Neural Network Architecture for Hybrid Network-On-Chip using Scalable Spiking for Man Machine Interface

S. Santhosh Kumar\textsuperscript{1*}, S. Vidhya\textsuperscript{2} and M. M. Shanmugapriya\textsuperscript{3}

\textsuperscript{1}Department of Mathematics, KG College of Arts and Science, Coimbatore – 641035, Tamil Nadu, India; santhoshkumar.s@kgcascom
\textsuperscript{2}Department of Information Technology, KG College of Arts and Science, Coimbatore – 641035, Tamil Nadu, India; vidhya.s@kgcas.com
\textsuperscript{3}Department of Mathematics, Karaunagaram University, Coimbatore – 641035, Tamil Nadu, India; kumaths22@gmail.com

Abstract

Hardware-based computer vision accelerators are going to be an important part of future mobile devices to satisfy the low power and data processing demand. In order to comprehend a high power potency and high turnout, the accelerator design will be massively parallelized and tailored to vision process that is a plus over software-based solutions and all-purpose hardware. In this research Spiking neural networks (SNNs) arrange to emulate scientific discipline within the class brain supported neurons parallel arrays that communicate through spike events. The opportunity to perform embedded neuromorphic circuits is supplied by SNNs, with low power consumption and high correspondence in comparison with the normal laptop paradigms of John von Neumann. Even so, the poor property and modularity shortage as shown in ancient neuron cell interconnect implementations supported shared bus topologies is barring climbable hardware operations of SNNs. In order to effectively apply SNN traffic patterns and neighborhood among neurons in the current design the Hybrid Network on Chip (H-NoC) design integrates a spike traffic compression technique, thus dropping traffic overhead and up turnout on the network provides world traffic hundreds to sustain turnout underneath bursting activity. The planned system reduces overhead and improves the performance through native routing of the neuron cell facilities that are the gifts within constant tile facility. This will increase the potency of the system. The scalability of the adopted H-NoC approach under completely different situations is shown by analytical results show, while synthesis and simulation analysis reveal, area of low-cost, and delay for each cluster severally. This methodology finds its application in various sector such as medical image processing and bio signal processing.

Keywords: Hybrid Network, Hybrid Network on Chip, Neural Network, Scalable Spiking, Spiking Neural Network

1. Introduction

Spiking neural network models. SNNs arrange to emulate scientific discipline within the class brain supported immensely neurons parallel arrays that communicate by spike events. In order to process the information, the delay of spikes, the topology, and colligation weights are used by SNNs. A spiking neuron cell contains a cell body with several input branches called dendrites, which transfer info from diverse neurons. The neuron cell output or axon transfers info to different neurons in the sort of spikes. Spikes are conveyed between neurons through weighted connections called synapses. Associate degree output spike is produced by a neuron cell when its post-synaptic response surpasses a firing threshold price. This recurring method permits tasks equivalent to organization or pattern recognition to be accomplished. Furthermore, SNNs display the power to the Network-On-Chip (NoC) paradigm was introduced in as a promising resolution to unravel the on-chip communication issues intimate in modern Multiprocessor...
Neural Network Architecture for Hybrid Network-On-Chip using Scalable Spiking for Man Machine Interface

Indian Journal of Science and Technology

1.1 Back Ground and Contribution of Superior Computer Approaches

Currently, the Blue Brain project\(^1\) is assumed as the top of the HPC trend regarding SNNs. Presently, the Blue Brain project is in a state to simulate up to 108 forthright neurons or up to 104 awfully advanced neurons, (in this condition, the “simplicity” is dependent to the details of particle channel level, besides rules of native and world synaptic plasticity planned for each neuron). The simulation situation is maintained on the IBM Blue Gene/L, a HPC that proposes up to 360tera FLOPS, by using 8,192 PowerPC CPUs, every running at 700 megahertz\(^2\) and prepared in a torus interconnection network\(^3\). Even so, the scalability of the High performance computing approach (HPC) arises at the expense of a very high power consumption budget (typically, within the order of many kilowatts) and a low level of correspondence that limits time period performance networks that are large-scale.

1.2 Back Ground and Contribution of Full-Custom Style Approaches

Altogether, HPCs are too slow to perform huge simulations of SNNs and do not balance efficiently to support future networks. Therefore, alternative approaches equivalent to full-custom hardware executions are explored in order to beat the scalability drawback.

