A ML-based Runtime System for Executing Dataflow Graphs on Heterogeneous Processors

Extended Abstract

Subho S Banerjee, Arjun P. Athreya, Zbigniew Kalbarczyk, Steven Lumetta, Ravishankar K. Iyer
University of Illinois at Urbana-Champaign, Urbana IL.

CCS CONCEPTS

• **Computing methodologies → Reinforcement learning**; • **Hardware → Hardware accelerators**; • **Software and its engineering → Operating systems**;

KEYWORDS

Accelerator, FPGA, GPU, Scheduling, Genomics

ACM Reference format:
Subho S Banerjee, Arjun P. Athreya, Zbigniew Kalbarczyk, Steven Lumetta, Ravishankar K. Iyer. 2018. A ML-based Runtime System for Executing Dataflow Graphs on Heterogeneous Processors. In *Proceedings of SoCC '18: ACM Symposium on Cloud Computing, Carlsbad, PA, USA, October 11–13, 2018 (SoCC '18),* 2 pages.
https://doi.org/10.1145/3267809.3275474

This paper briefly describes the design a system that addresses the challenge of scheduling distributed data-analytics workloads on heterogeneous processing fabrics (which include CPUs, GPUs, FPGAs and ASICs) in cloud-based, dynamic, multi-tenant environments. To demonstrate the capabilities of the proposed system, we use a computational genomics workflow that addresses the growing need for rapid genomic analyses in hospital environments [6].

Accelerators such as GPUs, FPGAs and other ASICs have become an integral component of modern datacenters [5]. However, several key issues remain for extensive user adoption of these technologies: (1) identifying acceleration targets and designing accelerators for distributed applications; (2) composing these accelerators using data-flow graphs (DFGs) to build applications; and (3) optimally orchestrating, deploying and executing these DFGs in a dynamically changing environment in order to maximize hardware performance. While identifying and designing accelerators is an application/domain specific problem, we present a system that allows users to deploy large DFG based applications on heterogeneous processing fabrics by designing a scheduler that can operate these accelerators in a dynamically changing environment. The system is able to extract the maximum possible performance from the heterogeneous system, while maintaining the DFG-based high-level user abstraction of modern data-analytics frameworks. To understand the value of Symphony, it is important to understand the value of accelerators in datacenters.

Accelerators are typically custom designed to support targeted applications. Our previous work has addressed these design points for computational genomics applications, where we have identified a small number of computationally intensive *kernels* that are reused across different genomics analyses [1], and have built CPU- and GPU-optimized implementations of these kernels as well as custom hardware accelerators on FPGAs [1–4]. These provide significant improvements in runtime and further savings in the energy efficiency of the applications (directly related to the operational costs of running the application). In addition to scheduling tasks to these accelerators, the proposed system provides support to multiplex many accelerators on the same FPGA by managing the common IO- and board-specific-services, as well as reconfiguring the FPGA when an accelerator is no longer reacquired.

At its core, the proposed system uses a data-driven strategy of integrating real-time measurements of performance counter events, prior knowledge about workloads, and information about system architecture and interconnect topology to infer system state (i.e., resource utilization) information to schedule tasks to processors. The scheduling algorithm is a fusion of two machine learning (ML) techniques. First, real-time performance counter measurements are used to infer hidden system resource utilization (i.e., utilizations that cannot be directly measured) using Bayesian statistics on Factor Graphs (FGs). A FG is the most general Bayesian model that integrates probabilistic (i.e., data-driven insights from profiling) and algebraic (i.e., prior knowledge about system

SoCC '18, October 11–13, 2018, Carlsbad, PA, USA

ACM ISBN 978-1-4503-6011-1/18/10…$15.00
https://doi.org/10.1145/3267809.3275474

architecture) relationships between the hidden and observed variables. Second, the inferred system state along with user DFG encoding are fed into a deep reinforcement learning (DRL) model to make intelligent scheduling decisions to optimize an utility function corresponding to runtime or the operational costs of executing a DFG. The compelling value of this method (and its two constituent ML models) is that it allows for the deployment of accelerator-centric DFG in dynamically changing timeshared environments. Incorporating these ML models allows us to (1) produce scheduling policies that are specific to system-architecture and dynamic workload characteristic; and (2) bootstrap the DRL based scheduling policy using samples generated from the generative FG model. The technique improves overall performance, resource utilization, enables fine-grained resource sharing and reconfiguration.

We demonstrate the efficacy of the proposed system for the *variant calling and genotyping analysis* [7] on human genome datasets appropriate for clinical use. This represents a complex analysis where the aforementioned kernels are composed into several applications, and the applications are composed into a workflow. The proposed system reduces the total time required to complete the benchmark from 73 hours (CPU baseline) to under 1.2 hours (61×) using our FPGA-based accelerators, using NVIDIA K80 GPUs further reduces the runtime to under 40 minutes (109×). This corresponds to a 210× reduction in performance per unit energy terms.

REFERENCES

[1] Subho S. Banerjee, Arjun P. Athreya, Liudmila S. Mainzer, C. Victor Jongeneel, Wen-Mei Hwu, Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer. 2016. Efficient and Scalable Workflows for Genomic Analyses. In *Proceedings of the ACM International Workshop on Data-Intensive Distributed Computing (DIDC '16)*. ACM, New York, NY, USA, 27–36. https://doi.org/10.1145/2912152.2912156

[2] Subho S. Banerjee, Mohamed El-Hadedy, Jong Bin Lim, Daniel Chen, Zbigniew T. Kalbarczyk, Deming Chen, and Ravishankar K. Iyer. 2017. ASAP: Accelerated Short Read Alignment on Programmable Hardware (Abstract Only). In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, FPGA 2017, Monterey, CA, USA, February 22-24, 2017*. 293–294. http://dl.acm.org/citation.cfm?id=3021796

[3] Subho S. Banerjee, Mohamed El-Hadedy, Jong Bin Lim, Zbigniew T. Kalbarczyk, Deming Chen, Steven S. Lumetta, and Ravishankar K. Iyer. 2018. ASAP: Accelerated Short-Read Alignment on Programmable Hardware. *CoRR* abs/1803.02657 (2018). arXiv:1803.02657 http://arxiv.org/abs/1803.02657

[4] Subho S. Banerjee, Mohamed El-Hadedy, Ching Y. Tan, Zbigniew T. Kalbarczyk, Steven S. Lumetta, and Ravishankar K. Iyer. 2017. On accelerating pair-HMM computations in programmable hardware. In *27th International Conference on Field Programmable Logic and Applications, FPL 2017, Ghent, Belgium, September 4-8, 2017*. 1–8. https://doi.org/10.23919/FPL.2017.8056837

[5] Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, Michael Haselman, Scott Hauck, Stephen Heil, Amir Hormati, Joo-Young Kim, Sitaram Lanka, James Larus, Eric Peterson, Simon Pope, Aaron Smith, Jason Thong, Phillip Yi Xiao, and Doug Burger. 2014. A Reconfigurable Fabric for Accelerating Large-scale Datacenter Services. In *Proceeding of the 41st Annual International Symposium on Computer Architecuture (ISCA '14)*. IEEE Press, Piscataway, NJ, USA, 13–24.

[6] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. 2015. Big Data: Astronomical or Genomical? *PLOS Biology* 13, 7 (jul 2015), e1002195. https://doi.org/10.1371/journal.pbio.1002195

[7] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. 2013. *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.* John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43