Nano Communication Networks 17 (2018) 21–29

Contents lists available at ScienceDirect

Nano Communication Networks

journal homepage: www.elsevier.com/locate/nanocomnet

A Spiking Neural Network implemented with Single-Electron Transistors and NoCs

Beatriz dos Santos Pês [a,*], Janaina Gonçalves Guimarães [b], Elder Oroski [c], Marlio José do Couto Bonfim [d]

[a] Federal Institute of Parana (IFPR), Campus Campo Largo, 83601-190, PR, Brazil
[b] Department of Engineering, University of Santa Catarina (UFSC), Campus Blumenau, 89036-256, SC, Brazil
[c] Department of Electrical Engineering, University of Parana (UTFPR), Campus Curitiba, 80230-901, PR, Brazil
[d] Department of Electrical Engineering, University of Parana (UFPR), Campus Politécnico, 81530-900, PR, Brazil

ARTICLE INFO

Article history:
Received 15 March 2018
Received in revised form 9 May 2018
Accepted 9 June 2018
Available online 19 June 2018

Keywords:
SNNs
NoC
Spiking neurons
ANNs
Nanoelectronics

ABSTRACT

This work proposes a Spiking Neural Network, SNN, based on a nanoelectronic spiking neuron – as building block – and a 2D-mesh network-on-chip, NoC — as interconnect architecture. The SNN obtained from the NoC is an alternative for high density architectures, providing reconfigurability, high scalability and low power consumption. A look-up table router was used to connect all units. Further on, the eXclusive-OR, XOR, benchmark problem was used to validate the functionality of the nanoelectronic SNN.

1. Introduction

For the last 50 years, MOS, Metal Oxide Semiconductor, has been the main technology in electronic industry. A great advantage of MOS transistors is that their physical dimensions can be easily reduced [1]. Miniaturization is a tendency on electronic devices due to many performance improvements that may be reached through it. Nanoelectronics emerged as an alternative technology, capable of providing very small scale devices. Nanoelectronic devices, such as SET, Single-Electron Transistor, are ruled by quantum principles [2,3]. This feature restrains charge transport to occur through tunneling events. The current, then, consists of few electrons [4]. Thus, SET has great current control which reduces power dissipation [5]. High frequency operation, low power consumption and scale make nanoelectronics a promising technology [6].

Artificial Neural Networks, ANNs, were developed aiming the emulation of biological neural systems behavior. Some unique features of the biological brain, such as high level of parallelism, fault tolerance and great capability of data processing are desirable in electronic circuits [7]. The large variety of applications illustrates the success obtained by ANNs. They can be applied in financial predicting [8] and biomedical field [9], among others high complex tasks [10,11].

In order to obtain a model closer to the biological neural system, networks of spiking neurons began to be considered [12]. These networks interact through pulses and, therefore, are more resembling to biological systems than traditional ANNs [13,14]. SNNs present potential for building highly dense, massively parallel and fully interconnected systems [7]. Such networks would have great capability of data processing and, due to the high level of parallelism, good fault tolerance [12]. SNNs are widely used in the community of computational neuroscience [14]. However, hardware implementation of a system with all those features is challenging [12].

Highly dense systems must be, at some level, fault tolerant in order to present reliability [12]. In order to build those robust circuits, researchers looked for a reconfigurable way of implementing SNNs. In this sense, FPGAs seemed the obvious choice [15].

When the firsts FPGAs were introduced, in 1980, the goal was to build large circuits with high performance [16]. On a summarized description, a FPGA consists of various independent programmable blocks which can be interconnected to create larger functions [17]. Despite the reconfigurable feature and the inherent parallelism, FPGAs are not able to support the high density interconnection of SNNs [12]. In fact, mapping SNNs using logic blocks limits the number of neurons in the network, since these blocks are not efficient regarding power consumption and scale [18]. Besides, the

* Corresponding author.
E-mail addresses: beatriz.santos@ifpr.edu.br (B.d. Santos Pês), janaina.guimaraes@ufsc.br (J.G. Guimarães), oroski@utfpr.edu.br (E. Oroski), marliob@eletrica.ufpr.br (M.J.d.C. Bonfim).

https://doi.org/10.1016/j.nancom.2018.06.001
1878-7789/

routing structure in FPGAs do not support the high number of interconnections needed to build a dense SNN [19].

Shared wires interconnection was also considered to build hardware implemented SNNs. Although the topology of a shared wired seems a good choice to emulate the interconnection between neurons, this attempt also proved inefficient [7]. For a fully interconnected network, the number of wires is proportional to the product between the number of neurons of the presynaptic layer and the number of neurons of the postsynaptic layer. Thus, the occupied area grows exponentially, depending on the number of neurons in the network [7]. It was clear at that time that a new interconnection paradigm was needed. The search for a solution capable of supporting high density of interconnection, presenting good yield and scalability lead to Networks-on-Chip, NoCs [20]. NoCs are formed by interconnection matrices, processing elements and routing elements. The routing element is able to reduce the number of interconnections. Since a NoC is inherently redundant, fault tolerance is also addressed by this paradigm [21].