The Neurogrid project\(^4\) designed a mixed-signal desktop mainframe computer for neuroscientists. It makes use of analog computation in order to emulate ion-channel activity and an electronic communication theme for supporting colligation connections. Neurocore as the most building block is in a situation to accommodate an entire of sixty five, 536 quadratic integrate-and-fire neuron cell models, associate degree. It makes use of an external Field Programmable Gate Array (FPGA) and bank of Static Random Access Memory (SRAMs) intended for electronic communication among neighboring neuro cores. The Neuro grid platform will not allow the software system flexibility to be shown by the HPCs even if Neuro grid shows sensible power consumption of five W for emulation of 1M neurons by using a grid of neurocores. It conjointly integrates a restriction on the most number of different neurons in each layer (up to a pair of,175 neurons) that is unable to supply biological real time, because of the firing rate for worst-case situation drops out of up to 16.9 spikes per second (i.e., 59:9 ms interspike interval)\(^5\).

The sides seem to be\(^6\) designed as a mixed-signal, with high density hardware neural specification that support a mix of associate degree beside neurons and a scheme of digital multi-layer bus communication; all of these located on an uncut wafer. The core process building block of sides includes the High Input Count Analog Neural Network Chip (HICANN), that covers up to 512 Adaptive Exponential Integrate-and-Fire Neuron Cell Models (AdEx)\(^7\). A full wafer will include 384 HICANN chips, resulting in a total of 196,608 neurons per wafer.

This paper gifts a sophisticated gradable intelligence agent design for SNN implementations by the native...
routing of the neuron cell facilities that are present within constant tile facility. The most building blocks the prevailing work the packet from neuron cell facility zero to different neuron cell facilities happens through the tile and cluster facilities. The communication between 2 neuron cell facilities in same tile facility conjointly takes place through cluster facility that is associate degree unwanted method that will increase the overhead of the system. to cut back this overhead we have a tendency to introduce associate degree increased tile facility that not solely will is basic work as mentioned in base work, however conjointly the native routing of the neuron cell facilities that are the gifts within constant tile facility. This will increase the system strength.

The prearranged H-NoC design is effective supported associate degree Register Transfer Logic RTL-level implementation, while area ,memory and delay analysis are accomplished by exploitation ISE machine .Based on the results there are vital scalable characteristics like high turn-out and low power/area footprint that make it suitable for large-scale SNN-embedded implementations. In initial version of this study there exist on-chip communication levels and consequently the traffic compression technique for SNNs. Further planned system through native routing of the neuron cell facilities that are gift within constant tile facility and the results from the analysis of the gradable approach and traffic compression in terms of area/memory and Placement and Routing delay details are described using Virtex 7 Low Voltage FPGA using Plan ahead 13.2 version.

2. Materials and Methods

2.1 Typical H-NoC for Spiking Neural Networks

An advanced gradable intelligence agent design for Spiking Neural network implementations by combining star-mesh topologies for higher exploitation of the one-to-many multicast communication among neurons among the networks. The most building block of the prearranged H-NoC design is assumed as the cluster facility, anywhere there are some interrelated supported neurons hierarchical data structure, exploitation associate degree array of low and high-level intelligence agent routers for supporting the native (intracluster facility) and world (intercluster facility) networks between neurons. There is H-NoC design that specializes in the essential ideas and therefore the explanation behind the the H-NoC design specifications and style. The interior structure of the H-NoC design is shown in Figure 1.

It maintained 3 modules as follows:
1) Neuron Cell facility
2) Tile facility
3) Cluster facility

2.2 Neuron Cell Facility

In the first H-NoC level of the design there is the neuron cell facility in the hierarchy rock bottom. It joins four spiking neural cells (n = 4) exploiting the alleged node router that changes the received knowledge from the neural cells into packets of intelligence agent knowledge. The main level accommodates an entire of four neuron cell facilities (m = 4). Therefore, a complete of sixteen neural cells (i.e., m x n) are comprised at rock bottom of the HNoC design. The node router provides the communication infrastructure and incorporates the mandatory specialized hardware modules to transfer and receive spike events to/from any of the four neural cells allotted in every neuron cell facility. This node router conjointly works as a entrance to succeed in different neurons settled either within constant or in different neighboring cluster facilities. As a result of the target neurons knowledge encoded on the packets of the intelligence agent, on arrival of 1 packet, each initial node router has to authenticate the kind of on-chip communication through checking its header info, and then accomplishes an easy comparison
among the target address contained on the packet of the intelligence agent and therefore the native node router address. If there exists a match between them, the router absorbs the packet of the intelligence agent knowledge, otherwise the packet is either rejected or bypassed to another router. If the node router obtains a spike event from a neural cell, by the spike wrapper module initiates a packetization method. There is an application of the spike wrapper module if the input ports are used so that the single or multiple spike events would be born-again into a legitimate intelligence agent knowledge packet based on its sort. The packet is prepared in three fields of header, supply address, and target address.

