An exploration of neuromorphic systems and related design issues/challenges in dark silicon era

Mudit Chandaliya, Nitin Chaturvedi, S Gurunarayanan
Birla Institute of Technology and Science
Department of Electrical and Electronics Engineering
Pilani, India, 333031
E-mail: f2015484, nitin80, sguru@pilani.bits-pilani.ac.in

Abstract. The current microprocessors has shown a remarkable performance and memory capacity improvement since its innovation. However, due to power and thermal limitations, only a fraction of cores can operate at full frequency at any instant of time irrespective of the advantages of new technology generation. This phenomenon of under-utilization of microprocessor is called as dark silicon which leads to distraction in innovative computing. To overcome the limitation of utilization wall, IBM technologies explored and invented neurosynaptic system chips. It has opened a wide scope of research in the field of innovative computing, technology, material sciences, machine learning etc. In this paper, we first reviewed the diverse stages of research that have been influential in the innovation of neurosynaptic architectures. These, architectures focuses on the development of brain-like framework which is efficient enough to execute a broad set of computations in real time while maintaining ultra-low power consumption as well as area considerations in mind. We also reveal the inadvertent challenges and the opportunities of designing neuromorphic systems as presented by the existing technologies in the dark silicon era, which constitute the utmost area of research in future.

1. Introduction
The last few years, witnessed successful scaling of MOS dimensions for several technology generations. However, reduction in the MOS dimensions in the present-day fabrication technologies have resulted to unexpected new problem for the chip designers know as dark silicon.

This phenomenon restricts the ever increasing portion of a chip's transistor to remain on, they remain dark in order to keep check on power budget of a chip and restrict the system designers to utilize the extra resources that comes with every new technology generation (Moore's Law). It is probable that we have attained the end of Dennard scaling and furthermore Moore's law. Figure 1 illustrates speed-up projections for last few decades as well as future trends.

In the past, clock frequencies of a single microprocessor have shown significant improvements, whereas subsequent improvement in microprocessors based on von Neumann computer architecture have been achieved by means of multiple parallel cores (parallelization). The current computing systems are based on classical von Neumann Architecture which have one or multiple cores (processing units) and access both data and program using the same shared resources. Since, the processor speed is much faster than memory, the processor stays in the idle state for most of the time. Also, there are many applications in which parallelism is difficult and these architectures are not compatible for it. Moreover, the complexity of the future application continues to increase along with the size of data (distributive), this presents the inconsistency between the massive future real time applications and what conventional
von Neumann architecture can offer. The main performance bottleneck of the von Neumann architecture is mainly due to increased performance gap between processing core and memory also known as memory wall problem, which motivated the researchers from different communities to propose novel computing architectures.

**Figure 1.** Speedup Projections.

Different research organizations have revealed that valuable insights can be gained from looking into the human brain. Figure 2 illustrates virtual mapping between current state of art technology with brain. This shift in new paradigm for computing is motivated by the structure of human brain and it is efficient in performing complicated tasks rapidly without pre-programming with accuracy and consumes very less energy. This has motivated another field of study focused at the growth and expansion of the neurosynaptic computing systems that could possibly copy the cerebrum's processing proficiency, its size and power. These novel architectures utilizes extensive parallelism and distributed description of data to facilitate novelty to address current unmanageable problem, mainly the reduction of the gap between computational resources and the data storage components present in conventional von Neumann architectures.

In this paper, we present an overview and challenges of the brain-inspired microprocessor architectures and provides a survey of different approaches being followed for developing artificial neural processing systems which exhibits substantial performance observed in biological systems.
2. Related Work
The interest in neuromorphic computing systems was first seen in late eighties, when Douglas et al. [9] proposed to use the strengths of CMOS VLSI technologies to construct a diverse set of neural analogues, starting from a single synapses, to a set of sensors and simple system. Thereafter, a number of small research labs in the world started working on VLSI model of active cochlea [10], oculomotor system [11]. Recently, an increasing number of researchers from both academia and industry are working on the design of neuromorphic chips [13]. Several researchers reported different new materials for constructing nano device that can mirror some of the properties found in bio-synapses. These neuromorphic computing systems are build using the basic design principles of bio-neurological systems [9].

Moreover, these system are designed using advanced mixed analog/digital circuits fabricated using VLSI processes. Analogous to the neurological system, the neuromorphic computing systems utilizes asynchronous, event driven, energy efficient and fault tolerant techniques [5]. The main difference between neuromorphic systems and traditional computing systems (CMP) is in the way, the memory is organized and used. The brain inspired computing presents an interesting solution for the development of modified von Neumann architectures. Merolla et al. [8] designed a digital neurosynaptic processor with the aim of implementing as much as neurons present in a cat brain at IBM SyNAPSE project. A group of researchers works on the comparative studies between neuroscience and neural network based machine learning. For example, Serre et al. presented HMAX, a visual cortex inspired neural network (rate-based, non-spiking, neurons) and demonstrated to be competing with stable machine- learning models. Vogelstein et al. [9] used basic lookup table synapses to reproduce biological neurons in VLSI process.

