THE ERA OF NEUROSYNAPTICS: NEUROMORPHIC CHIPS AND ARCHITECTURE

Siddhartha Agarwal
Divyanshi Rastogi
Aayush Singhal
Department of Electronics and Communication
JSS Academy of Technical Education NOIDA, India

Abstract

Since its invention the modern day computer has shown a significant improvement in its performance and storage capacity.However, most of the current processor cores remain sequential in nature which limit the speed of computation. IBM has been consistently working over this and with the launching of neurosynaptic chips, it has opened a new gateway of thought process. This paper aims at reviewing the various stages and researches that have been instrumental in the overall development of neuromorphic architecture which aims at developing flexible brain like structure capable of performing wide range of real time computations while keeping ultra-low power consumption and size factor in mind. Inspired by the human brain, which is capable of performing complex tasks rapidly and accurately without being programmed and utilizing very less energy, TrueNorth chips tends to mimic the human brain so as to perform complex computations at a faster pace. This has inspired a new field of study aimed at development of the cognitive computing systems that could potentially emulate the brain's computing efficiency, size and power.The paper also aims to highlight the inadvertent challenges of neuromorphic architecture as posed by the prevailing technologies which are a major field of research in near future.

Keywords: Neurosynaptic, ultra-low-power, TrueNorth, Cognitive Computing

Introduction

Human Brain- synonymous with the central processing unit of a computer , consists of a wide network of neurons and attached dendrons which on contraction and expansion help in transferring of information from one part of body to another. This very transfer is done by Synapse, which means ‘conjunction’. Thus, at a synaptic site, signal passing neuron comes in

close contact with the target which is rich in extensive array of molecular machinery.

In actual, the biological neural systems perform wide range of real time computations and tasks such as pattern recognition, sensory reconstruction carried out by these low power, dense neural circuits which within metabolic constraints are more efficient that traditional computers.

The basic computational unit of this system is the neurons that communicate with each other through the generation and modulation of spike trains where it may be an all-or-nothing pulse.

The human brain consists of a staggering number of over 100 billion neurons and over 1 trillion synapses.

The implementation of such a complex system is feasible through use of supercomputers but power and space has always been a constraint, preventing its usefulness in mobile systems for real time applications. Thus, mimicking human brain so as to take a big leap towards bionic systems has been a major concern over decades. With the aim of producing intelligent memory cells, materials such as chalcogenides need to be crafted which can sustain non-Boolean Algebra,thereby offering multi stable states. The memories so created are called 'cognitive memory' and forms the very basis of OCD's.

Ovonic computational devices (OCD) have analogous functions to neurons in brain and their synapses and they tend to offer same plasticity as organic molecules associated at neurosynaptic sites. Their very composition comprises of nano-dimensional amorphous structures like chalcogenide glasses offering high optical quality, which can be used to mimic brain like computations.

But in order to improve the switching times and reduce power consumption per synaptic event, the neuromorphic architecture came into being, which was capable of running spiking neural networks in compact and low power hardware. This neuromorphic architecture used analog circuits for biological components and digital asynchronous circuits for the communication of spike events. Inspite of its compactness and reduced power consumption , its sensitivity to process variations, ambient temperature and noisy environment posed a challenge in configuring circuit that could operate under a wide array of external parameters. This limited correspondence between analog implementations and neural algorithm was an obstacle to algorithm development and deployment . Even the lack of addition of high density capacitors and sub-threshold currents in analog implementation made it more complex and unreliable. Thus, implementation of neuromorphic architecture using discrete time, low power event driven circuits provided a path to both area and power efficient architecture, capable of one-to-one correspondence with the simulator.

This breakthrough came as a part of IBM's launch of neurosynaptic chips, which opened gateways to a new thought process.

