2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)
4-7 March 2018
Las Vegas, Nevada, USA

Energy-Efficient Multipath Ring Network for Heterogeneous Clustered Neuronal Arrays

Andrei Ardelean, Amir Zjajo, Sumeet Kumar and Rene van Leuken*

Abstract— **Simulating large spiking neural networks with a high level of realism in a FPGA requires efficient network architectures that satisfy both the resource and interconnect constraints, as well as the changes in traffic patterns due to learning processes. In this paper, we propose a dataflow architecture based on a multipath ring topology that offers traffic shaping capabilities, and high energy-efficiency for the neuron-to-neuron communications.**

I. Introduction

Spiking neural networks (SNN) have become more attractive in current scientific research due to their ability to closely mimic biological neural behavior, such as encoding information in spike timing, amplitude and train patterns [1]. Unfortunately, this high level of realism comes at the cost of substantial computational effort and high throughput demands. In addition, as the neural network undergoes a learning process, the neuron communication scheme changes, which can result in low efficiency of the overall system if the physical network implementation is not adaptive.

The neural networks exhibit a high level of parallelism; consequently, field programmable gate arrays (FPGAs) offer a very suitable platform for SNN simulation [2]. FPGAs also benefit from the flexibility of being able to change both the neuron models and the network topology. However, the FPGA interconnect power and delay limitations [3] add another design challenge. To mitigate this, the neurons, along with memory blocks and control logic, are grouped into units (referred to as *clusters*) [4]. Throughput in the interconnect network and the resource costs are reduced by shared memory blocks and by sharing computing hardware for intensive operations (i.e. exponentiation in the extended Hodgkin-Huxley model [5]), respectively. Nevertheless, the need for efficient dataflow architecture between clusters persists.

In this paper, we propose a *multipath ring* (MPR) topology, which we characterize in terms of traffic distribution and energy-delay product (EDP). Furthermore, to reduce strain on the network even more, we separate the neuron-to-neuron communication from the configuration and input/output traffic using two separate physical layers with customized topologies. The contributions can be summarized as follows: *i*) traffic distribution estimation in a MPR topology interconnect network through an abstract mathematical method, *ii*) MPR EDP estimation and comparison with other low power/latency topologies using an adaptive model, and *iii*) a customized physical extraction, insertion and configuration layer for setup and input/output data transfers, separate from the neuron-to-neuron communications.

*All authors are with the Circuits and Systems Group, Delft University of Technology, Delft, 2628 CD, The Netherlands (e-mail: a.ardelean@student.tudelft.nl).

II. Two layered Network structure

The focus of our study are biophysically (electrochemically) accurate models of biological systems, such as the ones using the Hodgkin-Huxley formalism. Due to the extended Hodgkin-Huxley neuron model [6] complexity, in addition to initial dendrite potentials, up to 19 chemical parameters [4] are required to configure each simulated neuron. For resource optimization purposes, multiple neurons are associated to a cluster and, as a consequence, a very large number of double or single precision data packets are sent to each one of them. If the common routing scheme based on routing tables is used, in addition to the increase in router complexity an overhead will be present (the address of each data packet would have to be checked even though the target cluster is the same). To reduce this overhead, a channel-based scheme is proposed where the routers have the ability to be configured for a specific destination cluster to which all data packets will be automatically routed.

Fig. 1 a) and Fig. 1b) show the two-step method used to send configuration data to a cluster. At first, a setup packet is sent through the network, in which each router select a specific output port, depending on a bit in the packet correlated to the router position in the tree. A global control signal indicates the end of the setup phase, after which the resulted channel will remain in place, and each data packet that follows will be sent to the targeted cluster. The similar routing method is used for sending necessary input stimuli to the neurons during simulation. The output communication is always directed towards the exterior of the system, and there is no need for a routing table, so the routers could be simplified to just handle arbitration in the case of conflicts between packets. Consequently, a binary tree network is sufficient as the network topology for the extraction, insertion and configuration layer.