### 2.3 Tile Facility

The second level of the hierarchy is the tile facility is that in order to cluster four neuron cell facilities uses the alleged tile router in a very network topology. Similar to the node router, every tile router supports multicast, unicast, associate degree broadcast communication protocols; this together with the network topology accustomed interconnect the four neuron cell facilities offer an economical thanks to accomplish the traffic demand of the neurons allotted in the middle level of the HNoC design\(^ {21,22} \). The tile router contains 2 crossbar switches. The downstream knowledge path receives intelligence agent knowledge packets returning as of the cluster facility. These packets of intelligence agent knowledge are then directed to multiple or individual neuron cell facilities at persistent time, reckoning conjointly on the kind of on-chip communication protocol and on the particular target position and already nominal on the info packet header.

On the other hand, intelligence agent knowledge packets is received by the upstream knowledge path receives from the neuron cell facilities, anywhere intelligence agent knowledge packets are either the intelligence agent knowledge packets is deviated off the upstream path to the downstream knowledge path or sent to the higher level at intervals the planned hierarchy, i.e., the cluster facility in order that any of the four neuron cell facilities gathered at intervals. Constant tile facility will absorb the intelligence agent knowledge packets.

### 2.4 Cluster Facility

Lastly, there is the cluster facility at the hierarchy highest level, associate degree it teams four tile facilities exploitation an adaptive router remarked for the cluster router. The cluster router allows the exchange of spike events among any of the sixteen neurons in the neurons cluster facility settled in neighboring cluster facilities. The cluster router incorporates associate degree adaptive routing theme that facilitates router adaptation per spike traffic hundreds, up router turnout per the traffic behavior bestowed through the network. It jointly provides multicast communication and broadcast with every cluster facility. In the current design, the gradable Network on Chip (H-NoC) design integrates a spike traffic compression technique in order to take advantage of SNN traffic patterns and neighborhood among neurons, and to reduce traffic overhead and up turnout on the network provides world traffic hundreds to sustain turnout underneath detonating activity. The planned system reduces overhead and improves the performance through native routing of the neuron cell facilities that are the gifts within constant tile facility.

### 2.5 Enhanced H-NoC Design for Spiking Neural Networks

This planned gradable intelligence agent design for SNN implementations by the native routing of the neuron cell facilities that are gift within constant tile facility. The most building block the prevailing work the packet from neuron cell facility zero to different neuron cell facilities happens through the tile and cluster facilities. The communication between 2 neuron cell facilities in same tile facility conjointly takes place through cluster facility that is associate degree unwanted method that will increase the overhead of the system. To cut back this overhead we have a tendency to introduce associate degree increased tile facility that is not solely the basic work as mentioned in base work, however conjointly the native routing of the neuron cell facilities that are gifted within constant tile facility. This will increase the potency of the system. As shown in the Figure 2.

---

**EXISTING WORK RESULTS**

|                |          |
|----------------|----------|
| LU\(\text{Ts}\) | 1207     |
| BELS           | 1209     |
| Delay          | 2.436ns  |
| Minimum input arrival time before clock | 2.096ns |
| Maximum output required time after clock | 0.730ns |
| Memory         | 240104 kilobytes |
| Flip Flops/Latches | 1291   |
The planned H-NoC design is valid supported associate degree RTL-level implementation, while memory, area and delay analysis are accomplished exploitation ISE machine. Results show vital climbable characteristics such as high turnout and low power/area footprint that build it appropriate for large-scale SNN-embedded implementations.

3. Results and Discussion

3.1 Xilinx/ISE Simulations and Exactitude RTL of Mentor Graphics

The planned style and its corresponding blocks are represented exploitation structural verilog and synthesized using Xilinx Synthesis Tool (XST), net PACK version thirteen.2 and exactitude RTL of Mentor Graphics.