The world of synapses

The innovation heralded with development of neurosynaptic core which had 256 integrate-and-fire neurons,1024 axons as the input which meant 1024X256 synapses in 4.2mm2 of silicon using a 45nm SOI process. This meant the innovation successfully was able to achieve the ultra-low energy consumption 1) At the circuit level by the usage of an asynchronous design where the switching in the circuit took place while performing neural updates;2) At the core level by using a crossbar memory implementation of a 256 neural fanout in a single operation was done; and 3) At architecture level by restricting core to core communication to only the spiking events which occurred sparsely in time .

Since the implementation was purely digital, the resultant was reliable and deterministic operation that helped in achieving one –to – one correspondence with a simulating software. It not only made the core readily scalable but also provided a platform for performing a wide range of real-time computations. The VLSI implementation - referred to as 'neuromorphic chips' overcame the area and power constraints of its biological counterpart. This facilitated a wide range of real time application that involved machine learning and perception. The discrete timed circuits hence provided an alternative solution to the constraints posed by the analog circuits. However, the parallel and event driven processing does not naturally fit the sequential processing model of the traditional computers.

As a result communication of spikes between physically separated processor and memory requires large bandwidth resulting in high power consumption and limited scaling. This can be taken care of by usage of a crossbar between memory and computation.

The asynchronous design methodology however is a natural fit to the distributed processing of neurons ensuring power dissipation levels of inactive parts in the system are kept minimum. The implementation although extremely robust remains operational under a wide range of process, voltage and temperature variations overcoming the earlier discussed constraints of immobility.

The neurosynaptic core basically has:

- Asynchronous circuits mimicking central elements of the neural system
- Computation , memory and communication integrated into one architecture
- Asynchronous communication accommodating the architecture
- Synchronization mechanism that maintains one-to-one correspondence

The prototype comprises of a single core with 256 digital leaky-integrate -and-fire neurons, 1024 inputs, and 1024x256 programmable binary synapses with a SRAM crossbar array. The entire core fits in a 4.22mm2 footprint in IBM's 45nm SOI process and consumes up to 45 pJ per spike.

Architecture and operation

Neurons: Physical Structure and Function

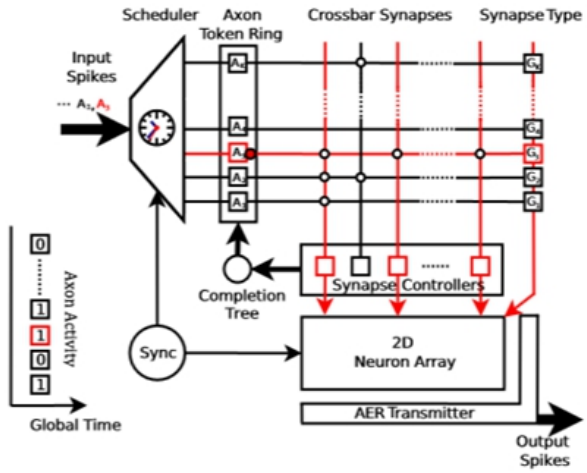

Fig. 1 Architecture of Neurosynaptic core

The individual neurons and the interconnecting synapses and their electro physical properties contributes to the computational power to the brain like networks. The neuron model forms the basis of the leaky integrate-and-fire model due to its ability to capture the behavior of real neurons in a range of situations and its efficiency.

The axons and synapses interconnect the various neurons. The axons correspond to the output of neurons in the same core or maybe driven by an external driver. If S_{ji} represents connection between axon *j* and neuron *i* and G_i characterizes the type of axon that can take different values indicating whether it's a strong excitatory, weak excitatory or inhibitory synapses that' axon forms with neuron it connects to and the neuron is parameterized by a leakage current L, a spike threshold theta, and three synapse weighs corresponding different axon types then neuron states are updated at each time step according to external input and interconnectivity. At any time t the voltage V_i [t] represents the neuron *i* and the activity bit A_i [t] represents the axon *j*:

The neurons voltage is updated at each time step after subtracting a leakage current from its voltage and integrating synaptic input from all the axons:

When this voltage level exceeds the specified threshold a spike is produced by the neuron and resets the voltage level. The negative voltages are also clipped back to 0 at the end of each time step.

