Accelerator Architectures for Machine Learning (AAML)

Syllabus

Tsung Tai Yeh
Department of Computer Science
National Yang-Ming Chiao Tung University

Course overview

- Instructor: **Tsung Tai Yeh**
- TA team+:
  - Kyle Lin;
  - Wei-Hsiang Lu;
- Lecture: T34
- Location: ED-302
- Office Hour: 9 – 10 am Thursday
- My Office: EC 516
- Course web site:
  - https://reurl.cc/axR9Xl

Course website QR Code

Discussion Forum

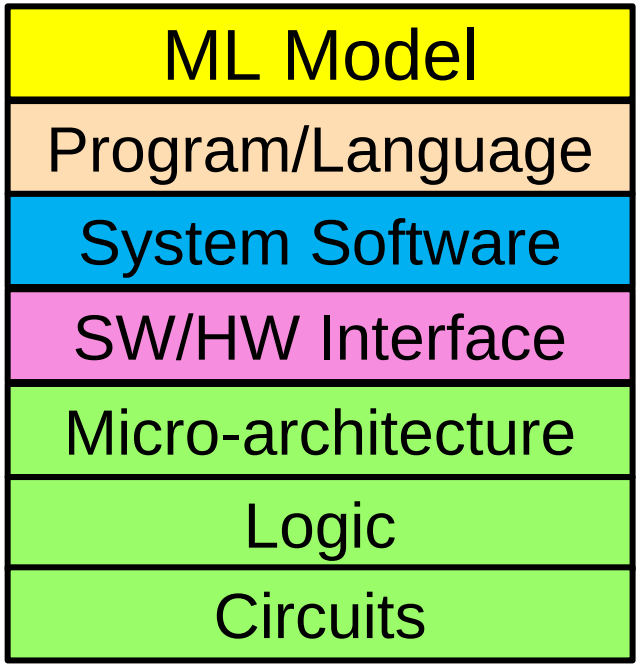
- Students should join our class discord discussion forum
- Discord forum
  - Course Announcement
  - Lab
  - Final Project

https://discord.gg/jQHyfFeD

Discord Forum QR Code

Course overview

- **Efficient Inference**
  - Basics of Deep Learning
  - Quantization + Model Pruning
- **AI Accelerator**
  - Digital/Analog AI Accelerators
- **Edge AI Acceleration**
  - TinyML Acceleration Architecture
- **Lecture + laboratory**
  - Class lecture + 5 labs + Final Project



Intended Lecture Outcomes (ILOs)

- AAML Course Intended Lecture Outcomes
  - **Understanding** the construction of DNN models
  - **Describing** details of AI accelerators
  - **Implementing** dataflow AI accelerator on Google CFU Playground
  - **Designing** AI accelerator to improve the performance of DNN models

What will you need to do in this course?

- Paper presentation (5%)
  - Groups of students present paper
  - Paper summary writing
- 5 Lab projects (55%) , Lab 1-2 (5%), Lab3-5 (15%)
  - Google CFU Playground
- 2 Quiz (20%)
- 1 Final Project (20%)
  - Optimize a Deep Neural Network Model on CFU Playground
  - Rule: 2 – 3 people/group

Prerequisites

- **Courses:**
  - Basic Programming , Computer Organization, Advanced Computer Architecture
- **You should:**
  - Basic understanding of computer architecture and digital logic design
  - Comfortable with programming in C/C++, Verilog and Python

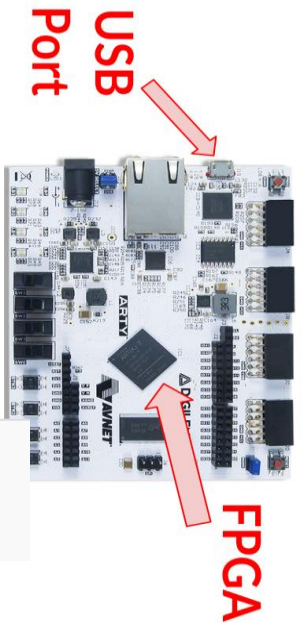
Lecture

- **Class lecture**
  - This lecture also covers three topics about AI accelerators and DNN models
  - Lecture (2 hours)– summarize course materials of each topic
  - Lab preview or paper presentation (1 hour)
  - Lecture materials have shown on the class website

