Course Syllabus

Edit

CS 742/842 High-Performance Computing on Emerging Architectures Syllabus

Course Objective

The CS 742/842 High-Performance Computing on Emerging Architecture course is about parallel programming with GPUs and is a project-based course. This course will introduce students to programming on GPUs using CUDA C/C++, PyTorch, and Triton. I will work with students to identify a project that involves programming numerically intensive computing on the latest GPUs from NVIDIA.

Programming GPUs

During the first part of the course, you will go over various course modules, watch videos, and complete homework assignments. Every module has a start and end date, and students are expected to complete the module by its end date. I will publish new modules a few days before their start date.

Course Project

In consultation with the instructor, students identify a project that involves a data-parallel computation suitable for GPUs. A typical project implementation consists of:

- A baseline C/C++ CPU implementation of the data-parallel computation that will also be used for validating GPU implementations
- An optimized CUDA implementation of the data-parallel computation
- A PyTorch/Triton implementation of the data parallel computation
- Evaluate various implementations

At the end of the semester, students are expected to do a presentation and demo of the project. Your grade in the project will depend on the successful completion of the project and the variety of techniques used in optimizing performance on the target hardware.

Grading Criteria

Your grade will be based on a total of 100 points with the following distribution:

- Homework Assignments: 25%
- Course Project: 75%

HPC Resources

For CUDA/Python/Triton programming on NVIDIA GPU, you will need access to ODU HPC resources.

ODU HPC URL: **https://www.odu.edu/ts/software-services/hpc (https://www.odu.edu/ts/software-services/hpc)**

Please visit the "Getting Started" module for more details on accessing ODU HPC resources. If your laptop has an NVIDIA GPU, you can also use that. You need to install the CUDA Toolkit to be able to use the GPU on the laptop. Please see:

https://docs.nvidia.com/cuda/cuda-installation-guide-microsoft-windows/index.html (https://docs.nvidia.com/cuda/cuda-installation-guide-microsoft-windows/index.html)

Reading Material

We will focus on the two programming environments. Here are the links where you will find good learning materials (I will be adding more links on the course website during the semester.

CUDA online learning resource:

https://www.olcf.ornl.gov/cuda-training-series/

PyTorch:

https://docs.pytorch.org/tutorials/beginner/pytorch_with_examples.html (https://docs.pytorch.org/tutorials/beginner/pytorch_with_examples.html)

Triton:

https://triton-lang.org/main/getting-started/tutorials/

Homework Assignments

Homework assignments will be posted on the course website under each module. Please submit your assignment by the specified due date.

Discussion Participation

Please post all technical questions and issues in the discussion forum. You are encouraged to both ask and answer technical questions there. I will monitor the forum and respond when needed. For personal matters, you may contact me directly by email.

Online Office Hours

Tuesday, 11:00 am - 1:00 pm

Students can reserve a meeting slot during online office hours using Calendar from the left menu.