Address Decoding

As mentioned earlier the heart of the core has a crossbar memory that forms the junction between axons and neurons. This array is configurable so as to be able to set up arbitrary networks in the systems, with the rows and column in the crossbar corresponding to an axon and input of a neuron respectively. Thus, for a network consisting of 1024 Synaptic inputs and 256 neurons, a 1024x256 crossbar synapses are obtained with a huge configuration space .With the help of address event representation [11] spiking events are sent to and from the core. AER transmitters and receivers play indispensable functions.

An AER transmitter[12] encodes the spiking activity by identifying locations of active neurons through a multiplexed channel leveraging the fact that bandwidth of wires is orders of magnitude larger than bandwidth of biological axons. A similar AER receiver takes charge of delivering the incoming spikes to the appropriate axon at a predefined time configured by a time scheduler block.

As the spikes are serviced sequentially address are decoded to the crossbar memory where the 256 synaptic connections of an active axon are read out in parallel.

Discrete –Time Operation

The figure explains the sequential operation taking place in the core. There are principally two phases involved: In the first phase of operation, initialized at the positive edge of the synchronization clock, address events and their time stamps are sent to the core to be received by the scheduler. This scheduler is responsible for evaluation of time stamps and assertion of the appropriate axons that go into a token ring. The units in the token ring that receive the active axons assert rows of the crossbar in mutually exclusive manner .Soon after the word line in a crossbar are activated the neurons that are connected to the axon receive input spike along with the information about the axon type. With the arrival of axon events the neuron voltage levels are updated and the whole phase is completed within the first half of the synchronization clock. This gives a precise margin for the neural updates for all the 1024 axon inputs.

In the second phase, the neuron respond to the negative edge of the clock.as soon as a negative peak is detected neurons whose voltage levels have exceeded the predefined threshold produce spikes in their output ports.

The spiking addresses are encoded by the AER transmitter and now again sent out to core sequentially. The second phase reaches its completion in the next half cycle of the clock. A 1 millisecond clock period would imply performance requirements for successful execution of the two phases is easily met.

The major advantage of breaking the whole operation of neural updates two phases is that the hardware is always in sync with the simulation with the end of each time step.

Event driven implementation

The architecture has the following concurrent processes using the communicating hardware processes [13]:

A. Scheduler

The packets received by the scheduler that maybe coming from the spiking neurons within the core or from outside to be delivered to the axons at specific times and at specific sites characterized by address contained in the axons and their associated spike time. In addition, the scheduler receives a global clock and global counter time. The scheduler decodes the packet and its corresponding location of delivery. The time specified in the packet is added to the axonal delay for the particular axon and then compares global counter time on

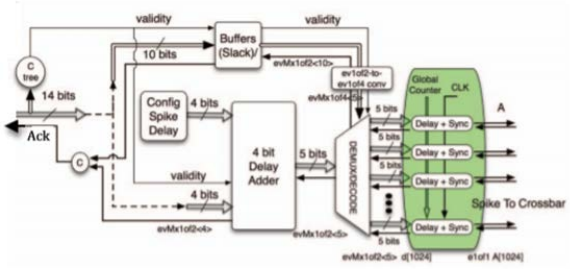

Fig.2 Internal structure of Scheduler

every tick of the clock. If the time matches to the clock a spike is delivered to the crossbar.

A. Axon Ring

With the detection of edge of the synchronization clock, scheduler picks up the axon lines with spikes. Each axon corresponds to the row in the crossbar memory. The dendrites correspond to the columns of the crossbar but since it can be a interconnection between various axons rows need to be essentially be asserted in mutually exclusive manner [14].On assertion of an axon, its server sends a request to the neighbour to pass a token. This action is carried by the multiple axon servers with each serve having an axon input

and word line of the crossbar at the output. The circulation of tokens between the servers ensures mutually exclusive access of the crossbar.

