DHyANA: A NoC-based Neural Network Hardware Architecture

Priscila C. Holanda, Cezar R. W. Reinbrecht, Guilherme Bontorin, Vitor V. Bandeira and Ricardo A. L. Reis
PPGC/PGMicro — Instituto de Informatica — Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre, Brazil
{pcholanda, cezar.reinbrecht, gbontorin, vvbandeira, reis}@inf.ufrgs.br

***Abstract*—Understanding and modeling the brain is one of the key scientific challenges in the twenty-first century, and a grown effort is rising on a global scale. Due to its high parallelism, the hardware implementation of large-scale spiking neural networks (SNNs) promises superior execution speed compared to sequential software approaches. Such systems can significantly benefit from the use of networks-on-chip(NoC), as they scale very well concerning area, performance, power/energy consumption, and overall design effort. We developed a hierarchical network-on-chip for a hardware SNN architecture to improve the communication and scalability of the system. The architecture was implemented in an Altera Stratix IV FPGA, and a logic synthesis was performed to evaluate the system, achieving an area of 0.23mm^2 and a power dissipation of 147mW for a 256 neurons implementation.**

***Index Terms*—Spiking Neural Network, Network-on-Chip, Hierarchy, Digital.**

I. Introduction

The brain is an amazing three-pound organ, and definitely one of the most complex and magnificent organs in the human body. It is formed by a network of more than 100 billion single nerve cells interconnected in systems that construct our perceptions of the world, fix our attention, and control the machinery of our actions [1]. The efforts put into the quest for knowledge of such a complex organ have raised significantly within the last years, what some call a global blossom of neuroscience research.

Such researches could bring several benefits to the human kind. From the ability to better understand and seek treatment for brain disorders to the expertise to model and develop intelligent systems, the motivations for implementing brain models are countless.

The development of Brain-Machine Interfaces (BMIs) and neuroprostheses, for instance, would help improve the quality of life of several people affected by neurological disorders, such as Alzheimer's disease. Still, to become possible, the realization of such prostheses requires the development of neuronal network models that are able to interact with biological neuronal cell assemblies, considering the intrinsic spontaneous activity of neuronal networks and understanding how to drive them into a desired state or to produce a specific behavior [2].

If one would have the need to simulate the interconnection topology of, say, a mammalian neocortex, the number of synapses per neuron should be on average between 2×10^3 and 2×10^4 [3], which is a significant limiting factor in the suitability of its implementation in hardware. In the context of a similar connectivity problem on System-on-Chip (SoC) design, where interconnect scalability is paramount, the concept of Network-on-Chip (NoC) was introduced. Within such concept, elements from traditional computer networking are employed in order to realize the communication of the hardware structure.

This paper proposes a new digital hardware architecture for a spiking neural network using hierarchical Network-on-Chip (NoC) communication.

II. Related Works

From supercomputer emulations to full-custom designs, there are a number of brain models found in literature, and several reviews were published [4]–[6]. Focusing on designs that have implemented custom large-scale computing platforms, this section presents some recent neuromorphic efforts.

The SpiNNaker project, from University of Manchester, aims to provide a platform for high-performance massively parallel processing by integrating a microprocessor-based system containing 18 ARM968 processor cores onto a single die [7]. Each of the systems nodes, consisting of a System-in-Package of the ARM cores microprocessor plus a 128Mbyte off-die SDRAM stacked on top of it, are interconnected by an NoC using six links wrapped into a triangular lattice.

The Stanford University Neurogrid [8]–[10] and the Heidelberg University BrainScales [11], [12] are mixed signal neuromorphic multi-chip systems. Both works, although different in many aspects, use analog computation to emulate neural dynamics, and digital communication schemes to support synaptic connections.

The EMBRACE project from University of Ulster and National University of Ireland [13], [14], and the ROLLS neuromorphic processor from University of Zurich and ETH Zurich [15] are low-power mixed-signal approaches. The EMBRACE architecture proposes an SNNs based on a hierarchical array of NoC routers, based on a hybrid star-mesh topology, while the ROLLS processor uses asynchronous digital logic circuits for setting different network congurations.

The IBM SyNAPSE project developed the TrueNorth chip, in which 4096 neurosynaptic cores are tiled in a 2-D array, containing an aggregate of 1 million digital neurons and 256 million synapses [16]. Its architecture also uses hierarchical communication, with a high-fanout crossbar for local communication and a network-on-chip for long-distance

978-1-5090-6113-6/16/$31.00 ©2016 IEEE

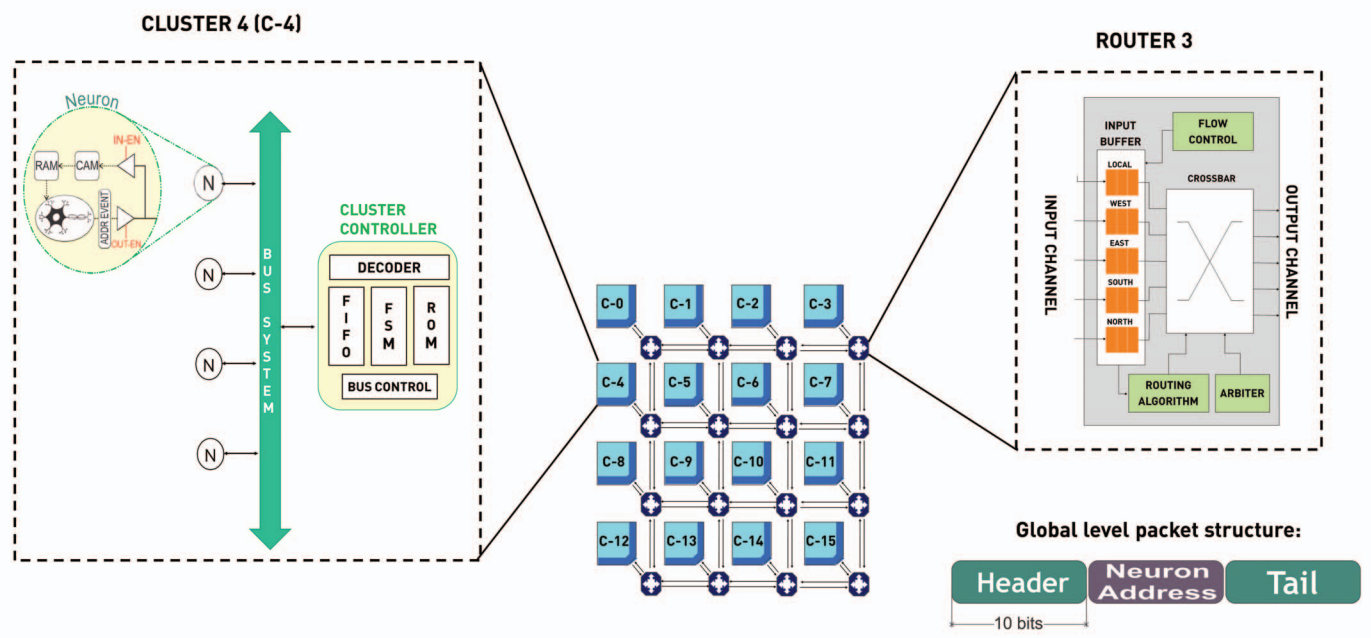

Fig. 1: DHyANA Architecture.

communication, and global system synchronization to ensure real-time operation.

III. DHyANA Architecture

This section presents the Digital HierArchical Neuromorphic Architecture (DHyANA). It is composed of a hierarchical NoC as the communication infrastructure and a digital Izhikevich neuron model as the processing element.

The hierarchical NoC is based on the HiCIT architecture proposed by Reinbrecht in [17]. Moreover, this hierarchical proposal has applicability for 3D integrated circuits, as presented in [18], being a strategic structure for new systems implementations. DHyANA hierarchical NoC is composed of two levels, the cluster and the global level. The processing elements, or neurons, are organized in groups and interconnected through a bus system, defining the cluster level. Moreover, these groups are interconnected trough a mesh-based NoC.

An example of this architecture is shown in Figure 1, where 16 clusters are interconnected through a 4x4 mesh NoC. Next subsections describe each communication level and the processing element.

A. Global Level

The inter-cluster communication has low activity during application execution comparing with intra-cluster communication. However, these few messages demand high throughput to handle the burst of spikes, to perform internal synchronization. It has been shown in works such as [19], [20] that SNNs can exhibit high communication locality. Therefore, a mesh-based NoC can provide sufficient throughput during these intermittent burst messages. Besides, it can provide flexibility, because it avoids bottlenecks in an unexpected inter-cluster communication, independently of message size or rate.

