ECE 277, FALL 2020
GPU Programming

Department of Electrical and Computer Engineering
University of California, San Diego

Instructor: Cheolhong An

Time and Location: TuTh 5:00-6:20pm, Zoom meeting

Contact: chan@eng.ucsd.edu
Office Hour: Fri 9:00 am - 10:00 am, Zoom meeting

Course TA: Jiawei Duan
Contact: jduan@eng.ucsd.edu
TA Hour: Wed 1:00 pm - 2:00 pm, Zoom meeting

Remote access to lab computers (EBU1-4309)

Go to https://guac.ucsd.edu and login if necessary.
Expand "EBU1 4309" folder and select your computer.

Objectives

This course introduces basic CUDA programming skills for parallel data processing. Topics cover parallel CUDA programming including efficient memory access and multithread programming under heterogenous CUDA programming environment. Especially, this course focus on hands-on learning by multiple example programs and labs including AI multiagents to learn a mine game environment.

Prerequisites

C/C++ programming skill (require hands-on experiences, ECE-15, ECE-17)
Computer architecture (ECE-30 or eqivalent)
Development environment (CMake, Visual studio)

Recommendations

Machine learning, Pytorch

Textbook

John Cheng, Max Grossman, Ty McKercher, "Professional CUDA C Programming", Wiley, 2014. ISBN: 978-1-118-73932-7

Lab Projects on GPU

1. Setup reinforcement learning environment
2. Single agent reinforcement learning
3. Multi-agent reinforcement learning

Midterm lab exam.

Accelerate the parallel Qlearning: Competition

Final project

Extend lab projects or speed up your domain problem using CUDA parallel processing. This is an individual project or a team project up to two.

Grading

- Take-home Quiz (35)
- 3 Homework (30)
- Midterm lab exam. (25)
- Final project. (10)

Course Outline

1. Course introduction
2. Introduction to GPU architecture

   Quiz: set up CUDA class lab projects

   (Ch.1, Ch.2 and Ch.3)

   class lab 1: hello GPU, thread, warp, threadblocks, etc.
3. GPU profile and debugging

   (Ch.10 implementation considerations)

   class lab 2: Nsight VS debugger and profiler, NVTX for joint CPU and GPU profile
4. GPU hierarchical memory: global memory

   (Ch.4 global memory, Ch.5 global memory access)

   case study: Single agent Qlearning
5. GPU hierarchical memory: global memory

   class lab 4: GMEM allocation and, aligned and coalesced access

   case study: Multiagent Qlearning

6. GPU hierarchical memory: shared memory

   (Ch.5 shared memory)

   class lab: SMEM allocation and bank conflict

   case study: Convolution, CuDNN TF, Pytorch and GEMM

7. Mid-term

   Accelerate the parallel Qlearning competition

8. Coarse-grained parallel programming: dma, zero copy, multi-stream

   (Ch.4 memory management, Ch.6)

   class lab: dma, zero copy, multi-stream, synchronization and events

9. Applicatoins

   Binding CUDA and Python: PyCUDA, Numba, and Pybind11

   ML environment: Pytorch C++ frontend and custom CUDA for CNN training and testing.

10. Final project presentations.

References

1. Anton Obukhov and Alexander Kharlamov, Discrete Cosine Transform for 8x8 Blocks with CUDA, Nvidia technical report, 2008

2. Victor Podlozhnyuk, Image Convolution with CUDA, Nvidia technical report, Sep. 2013

3. Michael McCool, Arch Robinson and James Reinders, Structured Parallel Programming : Patterns for Efficient Computation, Morgan Kaufmann, 2012

4. CUDA by Example
   `http://developer.download.nvidia.com/books/cuda-by-example/cuda-by-example-sample.pdf`

5. CUDA C BEST PRACTICES GUIDE
   `http://docs.nvidia.com/cuda/pdf/CUDA_C_Best_Practices_Guide.pdf`

6. NVIDIA, CUDA C PROGRAMMING GUIDE, Jun. 2017

7. Mark Harris, Introduction to CUDA C,NVIDIA Corporation, 2013,

8. CHRISTOPH ANGERER and JAKOB PROGSCH, CUDA OPTIMIZATION WITH NVIDIA NSIGHT VISUAL STUDIO EDITION, NVIDIA, 2016

9. NVIDIA, NVTX Library
   `http://docs.nvidia.com/gameworks/content/developertools/desktop/nsight/nvtx_library.htm`

10. NVIDIA, CUDA COMPILER DRIVER NVCC, Jan. 2017 http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf

11. NVIDIA, CUDA virtual assembly language (PTX) http://docs.nvidia.com/cuda/pdf/ptx_isa_5.0.pdf

12. NVIDIA, USING INLINE PTX ASSEMBLY IN CUDA, Feb. 2011

13. CURAND LIBRARY, programming guide, NVIDIA, 2017 http://docs.nvidia.com/cuda/pdf/CURAND_Library.pdf

14. Tuning CUDA Applications for Pascal http://docs.nvidia.com/cuda/pascal-tuning-guide/index.html#axzz4TM6Fo2tO

15. NVIDIA Nsight Visual Studio endition (CUDA Debugger/Profiler for Visual Studio) https://developer.nvidia.com/nvidia-nsight-visual-studio-edition

16. NVIDIA Nsight Eclipse edition (CUDA Debugger/Profiler for Eclipse) https://developer.nvidia.com/nsight-eclipse-edition

17. Whitepaper: GeForce GTX 1080 http://international.download.nvidia.com/geforce-com/international/pdfs/GeForce_GTX_1080_Whitepaper_FINAL.pdf NVIDIA