CS/EE 147 – GPU Computing and Programming

Introduction

Abbas Mazloumi
amazl001@ucr.edu

Thanks to Daniel Wong, Marcus Chow
Welcome!
About Me

• Abbas Mazloumi
• A 5th year Ph.D. student in CS
• Goal: working in academia
• Care for your health and safety
• Email: amazl001@ucr.edu
• Link: https://www.cs.ucr.edu/~amazl001/
About Me but Academics

• University of Tehran, Iran
  • MS Computer Architecture ’14

• University of California, Riverside
  • 5th year Ph.D. student
About Me but Research

Architectural & Algorithm solutions for high performance computing

Improving Interconnection communication between cores and memories
- power
- performance

Efficient running of graph processing algorithms on clusters of multiple machines:
- Each machine can be a CPU or GPU
Tell Me About You!
https://pollev.com/abbasmazloum123
What is this Course?
Course Goals

• Learn about calculating performance characteristics
• Understanding the motivation why the GPU is used at all
• How the motivation influenced the design of GPU architecture
• How the design of the architecture leads to specific programming paradigms
• How to program “Optimal” programs using these paradigms

• Technical subjects
  • principles and patterns of parallel algorithms
  • processor architecture features and constraints
  • programming API, tools and techniques
Logistics

• Course Website
  • [https://www.cs.ucr.edu/~amazl001/teaching/cs147/S21/](https://www.cs.ucr.edu/~amazl001/teaching/cs147/S21/)
  • Check often for announcements

• Assignments/Projects/announcements
  • iLearn (iLearn.ucr.edu)

• Discussion/Help
  • Piazza

• ENGR Account Setup
  [https://www.engr.ucr.edu/secured/systems/login.php](https://www.engr.ucr.edu/secured/systems/login.php)
Attendance/Grading

• Attendance
  • Do your best to attend all lectures.
  • However, you are still responsible for keeping up with recorded course lectures and assignments.

• Grade Breakdown
  • Homework/Labs: 25%
  • Midterm 1&2: 50%
  • Class Participation: 10%
  • Quizzes 15%
  • Extra Credit: 3%
Lab Policies

• 3 slip days
• 15% penalty per late day
• All labs/projects are due at the end of the due date (midnight)
• Projects should be uploaded to iLearn and github classroom
Contact

• Instructor: Abbas Mazloumi
  • Email: amazl001@ucr.edu
  • Homepage: https://www.cs.ucr.edu/~amazl001/
  • Office: Online
  • Office Hours: TBD

• TA: Mohammadeza Rezvani
  • Email: mrezv002@ucr.edu
  • Office Hours: TBD
Discussion for Wednesday

• Why do you think we teach this class at all? The reason why we use GPUs today
• Think about why you wanted to take this class? Deeper than course requirement
GPU Introduction
42 Years of Microprocessor Trend Data

Transistors (thousands)

Single-Thread Performance (SpecINT x 10^3)

Frequency (MHz)

Typical Power (Watts)

Number of Logical Cores

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp
Modern computer architecture is limited by:

- Process Technology / Transistor density
  (End of Moore’s Law)
- Power
  (End of Dennard Scaling)
- Temperature
General-purpose (Easier to program)

Single-core CPU | Multi-core CPU | GPU | FPGA | ASIC

Specialized

Energy-efficient
Examples of Specialization

Bitcoin Mining
Examples of Specialization

Microsoft Catapult
Examples of Specialization

Google TPU - Tensor Processing Unit
## Examples of Specialization

### GPUs - Supercomputers

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Rpeak (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>Summit - IBM Power System, IBM POWER9 22C 3.07GHz, NVIDIA V100, Dual-rail, Mellanox EDR Infiniband</td>
<td>2,282,544</td>
<td>122,300.0</td>
<td>187,659.0</td>
<td>8,856</td>
</tr>
<tr>
<td>2</td>
<td>National Supercomputing Center in Wuxi, China</td>
<td>Sunway TaihuLight - Sunway MPP, Sunway SW6010 286C 1.45GHz, Sunway NECPC</td>
<td>10,649,600</td>
<td>93,014.6</td>
<td>125,325.9</td>
<td>15,371</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA/LANL, United States</td>
<td>Sierra - IBM Power System, S2Z2L, IBM POWER9 22C 3.1GHz, NVIDIA V100, Dual-rail, Mellanox EDR Infiniband</td>
<td>1,972,480</td>
<td>71,610.0</td>
<td>119,193.6</td>
<td>11,378</td>
</tr>
<tr>
<td>4</td>
<td>National Supercomputer Center in Guangzhou, China</td>
<td>Tianhe-2A - TH-IVB-PEF Cluster, Intel Xeon E5-2650V2 12C 2.5GHz, TH Express-3, Matrix-2000, NUOT</td>
<td>4,991,740</td>
<td>61,444.5</td>
<td>100,678.7</td>
<td>18,482</td>
</tr>
<tr>
<td>5</td>
<td>National Institute of Advanced Industrial Science and Technology (AIST), Japan</td>
<td>Bridging Cloud Infrastructure (ABC) - PRIMERGY CX3350 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 Nvidia, Infiniband EDR, Fujitsu</td>
<td>391,680</td>
<td>19,880.0</td>
<td>32,576.6</td>
<td>1,169</td>
</tr>
<tr>
<td>6</td>
<td>Swiss National Supercomputing Centre (CSCS), Switzerland</td>
<td>Piz Daint - Cray XC30, Xeon E5-2680v3 12C 3.5GHz, Aries interconnect, NVIDIA Tesla P100, Cray Inc.</td>
<td>564,760</td>
<td>19,592.0</td>
<td>25,323.3</td>
<td>2,972</td>
</tr>
<tr>
<td>7</td>
<td>DOE/SC/Oak Ridge National Laboratory, United States</td>
<td>Titan - Cray XE6, Opteron 6274 14C 2.20GHz, Cray Gemini interconnect, NVIDIA K20e, Cray Inc.</td>
<td>540,640</td>
<td>17,590.0</td>
<td>27,112.5</td>
<td>8,359</td>
</tr>
<tr>
<td>8</td>
<td>DOE/NNSA/LANL, United States</td>
<td>Sequoia - BlueGene/Q, Power I93, 16C 1.4GHz, Custom IBM</td>
<td>1,972,844</td>
<td>17,172.2</td>
<td>28,123.7</td>
<td>7,810</td>
</tr>
</tbody>
</table>