Prerequisites

- Understanding of parallel programming, preferably OpenMP and MPI
- Enrolled MS/PhD student
- Proficiency in C/C++
- Some knowledge of Distributed Machine Learning (optional)

Motivation

The demand for artificial intelligence has grown significantly over the last decade and this growth has been fueled by advances in machine learning techniques and the ability to leverage hardware acceleration. However, in order to increase the quality of predictions and render machine learning solutions feasible for more complex applications, a substantial amount of training data is required. Although small machine learning models can be trained with modest amounts of data, the input for training larger models such as neural networks grows exponentially with the number of parameters. Since the demand for processing training data has outpaced the increase in the compute power of computing machinery, there is a need to distribute the machine learning workload across multiple devices, and thus transform a centralized approach into a distributed system. These distributed systems present new challenges to the HPC community.

Overview

Machine Learning workloads are often carried out on GPUs, which are very well suited for their computations. Recently, systems became available that consist of multiple GPUs attached to a multicore CPU (called multi-GPU nodes below). The MPI programming model is used for implementing distributed machine learning approaches across the nodes of High Performance Computing (HPC) machines, aka large compute clusters, such as Stony Brook’s SeaWulf. It can also enable the use of multiple GPUs inside a compute node and hence could be used to handle large ML problems that exploit all of the GPUs in a multi-GPU node. OpenMP is a popular programming model for creating code to run on multicore systems.
and any attached GPUs. We have already shown [1] that OpenMP task-based programs exhibit better scaling on a multi-GPU node than the corresponding MPI code. There is great interest in the HPC community regarding how OpenMP can be used to train distributed machine learning model on future architectures, especially on a multi-GPU node.

Project

This project will investigate how best to implement distributed machine learning models using OpenMP on multi-GPU nodes. The work will include implementation of a Master-Worker distributed training model using OpenMP tasking. We will compare the performance against an MPI implementation. The expected outcomes of the work are:

- A written report on the work done
  - This is potentially a research paper in a good conference
- Presentation to our research group
- The code
- Demo of implementations
- Ideas for future work.

For More Information

- Our group
  - https://you.stonybrook.edu/exascallab/
- https://www.openmp.org/
- https://eng.uber.com/horovod/
- https://horovod.ai/

Reference: