Computers and Electrical Engineering 45 (2015) 362–373

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

A model for communicating long synapses with guaranteed latencies on large neural networks ☆

Andres Gaona-Barrera [a,b], J. Manuel Moreno-Arostegui [a,*]

[a] Department of Electronic Engineering, Technical University of Catalunya, Barcelona, Spain
[b] Laboratory for Automation, Microelectronics and Computational Intelligence, Universidad Distrital, Bogota, Colombia

ARTICLE INFO

Article history:
Received 31 January 2014
Received in revised form 5 May 2015
Accepted 6 May 2015
Available online 27 June 2015

Keywords:
Circuit switching
Guaranteed latency
Network on chip
Randomly interconnected neuronal networks
Spiking neural network

ABSTRACT

This paper introduces a new approach for the implementation of randomly interconnected neural networks on hardware taking into account the length of the synapses. We divide the synapses into Long and Short according to the distance between the source and target neurons in a 2D mesh, and we demonstrate that it is possible to guarantee the latency of the Long synapses when they are routed through an additional layer which is based on hierarchical structures of Networks on Chip (NoC). The connection scheme consists in grouping neurons into four regions and communicating their sets of synapses between a pair of them, using circuit switching. In order to validate the interconnection scheme, we simulated the operation of this additional layer for two regions in a neuronal network with grid structure arrangement comprising 1.03×10^6 neurons, with a firing rate of 100 Hz and an average of 10^4 synapses per neuron. This pair of regions can support an average of 562 Long synapses per neuron, which is equivalent to managing 5% of the traffic generated by the grouped neurons, with the advantage of having the latency of the synapses guaranteed. A node of the one region has 30,528 neurons and operates with a throughput of 2.95 Millions of spikes per second (Mspk/s) approximately. In a complete operation, the additional layer has four regions and it would support 58 Mspk/s and 520,672 neurons of the network.

1. Introduction

Neurobiology has studied the physiological and topological processes of biological neural networks, particularly in mammals, in an attempt to explain or interpret the complex dynamic functions of the nervous system from the point of view of interactions among neurons and synapses [1,2]. However, many inner mechanisms of the brain are still a mystery, because the brain is a very complex system that performs massive parallel processing and it is highly intercommunicated. For example, it is estimated that the brain has billions of neurons, and each one can be communicated with thousands of other neurons located in different regions through synaptic connections. This also makes it difficult to establish the contribution of a particular neuron to the whole [3].

A synapse transmits an electrical stimulus (spike) generated inside a source neuron, towards one of the receptor neurons to which the source neuron is connected. That stimulus must be replicated and distributed throughout part of the neuronal structure or neuron population by means of multiple synapses in order to reach its final destination. A spike is produced when weighted stimuli received at the inputs of a neuron exceed the firing threshold of the neuron [4–8].

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Terrence Mak.
* Corresponding author. Tel.: +34 93 401 56 91; fax: +34 93 401 67 56.
E-mail addresses: andres.eduardo.gaona@upc.edu (A. Gaona-Barrera), joan.manuel.moreno@upc.edu (J.M. Moreno-Arostegui).

http://dx.doi.org/10.1016/j.compeleceng.2015.05.007
0045-7906/

A promising approach to understanding and simulating biological neuronal interactions is to use Spiking Neural Networks (SNN) on silicon, which permits to process large scale neuronal functions, and to do it on real time [9–11]. This is possible because they no longer use processing based on Von Neumann architectures, but replace them by customized solutions using parallel processing or NoC that increase the performance of this kind of systems [12–17].

Implementation of Large Scale SNNs (LSSNN) has evolved from flat structures in bus arrangement, to models in which a hierarchical layer arrangement enables neurons to be clustered. Bus based SNNs are not scalable because they need as many buses as the number of neurons in the network, and all the neurons that receive synaptic connections should be connected to each bus [18,19]. This makes bus based SNNs a clearly limited model. Other approaches use NoCs in mesh or torus topology in order to perform synaptic processes by communicating neuron clusters [19–22]. They have afforded densities of one million of neurons, and its routers yield throughputs up to 16 Gbps [14]. However, in these models it is not easy to estimate the latency of the synapses, because it varies according to the size and congestion of the NoCs. Additionally, the synapses between neurons that are in distant locations of the grid must travel through large amounts of routers, so their latency is penalized due to the normal NoC traffic.

In this paper we propose a modification to the synapse coding scheme, which is to divide synapses into Long or Short according to the number of routers that a packet must go through in order to reach its destination. The Long connections occur between distant neurons and are going to be routed through an additional layer, which will enable us to guarantee their latency. The connection between neurons uses two types of clusters, called nodes, and one circuit switching router to interconnect them. The routers that we use have a parallel structure and they provide bufferless routing and pipeline operation. The nodes associated to the target neurons are internally organized in an open ring topology of sub-nodes, where each sub-node contains a pair of registers in order to store or forward the packet with the synaptic information. The simulated network can support an average of 5.5% of the traffic of the neuron clusters comprised in the network, with maximum peaks of approximately 10% for the more distant neurons of the grid, with the guaranteed latency of all Long synapses. Embrace or other hierarchical architectures can be kept for the short [13,20]. However, their traffic load would be reduced by means of the additional layer for Long synapses.

This article is organized as follows: Section 2 presents a brief description of some relevant implementations of large-scale hardware neural networks, based on the switching interconnection scheme. Section 3 shows the long-synapses interconnection model proposed, and Section 4 describes the main components of the model from the hardware perspective. Section 5 evaluates the interconnection model considering an LSSNN comprised by one million neurons and 10^4 synapses per neuron. Finally, Section 6 explains the most relevant contributions of the interconnection model and Section 7 is dedicated to the conclusions of this work.

2. Related works

Several authors have demonstrated that using parallel hardware processing units together with customized communication structures in order to emulate large scale SNNs, implies benefits such as real time operation or more efficient scalability [23,24]. An example of the advantages of parallel versus sequential processing is provided by Software SNN models [25,26]. These models offer flexibility to modify the network architecture, but they are not easily scalable and are slow on large simulations. Regarding the communication infrastructure, several developments of SNNs on FPGA have shown that the arrangement in logic cells of this technology [9,27] and the organization of its interconnection matrix prevent managing large SNNs efficiently, due to strong constraints in the communication structure bandwidth and in the memory capacity required to perform synaptic processes.

Neuronal networks organized under flat [14,21,22,28] or hierarchical [13,20] NoC structures are able to support neuronal traffic, are scalable, and permits to manage the traffic generated on LSSNNs by using packet switching mainly [29,30]. Although the latency of the synapses is not constant when this scheme is used, the throughputs yielded are suitable to perform the synaptic processes required. The distribution of the synaptic connections in a mesh topology makes it difficult for neurons on the distant or opposite edges of the array to intercommunicate using a network with packet switching. This is due to drawbacks such as a high latency, congestion and strong granularity of the synapses in the NoC. Therefore, long synapses must be treated differently than short ones.

When circuit switching is conventionally used in order to connect two points on a network, a dedicated path is established between the transmitter and the packet receiver [31–35]. The circuit is released only when there is no more information to be sent from the transmitter. However, it has been shown that when multiple connections must be established on these types of NoC applications, the buildup of reserved links degrades the performance of the system, connection waiting times are cumulative and it is difficult to establish new circuits [32]. For those reasons, using circuit switching in order to perform neuron to neuron connections on LSSNNs is not convenient. Also, the high granularity of large neuronal models prevents this solution from being scalable.

This work proposes an additional layer for the packet switched mesh used in [13,20]. The new layer establishes a dedicated path between a neuron and a region of the mesh, which modifies the point-to-point connection paradigm on dedicated circuits. Only long connections will travel through the new layer using a circuit switching scheme that will guarantee the latency on these sorts of connections, which is not possible with the current implementations.