Lab

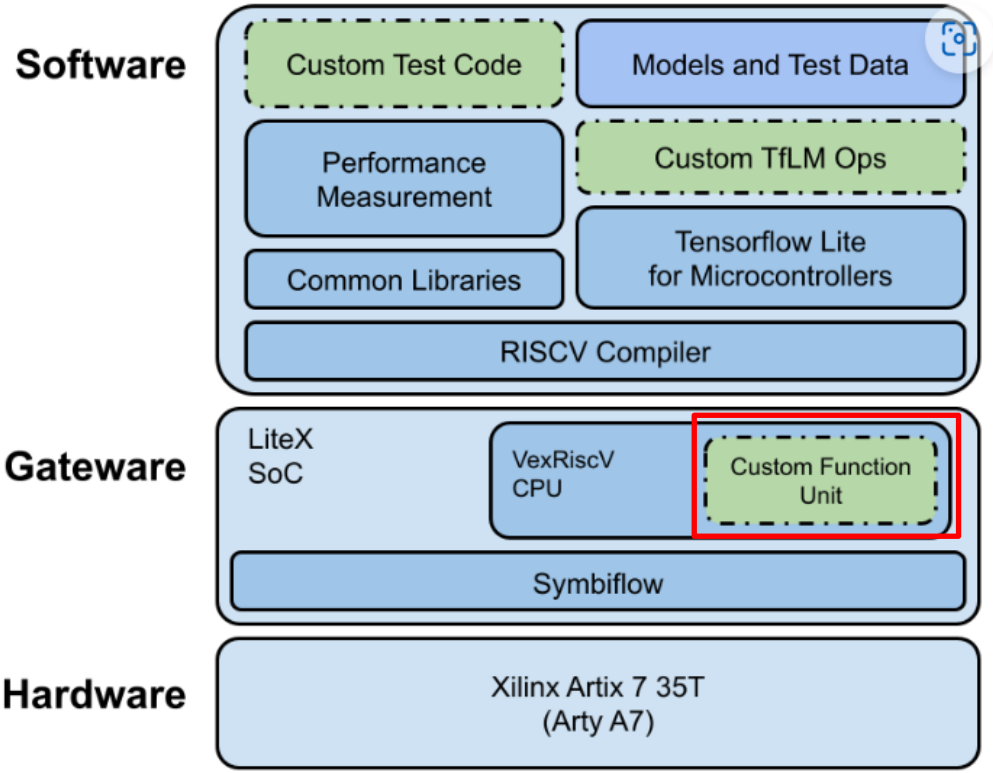
- **Platform**
  - **Google CFU Playground on Nexys A7-100T FPGA Board**
- Overview of AAML Labs
  - Build CFU + Run a model
  - CFU + (SIMD + Quantization)
  - Systolic Array Implementation (Verilog)
  - CFU + element-wise engine
  - Systolic Array + element-wise engine

https://nycu-caslab.github.io/AAML2025



CFU Playground

- **Google CFU Playground**
  - TensorFlow Lite for Microcontrollers (TFLM)
  - RISC-V CPU + Custom Function Unit
  - Simulation on FPGA



CFU Playground Overview

Lab

- One lab every two weeks
  - Lab 1-2 takes 5% each, 3-5 takes 15% each
  - Late submission:
    - 20% off for two weeks, 100% off for four weeks
- **Lab Demo**
  - Biweekly demonstration
  - Time: 12:00 – 1:00 pm on Thursday
  - Location: ED 302
  - TA will ask students questions about each lab assignment, 10% off if you do not answer TA's question correctly

Final Project

- The final project take **20%** score
- Problem:
  - How to optimize a Deep Neural Network Model on CFU Playground
  - Designing an AI accelerator to improve the performance of a DNN model by using CFU playground

