VOL. 14, NO. 6, MARCH 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A SURVEY ON REAL TIME PROCESSING WITH SPIKING NEURAL NETWORKS

Jeshmon K. Thomas and Harish Ram D. S
Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India
E-Mail: cb.en.d.ece16004@cb.students.amrita.edu

ABSTRACT

Neuromorphic computing is an emerging architecture to address the issues of parallel computing like energy efficiency, size and speed. In standard neural network the basic computation is matrix multiplication between inputs to the neurons and their weights. This type of heavy computation can be handled conventionally with high end Graphics Processing Units (GPUs) effectively. The current technologies like Parallel CPUs and GPUs can provide parallel computation, but by incurring heavy computation and power consumption overheads. This has motivated research in hardware implementation of spiking neural networks with silicon technologies. The hardware implementation of event driven networks can drastically reduce the computational load and hence the power consumption. This survey paper discusses the various hardware implementations of spike neural networks (SNNs) and how they address different issues related to parallel computation of neural network functions.

Keywords: neuromorphic computing, spiking neuron, spiking neural network, deep learning, neuro inspired hardware.

INTRODUCTION

Over the years, the field of computing has adapted several biological processes for designing architectures for complex computations. Artificial Neural Networks (ANNs) represent one such paradigm. Investigations into the way the human brain processes information led to the development of artificial neural networks. The Ideal structure of the biological brain for information processing is the basis of the research. The main applications of neural network model are pattern recognition, optimization, prediction and control. This is accomplished by preserving, modifying and creating data through different learning processes.

Several types of neuron models have been proposed. These models are mainly categorized into Conductance based models such as Hodgkin–Huxley (HH) model [1] and FitzHugh-Nagumo (FHN) neuron model [2]. The Morris-Lecar neuron [3] is a more precise biological neuron model. Spike based models represent the temporal behavior of the neuron. Izhikevich models [4], Integrate and Fire (IF), and the Leaky-Integrate and Fire (LIF) [5] are examples of Spike based models. Recent models such as [6] seek to investigate specific mechanisms that influence computation in neurons. Spiking neural network (SNN) approach in ANNs is more closely related to the biological neurons [7] with information encoding in the form of action potentials called spikes that last for one millisecond on an average [8]. Artificial SNNs process spikes from neurons that are encoded with information. Power consumption, communication constraints and large latencies are the main limitations in conventional ANNs which can be resolved by spike neural networks. The high connectivity between the neurons is responsible for the high parallelism. Basically, a neuron consists of four distinct components called synapses, dendrites, axon and soma as in Figure-1 [9]. Synapses provide the connection between two neurons, dendrites act as input node to the neuron for collecting the signals, soma process the signal from the dendrites and the axon caries the output spike to other neurons when the total inputs exceeds the threshold. The existing parallel computing technologies do not fully exploit the scope for reduction in the computation overhead inherent in some programs. For example, the frame-free processing of image information drastically reduces the computations by avoiding monotonous backgrounds and irrelevant features using an event driven computation approach [10]. The performance of the neural network is mainly based on the activity, not the size of the network. This helps the system to speed up the process by reducing the computation according to the requirement. Also it reduces the power consumption and is well suited for real time applications. The event driven network is highly energy efficient because in every time step there is no update of every neuron.

The exponential evolution of semiconductor technology from the discrete transistor to billion transistors in accordance with the Moore's law [11] has led to nanometer fabrication technologies. This drastic change in semiconductor fabrication technology initially led to the increase in the information processing capability, which has tapered off in recent times because of the limitations of physical laws. Thus providing a high degree of parallelism entails an increase in the number parallel processors and GPUs in the digital computing systems resulting in low energy efficiencies. Over half a century of research in the field of neuromorphic computing has

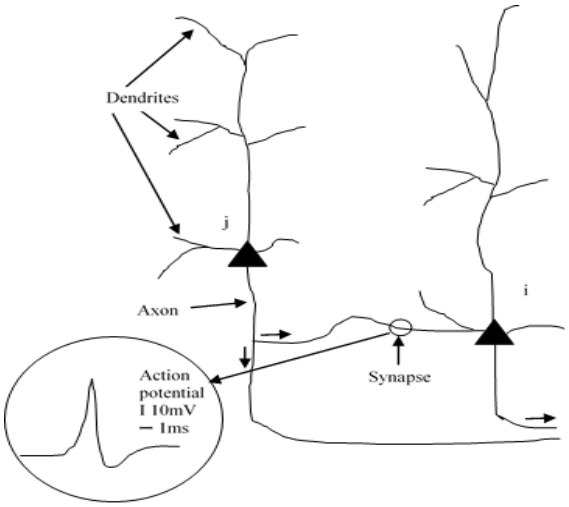

Figure-1. Neuron structure and signal transmission with action potential [9].

culminated in computing capabilities that match the processing prowess of biological brains that can perform massively parallel computations with a high degree of energy efficiency. These neuromorphic systems can be implemented in Field Programmable Gate Arrays (FPGAs) which are available in a variety of configurations. They are low-cost, easily reconfigurable and possess enough local memory. Though FPGAs are more power intensive compared to Application Specific Integrated Circuits (ASICs), FPGA vendors are continuously improving upon energy efficiency. They provide FPGA families tailor-made for power and energy efficient applications such as the Spartan family from Xilinx. ASIC implementations are also feasible if large production volumes are required. Researchers have also developed analog electronic circuits consisting of resistors, capacitors and transistors that have the same input-output relation of electrophysiological neural networks [12].

