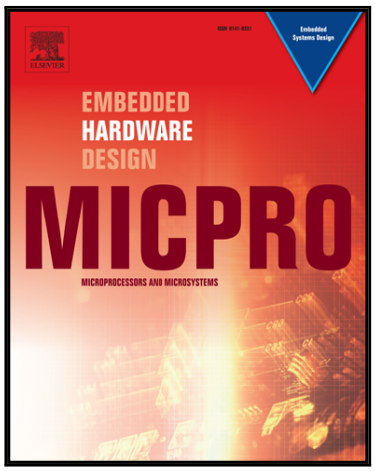

Energy Proportional Streaming Spiking Neural Network in a Reconfigurable System

Felipe galindo SANCHEZ , Jose Nunez-Yanez

PII: S0141-9331(17)30132-1
DOI: 10.1016/j.micpro.2017.06.018
Reference: MICPRO 2588

To appear in: *Microprocessors and Microsystems*

Received date: 27 February 2017
Revised date: 8 June 2017
Accepted date: 22 June 2017

Please cite this article as: Felipe galindo SANCHEZ , Jose Nunez-Yanez , Energy Proportional Streaming Spiking Neural Network in a Reconfigurable System, *Microprocessors and Microsystems* (2017), doi: 10.1016/j.micpro.2017.06.018

Energy Proportional Streaming Spiking Neural Network in a Reconfigurable System

FELIPE GALINDO SANCHEZ, University of Bristol
JOSE NUNEZ-YANEZ, University of Bristol

This paper presents a high-performance architecture for spiking neural networks that optimizes data precision and streaming of configuration data stored in main memory. The neural network is based on the Izhikevich model and mapped to a CPU-FPGA hybrid device using a high-level synthesis flow. The active area of the network is configurable and this feature is used to create an energy proportional system. Voltage and frequency scaling are applied to the processing hardware and memory system to deliver enough processing and memory bandwidth to maintain real-time performance at minimum power and energy levels. The experiments show that the application of voltage and frequency scaling to DDR memory and programmable logic can reduce its energy requirements by up to 77% and 76% respectively. The performance evaluation show that the solution is superior to competing high-performance hardware, while voltage and frequency scaling reduces overall energy requirements to less than 2% of a software-only implementation at the same level of performance.

Index terms: Hardware → Integrated circuits → Reconfigurable logic and FPGAs → Hardware accelerators, Hardware → Integrated circuits → Reconfigurable logic and FPGAs → Reconfigurable logic applications

1. INTRODUCTION

The ability to understand parts of the human brain and being able to perform large-scale simulations has allowed biologically inspired techniques in applications such as speech recognition, computer vision, natural language processing, drug discovery and pattern recognition, among others. The high computational costs of this processing has resulted in the development of high-performance solutions describing biologically realistic neuron models in hardware accelerated solutions using Field Programmable Gate Array (FPGAs) and Graphical Processing Units (GPUs).

In this paper we propose a fully connected feed-forward network using the Izhikevich's neuron model combined with a synapse model in which the post-synaptic current is modelled as a function of the synaptic conductance and the differential of the membrane potential and the reversal potential. Both models are described in C++ and then targeted to a Zynq MPSoC using high-level synthesis and implementation tools. The hardware uses an optimized synaptic weight representation and transmission using the AXI4-Stream as the primary data protocol. The current solution shows excellent scalable properties with configurations up to 250K neurons and 125M synapses possible in a single Zynq 7100 device. The performance evaluated on a Zynq 7020 device with the simulation of 28,900 neurons and 5M synapses results in speedups of 5 to 7 times in contrast with alternative computing solutions while using less than 2% of the energy of an Intel solution based on OpenCL.

This work is supported by the UK EPSRC under grants ENPOWER and ENEAC

Author's addresses: Felipe Galindo Sanchez, University of Bristol, Woodland Road, Bristol, BS8 1UB. j
and Jose Nunez-Yanez, Electronic Engineering Department, University of Bristol, Woodland Road, Bristol, BS8 1UB. j