The layer that routes all neuron-to-neuron communication is referred to as the *topological layer*. The multipath ring architecture offers high connectivity [7], flexibility and low implementation costs. Mathematically, this network topology belongs to the regular graphs family, having a constant nodal degree (in this case number of network router ports) of 4. The interior connection links reduce the average path length in the network by *skipping* over a number of nodes. This skip value provides flexibility to the topology by allowing the designer to shape the traffic, and distribute it through specific links.

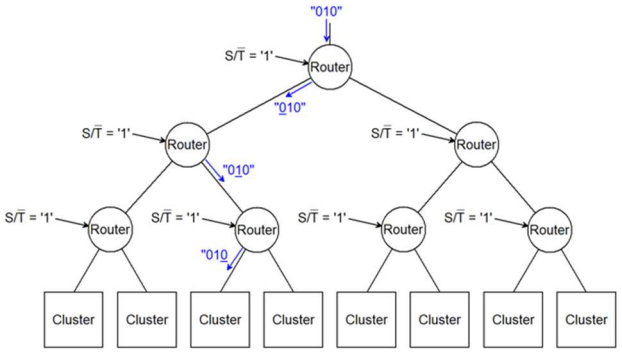

Figure 1. a) A configuration packet is sent and each router directs it according to a specific bit that is dependent on the router's position inside the network.

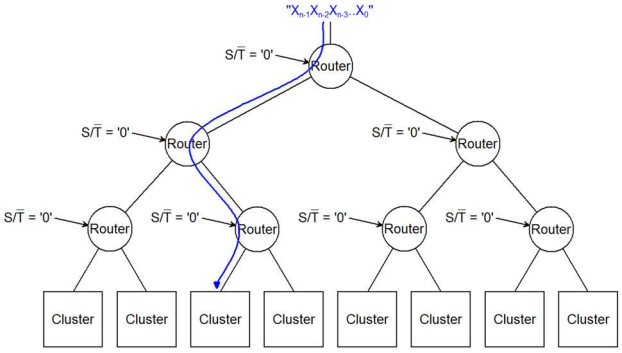

Figure 1. b) Once configured, the channel remains active and all data packets are routed to the same cluster without any further decision taking in the router.

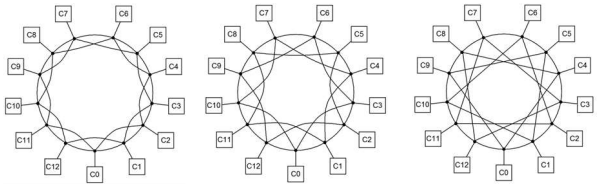

Figure 2. A multipath ring network with 13 clusters and a skip value of 1, 2 and 3.

Fig. 2 shows an example of a multipath ring network of size 13 with three different skip values. It should be noted that the MPR topology can accommodate any number of clusters without any redundancies, as opposed to a mesh (grid), or a two-dimensional toroid that may have extra routers if the number of clusters is prime. However, the size of the network sets a limit on the maximum skip distance that can be achieved: for example, a minimum number of 9 clusters is needed in order to have a balanced network with a skip of s3.

The downside of such a topology is having a relatively large average path length for very large networks [7]. However, neuron-to-neuron communications are a function of distance, with data locality being present i.e. neighboring neurons communicating more frequently than the ones further apart [1], thus, sharing memory blocks within clusters is an efficient technique for reducing throughput in the network. In addition, the number of clusters that can be implemented in currently available FPGA technology is limited, e.g. in [8] the cluster method averaging 6 clusters on a Xilinx Virtex 7 XC7VX550 is an optimal solution. As a result of this small network size, the aforementioned drawback of the MPR architecture does not materialize.