HARDWARE IMPLEMENTATION

A. GPUs and multi-core platforms

The human nervous system consists of nearly 10^{11} multi-input single output neurons and almost 10^{15} synapses. Communication between neurons occurs through generation of action potentials by electrochemical processes. These are asynchronous impulses that carry only the identity of the spiked neuron and the time of spiking through axons to post synapses at speeds-of a few milliseconds.

The implementation of these complex biological systems requires high performance computational architectures with approximately 10^{18} operations per second. Custom hardware accelerators are an ideal choice for achieving this scale of performance with optimal memory management and energy efficiency since they can exploit the parallelism inherent in hardware compared to von Neumann machines. Though GPUs exhibit significant degrees of data parallelism, they do not scale well for large number of neurons. They operate on batches of data instead of continuous streams and are not energy efficient [13]. The implementations with GPUs accelerate the computation compared to recent multicore computers and it can support significant amount of neurons without supercomputing technology. FPGAs (Field Programmable Gate Arrays) support real time simulation for limited resources [14]. The human brain project [14] is attempting to implement human realistic computing with help of high end GPUs for validation and tuning and FPGA for real time implementation.

B. Field programmable gate array (FPGA)

Modern FPGAs are more amenable for realizing large-scale neural networks with the availability of large numbers of logic gates and memory. Older generations of FPGAs could realize small neurons or small networks. The rapid strides made in FPGA technology have given rise to the possibility of implementing large networks using array architectures. A parallelized network of one million neurons has been realized with point to point connections [15].

Neural networks have emerged has one of the powerful tools for real time processing such as pattern classification, recognition, prediction and regression. In the last two decades, the high computational demands of neural networks have motivated researchers to explore and develop optimized hardware architectures. [16]-[19]. An FPGA Implementation of a large-scale neural network for pattern recognition by using Neural Engineering Framework (NEF) is reported in [20]. It is a standard three-layer feed forward network (input, hidden and output layers) constructed using subnetworks. The connections between the layers follow all-to-all connections with fixed random weights determined using pseudoinverse operation. The main aim of the work is not to achieve lowest test error but rather to develop a fast hardware pattern for real time tasks. The system is implemented with fixed-point numbers instead of floating point numbers to overcome the bottleneck of huge data storage requirement. Also to reduce the hardware cost the gray level of the input pattern is reduced to binary from a byte with negligible performance loss. Spike rates in NEF are used to calculate the weights. To implement the desired function, the low pass filter is used to sum the weighted output function to compute the firing rate. Effectively the neuron itself calculates the firing rate directly and becomes a non-spiking neuron. The high speed enables time multiplexing of neurons which will reduce the hardware resource overhead.

An event-driven Deep Belief Network (DBN) architecture called Minitaur with FPGA acceleration is proposed in [21] [22]. The accelerator is realized using Spartan-6 platform and consumes 200 out of 268 BRAMs as a cache for 1794 neurons and 647000 synapses distributed in four layers in 784-500-500-10 manner [13]. The neuron model in this work is divided into sub-models with Leaky Integrate and Fire model for soma, instantaneous synapse for input signal, and a fixed delay

axon for spike generation. This accelerator has 92% accuracy for MNIST handwritten digits' dataset [23] and 71% accuracy in 20 newsgroups classification data set [24]. The authors report that the error can be reduced by increasing the number of bits for weights and by improving the training methods of spiking neurons.

A modification in Minitaur called n-Minitaur with multi-modalities is reported to improve the computational speed compared to other deep neural networks (DNNs) implemented in FPGA [25]. It supports real time spikes from different input sensors related to the same data set and combines the output of these sensors for identification/classification applications and gives 98% accuracy with spiking DNNs.

The design of neuroprostheses, where the biological neuron tissues are replaced with artificial neuron tissues, requires accurate interface with innovative ideas to overcome the technological challenges to establish communication between living cells and artificial ones. Pani *et al*. [26] proposed an Izhikevich model [4] spiking neuron in Xilinx Virtex 6 FPGA which supports closed loop experiments.