DHyANA uses a 2D electrical mesh NoC with XY routing. The routers contain five ports, four to connect to the neighbor routers (north, east, south and west ports) and one to connect to the cluster (local port). Each local port connects to the cluster level, where a cluster controller component bridges both communication levels. The router input ports have buffers to handle high throughput messages, and a handshake to manage the data flow. Each of the router output ports has a crossbar, which switches the input according to the arbiter, which employs a Round-Robin scheduling algorithm.

The packets in the global level are composed of three parts: the Header; the Address-Event (AE); and the Tail. The header provides information of source and destiny to the routers, the AE contains the neuron address at the cluster, and the tail sets the end of the package. The structure of the packet can also be observed in Figure 1.

B. Cluster Level

A cluster is organized by computation complexity, communication requirement and functional relationship of IP cores [21]. Hence, inside these groups, the communication is very intense, independently of its behavior (high or low data rates). A solution that provides low latency is necessary to avoid a long waiting time. Besides, most communications on neural network systems are multicast messages (one node send to many). The bus system can provide good performance in such conditions with low cost in area and power. Its main limitation is the scalability, where the performance is limited by the number of nodes interconnected. Since this proposal uses clusters of 16 elements, the bus system is applied, in order to achieve better results in area and power. However, if future applications require bigger clusters, the cluster level will change to a crossbar-based solution, as presented by [17].

TABLE I: DHyANA and Related Work synthesis data

Ref.	**Project Name**	Neuron Model	A/D	Neurons	Power	Size	CMOS Process	Synapses per Neuron	Synapse Storage	Topology
[7]	**SpiNNaker**	IF or IZHI	DS	$16x10^3$	1W	$102mm^2$	130nm	$1x10^3$	TSM, SDRAM	Triangular Lattice*
[8]	**Neurogrid**	Quad. IF	A	983,040	5W	$168mm^2$	180nm	$6x10^9$	1-bit RAM	Star
[11]	**BrainScaleS**	Exp. IF	A	512	1kW	$430mm^2$	180nm	$112x10^3$	SRAM	Hierarchical Buses**
[13]	**EMBRACE**	IF	A	400	13.16mW	$0.587mm^2$	65nm	400	-	Hybrid (star-mesh)
[16]	**TrueNorth**	-	-	$1x10^6$	65mW	$430mm^2$	28nm	256	SRAM	Mesh
[15]	**ROLLS**	Exp. IF	A	256	4mW	$51.4mm^2$	180nm	256	1-bit bistable	-
This	**DHyANA**	IZHI	D	256	147mW	$0.23mm^2$	65nm	256	CAM, RAM	Hybrid (bus-mesh)

A = Analog, D = Digital, DS = Digital (Software), IF = Integrate-and-Fire, IZHI = Izhikevich. *folded into a toroid surface. **2D Torus (wafer). SpiNNaker, BrainScaleS, TrueNorth data per chip; EMBRACE data per cluster.

The cluster level communication is composed of three main parts: a bus system, a cluster controller, and the neurons.

1) Bus System: The bus system comprehends a handshake protocol and an Address Event Representation (AER) approach [22], [23], in which an address is assigned to each neuron cell in a chip, so that when a cell activity occurs, it is broadcast to all computational nodes within a defined region.

The AE packet is defined by the following premises: each neuron module is assigned an 8-bit address, in which the four least significant bits represent the neuron position within each cluster and the four most significant bits represent the cluster position within the system.

2) Cluster Controller: The cluster controller is a bridge between the mesh and the bus system. It modifies the content of the packets to translate the information involved. For this reason, a table with the NoC packets is stored in a local ROM memory, and are accessed for as many times as there are cluster connections for each particular neuron that spiked. Moreover, the controller has a FIFO to support high-throughput communications. As the communication behavior of a neural network system is very specific, the cluster controller uses a specific finite state machine (FSM) to manage the propagation of the spikes of any neuron.

The FSM contains eight states to provide the correct operations regarding spike propagation. When a spike occurs at any neuron, it sends the controller a request. Then, the controller reads its AE, and checks the ROM for the specific neuron's first position within the table. Such position contains a FLAG indicating how many clusters are connected to the neuron. So, for as many times as indicated, it reads the ROM and assembles the NoC packet.

C. Neuron Model

The neuronal ionic mechanisms which generate the action potentials are known today mostly due to the pioneering work developed by Hodgkin and Huxley in 1952 [24]. By researching the behavior of giant squid neurons, they developed a mathematical model which can reproduce all kinds of neurons with good precision in terms of shape of spike and complex firing activities [2], being the most biologically accurate model to date. Since then, a lot of mathematical models were developed, each varying in levels of complexity, computational intensity and biological accuracy, and a good overview of them can be seen in [25].

