Invited Online Talk

Introduction to JAX/XLA: Accelerating Large Language Model Training

Dr. Donglin Yang

Time: Wednesday, 05/29/2024, 3:00 PM — 4:00 PM (CST)

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

Email: dongliny@nvidia.com

Abstract: JAX is an open-source machine learning framework designed for transforming numerical functions for use in Python. Leveraging XLA, it compiles and executes machine learning models on GPUs with exceptional scalability and performance guarantees. This talk offers a concise introduction to both JAX and XLA. We'll explore JAX's ability to automatically differentiate native Python functions and delve into key optimizations within XLA, including kernel fusion, performance auto-tuning, and support for distributed training techniques like SPMD.

Bio: Dr. Donglin Yang is a deep learning software engineer at Nvidia, working on machine learning performance with a special focus on GPU. Before joining Nvidia, Dr. Yang received his Ph.D. in computer science from the University of North Carolina at Charlotte, and his thesis was on large-scale deep learning systems.