C. Application specific integrated circuits (ASIC)

Hardware implementation of spiking neural network with general processors not only limits the performance but also impacts the communication pattern in the brain which has a complex path to the target. In case of the FPGA, the main limitation is that a single FPGA is inadequate to implement very large networks consisting of large number of neurons. This has led to investigations into the feasibility of realizing single-chip neuromorphic systems using a target ASIC library. ASICs are typically faster and more area and energy efficient than FPGAs [27] [28]. However, FPGA prototyping is carried out to validate a design which will eventually be implemented as an ASIC. The evaluation of the system is possible only with multiple FPGA boards and serves to develop prototypes for the neurochips.

The spiking neural network architecture (SpiNNaker) inspired by the mammalian brain is a massively parallel million core architecture that aims to model real time large-scale spiking neural networks [29] [30]. The system blocks are realized with RISC processors like the ARM9 processor in each processing node. It suffers from the drawbacks of general multi-core processor architectures because it uses the same memory hierarchies and organization of conventional CPUs. The system is implemented with 18 ARM 968 processor cores for a node with 96kB of local memory, 128 MB of shared memory and packet router for each core. The project uses Address Event Representation (AER) encoding for communication of neural activity between neurons. It introduces lightweight multicast packet routing mechanism which is a modification of conventional AER. In this protocol significant number of small data packets are transmitted between interconnects SpiNNaker provides a fast simulation platform for large scale spiking neural networks. It can support different types of network topologies and synaptic weight modification including variety of neurons and learning methods.

SpiNNaker architecture is an optimized implementation of biological real time point neuron models which is not center to the accurate modeling of the complex biological neurons. This computer architecture supports the balancing of complex parallel computation with memory hierarchy and communication protocol compared to the other architectures by compromising development cost [30]

The advantages of spiking neural network, which is different from traditional artificial neural network, such as high performance and lower power consumption can be achieved through hardware implementation only. Darwin Neural processing unit (NPU) [31] introduces a System on Chip (SoC) approach which include a RISC CPU, local bus, 64 KB SRAM for membrane potential and refractory period, SPI flash, UART controller and SDRAM controller for weight and delay attributes of each synapse and connection topology of neural network with dedicated NPU, fabricated by standard 180nm CMOS technology. Darwin NPU achieved classification accuracy of 93.8% with 25 MHz clock speed and 0.16s average latency for handwritten digit recognition with Spiking Deep Belief Network (DBN).

IBM developed a cognitive computing system (TrueNorth) [32] - [35] for spiking neuron using ASIC gates inspired by brain's function, low power and compatible size. TrueNorth architecture is built with neurosynaptic cores and a scalable network topology. It is implemented for fixed point simple arithmetic like addition and multiplexing operations. The neurons have slow firing rate in real-time compared to new CMOS technology. This helps the system to reduce the cost and power by reuse of the arithmetic block, switching off the idle neurons and using the event driven nature of the neurons.

High Input Count Analog Neural Network with Digital Learning System (HICANN-DLS) [36] is a neuromorphic chip that employs LIF mixed signal neuron models with 65nm CMOS technology for both digital and analog cores. The system is highly energy efficient with minimum footprint. The system is tuned to minimum bias range and digitally switches off all the subcircuits. For the optimum power and area, the design uses MOS capacitor, dual supply and thin and thick transistors.

Systems with analog and digital hardware are used for efficient neuromorphic computing such as "cxQuad" multi-neuron chips [37]. The analog hardware usually implements the synapses and neurons and the digital circuits are used for communicating the spikes between neurons in asynchronous manner and for configuring the network topology in cxQuad. The designed neuromorphic system is realized in full custom manner including a dynamic visual sensor as input and output classifier.

DEVICE LEVEL HARDWARE

A. Memristor

Power consumption and high transistor count are the main challenges for the synaptic circuit implementation in CMOS technology. The memristor (combination of memory and resistor), which can emulate the function of the synapses by storing and modifying its resistance with respect to the time integral of the current through it [38], is one of the solutions for the above challenge. In [39] the authors propose a memristor bridge based synaptic scheme The proposed method takes into account the inherent spatial non-uniformity and non-idealities in the response of the memristor bridge synapse. A weight change rule based on an average firing rate [40] and memristor array [38] [41] are proposed synaptic devices.

In [42] the authors propose a neuromorphic hardware system with memristor array for synaptic connections, using modified spike timing dependent plasticity learning rule which requires two operation periods, integration and reset. The system includes a CMOS image sensor(CIS) and a signal processing unit (SPU) as input layer implemented with FPGA, a memristor array and leaky integrate and fire CMOS neurons as output layers. The memristor acts as synaptic weight of the neural network and its resistance changes according to the spike rate of input sections. The system gives 100% correct recognition for images without noise level but there is a limitation in the system for similar pixel value images.

The existing neuron architecture has several problems when it integrates with the memristors in large neural networks [43]. The integration of currents across thousands of memristors in parallel computing will cause summing current overhead [44]. The next challenge for memristor usage is synaptic learning signal of neural network which should be electrically compatible with two terminal memristors [45]. The basic advantage of nanoscale memristor is implementation of large number of synapses with high integration capability. This advantage will be neutralized if any accessory circuit is added for online learning of the network and can be avoided by the usage of compatible design which supports the same node of memristor. The nanoscale memristor implementations achieve a power efficiency of 97% for 10000 memristor synapses [43].

