SPEAKER: Prof. Hong Shen, Sun Yat-sen University, China

TITLE: Machine-Learning for Optimization: Coflow Scheduling in Cloud High-Performance Computing

ABSTRACT: Statistics shows that transmission delay of dependent data flows among computation tasks accounts for more than 50% of the total execution time of an HPC job. How to effectively schedule these data flows is therefore crucial for reducing the execution time of application jobs and increasing the profit of HPC systems, which is particularly important for managing shared HPC resources in a cloud computing environment. In this talk, I will address the problem of scheduling groups of parallel data flows, namely coflows, among computation tasks in HPC jobs. Because this problem even in its simplest case of single stage is known NP-hard, various greedy strategies, heuristic and machine learning based approaches have been proposed to obtain sub-optimal solutions. As an example of application of machine learning techniques for solving optimization problems, I will present our recent work in combining different machine learning models and techniques for effectively scheduling coflows. I will begin with an overview on machine learning and traditional optimization strategies for solving problems, our approaches of combining them in different settings, the coflow scheduling problem and its research status. I will then introduce our work on offline scheduling of single-stage coflows by combining meta-learning with DNN to balance the efficiency and fairness in terms of minimizing coflow completion time while ensuring the desired fairness among coflows. Next, I will present our work for online scheduling of multi-stage coflows, a more challenging but realistic situation in cloud HPC environments, by combining GNN, DRL and self-attention mechanism to improve the scalability and efficiency measured by job completion time. Finally, I will conclude this talk by showing some challenges in applying ML techniques in solving optimization problems and our future work.

SPEAKER BIO: Hong Shen received the BS degree from Beijing University of Science and Technology, MS degree from University of Science and Technology of China, and PhD degree from Abo Akademi University, Finland. He is Professor in Macao Polytechnic University, specially-appointed Professor in Sun Yat-sen University, and Adjunct Professor in University of Adelaide where he was a tenured Professor (Chair of Computer Science) for 15 years. With main research interests in parallel and distributed computing, privacy preserving computing and high performance networks, he has led numerous research centers and projects in different countries. He has published 400+ papers including over 100 papers in major international journals such as a variety of IEEE and ACM transactions. Prof. Shen received many honors and awards, and served on different roles in professional societies, journal editorial boards and conference committees.