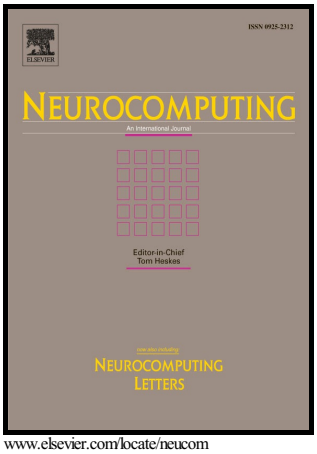

www.elsevier.com/locate/neucom

Energy Efficient Parallel Neuromorphic Architectures with Approximate Arithmetic on FPGA

Qian Wang, Youjie Li, Botang Shao, Siddhartha Dey, Peng Li

PII: S0925-2312(16)31121-3
DOI: http://dx.doi.org/10.1016/j.neucom.2016.09.071
Reference: NEUCOM17600

To appear in: *Neurocomputing*

Received date: 15 January 2016
Revised date: 11 June 2016
Accepted date: 2 September 2016

Cite this article as: Qian Wang, Youjie Li, Botang Shao, Siddhartha Dey and Peng Li, Energy Efficient Parallel Neuromorphic Architectures with Approximate Arithmetic on FPGA, *Neurocomputing* http://dx.doi.org/10.1016/j.neucom.2016.09.071

Energy Efficient Parallel Neuromorphic Architectures with Approximate Arithmetic on FPGA

Qian Wang[a], Youjie Li[a], Botang Shao[b], Siddhartha Dey[a], Peng Li[a]

[a]*Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843 USA*
[b]*Freescale Semiconductor, Inc., Austin, TX 78735 USA*

Abstract

In this paper, we present the parallel neuromorphic processor architectures for spiking neural networks on FPGA. The proposed architectures address several critical issues pertaining to efficient parallelization of the update of membrane potentials, on-chip storage of synaptic weights and integration of approximate arithmetic units. The trade-offs between throughput, hardware cost and power overheads for different configurations are thoroughly investigated. Notably, for the application of handwritten digit recognition, a promising training speedup of 13.5x and a recognition speedup of 25.8x are achieved by a parallel implementation whose degree of parallelism is 32. In spite of the 120MHz operating frequency, the 32-way parallel hardware design demonstrates a 59.4x training speedup over the single-thread software program running on a 2.2GHz general purpose CPU. Equally importantly, by leveraging the built-in resilience of the neuromorphic architecture we demonstrate the energy benefit resulted from the use of approximate arithmetic computation. Up to 20% improvement in energy consumption is achieved by integrating approximate multipliers into the system while maintaining almost the same level of recognition rate achieved using standard multipliers. To the best of our knowledge, it is the first time that the approximate computing and parallel processing are applied to FPGA based spiking neural networks. The influence of the parallel processing on the benefits of approximate computing is also discussed in detail.

1. INTRODUCTION

The human brain controls all our body movements, cognitive activities, emotions and other complex tasks. When it comes to complex tasks such as face recognition and language learning, a human brain can solve such problems with ease demonstrating much improved energy and space efficiency and show even better performance than supercomputers [1]. Brain-inspired computing has attracted much research interest, not only because of its application as a practical tool in areas such as pattern recognition, but also as a means of developing an understanding of mammalian brains and ultimately increasing our understanding of intelligence and consciousness.

Although most real world applications such as the processing of sensory inputs and pattern recognition can be realized by software models on Von Neumann machines, software simulation of complex biologically plausible models is intrinsically slow and may require tremendous energy consumption and space resources to solve these real-world problems. Brain-inspired neuromorphic hardware systems provide an appealing architectural solution to the above problems. They show good energy efficiency, potentially improved scalability and great suitability for pattern recognition problems. In addition, since one important property of neural networks is their parallel distributed nature, it is highly desirable to develop efficient parallel neuromorphic architectures for significantly acceleration. Meanwhile, the inherent error resilience and fault tolerance offered by brain-inspired architectures provide promising opportunities for leveraging approximate computing for additional energy and silicon area benefits.[2]-[10]

Traditionally, analog circuits are used to implement silicon neurons [11] [12]. However, they are difficult to reconfigure and intrinsically sensitive to process, voltage and temperature (PVT) variations [13][14]. For example, a same analog circuit design may probably demonstrate require different performance under different environments. In addition, large-scale integration of spiking neurons is hindered by the use of area consuming capacitors to keep synaptic weights [15]. The impact of PVT variations on the performance of digital neuromorphic designs is thoroughly investigated by [16], which provides guidance on the design of robust digital spiking neural circuits.

FPGAs offer great flexibility and reconfigurability for fast prototyping and hardware acceleration of software algorithms. To facilitate the application of SNNs in embedded systems and develop processing acceleration

Email addresses: qwangku@tamu.edu (Qian Wang), lyj2013apply@tamu.edu (Youjie Li), jackieshao2011@gmail.com (Botang Shao), sidhart.de@email.tamu.edu (Siddhartha Dey), pli@tamu.edu (Peng Li)