Adopting OpenCAPI for High Bandwidth Database Accelerators

Jian Fang[†], Yvo T.B. Mulder[†], Kangli Huang[†], Yang Qiao[†], Xianwei Zeng[†], H. Peter Hofstee[†*], Jinho Lee[*], and Jan Hidders[‡]

[†]Delft University of Technology [*]IBM Research [‡]Vrije Universiteit Brussel

[†]j.fang-1@tudelft.nl [†]{y.t.b.mulder, k.huang-5, y.qiao, x.zeng}@student.tudelft.nl [*]{hofstee, leejinho}@us.ibm.com [‡]jan.hidders@vub.be

1. INTRODUCTION

Due to the scaling difficulty and high power consumption of CPUs, data center applications look for solutions to improve performance while reducing energy consumption. Among different solutions, heterogeneous architectures utilizing both CPUs and accelerators, such as FPGAs, show promising results. FPGAs have more potential to achieve high throughput, low latency and power-efficient designs compared to a general-purpose processor. However, wide adoption of FPGAs is limited by the relatively low bandwidth between the CPU and FPGA, limiting applications mainly to computation-intensive problems.

Meanwhile, database systems have sought ways of achieving high bandwidth access to the data. One trend here is the increasing usage of in-memory database systems. This kind of system has higher data access speed than disk-based database systems, leading to a high data processing rate.

In order to leverage emerging heterogeneous architecture to accelerate the databases, we need to solve the interconnect bottleneck. A recent advancement is the introduction of the Open Coherent Accelerator Processor Interface (OpenCAPI) [1], which provides a significant increase in bandwidth compared to the current state-of-the-art (PCIe gen 3). This change requires re-evaluation of our current design methodologies for accelerators.

This abstract presents our ongoing work towards a heterogeneous architecture for databases with high memory bandwidth connected FPGAs. Based on this architecture, three accelerator design examples that have promising throughput and can keep up with the increased bandwidth are proposed.

2. SYSTEM ARCHITECTURE OVERVIEW

Modern systems consist of one or more CPUs that contain multiple cores and are coupled with DRAM. Accelerators are most commonly connected to the host using PCIe. The widely adopted PCIe gen 3 interconnect attains a bandwidth of roughly 8Gb/s per lane. In contrast, OpenCAPI enables a low latency, high bandwidth interconnect by using 25Gb/s differential signaling. Aggregating OpenCAPI bandwidth can rival or exceed the bandwidth of DDR memory, making OpenCAPI-attached accelerators candidates for bandwidth-limited applications.

H²RC 2017 Denver, Colorado, USA

3. ACCELERATOR DESIGN EVALUATION

Three high-bandwidth streaming accelerators for database queries have been studied: decompress-filter, hash-join and merge-sort. Each has different buffering requirements, which are challenging at this speed. Requirements vary over having to hide latency versus the number of read ports.

The critical path of a merge-sort occurs in the last pass. It merges multiple sorted streams into a final in-order stream. To guarantee a high output rate in the last pass demands a strong merge engine that has a stable and high processing rate. At the same time, this type of sorter requests data from different streams randomly. Consequently, an obvious but tough challenge is how to hide the high access latency of main memory with uncertain stream entry requests.

Hash-join is a memory-bound application. Increasing the bandwidth helps to reduce the data transfer time. However, not all the bandwidth can be fully utilized due to the low locality of the data and multiple passes of data transfers. To keep up with this high bandwidth, an algorithm with fewer passes of data access and higher throughput for each pass of data processing is required.

In contrast, the Parquet decompress-filter is computation-bound. The main bottleneck comes from the dependency between the interpretation of tokens that are adjacent. To have a high processing rate for decompressing one stream demands special designs to avoid or solve the dependency. Another possibility for high throughput is to use multiple identical, but small, decompression engines. However, excellent arbitration is required to schedule and balance between the different engines.

4. REFERENCES

[1] J. Stuecheli. OpenCAPI™ - A New Standard for High Performance Memory, Acceleration and Networks. HPC Advisory Council - Swiss Conference 2017, 2017.