Welcome !

Domain Specific Architectures
CSCE 4013/5013

David Andrews
Rm 527 JBHT
dandrews@uark.edu
CSCE University of Arkansas

Office Hrs: 3:00 – 4:00 MWF
Class website:
https://hthreads.github.io/classes/#eecs-5013-domain-specific-accelerators

Course Overview

- What is a Domain Specific Architecture ?
- What we study during the course
- What will be your involvement
- How you will be graded

What is a Domain Specific Architecture ?

A Hardware Architecture: Designed for a Specific Domain of Applications. Examples:

- *Graphics*
- *Image Processing*
- *Deep Learning*

General Purpose Architectures: Designed to be flexible enough to do everything but not optimal for anything ☺

- What was first GP uProcessor ?
- Turing Complete ?

What is a Domain Specific Architecture ?

Generally Accompanied by a Domain Specific Language:

Python, OpenCL

Pytorch, Tensorflow

Allows expression of the common types of parallelism within the domain

-AI domain dominated by Large Matrix Operations:

i.e., (SIMD) Data level Parallelism

DSL's are good at what they are targeting but are not general purpose

What is a Domain Specific Architecture ?

Domain Specific Accelerators exploit four main techniques to get performance and efficiency:

1. Data Specialization: Specialized ops on domain specific data types. Can do in one cycle what may take tens of cycles on GP computer.
2. Exploit Parallelism: Match what is available in the application:
    1. Locality of reference is key
    2. global memory references severely degrade performance

What is a Domain Specific Architecture ?

3. Local and Optimized Memory: Store highly used data structures in small high bandwidth memories close to processing units.

Increase Energy Efficiency

Decrease Processing Latency

4. Reduced Overhead: Specialized hardware and Languages decrease overhead of program interpretation and reduces #instructions.

GP Proc expends ~90% of energy on overhead:

<IF, ID, Data Supply, control>

How Important is Memory Design?

	Unit	Area (mm^2)	(%)	Power (W)	(%)
GACT	Logic	17.6	20.5	1.04	23.6
	Memory	68.0	79.5	3.36	76.4
D-SOFT	Logic	6.2	1.8	0.41	4.4
	Memory	320.3	98.2	8.80	95.6
EIE	Logic	2.8	6.9	0.23	40.3
	Memory	38.0	93.1	0.34	59.7

Area and Power of most accelerators dominated by Memory

-Performance often memory limited

Accelerator Costs

Op	Energy	Area
8-bit Add	10 fJ	4 um^2
Small (8 Kbyte) SRAM Local	50 fJ/bit	.013 um^2 per bit
Larger (100 MB) SRAM Local	.7 pJ/bit	
Global memory	4 pJ/bit	
Local Comm (on Chip)	100fJ/bit-mm	Linear Increase
Global Comm (off Chip)	10 pJ/bit	

The Big Three:

- Graphics Processing Units (GPUs)
  - NVIDIA ~88%. (~98% of data center market)
  - AMD ~12%
  - Intel ~0%
- Field Programmable Gate Arrays (FPGAs)
  - Xilinx -> AMD
  - Altera -> Intel -> Altera (split being completed)
- Application-Specific Integrated Circuits (ASICs)
  - Google: Tensor Processing Unit (TPU)
  - Microsoft: Athena this year
  - Amazon Web Services:
  - Some Interesting Startups: Cerebras, Groq

What will we study ?

Review of key concepts and technology trends

Array Processors/Systolic Arrays

Processor near/in Memory architectures

Case Studies

Crystal Ball gazing:

What is Your Responsibility?

Advanced Senior Level/Graduate Class: Topics and technologies continue to develop. Materials are from Conferences/Journals and not textbooks.

-Attend Class!

-Read papers before we discuss in class

-Attend class!

-Come prepared to engage in discussions

-Attend Class!

How will you be graded ?

Presentations:	30%
Quizzes:	30%
Participation:	10%
Final Project:	30%