One of the simplest models is the Leaky Integrate-and-Fire (LIF), which basically idealizes a neuron as having Ohmic leakage current and a number of voltage-gated currents deactivated at rest [26], at the cost of being not as nearly biologically plausible.

Another commonly used neuron mathematical model was proposed by Izhikevich [27], which stands in the middle ground between complexity and biological plausibility scale. In fact, although not as simple as the LIF, the Izhikevich model can mimic various nonlinear responses of biological neurons, making it almost as versatile as the Hodgkin-Huxley model at a fraction of its computational cost, being thus the model of choice for this work.

DHyANA uses a digital, low latency Izhikevich neuron model on hardware, as proposed in [28], in which a highly parallel and combinational circuit was developed. This choice, however, comes without loss of generality. Thus, at any circumstance, whenever another neuron model is more suitable for a certain application, it can be incorporated into the system.

IV. Experimental Results

Without loss of generality, a 4 x 4 NoC with a varying number of neurons in each cluster has been described in mixed-language, VHDL and Verilog HDL. The design was evaluated based on hardware simulations and in FPGA for real-time testing.

A network of 208 neurons (4 x 4 x 13) was implemented in an Altera Stratix IV FPGA (EP4SGX230KF40C2), using 99% Logic Utilization (154,826 ALUTs and 27,043 Dedicated logic registers). The maximum clock speed for the network was 59.62 MHz. For comparison, the same FPGA chip can be filled with up to 364 neurons only (without the memory cells) [28].

Additionally, a 65nm logic synthesis was performed for 256 neurons (4 x 4 x 16) to evaluate and better compare DHyANA with similar projects. Results can be observed in Table I.

V. Conclusion

In this work, a new digital neuromorphic device architecture was proposed, using hierarchical and network-on-chip approaches for neuron communication.

After successful simulation test, the architecture was implemented in FPGA, and a logic synthesis was performed. Both results showed a good area/power ratio.

More testing will be able to demonstrate DHyANA scalability. It is expected for the proposed device to be scaled enough to allow the simulation of neural networks with biologic realism. Then, applications within the areas of neuroprostheses or intelligent systems will be performed, and the systems functionality further proved.

Acknowledgment

This work is funded by the Federal Agency for Support and Evaluation of Higher Education of Brazil (CAPES), the National Council for Technological and Scientific Development (CNPq).

References

[1] E. R. Kandel, J. H. Schwartz, T. M. Jessell, and S. Mack, Eds., *Principles of neural science*. New York, Chicago, San Francisco: McGraw-Hill Medical, 2013. [Online]. Available: http://opac.inria.fr/record=b1135227

[2] M. Ambroise, T. Levi, Y. Bornat, and S. Saighi, "Biorealistic spiking neural network on fpga," in *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, March 2013, pp. 1–6.

[3] C. Johansson and A. Lansner, "Towards cortex sized artificial neural systems," *Neural Netw.*, vol. 20, no. 1, pp. 48–61, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2006.05.029

[4] H. de Garis, C. Shuo, B. Goertzel, and L. Ruiting, "A world survey of artificial brain projects, part i: Large-scale brain simulations," *Neurocomputing*, vol. 74, no. 13, pp. 3 – 29, 2010, artificial Brains. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231210003279

[5] C. Eliasmith and O. Trujillo, "The use and abuse of large-scale brain models," *Current Opinion in Neurobiology*, vol. 25, pp. 1 – 6, 2014, theoretical and computational neuroscience. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095943881300189X

[6] A. S. Cassidy, J. Georgiou, and A. G. Andreou, "Design of silicon brains in the nano-cmos era: Spiking neurons, learning synapses and neural architecture optimization," *Neural Networks*, vol. 45, pp. 4 – 26, 2013, neuromorphic Engineering: From Neural Systems to Brain-Like Engineered Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608013001597

[7] S. Furber, D. Lester, L. Plana, J. Garside, E. Painkras, S. Temple, and A. Brown, "Overview of the spinnaker system architecture," *Computers, IEEE Transactions on*, vol. 62, no. 12, pp. 2454–2467, Dec 2013.

