Approximate Computing and Microfluidic Cooling for Enhanced Machine Learning

Hardik Sharma * William Wahby* Thomas Sarvey Muhannad S. Bakir Hadi Esmailzadeh
Georgia Institute of Technology
{hsharma, wwahby3, tsarvey}@gatech.edu, muhannad.bakir@mirc.gatech.edu, hadi@cc.gatech.edu

Abstract

The aggressive pace of Moore's law has dramatically increased both the performance and the number of transistors available as computational resources in modern integrated circuits, but heat dissipation has emerged as a key limiter of high performance computing systems. In order to address this challenge, we will combine advanced heat sinking technologies with novel approximate computing techniques to develop an ultrahigh performance platform for machine learning applications. Typically, computational systems are limited to simply trading performance for heat dissipation, but approximate computing enables the exploration of a third dimension: computational precision. We will implement Deep Neural Networks (DNNs) on a Stratix V FPGA modified to support an on-die microfluidic heat sink as a thermal-computational testbed to investigate the tradeoffs between performance, computational precision, and heat dissipation in machine learning algorithms.

1. Introduction

The end of Dennard scaling has disrupted the cadence of improvements in the performance and energy-efficiency of general purpose computing platforms though traditional CMOS scaling. Although the transistor count on modern ICs can still sustain exponential growth governed by Moore's law, the increasing gap between transistor count and energy efficiency has led to dark silicon [1, 2]. Ultimately, computing performance is significantly limited by thermal considerations, leading to the development of advanced heat sinking technologies in order to eke out additional performance from the same hardware. Recently the concept of computational sprinting was introduced to increase the throughput and performance for highly-parallel workloads by temporarily exceeding the chip's thermal envelope [6, 7]. Unfortunately, computational sprinting does not address the need for higher sustained computing performance, leading to a demand for alternate approaches which can deliver sustained improvements to both energy-efficiency and heat removal. We propose to combine architectural and hardware approaches to address both of these needs simultaneously.

Advanced heat sinking technologies effectively increase the thermal budget for computation. Microfluidic cooling has recently been demonstrated as an attractive option for aggressive cooling of microelectronic devices [9, 12]. From an architectural standpoint, approximate computing has emerged as an attractive approach to increase the throughput in a given computing envelope and form factor using techniques like limited precision computation [11], and approximate storage [4, 8]. Singly, either of these approaches could significantly enhance computational performance; we anticipate that the combination of these approaches will yield ultra high performance on imprecision-tolerant parallel workloads.

* These authors contributed equally to this work.

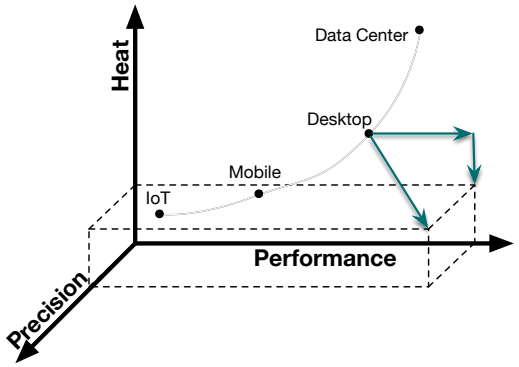

Figure 1. Performance, precision, and heat generation are intrinsically linked. Future high performance systems must intelligently balance each of these factors.

We will implement a novel approximate precision computing framework on a microfluidically-cooled FPGA to enable ultra-high performance on machine learning workloads. Machine learning workloads are ideal testbeds for approximate computing methods, as they are tolerant of relatively high levels of imprecision. Reducing computational precision will enable the FPGA to be partitioned into multiple parallel processing units, enabling ultra-high throughput, while the microfluidic heat sink will enable full utilization of the additional processing cores without exceeding the thermal limits of the system. High performance systems must navigate tradeoffs between performance, precision, and heat generation, as illustrated in Fig. 1; combining microfluidic cooling and approximate computing will lead to systems which can intelligently manage the tradeoffs between these conflicting design criteria.

