Chen WG. Graph accelerators—A case for sparse data processing. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(2): 243-244 Mar. 2024. DOI: 10.1007/s11390-024-0003-0

Graph Accelerators—A Case for Sparse Data Processing

Wen-Guang Chen (陈文光), *Fellow, CCF, Member, ACM*

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

E-mail: cwg@tsinghua.edu.cn

Graph is a powerful sparse data structure that intuitively represents entities and their relationships. Classic graph traversal algorithms such as Breadth-First Search (BFS), Single-Source Shortest Path (SSSP), PageRank, and Weakly Connected Components (WCC) have extensive applications in social network analysis, risk management for finance, and recommendation systems. However, graph processing in CPUs and GPUs is not very efficient due to its irregular memory accesses.

Many people have proposed software approaches to speed up graph processing, such as PowerGraph, PowerLyra, and Shentu, which address load imbalance issues by replicating high-degree vertices. XStream and GridGraph attempt to improve memory access locality by scanning the edge list of graphs while localizing the range of vertices accessed in a stage. Ligra and Gemini provide adaptive dual compute modes (bottom-up and top-down), which are particularly effective for BFS-like algorithms such as BFS and SSSP.

However, pure software approaches have their limitations, and it is desired to see how hardware could be employed to accelerate graph processing. This cover article shows a series of accelerator designs for graph processing. The first technique proposed is to leverage the characteristics of graph traversal algorithms, whose operations on vertices are commutative and associative. A hardware architecture is proposed to update vertices' status in parallel to significantly reduce pipeline stalls originally caused by conflicting updates on vertices. Additionally, to fit the high memory bandwidth provided by HBM, a distributed on-chip memory hierarchy is proposed to allow more processing units to be integrated on chip. Furthermore, the article discusses an FPGA-based graph update library, which can efficiently process dynamic graphs, i.e., graphs that are changing.

The article further discusses several emerging hardware/software co-design issues for graph processing, such as processing hypergraphs and heterogeneous graphs, and using PIM (Process-In-Memory) architecture to accelerate graph processing.

There are a few interesting issues beyond the article itself. For graph processing, the current accelerators are focusing on graph traversal algorithms. It would be desirable to explore hardware accelerator architectures for graph mining and graph neural networks, supporting property graph models and graph query languages such as GQL, which are used extensively in the industry.

Finally, while the article gives a successful case for graph processing, it is motivating to ask whether the techniques invented for graph processing could apply to a broad range of sparse computing. The economics of the hardware industry favor massive amounts. To compete with general architectures such as CPUs and GPUs, it would be essential for a specialized architecture to reach a critical mass to stay in the market. While graph computing is still a niche market, the spectrum of sparse processing, from scientific computing to sparse neural networks, may have a much larger market share to survive a specialized hardware architecture. I believe the article encourages readers to think about these questions thoroughly.

Perspective

For Cover Article: Liao XF, Zhao WJ, Jin H *et al.* Towards high-performance graph processing: From a hardware/software co-designper-spective. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 39(2): 245-266 Mar. 2024. DOI: 10.1007/s11390-024-4150-0

Wen-Guang Chen is a professor in the Department of Computer Science and Technology, Tsinghua University, Beijing, where he has been teaching since 2003. He received his B.S. and Ph.D. degrees both in computer science from Tsinghua University, Beijing, in 1995 and 2000, respectively. His research interest is in parallel and distributed computing. He is a CCF fellow and a CCF distinguished speaker, and an ACM member and the member at charge of ACM China Council. He has served in program committees of a variety of major conferences in the parallel and distributed computing area, including PLDI, PPoPP, OSDI, SC, EuroSys, CGO, IPDPS, APSys, and ICPP.