[8] S. Choudhary, S. Sloan, S. Fok, A. Neckar, E. Trautmann, P. Gao, T. C. Stewart, C. Eliasmith, and K. Boahen, "Silicon neurons that compute." ser. Lecture Notes in Computer Science, A. E. P. Villa, W. Duch, P. rdi, F. Masulli, and G. Palm, Eds., vol. 7552. Springer, 2012, pp. 121–128.

[9] B. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. Chandrasekaran, J.-M. Bussat, R. Alvarez-Icaza, J. Arthur, P. Merolla, and K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[10] P. Merolla, J. Arthur, R. Alvarez, J.-M. Bussat, and K. Boahen, "A multicast tree router for multichip neuromorphic systems," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 61, no. 3, pp. 820–833, March 2014.

[11] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, May 2010, pp. 1947–1950.

[12] D. Schmidt, "Automated characterization of a wafer-scale neuromorphic hardware system," Master's thesis, University of Heidelberg, Germany, 2014.

[13] S. Carrillo, J. Harkin, L. McDaid, S. Pande, S. Cawley, B. McGinley, and F. Morgan, "Hierarchical network-on-chip and traffic compression for spiking neural network implementations," in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on*, May 2012, pp. 83–90.

[14] S. Carrillo, J. Harkin, L. McDaid, F. Morgan, S. Pande, S. Cawley, and B. McGinley, "Scalable hierarchical network-on-chip architecture for spiking neural network hardware implementations," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 12, pp. 2451–2461, Dec 2013.

[15] N. Qiao, H. Mostafa, F. Corradi, M. Osswald, F. Stefanini, D. Sumislawska, and G. Indiveri, "A re-configurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses," *Frontiers in Neuroscience*, vol. 9, no. 141, 2015.

[16] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, B. Taba, M. Beakes, B. Brezzo, J. Kuang, R. Manohar, W. Risk, B. Jackson, and D. Modha, "Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 34, no. 10, pp. 1537–1557, Oct 2015.

[17] C. R. W. Reinbrecht, "Desenvolvimento e avaliao de redes-em-chip hierrquicas e reconfigurveis para mpsocs," Master's thesis, Universidade Federal do Rio Grande do Sul, 2012.

[18] D. Matos, C. Reinbrecht, T. Motta, and A. Susin, "A power-efficient hierarchical network-on-chip topology for stacked 3d ics," in *2013 IFIP/IEEE 21st International Conference on Very Large Scale Integration (VLSI-SoC)*, Oct 2013, pp. 308–313.

[19] A. Kumar, S. Rotter, and A. Aertsen, "Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding," *Nat Rev Neurosci*, vol. 11, no. 9, pp. 615 – 627, 2010. [Online]. Available: http://www.nature.com/nrn/journal/v11/n9/suppinfo/nrn2886_S1.html

[20] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268 – 276, 2001.

[21] X. Leng, N. Xu, F. Dong, and Z. Zhou, "Implementation and simulation of a cluster-based hierarchical noc architecture for multi-processor soc," in *IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005.*, vol. 2, Oct 2005, pp. 1203–1206.

[22] M. F. Mahowald, "Vlsi analogs of neuronal visual processing: a synthesis of form and function," Ph.D. dissertation, Calif. Univ. Pasadena, Pasadena, CA, 1992, presented on 12 May 1992. [Online]. Available: https://cds.cern.ch/record/253521

[23] M. A. Sivilotti, "Wiring considerations in analog vlsi systems, with application to field-programmable networks," Ph.D. dissertation, Pasadena, CA, USA, 1991, uMI Order No. GAX91-37292.

[24] A. Hodgkin and A. Huxley, "Propagation of electrical signals along giant nerve fibres," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 140, no. 899, p. 177183, Oct 1952.

[25] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *Neural Networks, IEEE Transactions on*, vol. 15, no. 5, pp. 1063–1070, Sept 2004.

[26] ——, *Dynamical systems in neuroscience : the geometry of excitability and bursting*, ser. Computational Neuroscience. Cambridge, Mass., London: MIT Press, 2007. [Online]. Available: http://opac.inria.fr/record=b1125242

[27] ——, "Simple model of spiking neurons," *IEEE Trans. Neural Networks*, pp. 1569–1572, 2003.

[28] V. Bandeira, V. Costa, G. Bontorin, and R. Reis, "Low latency fpga implementation of izhikevich-neuron model," in *International Embedded Systems Symposium (IESS 2015)*, 2015.