CSYE7105 HW4

Instructor: Dr. Handan Liu

2020-12-06

The following parts are required to be written in python file and run in parallel mode on the Discovery Cluster.

Part 1: 60 points

1. Please list the general idea of a CUDA workflow. (hint: 6 steps) [6 pts]
2. Please indicate which types of memory are on-chip and off-chip in GPU memory? [6 pts]
3. Why deep neural network training is slow? List three or more main reasons. [6pts]
4. List parallel methods for DNN training on a single machine and multiple machines. [5pts]
5. What type of communication of MPI is used for data parallel deep learning? And give 2 MPI_ functions for instance. [6pts]
6. In deep learning, how many dimensional tensors are used to express the time series, images and videos, respectively? [6pts]
7. In Pytorch, what method(s) is used to move tensors between CPU and GPU? [2pts]
8. If we want to store 64 RGB images of 128x128 pixels, write this tensor. [2pts]
9. What type(s) of neural networks are good to use data parallelism? [4pts]
10. Under what circumstances can we develop model parallelism for deep learning training on multiple GPUs? [4pts]
11. What is profiling? Why should I profile my code? [6pts]
12. How many types of precision mixed calculations does the GPU provide? [3pts]
13. Using AMP in Pytorch, what are the roles of the GradScaler and autocast? [4pts]

Part 2: 20 points

Compute A^n + B^n on 2 GPUs. Here, n=10, A and B are 10000 x 10000 matrices.
Please print out the computation time on 1 GPU and 2 GPUs.
You can implement by using PyTorch or TensorFlow.

Part 3: 20 points

Given the original code *pt_modelparallelism.py*. It is parallelized on 2 GPUs with model parallelism. Please split the layers so as to implement model parallelism on 4 GPUs.
Note that only the model parallelism is practiced here, without considering the real performance (maybe very poor).

Review and Grade:

TA will review the homework including run the code on Discovery. TA will grade for my reference. No review time in person or on Zoom.

Submission format: compact all files as a tarball with your name for the convenience of TA's review.
Submission through Canvas
Deadline is by the end of December 13th.