B. Resistive random access memory (RRAM)

The complex computations in neural network increases the number of synapse circuits and hence the footprint. There is also a corresponding increase in power and memory overheads. The current developments in nanotechnology have led to the invention of a new resistive switching device called Resistive Random Access Memory (RRAM) which offers an alternative for implementation of biological synapses [41]; The RRAM is a low power device with a capability of multi-bit storage. Its main limitation is the presence of parasitic current paths in passive arrays which cause interference between cells. This can be overcome by selection of devices like metal insulated transition devices.

C. Conductive bridge random access memory (CBRAM)

The resistive non-volatile memory technology like Conductive Bridge Random Access Memory (CBRAM) [46] consumes ultra-low power and it can be used for implementing binary synaptic circuit with high accuracy in a neuromorphic system. The compatibility with CMOS, ease of fabrication, low power and scalability gives superiority for CBRAM in synaptic circuit implementation.

D. Leaky integrate fire neuron based on floating gate integrator (FG-LIF)

The artificial neuromorphic system always tries to imitate the biological brain system with synaptic and neuron circuits communicating with each other by spikes. The postsynaptic neurons integrate the spikes potential from presynaptic neurons until the potential reaches the threshold for firing. This leaky integrate and fire model for neurons is widely used in artificial spiking neural networks. The integration of spikes can be implemented through capacitor based integrators [47] - [49] which introduces time delay. LIF neurons based on Floating Gate integrator (FG-LIF) [50] have also been proposed instead of capacitor integrators. The Charging and discharging mechanism in FG is independent of area, width and thickness of tunnel barrier. This will help the scaling of the circuit and the reduction in power consumption.

METRICS

A. Speed

The speed of the biological neuron is very less approximately 200,000 times, compared to FPGA implemented neuron. So a real neuron can time multiplex to many virtual neurons [51]. Data storage is the actual limitation of this time multiplexing approach. In FPGAs limited storage capability of on-chip SRAMs will be one of the main challenges to data storage. When considering the speed of FPGA use of off-chip memory is possible but difficult because of the bandwidth of the memory. To overcome this difficulty an address buffer for virtual neurons is used and from the address buffer the weights of virtual neuron are calculated with the help of fixed random generator instead of storing weight of virtual neuron [18]. In this work, the neuron is designed based on conductance and randomized parameters in a highly effective way. The computational speed of the neural network is improved in n-Minitaur which is based on spiking DNN with fusion of multi sensors [25].

B. Memory

The event-based neural network's performance is limited during the spike generation because of the memory intensive nature of the operation. The operations that are memory intensive during spike generation are receipt neuron determination and the weights of each neuron.

Instead of using lookups for neuron connection, rule-based neural network connections are effective like DBNs, [52] Restricted Boltzmann Machines (RBMs) [53] and multilayer perceptron for optimum memory usage. DBNs is a deep layered network developed from two layer RBM. The input of the visible layer is the output of the hidden layer of two layer RBM with Contrastive Divergence unsupervised learning [54] of layers one after other.

The next challenge in neural network acceleration is managing postsynaptic currents because of large number of output connections. One solution for this is to store the spike source address instead of storing the destination address and determine the destination by rule-based connections [13]. The next problem in event driven system is effective usage of caches in neuron weight and states for real time applications by exploiting only significant inputs. For example, instead of frame based processing of video data, frames with local intensity changes in the video data alone are considered for event based spike generation as in [55].

The RBM used in DBN tolerate memory errors in the time of pre-training and post-training [56]. The fault tolerance is analysed in [57] by fault models like stuck at 0 and stuck at 1 which changes the parameters such as weights and bias of neurons, by fault injection in weights and biases, and by bit flipping.

C. Size

The scalability of neuromorphic system implemented in silicon is limited by several factors such as silicon technology, learning circuits for neurons, memory and the inter neuron communication. Seo et.al.[57] proposed a system with transposable SRAM arrays that shares the circuits for learning algorithm and extend only according to the requirement. The system employs novel crossbar switches for communication, and robust neuron circuits. It is difficult to implement the all to all connectivity between layers of neural network in hardware due to the requirement of significant hardware resources with increase in neurons of layers. This will limit the size and flexibility of networks.

One way to alleviate this issue is by a receptive field (RF) approach based neural network [58], [59]. In RF as shown in Figure-2, only a limited spatial range of inputs is considered for the tasks assigned to the network at a time. This approach with fixed amount of neurons in each layer gives the same accuracy of Extreme Learning Machines (ELMs) which have all-to-all connectivity between layers [58]. Still there is a difficulty in implementing hardware for the RF approach to determining the size, position and shape of RF randomly in a flexible manner. The improved RF approach with vectored input from sequentially indexed region from input image using shift registers helps to reduce this problem. An SRAM based implementation of RF, which has smaller footprint compared to logic gates, yields a better accuracy than ELM approach [60]. This gives an average accuracy of 96.22% with MNIST dataset.