III. Multipath Ring Topology Analysis

Due to the high complexity involved in the analysis of very large neuron networks, a number of assumptions were made that simplify traffic and EDP estimations at the cost of using abstract descriptions for network behavior: *i*) a neuron contributes to the network traffic with an output rate, which depends on the model complexity, and is known at the beginning of the simulation, *ii*) all data packets are of equal but arbitrary size and *iii*) a data packet is sent in the correct direction when it encounters a router based on a set of static rules.

Although these assumptions appear to simplify the problem, they allow us to identify the critical links in the network, and to examine the effect changing neuron placement has on the network traffic. In the used routing scheme, a router will prioritize the path with the lowest number of hops between the source and destination i.e. if it can, the packet will be sent on an interior link first. To make the analysis communication-protocol independent, we assign all the data packets of equal but arbitrary size. Consequently, we can examine the influence a cluster has on the network in the form of a message injection rate α, which is measured in packets/simulation steps equal to the average output rate of the neurons inside the cluster.

A. Traffic Estimation

We estimate the traffic based on three parameters: α_i – number of distinct outgoing messages from cluster i, β_{ij} – number of messages from cluster i that are destined for cluster j, and η – average percentage of outgoing messages common to all destination clusters. All parameters are calculated after the neuron communication scheme and placement in the physical network are known. The output estimation values are the throughput of the exterior (φ_i for counter-clockwise and φ'_i for clockwise), and interior (θ_i and θ'_i) MPR connections at every cluster/router i. Assuming a multipath ring of size N and skip s, for every cluster n the following equation stands:

$$\Sigma_{0\le i<N}\beta_{i,n} = \varphi_{(n-1)modN} + \varphi'_{(n+1)modN} + \\ + \theta_{(n-s-1)modN} + \theta'_{(n+s+1)modN} \quad (1)$$

where we consider $\beta_{n,n}=0$. For example, in the case of the network shown in Fig. 3 that has $N = 5$ and $s = 1$, applying (1) results in the underdetermined system of equations (2).

$$\begin{cases} \beta_{1,0}+\beta_{2,0}+\beta_{3,0}+\beta_{4,0}=\varphi_4+\varphi'_1+\theta_3+\theta'_2 \\ \beta_{0,1}+\beta_{2,1}+\beta_{3,1}+\beta_{4,1}=\varphi_0+\varphi'_2+\theta_4+\theta'_3 \\ \beta_{0,2}+\beta_{1,2}+\beta_{3,2}+\beta_{4,2}=\varphi_1+\varphi'_3+\theta_0+\theta'_4 \\ \beta_{0,3}+\beta_{1,3}+\beta_{2,3}+\beta_{4,3}=\varphi_2+\varphi'_4+\theta_1+\theta'_0 \\ \beta_{0,4}+\beta_{1,4}+\beta_{2,4}+\beta_{3,4}=\varphi_3+\varphi'_0+\theta_2+\theta'_1 \end{cases} \quad (2)$$

As the system in (1) is underdetermined, i.e. the system has $4N$ unknown variables and only N equations, we use superposition to find the solution. The first step for solving (2) is given in (3), where only the effects of cluster C0 are

considered. After repeating this step for each cluster, the results are combined and solution (4) extracted.

$$\begin{cases} \varphi_0=\beta_{0,1} & \varphi'_0=\beta_{0,4} & \theta_0=\beta_{0,2} & \theta'_0=\beta_{0,3} \\ \varphi_{1,2,3,4}=0 & \varphi'_{1,2,3,4}=0 & \theta_{1,2,3,4}=0 & \theta'_{1,2,3,4}=0 \end{cases} \quad (3)$$

$$\begin{cases} \varphi_0=\beta_{0,1} & \varphi'_0=\beta_{0,4} & \theta_0=\beta_{0,2} & \theta'_0=\beta_{0,3} \\ \varphi_1=\beta_{1,2} & \varphi'_1=\beta_{1,0} & \theta_1=\beta_{1,3} & \theta'_1=\beta_{1,4} \\ \varphi_2=\beta_{2,3} & \varphi'_2=\beta_{2,1} & \theta_2=\beta_{2,4} & \theta'_2=\beta_{2,0} \\ \varphi_3=\beta_{3,4} & \varphi'_3=\beta_{3,2} & \theta_3=\beta_{3,0} & \theta'_3=\beta_{3,1} \\ \varphi_4=\beta_{4,0} & \varphi'_4=\beta_{4,3} & \theta_4=\beta_{4,1} & \theta'_4=\beta_{4,2} \end{cases} \quad (4)$$