The logical routing will be discovered from the obtained Place and route result from the FPGA Editor possibility in Xilinx synthesizer. It would discover that concerning areas for the targeted FPGA is roofed for the implementation of this technique. The CLB’s are connected in cascade manner to get the practicality for the designed system to confirm that the hardware implementation works properly, simulation check was performed exploitation I-Sim (O.76.xd). As shown within the Table 1 and 2.

Table 1. Parameter comparison of existing work

| PAD REPORT  |                     |
|-------------|---------------------|
| Total REAL time / completion | 8 mins 41 secs |
| Total CPU time / completion    | 8 mins 24 secs |

Table 2. Parameter comparison of proposed work

| PROPOSED WORK RESULTS |                     |
|-----------------------|---------------------|
| LUTs                  | 1048                |
| BELS                  | 1050                |
| Delay                 | 1.905ns             |
| Minimum input arrival time before clock | 1.514ns |
| Maximum output required time after clock   | 0.730ns |
| Memory                | 269672 kilobytes    |

3.2 Impact of the Planned Flow on Peak Memory Usage, Delay and Area

In this paper, the traditional approach and therefore the planned techniques are analyzed supported the value perform of alluvium and router. As shown within the Table 1 the quantity of LUT’s, memory usage and delay are reduced in planned flow because of less consumption of planned style.

The planned technique outperforms the traditional design in terms of placement and routing that has shown within the Table 1 and 2 shows the planned technique reduced the PAR delay upto ten in comparison to standard technique.

In terms of overhead, since the traditional approach and therefore the planned technique solely modification the location and routing of the look, because the usage of the CLB (configurable logic blocks- BELS/LUT’s i.e. Basic logical part and appearance up table) varies that provides the overhead and delay lesser than existing approach. Additionally, no unapproachable CLBs are reported by the initial technique and therefore the planned technique.
that helps to beat the limitation of the initial approach. Hence, the traditional approach and therefore the planned technique sustain CLB overhead. And the optimized parameter comparison is shown in the Figure 3.

4. Conclusion

The writers have designed a novel gradable intelligence agent plan that allows cluster of neurons to be imposed on hardware exploitation a gradable array of intelligence agent routers. The planned increased SNN for H-NoC design exploitation tile facility is scalable; it offers standard on-chip communication infrastructure for interneuron property, and a spike traffic compression technique to cut back traffic overhead. The planned H-NoC design was represented at the RTL-level implementation then thoroughly evaluated supported hardware simulations employing a big variety of artificial spike traffic situations. Simulation results demonstrate that the planned H-NoC approach offers a high, vital performance in terms of area, delay and temporal order. Results are auspicious and create the incentive to continue exploiting the paradigm of intelligence agent as a way to beat the interconnection issues for large-scale hardware SNN executions. High-throughput design is necessary though having associate degree economical, and additionally it is very significant to have a balance between inflated performance based on area, delay and timings attained. Therefore, during this direction, the designed H-NoC plan may be a leap forward.