2. Proposed System

We propose to combine the architectural and the hardware approaches discussed in the subsequent sections to enable an ultra-high performance machine learning platform. We will implement the dynamic machine learning framework onto an Altera Stratix V FPGA, which has been modified with an integrated microfluidic cooler, as shown in Fig. 2 [9]. Based on previous testing, we expect to be able to fully utilize the FPGA resources for parallel computations, as the microfluidic cooling solution enables all of the computational resources to be used without exceeding the thermal limits of the chip. Using this testbed we will fully characterize the tradeoffs between precision, parallelism, energy efficiency, and heat generation. In the following subsections, we describe each component of the proposed system in further detail.

2.1 Framework for Accelerating Deep Neural Networks

Deep Neural Networks (DNNs) are compute intensive machine learning workloads that can benefit significantly from acceleration.

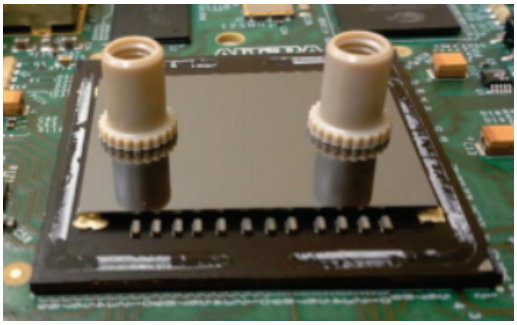

Figure 2. Altera Stratix V die with integrated microfluidic heat sink. Figure from [9]

We propose a framework based approach for accelerating DNN workloads using FPGAs. Instead of designing an accelerator for a particular DNN model, we use DNNWEAVER, a framework that generates synthesizable accelerators tailored to the specified DNN and the target FPGA.

Programming interface. We leverage the commonalities between computations in a DNN to provide a coarse-grained ISA and expose this ISA to software. With DNNWEAVER, the programmer can specify a DNN model using high level abstractions in Caffe format [3].

Instruction set architecture design. To minimize the overhead of the von Neumann execution model (instruction fetch, decode, etc.), we choose a macro data flow virtual machine for DNNWEAVER and expose its ISA to the software. The instructions for this virtual ISA are translated to microcodes and state machines, and are embedded in the final design for static scheduling. Static scheduling greatly simplifies the hardware and provides improved energy efficiency.

Template based accelerator architecture. DNNWEAVER uses hand-optimized templates and automatically generates a synthesizable accelerator optimized for the specified DNN and target FPGA platform. As depicted in Figure 3, DNNWEAVER's templates provide a clustered hierarchical architecture for acceleration. The key features of the templates are *customizability* and *scalability*. The template architecture is clustered into independent Processing Units (PUs), each consisting of several Processing Engines (PEs) as depicted in Figure 3. The clustered hierarchy provides scalability through data locality within PUs and by using an untied bussing fabric among PUs.

Approximation for increased parallelism. DNNWEAVER provides the flexibility to vary precision in both computation as well as storage. By lowering precision, the framework reduces the resources consumed by both computational and storage components in each Processing Unit. Hence, by lowering precision, DNNWEAVER can accommodate more Processing Units within the resources available in an FPGA to achieve higher parallelism.

2.2 Microfluidic Cooling

High performance computational systems experience significant thermal challenges. Typically, extracting additional performance from high performance hardware requires additional power consumption, but air-cooled computing devices are frequently thermally-limited, leading to the increasing prevalence of dark silicon, or on-chip computing resources which cannot be fully utilized due to thermal limits. In order to address the thermal challenge, heat removal technologies have been significantly explored in the literature, ranging from thermoelectric cooling, passive cooling with novel high-conductivity materials, and liquid cooling [5]. Of the

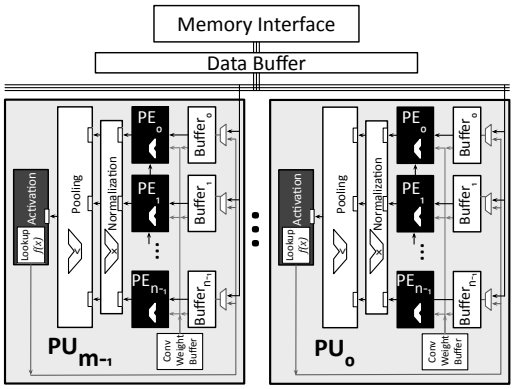

Figure 3. Overview of the hierarchical template design. The template accelerator is clustered and divided into Processing Units (PUs), comprised of multiple Processing Engines (PEs).