After request for passing the token is asserted by the server the request propagates through the token ring and passes along the requesting server. Upon receiving the incoming token the corresponding row of the crossbar is asserted.

All this paved way for IBM's Neurosynaptic chip popularly known as TrueNorth, which has become popular on account of its efficiency and speed.

The era of neurosynaptic chips: truenorth architecture

Neuro synaptic computation chip, TrueNorth is an entirely new class of processor chip that mimics human brain both in complexity and functioning. The chip has totally revolutionized the cognitive computing principles prevailing, by the extensive network of components mounted and associated power required for the same. Inspired by the hypothesis, that cerebral cortex comprises of repeating canonical cortical microcircuits, the neurosynaptic chip comprises of a tileable on-chip-two-dimensional mesh network of 4,096 digital, distributed neurosynaptic cores. It comprises of 5.4 billion transistors, making it second largest CMOS chip in the world. It has over 400 million bits of local on chip memory to store synapses .It supports hierarchical communication, with on-chip message spike routing network followed by local fanout crossbar, thereby reducing network traffic. The first ever chip to use digitally driven mixed asynchronous and synchronous neuromorphic network, reduces neuron switching by 99% of average. The design scaled down to 28nm, consumes power as low as desktop computer; 26pJ per synaptic event, lowest ever recorded. The TrueNorth architecture is modular and non Von Neumann, consisting of scalable neurosynaptic cores, each consisting of neurons, dendrites, synapses and axons. With the existent CMOS technology, it tends to capture the essence of neuroscience both in function and complexity. Though mimicking the right side of the brain, which is comparatively slower than left side, the TrueNorth architecture is yet 388 times slower than real time.

Truenorth working and architecture

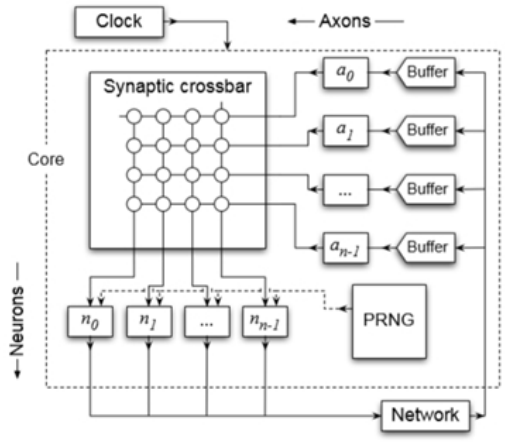

Fig. 3 Conceptual architecture of Neurosynaptic core

Replication of brain like mechanism consists of a Neurosynaptic core [1], [2], [15], [16] connected to the other cores in the system via some communication network. The simulated Neurosynaptic core consists of 256 axon, 256 dendrites feeding to 256 neurons and a 256x256 binary crossbar synaptic array. A neuron, characterised by configurable parameters, on any Neurosynaptic core, operating in parallel can connect to an axon on any core in the network in semi-synchronous fashion. Neurons are digital integrate-leak-and-fire circuits which send a spike message to the connecting axon.

Each TrueNorth core receives a clock operating at 1000Hz which is employed to discretize neuron dynamics in time steps of 1 millisecond. Post to this event, the core cycles through each of its axon, to check if the axon's buffer has a spike ready for delivery at the current time step. The synaptic value on the horizontal synaptic line is transferred from pre to post synaptic site. On such examination, if a particular axon-neuron pair carries a non-zero synapse value, the neural membrane's potential is incremented by (possibly stochastic) weight corresponding to the axon type. In similar fashion, a configurable stochastic leak is applied by each neuron and the finally that neuron fires a spike whose membrane potential has exceeded its threshold. The so generated spike travels to corresponding target axon, which schedule the spike the delivery at a future time step in its buffer. The communication channel plays an indispensable role by bridging axon of one neuron to dendrites of other, thus neuron axon connection communication while synapses represent memory. The entire cycle repeats until the tick of slow clock arrives. Thus, synaptic state never leaves the core only the spikes do so, thereby driving the communication network solely by spike events. The

TrueNorth quite successfully amalgamates computation and memory, thereby breaking the Von-Neumann bottleneck.

