"Trillion-parameter scale model training and inference with DeepSpeed"

Dr. Ammar Ahmad Awan
Senior Researcher at Microsoft Corporation

Abstract: In this talk, Dr. Awan will cover a broad set of challenges and solutions related to efficient training and inference of large DL models at scale. He will provide an overview of the open-source DeepSpeed library that is being developed at Microsoft Corporation to enable trillion-parameter scale model training and inference on current as well as next-generation of HPC and AI platforms including Microsoft Azure and explain how DeepSpeed enables memory-efficient training and inference of Transformer-based large language models (LLMs) like Chat-GPT, GPT-3, and Megatron-Turing NLG 530B. These two topics: 1) ZeRO-Infinity and 2) DeepSpeed Mixture of Experts (MoE) will be discussed in detail and he’ll conclude his talk with a brief overview of the software architecture of DeepSpeed and its training and inference APIs.

Bio: Ammar Ahmad Awan is a Senior Researcher at Microsoft working on the DeepSpeed library with Yuxiong He and the DeepSpeed team. He is the lead developer of the DeepSpeed Mixture of Experts (MoE) system that supports both training and inference of MoE models at scale. He received his B.S, M.S, and Ph.D. degrees in Computer Science from the National University of Science and Technology (NUST), Pakistan, Kyung Hee University (KHU), South Korea, and The Ohio State University, respectively. His current research focus lies at the intersection of high-performance systems and large-scale training and inference of deep learning (DL) models. He previously worked on a Java-based Message Passing Interface (MPI) and nested parallelism with OpenMP and MPI for scientific applications. He has published several papers in conferences and journals related to these research areas. He actively contributed to various projects like MVAPICH2-GDR (High Performance MPI for GPU clusters, OMB (OSU Micro Benchmarks), and HiDL (High Performance Deep Learning) during his graduate studies at OSU. He is the lead author of the OSU-Caffe framework (part of HiDL project) that allows efficient distributed training of Deep Neural Networks.

Wednesday, July 19, 2023 at 11:00 AM ET

Join Zoom Meeting

https://osu.zoom.us/j/99365919979?pwd=OEVGdjFXc2E3TmVEMUNNYmkycXdCdz09&from=addon