In the case of such a small network, the resulting throughput values consist of only one β parameter, and as such they are exact and no scaling with η is required. However, for bigger networks, the solution is more complex, as in (5) where N=13 and s=2. Note that results from the same cluster are scaled with $(1-\eta)$.

$$\begin{aligned} \varphi_0 &= (1-\eta)(\beta_{0,1} + \beta_{0,2} + \beta_{10,1} + \beta_{10,2}) + \beta_{9,1} + \beta_{12,1} \\ \varphi'_0 &= (1-\eta)(\beta_{0,11} + \beta_{0,11} + \beta_{3,12} + \beta_{3,11}) + \beta_{1,12} + \beta_{4,12} \\ \theta_0 &= (1-\eta)(\beta_{0,3} + \beta_{0,4} + \beta_{0,5} + \beta_{0,6}) + \beta_{10,3} \\ \theta'_0 &= (1-\eta)(\beta_{0,10} + \beta_{0,9} + \beta_{0,8} + \beta_{0,7}) + \beta_{3,10} \end{aligned} \quad (5)$$

B. EDP Estimation

The EDP estimate for a packet sent from cluster i to cluster j is defined as

$$EDP_{i-j} = E_{i-j} * D_{i-j} \quad (6)$$

where E_{i-j} is the average energy required to send the packet from source to destination, and D_{i-j} is the delay of the path. The two parameters are calculated as

$$\begin{aligned} E_{i-j} &= N_{router} * E_{router} + \Sigma E_{link} \\ D_{i-j} &= (N_{link} - 2) * D_{link} + 2 * D_{clust} + N_{router} * D_{router} \end{aligned} \quad (7)$$

where E_{router} and E_{link} are the average energies consumed in the router and link segment as the packet passes through, with ΣE_{link} representing the total energy spent in the links, and N_{router} and N_{link} are the number of routers and links in the entire path, respectively.

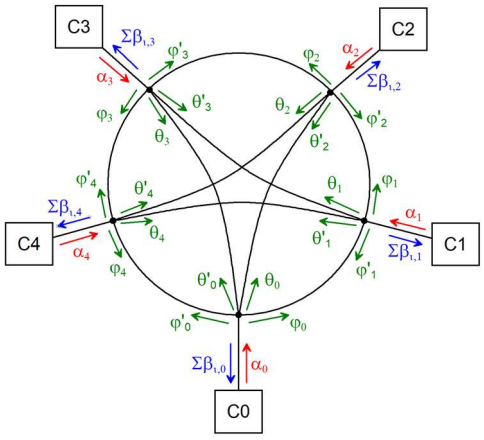

Figure 3. An example multipath ring network of size 5 with metrics.

Three elements influence the delay: D_{router}, the average delay through a router from one network port to the other, D_{clust} which is the delay from a cluster to its dedicated router, and D_{link}, the delay through a link segment. After implementing the design of choice for the network components, the delay parameters are extracted from synthesis reports, while the energy consumed is estimated using Xilinx XPE. All delay parameters, as well as E_{router}, can be changed accordingly. E_{link} is considered dependent on the link segment's length, and as such, has different values depending on the types of links in the network (e.g. the internal links in the MPR topology are physically longer than the external ones). The effect traffic has on the global EDP can be studied through (8), where EDP_x refers to the value resulted from (6) if an index difference of x is considered.

