Ph.D. DISSERTATION DEFENSE

Candidate: Anil Gaihre
Degree: Doctor of Philosophy
School/Department: Charles V. Schaefer, Jr. School of Engineering and Science /Electrical and Computer Engineering
Date: Thursday, November 9, 2023
Time/Location: 11 AM / https://stevens.zoom.us/j/9468514606
Title: Accelerating Data-Analytical Algorithms on GPUs

Chairperson: Dr. Zhuo Feng, Department of Electrical and Computer Engineering, Charles V. Schaefer, Jr. School of Engineering and Science

Committee Members: Dr. Hang Liu, Department of Electrical and Computer Engineering, Rutgers University
Dr. Xiaoye S. Li, Computational Research Division Lawrence Berkeley National Laboratory
Dr. Min Song, Department of Electrical and Computer Engineering, Charles V. Schaefer, Jr. School of Engineering and Science
Dr. Lei Wu, Department of Electrical and Computer Engineering, Charles V. Schaefer, Jr. School of Engineering and Science
Dr. Yue Ning, Department of Computer Science, Charles V. Schaefer, Jr. School of Engineering and Science

ABSTRACT

The surge in digital technologies and internet usage has resulted in an influx of data from diverse sources. Traditional data processing methods and computing resources prove inadequate in handling this vast and intricate data. In response, Graphics Processing Units (GPUs) have emerged as a solution, leveraging their parallel processing capabilities. This dissertation focuses on optimizing key data processing algorithms on GPUs.

In Chapter 2, we introduce Dr. Top-*k*, a delegate-centric top-k system for GPUs, presenting three contributions. Firstly, we devise a comprehensive delegate-centric framework, including the maximum delegate, delegate-based filtering, and β delegate mechanisms, reducing top-k workloads by over 99%. Secondly, we rigorously analyze and determine an optimal subrange size, for performance. Thirdly, we introduce four key system optimizations, enabling fast multi-GPU top-k computation.

Chapter 3 presents XBFS, a Breadth First Search (BFS) traversal on GPUs to cope with the nondeterministic characteristics of BFS with the following three techniques:

First, XBFS adaptively exploits four either new or optimized frontier queue generation designs to accommodate various BFS levels that present dissimilar features. Second, after observing that the workload associated with each vertex is not proportional to its degree in the bottom-up phase, we design three new strategies to better balance the workload. Third, XBFS introduces the first truly asynchronous bottom-up traversal which allows BFS to visit vertices for multiple levels at a single iteration with both theoretical soundness and practical benefits.

Chapter 4 introduces GSOFA, the first GPU-based scalable symbolic factorization algorithm. We propose a fine-grained parallel symbolic factorization algorithm optimized for the GPU's Single Instruction Multiple Thread (SIMT) architecture. We develop SIMT-friendly supernode detection to balance workloads, minimize communication, and maximize GPU resource utilization. Additionally, we employ a three-pronged optimization strategy to alleviate excessive space consumption issues.

Chapter 5 introduces parallel refinement methods in fill-reduction ordering for sparse direct solvers on a GPU. The ordering process is time-consuming, and currently, GPU implementations are not available. The challenge lies in the sequential nature of the refinement process when utilizing the multi-level nested-dissection approach in GPU implementation. This chapter introduces parallel VFM refinement algorithm and Independent Set based algorithm suitable for GPU.