Accelerating Data Center Applications with Reconfigurable DataFlow Engines

Sakil Barbhuiya
Queen's University Belfast
sbarbhuiya01@qub.ac.uk

Yun Wu
Queen's University Belfast
yun.wu@qub.ac.uk

Karen Murphy
Queen's University Belfast
k.l.murphy@qub.ac.uk

Hans Vandierendonck
Queen's University Belfast
h.vandierendonck@qub.ac.uk

George Karakonstantis
Queen's University Belfast
g.karakonstantis@qub.ac.uk

Dimitrios S. Nikolopoulos
Queen's University Belfast
d.nikolopoulos@qub.ac.uk

ABSTRACT

Energy-efficient, programmable accelerators are an important value proposition for pushing the limits on the computation capacity and density of future data centers. Integrating the accelerators in the software stacks of cloud-based data analytics frameworks is, however, an open issue. The goal of this work is to achieve seamless integration of accelerators in data analytics programming frameworks.

Keywords

Heterogeneous computing, dataflow engines, data analytics

1. INTRODUCTION

In excess of 3,000 exabytes of data have been pushed through data centers in 2013[1]. This number is expected to increase at a 33% CAGR growth rate. An important challenge for future data centers is scaling them up to increasingly higher data volumes and compute capacity, while meeting stringent energy budgets.

Programmable accelerators based on Field Programmable Gate Arrays (FPGAs) or Graphics Processing Units (GPUs) are a promising value proposition to achieve energy-efficient computation at latencies, throughput and density that exceed that of general-purpose processors. The aim of this work is to enable data analysts to seamlessly utilize hardware accelerators by using traditional data center programming frameworks such as MapReduce, Storm and Spark.

2. APPROACH

Programmable accelerators have not yet been integrated well in the software stacks of multi-core programming frameworks used in cloud-based data analytics. The lack of clean integration is due in part to the semantic gap between accelerators and general-purpose processors (e.g., differences in the memory model), as well as lack of resource management support (e.g., for allocation and scheduling of tasks on accelerators). While the high-performance computing community has adopted hybrid programming models, this solution is highly undesirable for virtualized accelerators in the cloud. As such, we adopt a library of implementations of algorithms on accelerators and enable them to be called from high-level programming frameworks.

Key challenges to address in this work are to (i) hide the accelerator from the programmer by presenting it as a library function, embeddable in query processing, data processing or aggregation tasks; (ii) extend the runtime systems of high-level analytics languages to handle efficiently scheduling, communication, and synchronization with programmable accelerators; and (iii) improve the performance robustness of analytics written in high-level languages against performance artefacts of virtualization.

3. INITIAL RESULTS

We have characterized the overhead of off-loading tasks from Spark onto Maxeler DataFlow Engines (DFE). We have measured the performance of linear correlation when executing on the DFE compared to executing in Spark on 8 CPUs. We have measured a speedup of 46–374x for data set sizes ranging from 100 to 10 M pairs of floats. Moreover, we have measured that moving data between the Java Virtual Machine and the DFE has about 3% overhead over execution of the DFE kernel, which makes this overhead practically negligible. In contrast, allocating and configuring the DFE takes about 3.2 seconds, and bears significant overhead.

4. CONCLUSION

Accelerators can make a step change in the energy-efficiency and density of data centers. Seamless integration of accelerators in data center programming frameworks is however an open question, which will be addressed by this work.

5. ACKNOWLEDGMENTS

This research is supported in part by the European Community under the VINEYARD (Horizon2020, grant 687628) and ASAP projects (FP7, grant 608224).

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019. White Paper

Heterogeneous High-Performance Reconfigurable Computing '16 Salt Lake City, Utah USA

DOI: 10.475/123_4