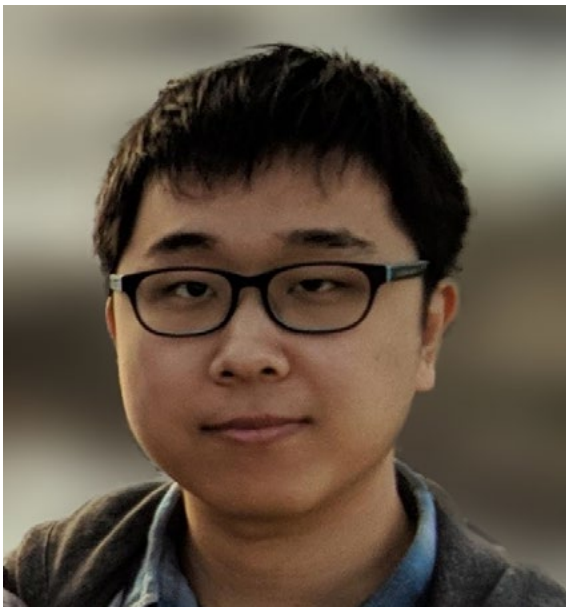
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



PEIFENG YU

Application-Aware Scheduling in Deep Learning Software Stacks

Thursday, May 5, 2022
10:00am – 12:00pm
Virtual

ABSTRACT: Deep learning (DL) has pervaded many areas of computing due to the confluence of the explosive growth of large-scale computing capabilities, availability of datasets, and advances in learning techniques. However, DL infrastructure is still in its early stages and there is a mismatch between the hardware, the software stack, and DL applications. Despite the emergence of new unique hardware and use cases, the software stack that abstracts and schedules these hardware resources remains largely unchanged. Furthermore, user-defined performance metrics urge better schedulers tailored to applications' specific needs. Motivated by the mismatch, this dissertation revisits the system design across the stack, focusing on the synergy between schedulers and application/system-specific information.

At the bottom level, the adoption of specialized hardware like GPU poses challenges to efficient usage. Due to the lack of OS arbitration, applications usually assume exclusive access, making the otherwise underutilized resource unusable for other jobs on the same host. We design Salus to leverage DL-specific usage patterns to provide two missing primitives for GPU: fast job switching and memory sharing.

Even with an efficient execution platform, it is still not trivial to harvest its full potential for higher-level applications.

With the proliferation of distributed computing, hyperparameter tuning generates many small interdependent training trials. It suffers from poor resource usage due to the oblivion of advanced execution strategies. We propose Fluid as a generalized hyperparameter tuning engine that uses a water-filling approach to make the best use of resources both at intra- and inter-GPU granularity.

Model inference serving also requires careful scheduling to achieve tight latency guarantees and maintain high utilization. With the rise of dynamic models, data-dependent networks become less predictable by a single, point estimation of the true running time. With Orloj, we show that probability distributions of execution times can be exploited for scheduling in the presence of SLA constraints.

In this dissertation, we combine application/system-specific information with scheduler design as a means of supporting new hardware and applications with heterogeneous performance objectives, in the hope that our crude work may be a basis for a more efficient DL stack.

CHAIR: Prof. Mosharaf Chowdhury