5. References

1. Dally WJ, Towles B. Route packets, not wires: on-chip interconnection networks. Proceedings of the 38th Design Automation Conference, Las Vegas, Nev, USA: June 2001. p. 684–9. Crossref
2. Design Automation Conference, 2001. p. 684–9.
3. Theocharides T, Link G, Vijaykrishnan N, Invin M, Srikantam V. A Generic Reconfigurable Neural Network Architecture as a Network on Chip. Proc. IEEE Int’l SOC Conf., 2004 Sep. p. 191–4. Crossref
4. Emery R, Yakovlev A, Chester G. Connection-Centric Network for Spiking Neural Networks. Proc. Third ACM/IEEE Int’l Symp. Networks-on-Chip, 2009 May. p. 144–52. Crossref
5. Harkin J, Morgan F, McDaid L, Hall S, McGinley B, Scaly. A Reconfigurable and Biologically Inspired paradigm for Computation Using Network-on-Chip and Spiking Neural Net-works. Int’l J. Reconfigurable Computing. 2009; 1–13. Crossref
6. Carrillo S, Harkin J, Media L, Pande S, Cawley S, McGinley B, Morgan F. Advancing Interconnect Density for Spiking Neural Network Hardware Implementations Using Traffic-Aware Adaptive Network-on-Chip Routers. Neural Net-works. 2012 Sep; 33:42–57. Crossref PMid:22561008
7. Indiveri G, Linares-Barranco B, Hamilton TJ, Schaik VA, et al. Neuromorphic Silicon Neuron Circuits. Frontiers in Neuroscience. 2011 Jan; 5:73. Crossref PMid:21747754 PMCid:PMC3130465
8. Strogatz SH. Exploring Complex Networks. Nature. 2001 mar; 410(6825):268–76. Crossref PMid:11258382
9. Watts DJ, Strogatz SH. Collective Dynamics of ‘Small-World’ Networks. Nature. 1998 Jun; 393(6684):440–2. Crossref PMid:9623998
10. Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger DR, Smoore W, Bullmore ET. Efficient Physical Embedding of Topologically Complex Information Processing Networks in Brains an Computer Circuits. PLoS Computational Biology. 2010 Apr; 6(4):e1000748. Crossref PMid:20421990 PMCid:PMC2858671
11. Markram H. The Blue Brain Project. Nature Rev. Neuroscience. 2006; 7(2):153–9. Crossref PMid:16429124
12. Blumrich M, Chen D, Coteus P, Gara A, Giampapa M. Design and Analysis of the Bluegene/L Torus Interconnection Network. IBM Research Report RC23025 (W0312-022),2003; 23025.
13. Schemmel J, Fieres J, Meier K. Wafer-Scale Integration of Analog Neural Networks. Proc. IEEE Int’l Joint Conf. Neural Networks. 2008 Jun; 431–8. Crossref
14. Bruderle D, Petrovici MA, Vogginger B, Ehrlich M, et al. A Comprehensive Workflow for General-Purpose Neural Modeling with Highly Configurable Neuromorphic Hardware Systems. Biological Cybernetics. 2011 May; 104(4/5):263–96. Crossref PMid:21618053
15. Carrillo S, Harkin J, McDaid Lj, Morgan F, Pande S, Cawley S, McGinley B. Scalable Hierarchical Network-on-
Chip Architecture for Spiking Neural Network Hardware Implementations. IEEE Transactions on Parallel And Distributed Systems. 2013 Dec; 24(12). Crossref
16. Diehl PU, Pedroni B, Cassidy A, Merolla P, Neftci E, Zarrella G. Truehappiness: Sentiment analysis on truenorth. arXiv, 2016.
17. Cao Y, Chen Y, Khosla D. Spiking deep convolutional neural networks for energy-efficient object recognition. International Journal of Computer Vision. 2014; 1–13
18. D. Neil, Liu S-C. Minitaur, an event-driven fpga-based spiking network accelerator. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2014; 22(12):2621–8. Crossref
19. Stromatias E, Neil D, Pfeiffer M, Galluppi F, Furber SB, Liu S-C. Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms. Frontiers in neuroscience, 2015; 9. Crossref
20. Esser SK, Appuswamy R, Merolla P, Arthur JV, Modha DS. Backpropagation for energy-efficient neuromorphic computing. Advances in Neural Information Processing Systems. 2015; 1117–25.
21. Li X, Roth D. Learning question classifiers. Proceedings of the 19th international conference on Computational linguistics. 2002; 1:1–7. Crossref
22. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
23. Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. Advances in Neural Information Processing Systems. 2014; 2177–85.
24. Li Y, Xu L, Tian F, Jiang L, Zhong X, Chen E. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. 2015. [27] Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015; 61:85–117.
25. Merolla PA, Arthur JV, Alvarez-Icaza R, Cassidy AS, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science. 2014; 345(6197):668–73. Crossref PMid:25104385
26. Indiveri G, Chicca E, Douglas R. A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. IEEE Transactions on Neural Networks. 2006 Jan; 17(1):211–21. Available from: Crossref PMid:16526488
27. Khan M, Lester D, Plana L, Rast A, Jin X, Painkras E, Furber S. Spinnaker: mapping neural networks onto a massively parallel chip multiprocessor. IEEE International Joint Conference on Neural Networks, IJCNN, 2008. IEEE, p. 2849–56. Crossref
28. Benjamin BV, Gao P, McQuinn E, Choudhary S, et al. Neurogrid A mixed-analog-digital multichip system for large-scale neural simulations. Proceedings of the IEEE. 2014; 102(5):699–716. Crossref
29. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11):2278–324. Crossref