THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING SPEAKER SERIES

PRESENTS

Boosting Performance and Scalability of Distributed Deep Learning Systems via Efficient Data Management

Dr. Dingwen Tao
Associate Professor of Intelligent Systems Engineering,
Indiana University, Bloomington

Monday**, 18 September 2023,** 9:55 AM

Room: **Zoom** (Meeting ID# 976-269-9678; Passcode: K91Bwy https://zoom.us/j/9762699678?pwd=RUp5ZmN3cHUyQ1FvUExVQjVsc1hVUT09)

LECTURE ABSTRACT

Deep learning (DL) has revolutionized various fields, from computer vision to natural language processing, delivering unprecedented accuracy across numerous applications. As both the scale of datasets and the complexity of models increase, distributed training becomes crucial to meet computational demands and expedite training. Yet, the efficiency of distributed deep learning systems frequently faces obstacles related to data management, such as limited memory capacity, sluggish data loading, and high communication overheads. In this talk, I will delve into three of our recent innovations: (1) I will introduce a method that leverages compression techniques to significantly reduce memory usage during the training of convolutional neural networks. This strategy allows for training more extensive and sophisticated models with larger batch sizes without demanding additional hardware resources. (2) I will present a novel data-loading framework for distributed training designed to overcome the significant performance bottleneck encountered when accessing random data samples from vast datasets, ranging from hundreds of GBs to TBs. (3) I will delineate a compression-centric technique that drastically cut communication overheads in operations such as alltoall. This leads to a swifter convergence without compromising accuracy, notably evident during the training of Deep Learning Recommendation Models (DLRMs).

SPEAKER BIOSKETCH

Dr. Dingwen Tao is an associate professor at Indiana University Bloomington, where he directs the High-Performance Data Analytics and Computing Lab. He received his Ph.D. in Computer Science from University of California, Riverside in 2018 and B.S. in

Mathematics from University of Science and Technology of China in 2013. He is the recipient of various awards including NSF CAREER Award (2023), Amazon Research Award (2022), Meta Research Award (2022), R&D100 Awards Winner (2021), IEEE Computer Society TCHPC Early Career Researchers Award for Excellence in HPC (2020), NSF CRII Award (2020), and IEEE CLUSTER Best Paper Award (2018). He is serving on the Technical Review Board of IEEE Transactions on Parallel and Distributed Systems. He served as the Program Chair of IEEE ScalCom-2021, DRBSD-9, and IWBDR workshops. He is also a reviewer, program committee member, or session chair of major HPC venues, such as SC, HPDC, ICS, IPDPS, CLUSTER, ICPP, CCGrid, and HiPC.