ECE 587 – Hardware/Software Co-Design
Lecture 18 Hardware Accelerators and CUDA

Professor Jia Wang
Department of Electrical and Computer Engineering
Illinois Institute of Technology

March 24, 2025

Outline

Hardware Accelerators

Reading Assignment

- This lecture: Hardware Accelerators and CUDA
- Next lecture: Neural Networks and Systolic Array

Outline

Hardware Accelerators

Hardware Acceleration

- (Much) better performance, performance per cost and/or per power/energy than general purpose processors.
  - On specific applications: data analytics, deep learning, bioinformatics, etc.
  - In specific environments: cell phone, cloud, data center, etc.
- (Much) less NRE cost and shorter time-to-market than ASIC.
  - Use commercial off-the-shelf hardware platform.
  - Provide flexibility in functionality via software.

Hardware Accelerators

- Independent accelerator
  - A standalone device with its own processor and memory that executes specifically made binaries.
  - Communicate with main processors (host) through a device bus or via networking interfaces.
  - E.g. GPUs.
- Coprocessor-based accelerator
  - Integrate with host processors via internal bus and execute special instructions.
  - Access shared host memory via coherence protocols.
  - E.g. Rocket Custom Coprocessor (RoCC) accelerators.
- Considerations
  - Programming model: explicit data transfer vs. shared memory, scratch pad vs cache, etc.
  - Memory heirarchy optimization
  - How to interconnect accelerators to fit larger neural network models?