The weights learned with double floating point precision in the time of training can be converted to lower floating point precision to reduce the hardware requirement [61]. According to IEEE 754 the double floating point precision requires one bit for sign, 52 bits for fraction and 11 bits for exponent. By holding the integer part constant, the synaptic weights, bit precision is reduced to one bit which is capable of handling the maximum and minimum weight values. The lower precision weights are calculated from trained double precision floating point weights in [61] as

$$W_L = round\left(2^f . W_H\right) 2^{-f} \quad (1)$$

Where W_L is the lower precision representation, W_H is the double precision floating point weights and f is the lower precision resolution.

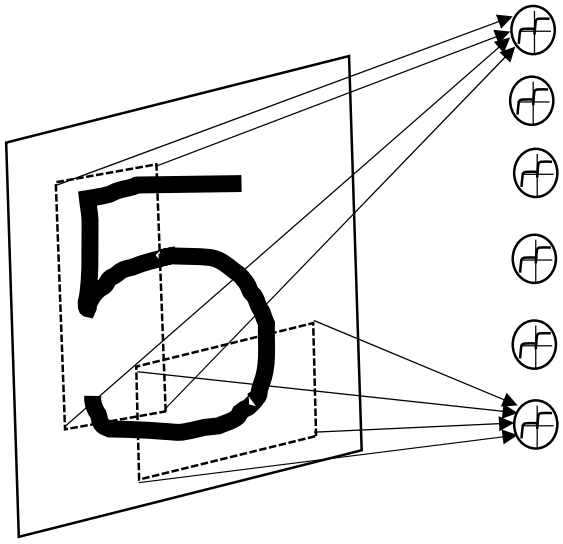

Figure-2. Illustration of RF Approach for Neural Network: The hidden layer receives a small rectangular field of original image. This figure adapted from [58].

One of the recent critical issues in semiconductor technology is the scaling limit of the device. The devices like memristor and RRAM require less physical space compared to silicon transistors. The area required for five titanium oxide memristors is less compared to single transistor which is compatible with CMOS technology. Reduction in the area of transistor is difficult, whereas the savings in area of memristors is obvious [38]. The Floating Gate LIF neuron based spike integration depends on tunnel barrier property which is area independent (unlike capacitor based integrators) thereby resulting in high scalability for the neural network implementation [50].

D. Communication

The communication in the brain is very complex and occurs in the form of small packets of information to a large number of destinations. The communication in computers on the contrary happens with large data packets. This limits the conventional communication

methods for implementation of communication in SNNs. The SpiNNaker architecture proposed optimized communication infrastructure with large number of small data packets using a simple packet protocol called SpiNNaker Datagram Protocol (SDP) [29]. The spikes from and to the neuron are communicated using Address-Event Representation (AER) format. The spikes are communicated as AER packets with identity of neuron and time of the spike generation. Darwin NPU [31] system uses AER and reduces the power consumption by enabling the NPU logic functions only when packets are received.

E. Computation

The inherent parallel computing of neural network limits the implementation with CPUs and requires dedicated architectures. The event driven network controls the computation by the events in the neurons, The Minitaur accelerator pre-processes the data to reduce the computation time by applying frequency inverse document frequency transform which controls the frequently used words and identifies uncommon words [13].

One of the drawbacks of DBN is the requirement of multiple layers and the connections between these layers. This makes the system computationally intensive. An alternative approach for this problem explores an unconventional method based on stochastic unary computation for DBN [62]. Stochastic computation helps to convert complex computing into simplified bitwise logic functions. It uses the statics of binary numbers present in the input with the help of uniformly distributed random numbers and comparators.

F. Noise

Noise is uniformly distributed. The noise will lead to decision errors in the network. This can be reduced by increasing the input data from the sensors with a penalty of increased complexity. The memristor based synaptic connections introduces decision errors due to the non-uniformity in the memristance due to process variations [42]. The hardware implementation with integrated circuits also produces noise which creates irregular spikes. The different types of noises simulated with FG LIF neuron circuits and the inter spike interval between adjacent spikes can be fit in to Gamma distribution with parameters similar to biological neurons [50]. The effect of noise produces random spikes. The RBM weights have an ability to denoise by extracting the significant features of the data and propagating through the layers by Deep Belief Networks [13], [61].

G. Power

CMOS implementations of synaptic circuits need large numbers of transistors and hence the power requirement of the system is very high as a consequence. The event driven neurons consume more energy when they produce spikes. In a given period of time, higher the spike activity, the more power consumed by the neurons. The power consumption of neural network depends on several factors like CMOS technology, firing rate and spike width of neuron.