Thus to set wave for TrueNorth a parallel functional simulator, Compass was developed, which has one-to-one equivalence to functionality of TrueNorth. Compass is an architectural simulator which is multi-threaded, massively parallel and highly scalable, which can simulate under a soft real-time constraint. It incorporates several innovations in communication, computation and memory.

Compass neurosynaptic chip simulator

Compass, which is implemented using a combination MPI library calls and Open MP threading primitives, is an architectural simulator for TrueNorth cores. Various threads within a process independently simulate synaptic crossbar and neuron behaviour in a semi-synchronous manner. Compass has got an inherent tendency to minimize communication overheads between pair of processes into a single MPI message. As Compass executes a process map is used to identify all destination on remote processes. During a tick, as all neurons integrated, leaked and fired, the process aggregate all spikes and uses a single MPI send call to transmit each buffer, within a TrueNorth core.

Thus compass is indispensable for performing the following functions:

1) Regression testing to verify TrueNorth correctness
2) Dynamic behaviour study of TrueNorth cores
3) Hypothesis testing, verification and iteration of neural cores
4) Approximation of power consumption
5) Bench marking inter-core communication topologies
6) Demonstrating application in the field of optic flow, sensor integration, real time motor control (robotic navigation), multi-modal image audio classification and spataio-temporal feature expansion.

Parallel compass compiler

In order to translate a compact of functional regions of TrueNorth cores into explicit neuron parameter, PCC is extensively exploited. Unlike the conventional Compass simulator, PCC tends to minimize MPI message counts by assigning TrueNorth cores in the same functional region as few Compass processors. This shared memory concept helps in handling intra-region spiking, thereby increasing the speed manifold. This helps to create lower level core parameter from higher level. Though complex in functionality, computation in terms of neuron count and synaptic connection count is enhanced.

Challenges to neurosynaptics

1) Despite being fast and reliable, as the neural network grows in size, delays in switching time, howsoever small (as may be) are inevitable.

2) As a mesh of dendrons and axons are interconnected, there are probable chances of reaching of information to wrong location because of leakage current/ voltages at the Neurosynaptic site.

3) The architecture is based on non- Von Neumann architecture, thereby posing a problem of interfacing it with other devices which are predominantly Von Neumann.

4) The technology is costly and very complicated. Moreover, it is 388 times slower than Human brain.

Conclusion

Hailing towards Bionic computations and brain like systems has been a matter of great concern over decades. The advent of Neurosynaptic era was with IBM's TrueNorth. The foundations for the same were laid by Ovonic Computational Devices which comprised of chalcogenides, which could retain non-Boolean Algebra via multi stable states. The message transfer occurs in three respective phases, Synapse phase, Neuron phase and Network phase, which have been successfully simulated using Compass and PCC as simulating tools. The field is ever growing and there are various challenges like that of switching times and interfacing problems which have to tackled in the long run before the chips can be made available commercially.

Refrences:

P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. Modha, "A digital neurosynaptic core using embedded crossbar memory with 45pj per spike in 45nm," in Custom Integrated Circuits Conference (CICC), 2011 IEEE, sept. 2011, pp. 1 –4.

J. Seo, B. Brezzo, Y. Liu, B. Parker, S. Esser, R. Montoye, B. Rajendran, J. Tierno, L. Chang, D. Modha, and D. Friedman, "A 45nm cmos neuromorphic chip with a scalable architecture for learning in networks of spiking neurons," in Custom Integrated Circuits Conference (CICC), 2011 IEEE, sept. 2011, pp. 1 –4.