$$EDP = \sum_{i=0}^{N-1} \left(\varphi_i + \varphi'_i\right) \cdot EDP_1 + \sum_{i=0}^{N-1} \left(\theta_i + \theta'_i\right) \cdot EDP_{s+1} \quad (8)$$

IV. Experimental Results

Four commonly used neural network communication patterns were considered when studying traffic through the multipath ring: feed forward, layered full lateral inhibition, layered neighbor lateral inhibition, and complete hidden layer. 210 neurons were randomly placed in a multipath ring network consisting of 9 (the minimum size for which a skip of 3 is possible) up to 25 clusters for three different skip values (the number of logic cells available in a medium to top of the range Xilinx Virtex UltraScale is roughly four times as large as the one in the device used in [8]). In total, 1000 iterations were run for every MPR network configuration, and the traffic measurements were then averaged. The total number of neurons chosen can be easily divided amongst various numbers of equal sized clusters. The overall result characteristics would not change, if a larger number of neurons were used in the same number of clusters.

For EDP estimates, the network size was increased to 32 clusters to allow for a wider comparison range, and a message originating from cluster C0 was assumed to travel through the network. Three MPR networks of varying skip were compared with another low power design: *inverse clustering with mesh* [9] along with three benchmarking topologies: *binary tree, inverse clustering* and *mesh*. Static routing was assumed in all cases, the previously mentioned scheme for MPR, a shortest path approach for binary tree, inverse clustering and inverse clustering with mesh, and XY routing for mesh. All simulations are performed in MatLab.

The estimates for the maximum exterior and interior counter-clockwise traffic for the feed forward network are shown in Fig. 4. Due to the symmetric nature of the multipath ring, the clockwise traffic is identical. Also, due to the very large number of neurons in the hidden layer of the connection scheme, as opposed to the ones in the input and output ones, as well as the small cluster size (24 neurons at most), the average traffic values for the four connection schemes are almost identical. The results indicate that traffic is heavily influenced by changes in the skip. We can observe that increasing the skip redirects the traffic through the exterior links, and that there is an optimal value for which the network is balanced.

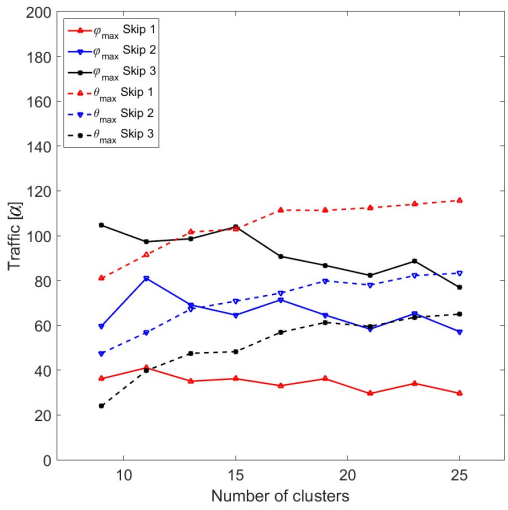

Figure 4. Maximum counterclockwise external (φ) and internal (θ) traffic estimates for Feed Forward.

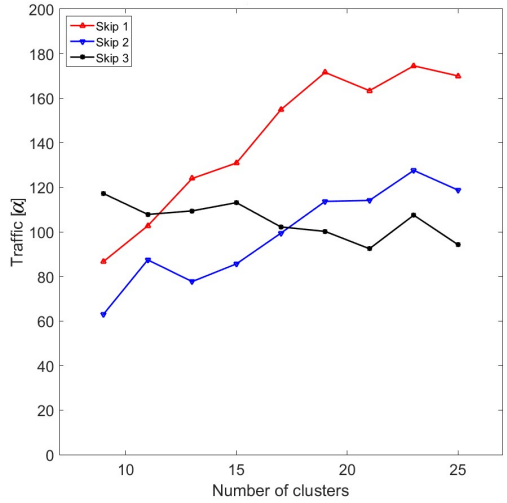

Figure 5. Average difference between maximum and minimum throughput values in the network for Feed Forward.