The SpiNNaker project uses SpiNNaker application programming interface (spin1API) that reduces the power requirement by using interrupts and queue structure for the event-driven programming model [29], [30]. In True North design, the power consumed by global clock circuits, and collocated memory was eliminated. True north chips also have reduced datapaths and implement sparse distributed memory [35]. The chip manufactured with low power 28nm CMOS technology has reduced static power consumption.

The high scalability of FG LIF neuron due to its tunnel barrier property and the subthreshold operation in floating gate transistor, achieves low power consumption less than 30pW [50]. The passive devices like memristors as synaptic connections also act as alternatives for power reduction [38], [41], [42]. The synapses with RRAM technology reduce power consumption by short spikes and burst mode operation [63]

CONCLUSIONS

The neuromorphic system with spiking neurons has the advantages of event driven nature, high parallelism, energy efficiency and close relationship with the biological neurons. The potential of artificial spiking neurons is still not completely utilized especially in the area of training and learning. In this paper we have discussed the different hardware approaches and novel devices for the implementation of spike based neuromorphic systems. The major issues that exist in neuromorphic hardware and solutions proposed that address these issues have also been investigated. The extensive computation overhead of neuromorphic systems is a major concern for real-time applications. Future research is expected to focus on scaling of the network architectures with effective use of time multiplex neurons to minimize memory storage and the bandwidth requirement of off-chip memory.

REFERENCES

[1] L. Hodgkin and A. F. Huxley. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. 117(4): 500-544.

[2] R. FitzHugh. 1961. Impulses and physiological states in theoretical models of nerve membrane. J. Biophys. 1(6): 445-466.

[3] Morris and H. Lecar. 1981. Voltage oscillations in the barnacle giant muscle fiber. J. Biophys. 35(1): 193-213.

[4] M. Izhikevich. 2003. Simple model of spiking neurons. IEEE Trans. Neural Networks. 14: 1569-1572.

[5] A. Burkitt. 2006. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. Biol. Cybern. 95(1): 1-19.

[6] Shyam Diwakar, Jacopo Magistretti, Michell Goldfarb, Giovanni Naldi, Egidio D' Angelo, Axonal Na+ channels ensure fast spike activation and back-propagation in cerebellar granule cells, J Neurophysion, 101, 2009, 519-32.

[7] Paugam-Moisy, S. Bohte. 2010. Computing with spiking neuron networks. In Handbook of Natural Computing, New York: Springer-Verlag.

[8] R. Kandel, J. H. Schwartz and T. M. Jessel. 2000. Principles of Neural Science. 4th ed. New York, NY: McGrail-Hill Health Prof. Division.

[9] W. Gerstner, W. M. Kistler. 2002. Spiking Neuron Models- Single Neurons, Populations, Plasticity, Cambridge University Press.

[10] T. Delbruck. 2008. Frame-free dynamic digital vision. Proceedings of Intl. Symp. on Secure-Life Electronics Advanced Electronics for Quality Life and Society. *pp. 21-26.*

[11] E. Moore. 1998. Cramming more components onto integrated circuits. Proc. IEEE. 86: 82-85.

[12] F. Castaños, A. Franci. 2017. Implementing robust neuromodulation in neuromorphic circuits. Neurocomputing. 233: 3-13.

[13] Neil, S.-C. Liu. 2014. Minitaur an event-driven FPGA-based spiking network accelerator. IEEE Trans on Very Large Scale Integration (VLSI) Systems. 22(12): 2621-2628.

[14] Florimbi *et al*. 2016. The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells. Microprocessors and Microsystems.

[15] A. Cassidy, A. Andreou, J. Georgiou. 2011. Design of a one million neurons single FPGA neuromorphic system for real-time multimodal scene analysis. Proc. 45th Annu. CISS. pp. 1-6.

[16] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, G. Cauwenberghs. 2007. Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses. IEEE Trans. Neural Netw. 18(1): 253-265.

[17] T. Pfeil, A. Grübl, S. Jeltsch, E. Müller, P. Müller, M. A. Petrovici, M. Schmuker, D. Brüderle, J. Schemmel, K. Meier. 2013. Six networks on a universal neuromorphic computing substrate. *Front. Neurosci.* 7(11).

[18] R. Wang, T. J. Hamilton, J. Tapson, A. Van Schaik. 2014. An FPGA design framework for large-scale spiking neural networks. 2014 IEEE International Symposium on Circuits and Systems (IS CAS). pp. 457-460.

[19] JR de Oliveira Neto, JPC Cajueiro and J. Ranhel. 2015. Neural encoding and spike generation for Spiking Neural Networks implemented in FPGA. In Electronics, Communications and Computers (CONIELECOMP), 2015 International Conference on, pp. 55-61. IEEE.

[20] R. Wang, C. S. Thakur, T. J. Hamilton, J. Tapson, A. van Schaik. 2015. A neuromorphic hardware architecture using the neural engineering framework for pattern recognition. arXiv: 1507.05695.

