A Unified Comparison of Hardware Acceleration for Relational Data Analytics

Alireza Shateri

Computer Science
University of Toronto
alireza@cs.toronto.edu

1 Abstract

The high efficiency of domain-specific hardware has sparked substantial interest in adopting accelerators in data analytics systems. Among many choices, GPUs and FPGAs thrived as the most popular solutions due to their prevalent deployments in cloud data centers. Accelerators have been widely deployed in cloud data centers to offload a variety of infrastructure and application functionalities [2, 3, 4, 1, 5]. Generally, the high parallelism, e.g., tens of thousands of cores in recent GPUs, and memory speed, e.g., up to 1 TB/s bandwidth with High Bandwidth Memory (HBM), offered by accelerators can drastically improve the performance of data analytics tasks. Despite the tremendous efforts and promising results of exploiting these devices to accelerate a variety of data processing tasks, a fundamental question remains unaddressed: Given a relational data analytics task, which hardware is most suitable? The answer to this question is meaningful not only for practitioners to select accelerators for their target workloads, but also for researchers to rethink existing hardware approaches and design more effective optimizations. However, answering the question requires a systematic comparison between the hardware for data analytics.

In this talk, I present an empirical study that fulfills the purpose. We characterize GPUs and FPGAs and evaluate their benefits and limitations using a unified framework and a comprehensive set of data analytics operators. Our extensive comparison (across tasks, between hardware, and using metrics that consider performance, cost, and energy consumption) provides key insights into hardware acceleration for relational data analytics.

References

[1] Alibaba Cloud Community . A detailed explanation about alibaba cloud cipu. https://www.alibabacloud.com/blog/a-detailed-explanation-about-alibaba-cloud-cipu_599183, 2022.

[2] Amin Vahdat. Announcing trillium, the sixth generation of google cloud tpu. https://cloud.google.com/blog/products/compute/introducing-trillium-6th-gen-tpus, 2024.

[3] AWS. Aws nitro system. https://aws.amazon.com/ec2/nitro/, 2024.

[4] Daniel Firestone, Andrew Putnam, Sambrama Mundkur, Derek Chiou, Alireza Dabagh, Mike Andrewartha, Hari Angepat, Vivek Bhanu, Adrian M. Caulfield, Eric S. Chung, Harish Kumar Chandrappa, Somesh Chaturmohta, Matt Humphrey, Jack Lavier, Norman Lam, Fengfen Liu, Kalin Ovtcharov, Jitu Padhye, Gautham Popuri, Shachar Raindel, Tejas Sapre, Mark Shaw, Gabriel Silva, Madhan Sivakumar, Nisheeth Srivastava, Anshuman Verma, Qasim Zuhair, Deepak Bansal, Doug Burger, Kushagra Vaid, David A. Maltz, and Albert G. Greenberg. Azure accelerated networking: Smartnics in the public cloud. In *15th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2018, Renton, WA, USA, April 9-11, 2018*, pages 51–66. USENIX Association, 2018.

[5] NVIDIA. Nvidia data center gpus: The heart of the modern data center. https://www.nvidia.com/en-us/data-center/data-center-gpus/, 2024.