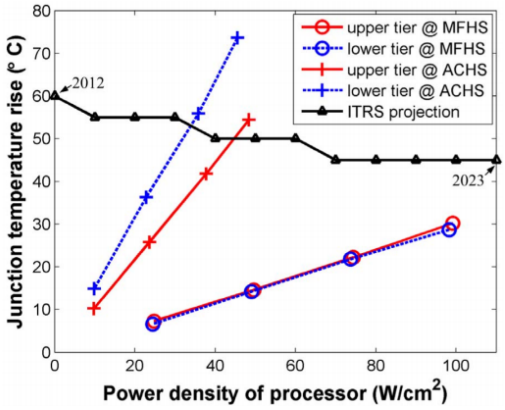

Figure 4. Thermal response of air-cooled (ACHS) and microfluidically-cooled (MFHS) two-die stacks. Measurements were performed with dummy dice incorporating heaters and thermometers to emulate high-power circuitry. Significant reduction in junction temperature is observed for the fluidically-cooled dice. Figure from [12]

various cooling approaches, liquid cooling is attractive due to its ability to deal with large steady-state heat fluxes. Microfluidic cooling, initially proposed in 1981 [10], represents the ultimate limit of liquid cooling, in which microchannels are etched on the back side of the silicon die, through which coolant can be pumped to remove heat directly from the die. Microfluidic cooling has been recently shown to be an attractive option for heat removal from high power devices, as shown in Fig. 4. Recently a microfluidic cooler was implemented for the first time in an active FPGA die, leading to significant throughput improvements and elimination of dark silicon for the evaluated benchmark design [9].

3. Conclusion

We have proposed using a microfludically-cooled FPGA as a functional thermal testbed to enable an ultra-high performance machine learning engine. Individually, aggressive cooling solutions such as microfluidic cooling and novel computing paradigms such as approximate computing can both deliver significant performance enhancements to high performance computing systems. We anticipate that the integration of these two technologies will unlock even greater levels of performance from the same hardware.

References

[1] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger. Dark silicon and the end of multicore scaling. In *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*, pages 365–376. IEEE, 2011.

[2] N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki. Toward dark silicon in servers. *IEEE Micro*, 2011.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn. Flikker: Saving refresh-power in mobile devices through critical data partitioning. In *ASPLOS*, 2011.

[5] R. Mahajan, C. pin Chiu, and G. Chrysler. Cooling a microprocessor chip. *Proceedings of the IEEE*, 94(8):1476–1486, Aug 2006. ISSN 0018-9219. doi: 10.1109/JPROC.2006.879800.

[6] A. Raghavan, Y. Luo, A. Chandawalla, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin. Computational sprinting. In *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*, pages 1–12. IEEE, 2012.

[7] A. Raghavan, L. Emurian, L. Shao, M. Papaefthymiou, K. P. Pipe, T. F. Wenisch, and M. M. Martin. Computational sprinting on a hardware/software testbed. *ACM SIGPLAN Notices*, 48(4):155–166, 2013.

[8] A. Sampson, J. Nelson, K. Strauss, and L. Ceze. Approximate storage in solid-state memories. In *MICRO*, 2013.

[9] T. Sarvey, Y. Zhang, L. Zheng, P. Thadesar, R. Gutala, C. Cheung, A. Rahman, and M. Bakir. Embedded cooling technologies for densely integrated electronic systems. In *Custom Integrated Circuits Conference (CICC), 2015 IEEE*, pages 1–8, Sept 2015. doi: 10.1109/CICC.2015.7338365.

[10] D. Tuckerman and R. Pease. High-performance heat sinking for vlsi. *Electron Device Letters, IEEE*, 2(5):126–129, May 1981. ISSN 0741-3106. doi: 10.1109/EDL.1981.25367.

[11] S. Venkataramani, V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan. Quality programmable vector processors for approximate computing. In *MICRO*, 2013.

[12] Y. Zhang, L. Zheng, and M. Bakir. 3-d stacked tier-specific microfluidic cooling for heterogeneous 3-d ics. *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, 3(11):1811–1819, Nov 2013. ISSN 2156-3950. doi: 10.1109/TCPMT.2013.2281605.