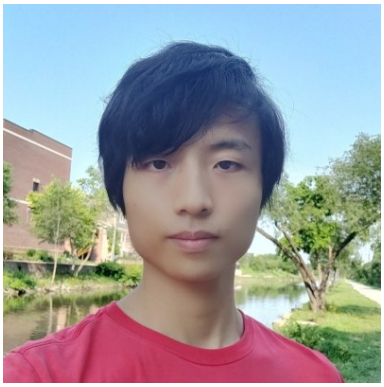
Invited Online Talk



Fast Inter-GPU Communication with NCCL for AI Training

Dr. Kaiming Ouyang

Time: Tuesday, 05/28/2024, 11:00 AM — 12:00 PM (CST)

Zoom Link: https://unt.zoom.us/j/86512602380

Email: kouyang@nvidia.com

Abstract: Nowadays GPU, as a compute accelerator, dominates AI training around the world, because of its high performance and excellent power efficiency. Due to the extremely large model size and compute complexity, multi-GPU and multi-node training have become mainstream, and communication among GPUs and nodes plays a critical role in AI training. To provide high-speed communication, we developed the Nvidia Collective Communication Library (NCCL) to help users better utilize the full potential of hardware and reach the maximal performance under all circumstances. Until now NCCL has been widely adopted by many AI frameworks such as Pytorch, TensorFlow, Caffe, NeMo and so on.

Bio: Dr. Kaiming Ouyang is a software engineer at Nvidia Corporation. He received his Ph.D. degree in computer science from the University of California, Riverside in 2022. He has 5 years of experience in large-scale parallel computing and communication and has contributed to the implementation and optimization of MPICH and Nvidia Collective Communication Library (NCCL). Recently, he has been working on NCCL to enable fast, scalable, and reliable machine learning training at an extreme scale.