. Hence auditors are not advised to attend lectures without the intent to enroll.

4.2. Grading

The following weighting scheme will be used to calculate the final letter grade:

Class participation – 4%
Quizzes – 6%
Assignments – Homework (24%) + Labs (21%)
Exams – 35% = prelim (15%) + final (20%)
Project – 10%

Please note that a student must at least satisfy the following minimum requirements in order to pass the course: (1) submit at least five assignments; (2) take both exams; (3) complete the final project. If a student fails to meet any of these criteria, then that student will automatically fail the course regardless of the actual numerical grade.

We use the Cornell University Grading System as the basis for determining grades.

4.3. Assignment Submission

For each assignment, students must submit the solution source code, analysis artifacts, and/or report to the CMSX system by 11:59pm Eastern Time on the due date. When a report is required, the report file must be in PDF format. **Late policy:** 3% off per assignment per day (e.g., late by three days means a 9% penalty); cannot be late by more than seven days.

4.4. Regrade Policy

All regrade requests must be submitted via a private post on Ed to the course staff. The request must state exactly what should be regraded and why. The regrade request has to be received within one week from when the grade in question has been posted.

4.5. Usage of Artificial Intelligence (AI)

The course policy on AI tool usage strategically leverages their strengths while maintaining academic integrity and individual learning. **AI tools may be used as a learning aid and for refining lab reports, but students are expected to complete the work of labs and homework assignments themselves. Limited use of AI tools for tasks such as clarifying concepts may be appropriate when it does not substitute for the student's own work.** For the final project, however, we will consider relaxing these restrictions to allow students to explore and make effective use of agentic AI tools for tasks such as design exploration and optimization, while still requiring students to understand, evaluate, and take responsibility for the resulting work.

4.6. Personal Conduct

The School of ECE expects students to be respectful and professional in all participation and communication. You are expected to maintain professional conduct and speech in all aspects of this course. Professional behavior demands you have a responsible and mature attitude in person and online. Disrespectful, unethical, and/or unprofessional behaviors will not be tolerated and can result in course failure and/or dismissal from the program.

4.7. Academic Integrity

Each student in this course is expected to abide by the Cornell University Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work.

The term "group" in this section refers to yourself if you work alone or to you and your partner in case of a group project. The work your group submits is expected to be the result of your group's effort only. The use of a computer in no way modifies the standards of academic integrity expected under the Cornell University Code of Academic Integrity. You are encouraged to study together and to discuss information and concepts covered in lecture with other students. You can give "consulting" help to or receive "consulting" help from such students. However, this cooperation should never involve one group having possession of a copy of all or part of the work done by some other group, including work from previous years. Should copying occur, both the student(s) who copied work from another student and the student(s) who gave material to be copied will automatically receive a zero on the work, and an extra penalty will be assessed, ranging anywhere from a deduction on the final grade to failure of the course and university disciplinary action. Please notice that this implies that at no time are you allowed to grant anyone but your group partner access to your computer files. Be sure to master the use of chmod and umask before starting to work on your projects. During exams, you must do your own work. Communication among students is not permitted during the exams, nor may you compare or borrow notes, copy from others, or collaborate in any way. You are strongly encouraged to read Cornell University's Code of Academic Integrity.

4.8. Accommodations for Students with Disabilities

In compliance with the university policy and equal access laws, the instructor is available to discuss appropriate academic accommodations that may be required for students with disabilities. Requests for accommodations are to be made during the first three weeks of the semester, except for unusual circumstances. Students are encouraged to register with Student Disability Services to verify their eligibility for appropriate accommodations.

4.9. Attestation

By registering for this class and accessing course materials through the course website, students agree to abide by University, College, Department, and Course policies.