香港浸會大學理學院
HKBU Faculty of Science

# DEPARTMENT OF COMPUTER SCIENCE

## PhD Degree Oral Presentation

| | |
|---|---|
| PhD Candidate: | Mr. Shaohuai SHI |
| Date | August 7, 2020 (Friday) |
| Time: | 2:00 pm – 4:00 pm (35 mins presentation and 15 mins Q & A) |
| Venue: | Zoom ID: 981 8675 2012<br>(The password and direct link will only be provided to registrants) |
| Registration: | https://bit.ly/sem-zm (Deadline: 5:00pm, 6 August 2020) |

***Communication Optimizations for Distributed Deep Learning***

## Abstract

The mini-batch stochastic gradient descent (SGD) algorithm and its variants are the most widely used algorithms in training deep neural network models. Nowadays it becomes a common practice to exploit multiple processors (e.g., GPUs or TPUs) to accelerate the training process using distributed SGD. However, the iterative nature of distributed SGD requires multiple processors to iteratively communicate with each other to collaboratively update the model parameters. The intensive communications easily become the system bottleneck and limit the system scalability. In this thesis, we study the communication-efficient techniques for distributed SGD to improve the system scalability and thus accelerate the training process.

First, we build a performance model with a directed acyclic graph to model the training process of distributed SGD and evaluate its accuracy with extensive benchmarks on existing state-of-the-art deep learning frameworks. Our benchmarking and modeling results point out that existing optimizations for the communication problems are sub-optimal, which we need to address in this thesis.

Second, to address the startup problem (due to the high latency of each communication) of layer-wise communications with wait-free backpropagation (WFBP), we propose an optimal gradient merging solution for WFBP, named MG-WFBP, that exploits the layer-wise property to overlap the communication tasks with the computing tasks and can be adaptive to the training environments.

Third, to make the high computing-intensive training tasks possible in GPU clusters with low-bandwidth interconnect, we investigate the gradient compression techniques in distributed training. The traditional top-k sparsification can compress the communication traffic with little impact on the model convergence, but suffers from a linear communication complexity to the number of workers. To address this problem, we propose a global top-k (gTop-k) sparsification algorithm that reduces the communication complexity to be logarithmic to the number of workers.

Lastly, to enjoy both benefits of the pipelining technique and the gradient sparsification algorithm, we propose a new distributed training algorithm, layer-wise adaptive gradient sparsification SGD (LAGS-SGD), which supports layer-wise sparsification and communication. The LAGS-SGD algorithm also supports optimal merged-gradient with sparsification (OMGS) to reduce the impact of the latency to further improve the system scalability. We also theoretically and empirically prove that the layer-wise gradient sparsification preserves the convergence properties.

***** ALL INTERESTED ARE WELCOME *****