[21] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, M. Pfeiffer. 2013. Real-time classification and sensor fusion with a spiking deep belief network. Front. Neurosci. Vol. 7.

[22] Kiselev, D. Neil, S.-C. Liu. 2016. Live demonstration: Event-driven deep neural network hardware system for sensor fusion. In Circuits and Systems (ISCAS), 2016 IEEE International Symposium on, pp. 452-452. IEEE.

[23] L. Deng. 2012. The mnist database of handwritten digit images for machine learning research. Signal Processing Magazine, IEEE. 29(6): 141-142.

[24] Lang. 1995. Newsweeder: Learning to Filter Netnews. Proc. 12th Int'l Conf. Machine Learning.

[25] Kiselev, D. Neil, S.-C. Liu. 2016. Event-driven deep neural network hardware system for sensor fusion. 2016 IEEE International Symposium on Circuits and Systems (IS CAS).

[26] Pani, P. Meloni, G. Tuveri, F. Palumbo, P. Massobrio, L. Raffo. 2017. An fpga platform for real-time simulation of spiking neuronal networks. Frontiers in Neuroscience. Vol. 11.

[27] S.M. Trimberger. 2015. Three Ages of FPGAs: A Retrospective on the First Thirty Years of FPGA Technology. Proc. IEEE. 103(3): 318-331.

[28] Andreas Ehliar, Dake Liu. 2009. An ASIC perspective on FPGA optimizations. Field-Programmable Logic and Applications (FPL). pp. 218-223.

[29] S. Furber, F. Galluppi, S. Temple, L. Plana. 2014. The SpiNNaker Project. Proceedings of the IEEE. 102(5): 652-665.

[30] S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, A. D. Brown. 2013. Overview of the SpiNNaker system architecture. IEEE Trans. Comput. 62(12): 2454-2467.

[31] Shen, D. Ma, Z. Gu, M. Zhang, X. Zhu, X. Xu, Q. Xu, Y. Shen, G. Pan. 2016. Darwin: a neuromorphic hardware co-processor based on spiking neural networks. Science China Information Sciences. 59(2): 1-5.

[32] Cassidy, Andrew S., Paul Merolla, John V. Arthur, Steve K. Esser, Bryan Jackson, Rodrigo Alvarez-Icaza, Pallab Datta *et al.* 2013. Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores. In Neural Networks (IJCNN), The 2013 International Joint Conference on, pp. 1-10. IEEE.

[33] A. Amir, P. Datta, W. Risk, A. S. Cassidy, J. A. Kusnitz, S. K.Esser, A. Andreopoulos, T. M. Wong, M. Flickner, R. Alvarez-Icaza, E. McQuinn, B. Shaw, N. Pass and D. S. Modha. 2013. Cognitve computing programming paradigm: A corelet language for composing networks of neuro-synaptic cores. in International Joint Conference on Neural Networks (IJCNN). IEEE.

[34] S. K. Esser, A. Andreopoulos, R. Appuswamy, P. Datta, D. Barch, A. Amir, J. Arthur, A. S. Cassidy, P. Merolla, S. Chandra, N. Basilico, S. Carpin, T. Zimmerman, F. Zee, M. Flickner, R. Alvarez-Icaza, J. A. Kusnitz, T. M. Wong, W. P. Risk, E. McQuinn, and D. S. Modha. 2013. Cognitive computing systems: Algorithms and applications for networks of neurosynaptic cores. In International Joint Conference on Neural Networks (IJCNN). IEEE.

[35] Akopyan *et al*. 2015. TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. IEEE Trans. Comput. Des. Integr. Circuits Syst. 34(10): 1537-1557.

[36] S.A. Aamir, P. Müller, A. Hartel, J. Schemmel, K. Meier. 2016. A highly tunable 65-nm cmos LIF neuron for a large scale neuromorphic system. Proc. 42nd ESSCIRC Conf. pp. 71-74.

[37] Indiveri, F. Corradi, N. Qiao. 2015. Neuromorphic Architectures for Spiking Deep Neural Networks. Electron Devices Meeting (IEDM) 2015 IEEE International. pp. 4.2.1-4.2.14.

[38] Kim, M. P. Sah, C. Yang, T. Roska, L. O. Chua. 2012. Neural synaptic weighting with a pulse-based memristor circuit. IEEE Trans. Circuit Syst. I. 59(1): 148-158.

[39] S. P. Adhikari, C. Yang, H. Kim, L. O. Chua, 2012. Memristor bridge synapse-based neural network and its learning. IEEE Trans. Neural Netw. Learn. Syst. 23(9): 1426-1435.

[40] D. Cantley, A. Subramaniam, H. J. Stiegler, R. A. Chapman, E. M. Vogel. 2012. Neural learning circuits utilizing nano-crystalline silicon transistors and memristors. IEEE Trans. Neural Netw. Learn. Syst. 23(4): 565-573.

