



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⁺:
- Lecture: Thursday 10:10 am– 12 pm
- Location: ED-301
- Office Hour: 9 – 10 am Thursday
- My Office: EC 516
- Course web site:
 - <https://reurl.cc/4Y6xK2>



Course website QR Code



Discussion Forum

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

<https://discord.gg/9UpRBeSgBn>

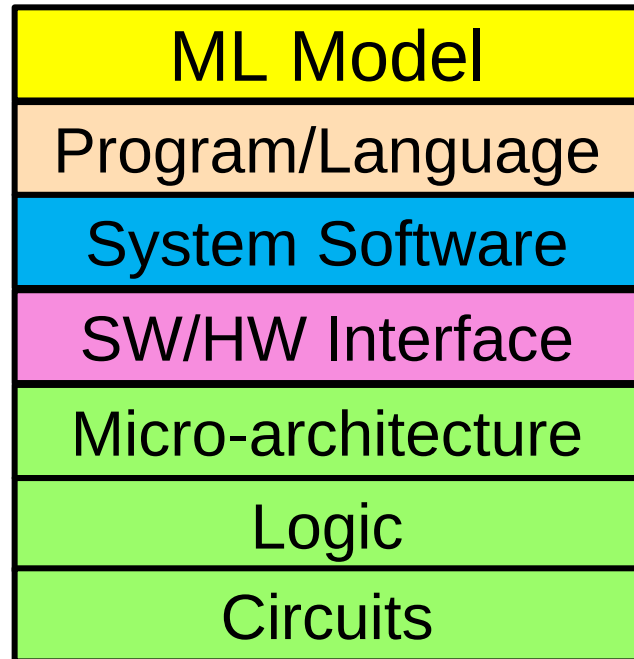


Discord Forum QR Code



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 TinyRISC-V SoC Platform
 - **Designing** AI accelerator to improve the performance of DNN models



What will you need to do in this course?

- Paper presentation (5%)
 - Paper summary writing
- Takeaway Questions (5%)
- 5 Lab projects (60%) , Lab 1-3 (10%), Lab4-5 (15%)
 - TinyRISC-V SoC Platform
- 1 Quiz (10%)
- 1 Final Project (20%)
 - Optimize a DNN Model on TinyRISC-V SoC Platform
 - 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

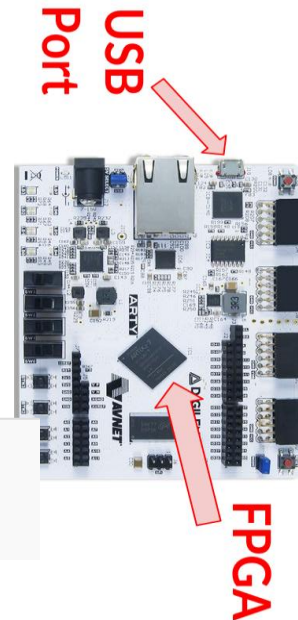
- **Platform**

- **TinyRISC-V SoC Platform on Arty A7-100T FPGA Board**

- **Overview of AAML Labs**

- Aligned AXI Data on SIMD
 - AXI4 Burst Reads for a SIMD
 - Systolic Array Dataflow
 - Systolic Array + element-wise engine
 - Co-Processor Cache Coherence

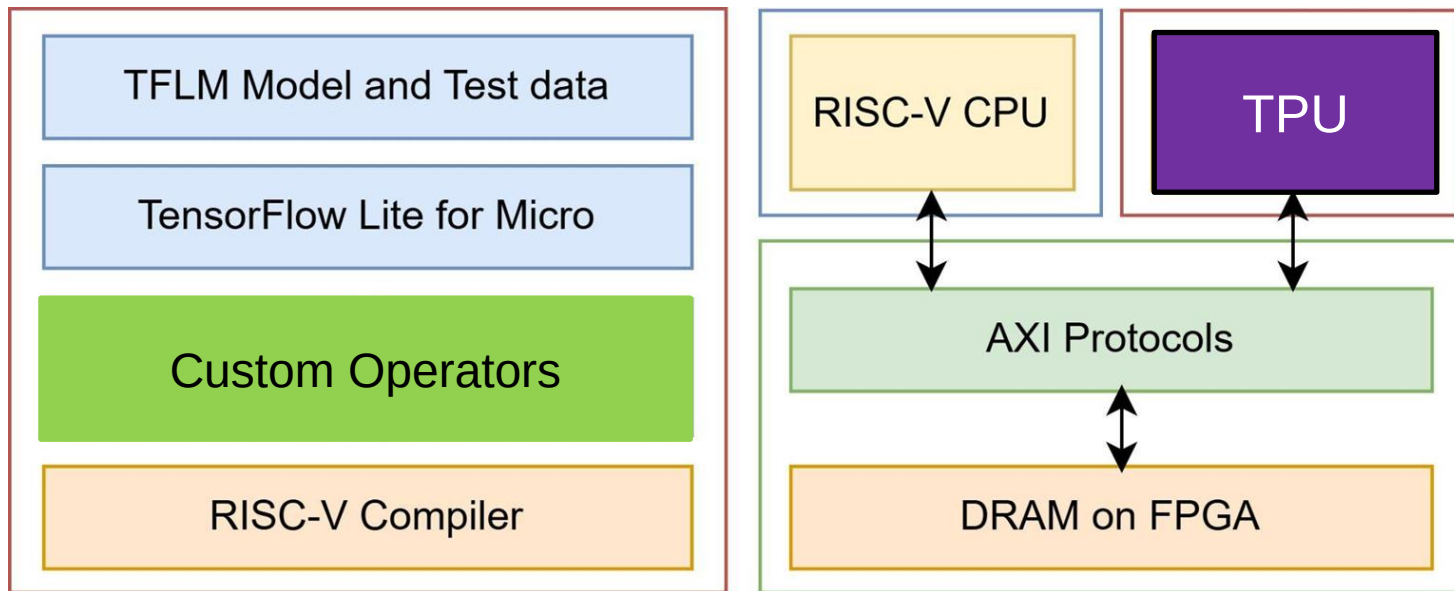
<https://nycu-caslab.github.io/AAML2026>