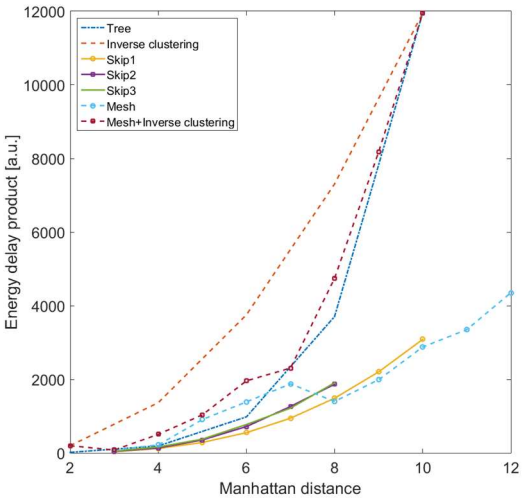

Figure 6. EDP of different network topologies of size 32 as a function of Manhattan distance

In this case, a skip of 2 is best suited for the analyzed size range, but tendencies in the plots suggest that for larger networks a skip of 3 would be more appropriate. This is confirmed in Fig. 5 as well, where the average difference between the largest and smallest throughput values for the feed forward network are shown. A skip of 1, however, is not recommended because it concentrates the traffic through interior links and for large networks this creates a large unbalancing effect.

Energy-delay product estimates, as a function of Manhattan distance (number of hops from source to destination cluster), are shown in Fig. 6. For small distances, the multipath ring has the smallest estimated EDP for all skip values. Tree and inverse clustering with mesh have an almost identical EDP, with the latter behaving marginally worse. As the distance increases, MPR with skip 1 becomes the best choice out of the three MPR topologies, with mesh becoming slightly more advantageous for large distances. Nevertheless, MPR with skip 2 and 3, which have almost the same EDP, have a smaller maximum distance (network diameter) than all the other topologies and remain the attractive solution.

V. Conclusion

In this paper, we proposed a layered structure for a clustered spiking neural network simulator to be implemented on FPGA, with a multipath ring as an energy-efficient topology for the neuron-to-neuron communication layer, and a simplified binary-tree for the configuration and input/output data. Traffic in networks of varying sizes was analyzed with the results showing that by changing the internal connections in the multipath ring, traffic can be shaped to either bring more balance in the network, or to concentrate through specific links. The energy-delay product estimates indicate that the multipath ring offers high energy-efficiency when compared with other low power architectures, and has the additional benefits of symmetry and physical simplicity.

VI. References

[1] W. M. K. W. Gerstner, Spiking Neuron Models: Single Neurons, Populations, Plasticity, Cambridge University Press, 2002.

[2] R. Wang et al, "An FPGA design framework for large-scale spiking neural networks," *IEEE ISCAS,* pp. 457-460, 2014.

[3] U. Farooq et al, Tree-based Heterogeneous FPGA Architectures Application Specific Exploration and Optimization, Springer, 2012.

[4] J. Christiaanse et al, "A real-time hybrid neuron network for highly parallel cognitive systems," *IEEE EMBC,* pp. 792-795, 2016.

[5] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology,* vol. 117, no. 4, pp. 500-544, 1952.

[6] J. d. Gruijl et al., "Climbing fiber burst size and olivary subthreshold oscillations in a network setting," *PLoS Computational Biology,* vol. 8, no. 12, pp. 1-10, 2012.

[7] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, no. 6684, pp. 440-442, 1998.

[8] J. Hofmann et al., "Multi-chip dataflow architecture for massive scale biophysically accurate neuron simulation," *IEEE EMBC,* pp. 5829-5832, 2016.

[9] V. George et al, "The design of low energy FPGA," *IEEE ISLPED,* pp. 188-193, 1999.