[41] S. Park *et al*. 2013. Nanoscale RRAM-based synaptic electronics: Toward a neuromorphic computing device. Nanotechnology. 24(38): 384 009.

[42] Chu *et al*. 2014. Neuromorphic hardware system for visual pattern recognition with memristor array and cmos neuron. Industrial Electronics. (99): 1-1.

[43] X. Wu, V. Saxena, K. Zhu. 2015. A CMOS spiking neuron for dense memristor-synapse connectivity for brain-inspired computing. Proceedings of the International Joint Conference on Neural Networks (IJCNN). pp. 1-6.

[44] Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, T. Prodromakis. 2013. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology.* 24(38).

[45] Zamarreño-Ramos, L. A. Camuñas-Mesa, J. A. Pérez-Carrasco, T. Masquelier, T. Serrano-Gotarredona, and B. Linares-Barranco. 2011. On spiketiming-dependent-plasticity, memristive devices, and building a selflearning visual cortex. Frontiers in Neuroscience. 5(March): 26.

[46] Suri et al. 2012. CBRAM devices as binary synapses for low-power stochastic neuromorphic systems:

Auditory (cochlea) and visual (retina) cognitive processing applications. Proc. IEEE Int. Electron Devices Meeting (IEDM). pp. 10.3.1-10.3.4.

[47] Arthur, K. Boahen. 2004. Recurrently connected silicon neurons with active dendrites for one-shot learning. Proc. IEEE Int. Joint Conf. Neural Netw. 3: 1699-1704.

[48] C. Bartolozzi, G. Indiveri. 2007. Synaptic dynamics in analog VLSI. Neural Comput. *19*(10): 2581-2603.

[49] G. Indiveri, B. Linares-Barranco, T. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S. C. Liu, P. Dudek, P. Hafliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, K. Boahen. 2011. Neuromorphic silicon neuron circuits. Front. Neurosci. 5(23).

[50] V. Kornijcuk, H. Lim, J. Y. Seok, G. Kim, S. K. Kim, I. Kim, B. J. Choi, D. S. Jeong. 2016. Leaky integrate-and-fire neuron circuit based on floating-gate integrator. Frontiers in neuroscience. Vol. 10.

[51] R. Wang, G. Cohen, K. M. Stiefel, T. J. Hamilton, J. Tapson, and A. van Schaik. 2013. An FPGA Implementation of a Polychronous Spiking Neural Network with Delay Adaptation. Frontiers in neuroscience. 7(February): 14.

[52] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck and M. Pfeiffer. 2013. Realtime classification and sensor fusion with a spiking deep belief network. Frontiers in Neuroscience. 7(178).

[53] R. Salakhutdinov, A. Mnih, G.E. Hinton. 2007. Restricted Boltzmann Machines for Collaborative Filtering. Proc. Int'l Conf. Machine Learning.

[54] G. E. Hinton, R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science.* 313(5786): 504-507.

[55] P. Lichtsteiner, C. Posch and T. Delbrück. 2008. A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J. Solid State Circuits. 43(2): 566-576.

[56] Ueyoshi, T. Marukame, T. Asai, M. Motomura, A. Schmid. 2016. Robustness of hardware-oriented restricted Boltzmann machines in deep belief networks for reliable processing. Nonlinear Theory and Its Applications. E7-N(3): 395-406.

[57] Seo, B. Brezzo, Y. Liu, B. D. Parker, S. K. Esser, R. K. Montoye, B. Rajendran, J. A. Tierno, L. Chang, D. S. Modha and D. J. Friedman. 2011. A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons. In IEEE Custom Integrated Circuits Conference (CICC). pp. 14.

[58] D. McDonnell, M. D. Tissera, T. Vladusich, A. van Schaik and J. Tapson. 2015. Fast, Simple and Accurate Handwritten Digit Classification by Training Shallow Neural Network Classifiers with the 'Extreme Learning Machine Algorithm. PLoS One. 10(8): e0134254.

[59] G.-B. Huang, Z. Bai, L. L. C. Kasun, and C. M. Vong. 2015. Local Receptive Fields Based Extreme Learning Machine. IEEE Comput. Intell. Mag. 10(2): 18-29.

[60] R. Wang, G. Cohen, C. S. Thakur, J. Tapson, A. Van Schaik. 2016. An SRAM-based implementation of a convolutional neural network. IEEE Biomedical Circuits and Systems Conference. pp. 1-4.

[61] Stromatias, D. Neil, M. Pfeiffer, F. Galluppi, S. B. Furber, S.-C. Liu. 2015. Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms. Frontiers in neuroscience. Vol. 9.

[62] Sanni, G. Garreau, J. L. Molin, A. G. Andreou. 2015. FPGA implementation of deep belief network architecture for character recognition using stochastic computation. Conference on Information Sciences and Systems.

[63] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z. Wang, A. Calderoni, N. Ramaswamy, D. Ielmini. 2016. Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning. Proc. IEEE Symp. VLSI Technol. pp. 1-2.