Enabling Effective Utilization of GPUs for Data Mamagement Systems

Xiaodong Zhang
Department of Computer Science and Engineering
The Ohio State University
Columbus, Ohio, USA
zhang@cse.ohio-state.edu

Abstract—The research and development of today's data management systems for both transaction processing and data analytics are driven by an inevitable technology trend, namely in-memory computing, and by an increasingly high demand of online applications from billions of concurrent users. This requires that a data management system must be able to provide both high throughput transactions and fast data analytics execution. GPU can play an important role in data management systems for both high throughput transactions and high performance data analytics. However, several fundamental challenges must be addressed before GPU can be integrated into data management production systems. First, the low bandwidth PCIe connection between CPU host memory and GPU device memory is a bottleneck to limit GPUs to achieve their full potential for high performance data analytics. Second, the mismatch between limited GPU programming capabilities and complex SQL query structures makes it hard to utilize GPUs to execute certain practically important queries including nested sub-queries and recursive queries. Finally, system support for resource management in GPUs is very limited, such as memory management and job scheduling, being unable to coordinate concurrent query executions, and causing a high system underutilization or overutilization on GPUs.

In this presentation, I will first make a strong case for GPUs to serve as special-purpose devices to greatly accelerate the operations of in-memory key-value stores. Specifically, I present the design and implementation of Mega-KV, a GPU-based in-memory key-value store system that achieves high performance and high throughput. Mega-KV can process up to 160+ million key-value operations per second. I will discuss the challenges we are facing to accept GPU as a regular member in our conventional computing ecosystem along with several proposed solutions.

Keywords—Data management, data analytics, Graphics Processing Unit (GPU), Key-value store, in-memory computing