Fall 2022: ME759 Final Project Proposal

Project Title: Comparison of Deep Neural Network Training using AMD HIP vs NVIDIA CUDA

Link to git repo for project: https://git.doit.wisc.edu/AGELII/repo759/-/tree/main/FinalProject759

Problem statement:

The end goal is to have implemented the same deep neural network model in both CUDA and HIP and benchmark them. By doing this, I will learn about HIP and its performance and easibility to program with compared to CUDA, that we've used for the last month on assignments.

Motivation/Rationale: explain why you chose to work on this project.

Within our class, we've been implementing many algorithms, such as matrix multiplication, in CUDA that have ties to machine learning, specifically the training of deep neural networks. As someone interested in machine learning, I wanted to some how incorporate it with high performance computing.

Explain how you contemplate going about it: indicate if you'll use GPU/OpenMP/MPI parallel computing, what libraries, etc. Indicate what algorithms/approaches you are considering.

I will be using GPUs, AMD and NVIDIA, for parallel computing with HIP and CUDA to implement and benchmark deep neural network training. Within the layers, I would like to add something we've done before, convolutions.

ME759 aspects the proposed work draws on: bulleted list, be brief

- CUDA to implement the network
- parallelism in training, using GPUs significantly speed up training time

Deliverables: what you expect to deliver on 12/14/2022, 9 PM: code, input files, tech report, etc.

I will submit all the code for the same deep neural network implementation in both CUDA and HIP, along with this, a tech report that showcases the benchmarking of training (a scaling analysis). I will also submit the data used to train and test the models.

How you will demonstrate what you accomplished: this is particularly important if what you do is a small piece of a bigger project that you will continue to pursue after wrapping up ME759.

This will all be in the report, the running times and scaling analysis for both HIP and CUDA implementations of training the same deep neural network.

Other remarks: say here anything else that you think Dan should be aware of and doesn't fall within any other category above.