Some implementations of SNNs with NoCs can be found in literature. With different neuron models, routing algorithms and NoC architectures, examples such as EMBRACE, EMulating Biologically-inspiRed ArChitectures in hardwarE [12], SpiNNaker [22] and FACETS, Fast Analog Computing with Emergent Transient Scale [23], can be mentioned. This work proposes the use of a single-electron spiking neuron model. The SET neuron model may provide gain concerning scale and power consumption. For that reason, in this work, the SET neuron model is the processing element of the NoC, the routing element is implemented through a Look-Up Table, LUT. In order to validate the network functionality, the eXclusive-OR, XOR, benchmark problem was used [12,13].

This work can be regarded as a feasibility study concerning the use of a SET neuron model as processing element in a NoC architecture aiming the implementation of a SNN. This paper can be considered as complimentary to existing work, since a SET neuron model is applied as processing unit of a SNN mapped in a 2D mesh topology. Besides, difficulties found during the deployment of this work may be helpful to future research in the area.

This paper is divided as follows: Section 2 presents a brief discussion concerning related works. Section 3 presents the main points on implementing SNNs and some models of nanoelectronic spiking neurons. Section 4 is dedicated to an introduction and main ideas concerning NoCs. Section 5 presents the processing unit of the NoC used in this work. Section 6 details the routing unit used in this work. Section 7 presents the simulation results. Section 8 presents some discussions concerning the results. Finally, Section 9 is dedicated to the conclusions.

2. Related work

Neuromorphic computation applies Very Large Scale Integration, VLSI, systems to mimic neuro-biological structures present in the mammalian nervous system [24]. Despite some criticism, neuromorphic computation offers an appealing alternative beyond the traditional von Neuman computer paradigm [14]. This section discusses some proposals published on the last years concerning SNNs.

Fidjeland et al. [25] proposed a platform for simulating Izhikevich model of spiking neurons [26]: NeMo. That platform uses high parallel graphics processing units, GPUs, to improve its performance. The management of spike transfer relies on an external memory. This proposal made use of FPGAs in order to implement the SNN.

FACETS project [23] uses the wafer-scale integration approach in order to address the complex interconnection pattern in dense networks. For this approach to be feasible, low power consumption and fault tolerance are required. BrainScaleS [27], the successor of FACETS, reported the implementation of 2400 analog neurons through the use of switched capacitor technology and communicating via an asynchronous event-driven bus.

SpiNNaker [28] is a hardware approach based on ARM9 cores, communication is made through a 2D torus NoC. Each NoC node is capable of simulate 1000 leaky-integrate-and-fire, LI&F, neurons [29] based on traditional CMOS technology, each one with 1000 synapses inputs.

EMBRACE [12] proposes a compact hardware SNN architecture. EMBRACE employs the partition of complex tasks into smaller and simpler subtasks executed in distinct neural modules. This paradigm is known as Modular Neural Network, MNN. The communication is on-chip, based on a customized array of NoC routers. The processing element of the NoC is a neural block, which can be programmed to operate as a spiking neuron.

In 2014, IBM developed TrueNorth [30,31], a digital neuromorphic chip. The interconnection pattern is implemented via a 2D mesh NoC and the neurocomputer is also based on traditional CMOS technology.

In 2016, Sengupta et al. [32] proposed a hybrid spintronic-CMOS implementation for a SNN. The main purpose was to address energy efficiency comparable to the human brain [33].

All works previously mentioned have one thing in common: the usage of CMOS technology. However, the mismatch between mechanisms involved in mammalian nervous system and MOS transistors impose a limit in the capability of emulating biological efficiency. This limit is specially strong concerning area and power efficiency [32]. Using EMBRACE as main inspiration, this paper proposes to employ a nanoelectronic SET neuron model to address scale and power consumption problems. Interconnections are implemented via a 2D mesh NoC. Further on, XOR benchmark was chosen due to its wide use in artificial neural networks proposals [12,13]. In this work, the results were obtained via transient analysis using LTSpice IV.

It is worth mentioning, though, that the pulsed paradigm is often considered in the implementation of logic gates [34,35]. Guerreiro et al. [36] described a spiking neuron capable of implementing logic gates with simpler architectures, shorter learning periods and faster processing when compared to traditional neurons. Regarding the comparison between SET and CMOS logic gates, Pal et al. [34] studied the performance of both technologies, the simulation results showed that SET gates consume less power and are capable of faster information processing.

Next section presents some analysis concerning hardware implementations of SNNs.

3. Hardware implementation of spiking neural networks

The processing capability and the fault tolerance of nervous systems can be credited to two main characteristics: the high number of processing units and the complex pattern of interconnection [12]. The human brain has about 10^{10} neurons connected through 10^{15} parallel branches [37].

SNNs can be defined simply as networks of spiking neurons [13]. Since they are suitable for building dense and parallel systems, researchers began to look at these networks as more powerful, computationally speaking, than traditional networks [12]. In SNNs, the information is encoded through the timing of pulses, the topology of the network and the synaptic weights [7].

In order to provide reasonable parallelism the SNN requires a large number of neurons [12]. That feature made physical implementation and simulation of SNNs become a problem. Standard simulation tools in regular computers are not able to simulate high density networks. In fact, software implemented networks become slower as the number of neurons grows [7]. Therefore, the hardware approach replaced the software approach in many researches [12]. It is worth mentioning that, although SNN software