課程名稱
人工智慧晶片設計與應用
AI-on-Chip for Machine Learning and Inference
3 學分

開課教師: 陳中和， 劉峻誠，耐能智慧業師 陳宇春

開課日期
108 學年第二學期 (2020, Spring semester)

Intended students: Undergraduate senior, graduate students, EE or CSE major.
(大學部 4 年級與碩士生)

Prerequisites:
Digital logic design
Computer organization
Familiar with C, or C++, or Verilog

Objectives:
To teach students, computer architecture and programming for machine learning and inference, neural net model quantization and optimization, edge AI accelerator design and AI application development, case study on contemporary machines.

Grading Policy:
Exam: 2 x 20% = 40 %
LABs and Projects: 60%

Class Handout:
AI-on-chip system overview
Machine learning hardware design
1D-PE design for convolution operation
Instruction level and thread level parallelism machines
OOO Processor, multi-threading
Data level parallelism machine, SIMD ISA
Vector, GPU architecture
Overview of CASLab GPU system
Architecture of CASLab GPU

Supplemental class materials:
Introduction of OpenCL
Introduction of CASLAB GPU Compiler

Reference Textbook:

1. Computer architecture, A Quantitative Approach, by John Hennessy and David Patterson, 5/6th edition
2. Deep Learning-Hardware Design (深度學習-硬件設計) 劉峻誠, 2020.

LABS:
LAB1: ML Tool Introductions and Installations (GPU)
LAB2: Implement Lenet-5 model in Tensor-flow (GPU)
LAB3: Kneron Accelerator Platform (KAP) and SDK (AI Accelerator)
LAB4: AI model on Kneron KAP (AI Accelerator)
LAB5: OpenVINO and Intel Movidius (AI Accelerator)
LAB6: OpenCL Exercises on CASLab GPU (GPU)

Projects:
P1: Implement 1D PE convolution accelerator
P2: Propose an application implementation using Kneron KAP and Intel Movidius

Class: Wednesday, 2:10 PM To 5 PM.
EEB classroom: 92277

3/4 (陳中和)
Lecture: Overview of AI-on-Chip (2 hours)

Lab1: ML Tool Introductions and Installations (one hour)

3/11 (劉峻誠 or 業師陳宇春)
Machine learning hardware design (2 hours)
LAB2: Implement Lenet-5 model in Tensor-flow (GPU) (one hour)

3/18 (陳中和)
Lecture: 1D PE design for convolution layer (3 hours)
Announcement of AI accelerator design Project P1.

3/25, 4/1　(劉峻誠 or 業師陳宇春)

Total 6 hours

Lecture: Case study, Introduction of Kneron NPU

LAB3: Kneron Accelerator Platform (KAP) and SDK (AI Accelerator)

LAB4: AI model on Kneron KAP (AI Accelerator)

Announcement of AI application implementation project P2.

4/8 (陳中和)

Lecture: Instruction level/thread level parallelism machine

4/15 (王宗業)

Intel OpenVINO and Intel Movidius (2 hours)

LAB 5: OpenVINO and Intel Movidius　(AI Accelerator) (one hour)

4/22 Exam 1 (陳中和)

4/29 (陳中和)

Lecture: OOO processor, MIMD processor

Advanced computer architecture for instruction level and thread level parallelism applications

5/6 and 5/13 (陳中和)

Total 6 hours

Lecture: Data level parallelism machine, SIMD ISA

Lecture: Vector, GPU architecture

Video for discussion

Accelerating AI: Past, Present, and Future by Krste Asanovic

https://www.youtube.com/watch?v=8n2HLp2gtYs

5/20: Exam 2 or Labs (陳中和), ISCSAs Conference

5/27, 6/3 (陳中和)

Lecture: CASLAB GPU, overview, architecture (2 hours)

LAB6: OpenCL Exercises on CASLab GPU (GPU) (one hour)

Lecture: CASLab GPU software stack, … (3 hours)

6/10 (劉峻誠)

Lecture: AI-on-Chip advanced topics (3 hours)

6/17 and 6/24 (陳中和，劉峻誠，業師陳宇春)

Total 6 hours

Student final project presentation

Final exam week (week 18)