K. Tanaka, T. Gotoh, K. Sugawara, J. Optoelectron. Adv. Mater.6(4), 1133, 2004.

A. Lorinczi, J. Optoelectron. Adv. Mater 5(5), 1081, 2003.

Y. Hamakawa, W. S. Kolahi, K. Hattori, C. Sada, H. Okamoto, J. Jap. Soc. MicrogravityAppl. 12, 27, 1995.

Wahid Shams-Kolahi, M. Kobayashi, H. Hanzawa, H. Okamoto, S. Endo, Y. Kobayashi, Y. Hamakawa, Jap. J. Appl. Phys. 35, 4713 (1996).M.W. Madea, et al.: The effect offour-wave mixing in fibres on

M. Popescu, F. Sava, A. Lorinczi, J. Optoelectron. Adv. Mater. 6(3), 887, 2003.

R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, "The cat is out of the bag: cortical simulations with 109 neurons, 1013 synapses," in Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, SC '09, (New York, NY, USA), pp. 63:1–63:12, ACM, 2009.

C. Mead, "Neuromorphic electronic systems," Proceedings of the IEEE, vol. 78, pp. 1629 –1636, Oct. 1990.

K. Boahen, "Neurogrid: emulating a million neurons in the cortex," IEEE international conference of the engineering in medicine and biology society, 2006.

M. A. Mahowald, VLSI analogs of neuronal visual processing: a synthesis of form and function. PhD thesis, Pasadena, CA, USA, 1992. UMI Order No. GAX92-32201.

N. Imam and R. Manohar, "Address-event communication using token-ring mutual exclusion," in ASYNC, pp. 99–108, IEEE Computer Society, 2011.

A. J. Martin, "Programming in VLSI: From communicating processes to delay-insensitive circuits," in Developments in Concurrency and Communication, UT Year of Programming Series (C. A. R. Hoare, ed.), pp. 1–64, Addison-Wesley, 1990.

A. J. Martin, "Distributed mutual exclusion on a ring of processes," Sci. Comput. Program., vol. 5, pp. 265–276, October 1985.

N. Imam, F. Akopyan, J. Arthur, P. Merolla, R. Manohar, and D. S. Modha, "A digital neurosynaptic core using event-driven qdi circuits," in ASYNC 2012: IEEE International Symposium on Asynchronous Circuits and Systems, 2012.

J. V. Arthur, P. A. Merolla, F. Akopyan, R. Alvarez-Icaza, A. Cassidy, S. Chandra, S. K. Esser, N. Imam, W. Risk, D. Rubin, R. Manohar, and D. S. Modha, "Building block of a programmable neuromorphic sub- strate: A digital neurosynaptic core," in International Joint Conference on Neural Networks, 2012.

R. Preissl, T. M. Wong, P. Datta, M. Flickner, R. Singh, S. K. Esser, W. P. Risk, H. D. Simon, and D. S. Modha, "Compass: A scalable simulator for an architecture for cognitive computing," in *Proceedingsof the International Conference for High Performance Computing,Networking, Storage, and Analysis (SC 2012)*, Nov. 2012, p. 54.

E. McQuinn, P. Datta, M. D. Flickner, W. P. Risk, D. S. Modha, T. M.Wong, R. Singh, S. K. Esser, and R.Appuswamy, "2012 international

science & engineering visualization challenge," *Science*, vol. 339, no.6119, pp. 512–513, February 2013.

A. S. Cassidy, P. Merolla, J. V. Arthur, S. Esser, B. Jackson, R. Alvarez-Icaza, P. Datta, J. Sawada, T. M. Wong, V. Feldman, A. Amir, D. Rubin, F. Akopyan, E. McQuinn, W. Risk, and D. S. Modha, "Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013.