TinyRISC-V SoC Platform

- TensorFlow + RISC-V CPU + TPU





Lab

- One lab every two weeks
 - Lab 1-3 takes 10% each, 4-5 takes 15% each
 - Late submission:
 - 10% off per week, 100% off for four weeks
- **Lab Demo**
 - Biweekly demonstration
 - Time: 12:00 – 1:00 pm on Thursday
 - Location: ED 301
 - 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 TinyRISC-V SoC Platform
 - Designing an AI accelerator to improve the performance of a DNN model by using TinyRISC-V SoC Platform



Paper Presentation

- **Paper Presentation**

- 7 papers, < 10 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 **6** times attendance, 1% score off when you less than 6 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
 - Conclusion
 - 3 Takeaway Questions



Paper Presentation Slide Template

Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture

MICRO '52, 2019
Proceedings of the 52nd Annual IEEE/ACM
International Symposium on Microarchitecture

Group 5

312552021 王培碩 312554031 俞浩君 412510020 高振翔
312551129 蔡政邦 312552010 許凱傑 311552067 劉承熙

1

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



Takeaways Questions

- **Takeaways Questions**

- Student needs to get involve in takeaways questions in each class
- Students need to fulfill **at least 13** times attendance, 1% score off when you less than 13 times attendance
- The takeaway question takes **5 %** of the total score



Schedule

Course Outline

Week	Date	Lecture Topics	Paper Report	Lab Deadline	Misc.
1	9/10	Basics of AI Accelerator [pdf]	Syllabus [Syllabus]		Build AI Silicon
2	9/17	Large Language Model			CFU Playground
3	9/24	RISC-V for AI	BitDecode, HPCA, 2026 [pdf]		
4	10/1	RISC-V VPU		Lab 1	
5	10/8	Tensor Processing Unit	Titan, MICRO, 2025 [pdf]		
6	10/15	Neural Processing Unit		Lab 2	
7	10/22	Emerging AI Accelerator	LiLo, HPCA, 2026 [pdf]		
8	10/29	GPU Architecture		Lab 3	
9	11/5	Midterm Exam	LATPC, MICRO, 2025 [pdf]		
10	11/12	GPU Tensor Core		Lab 4	
11	11/19	Sparse Accelerator	SegFold, ISCA 2026 [pdf]		
12	11/26	Chiplet Accelerator		Lab 5	
13	12/3	CIM Accelerator	Delegato, MICRO, 2025 [pdf]		
14	12/10	Invited Talk (UCR)			
15	12/17	Invited Talk (SNU)			
16	12/24	Final Project Due			



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

🔒 | <https://ermg.lib.nctu.edu.tw/cgi-bin/er/swlink.cgi>

註：離線下載（離線）須註冊個人帳戶才可使用。

2. 儲存/列印的頁數有限制，每本不一。
註：若已使用到上限，欲克服頁數限制，清除瀏覽器紀錄並開啟，即可重新計算頁數。

13 **EBSCOhost Interface**

EBSCOhost 為 EBSCO Publishing 公司於 1994 年所發展之線上資料庫檢索介面系統，主要提供綜合學科、商管財經、生物醫護、人文歷史、法律等期刊之電子全文資料庫，以及部分當今全球知名之索引摘要資料庫。
涵蓋資料庫如：ASP、BSC、CMMC....等，可個別檢索單一資料庫，亦可整合檢索多種（或全部）資料...[more](#)