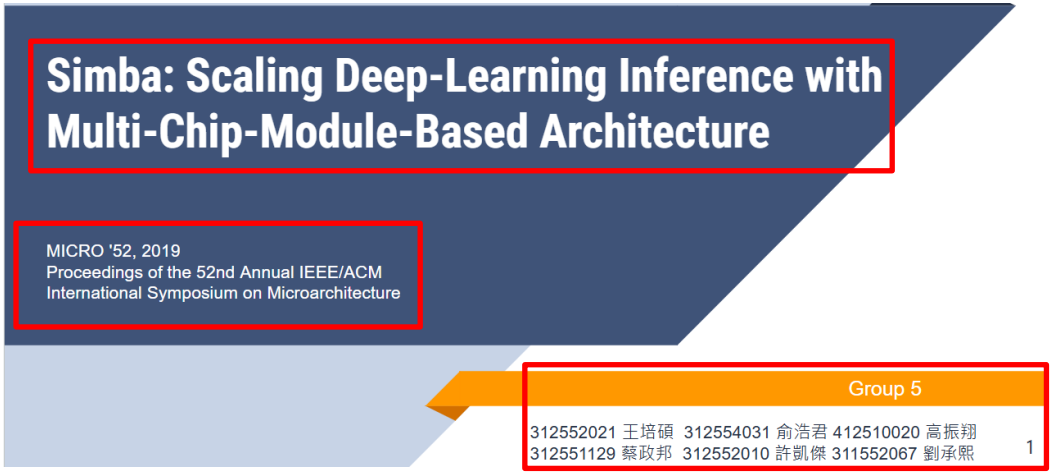
Paper Presentation

- **Paper Presentation**
  - 7 papers, 5 - 7 students are responsible for the presentation of one paper (30 – 40 mins)
  - Upload paper slide to discord “paper-collection” channel before the presentation
  - Peer review feedback form – students need to fulfill **5** times attendance, 1% score off when you less than 5 times attendance
  - Each paper presentation takes **5 %** of the total score

Paper Presentation Slide

- The paper presentation slide should include:
  - Paper Title
  - The origin of the paper and year
  - Name of presenters
  - Research problems
  - Contributions and outcome
  - Methodology
  - Evaluation

Paper Presentation Slide Template



Outline

- Introduction (Research problem)
- Research Background (Related work)
- Contribution
- Design
  - Designing SNAFU to Flexibility
  - Designing SNAFU to Minimize Energy
  - SNAFU-ARCH : a Complete ULP System W / Cgra
- Evaluation
- Conclusion

Schedule

Week	Date	Lecture Topics	Paper Report	Lab Deadline	Misc.
1	9/4	Basics of AI Accelerator[pdf]	Syllabus[Syllabus]		Build AI Silicon
2	9/11	Large Language Model			CFU Playground
3	9/18	Quantization		Lab 1	
4	9/25	Pruning and Sparsity	BitMOD, HPCA, 2025[pdf]		
5	10/2	Systolic Accelerator		Lab 2	
6	10/9	Digital AI Accelerator	FIGLUT, HPCA, 2025 [pdf]		
7	10/16	GPGPU Architecture		Lab 3	
8	10/23	Midterm Exam	GARDE, ISCA, 2025 [pdf]		
9	10/30	GPU Tensor Core	RTSpMSpM, ISCA, 2025 [pdf]		
10	11/6	Sparse DNN Accelerators		Lab 4	
11	11/13	Chiplet Accelerator	vAttention, ASPLOS 2025 [pdf]		
12	11/20	Analog ML Accelerator	floatap, MICRO 2024 [pdf]		
13	11/27	Machine Learning Compiler		Lab 5	
14	12/4	Emerging AI Accelerator	ATiM, ISCA 2025 [pdf]		
15	12/11	Final Exam			
16	12/18	Final Project Due			Invited talk (AMD)

Textbook

- Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, **Efficient Processing of Deep Neural Network**, Morgan and Claypool Publisher, 2020
- You can download the e-book from NYCU library through EBSCOhost E-book database within NYCU campus