However, the simple model of synapse was unsuccessful in identifying the plasticity of the bio-synapses, which is key to information storage and brain learning. Bartolozzi et al. [11] build a VLSI plastic synapse model, whereas Jackson et al.[12] did the same using phase change devices with increased synaptic density.
3. Overview of Neurosynaptic System Architecture

The human brain is capable of astonishing feats of perception, action and memory. There are a million neurons in a human body that make multitasking and memory management very efficient and fast. With the present scaling limitations and von Neumann bottleneck, there is a need for more efficient system hence the tremendous amount of research in neuromorphic system.

![Block Diagram of Neurosynaptic Architecture](image)

**Figure 3.** Block Diagram of Neurosynaptic Architecture [13].

3.1. Architecture

The idea is to integrate the memory and the computation for quick results. Figure 3 presents the block Diagram of a Neurosynaptic Architecture. The fundamental units of neurosynaptic system are the dendrons, axons and synapses. The dendrons communicate with axons through synapses. The system can have a digital or an analog implementation. The major drawbacks of an analog implementation are its sensitivity to PVT variations and the limited scalability. Hence the digital implementation is preferred. The conventional SRAM can be replaced by memristors for better performance. A very popular implementation is obtained by arranging memory elements in an $M \times K$ crossbar array to store the synaptic weight values. The core contains $M$ axons that are connected to $K$ dendrons through $M \times K$ binary valued synapses. The connection between the $j^{th}$ axon and the $i^{th}$ dendron is given by $W_{ji}$. The activity bit, $A_j(t)$, for the $j$th axon represents whether the corresponding neuron was fired in the previous
time step. The state of the \( j^{th} \) axon is denoted by \( G_j \) which can assume a value of 0, 1 or 2. These values distinguish the connection (eg. Excitatory or inhibitory). The synaptic weight between the \( i^{th} \) dendron and \( j^{th} \) axon is given by

\[
S_{i}^{G_j}
\]

Thus the input to the dendron is:

\[
A_j(t) \times W_{ji} \times S_{i}^{G_j}
\]

Assuming a membrane voltage of \( V(t) \), leak voltage \( L \) and threshold voltage \( T \), the membrane voltage of a neuron is updated at each time step as:

\[
V(t + 1) = V(t) + L_i + \sum_{j=1}^{M} A_j(t) \times W_{ji} \times S_{i}^{G_j}
\]

Whenever this membrane voltage of a neuron exceeds the threshold value, it generates a spike and its membrane voltage is reset to 0. The fundamental unit can be arranged in various ways to achieve the required architecture that mimics the brain.

3.2. **Deep Neural Networks**

Deep Neural Networks (DNN) are composed of multiple layers of neurons. There is a one to one correspondence between the neurons belonging to adjacent layers hence a two way communication is possible. There is a high fan in as well as fan out in this architecture as there is a feedback as well as a feed forward path between neurons from different layers. This architecture is typically used in wide range of classification problems. Convolution network are similar to DNN but they are composed of only feed-forward network. The inputs to a layer from the neurons of previous layer are used to compute the weighted sum which is the input for the subsequent layer. This connectivity pattern forms the convolution kernel.

4. **Neuromorphic Engineering Challenges**

In order to fully utilize the strength of brain-inspired neuromorphic system. We need to address the following challenges.

4.1. **Focused Effort to design Neuromorphic System**

Numerous years have passed after the first research publication on neuromorphic system [mead]. In the near past, only a few small groups started working on neuromorphic engineering inorder to utilize huge number of transistors in current technology node also called the dark silicon era, which restricts further performance improvement in chip multiprocessors. In dark silicon based CMPs, all the on-chip cores cannot be switched on together. The progress in neuromorphic engineering is slow as compared to conventional engineering and in the field of technology (for example: a large number of engineers are assigned to the development of GPUS and CMPS) due to which there is a shortage of technology infrastructure for neuromorphic circuit design.

4.2. **Limited Brain Understanding**

Another, major hurdle in the rapid development of neuromorphic systems is the limited knowledge and understanding of the functions of brain and neural computing. The complexity of human brain makes it almost impossible to mimic. This complexity comes due to the high number of neurons. The neural engineering research groups much learn and comprehend the art of building real time event-based sensors, motor-reactive systems. The research community should propose novel neural based techniques to link machine learning with neuromorphic systems that models cognitive abilities.
4.3. Multi-disciplinary Research
In order to achieve the aim of developing neuromorphic systems, it is extremely necessary for various researchers from diverse field of engineering to develop a multi-disciplinary groups. These groups together must explore and investigate the main principles of neural computing and develop techniques that serve as a platform to perform complex neural simulations.

4.4. Other Challenges
The existing computing platform are approximately 300 times slower than human brain. Moreover, the existing neural networks are fast and reliable but even a small network grows inevitably in size along with switching time delays. Another, important limitation is the leakage current at the neurosynaptic nodes due to the high density of interconnections which is responsible of transmitting information to the wrong node. Another issue is the interfacing of these neuromorphic architecture (modified von Neumann) to the existing computing devices with von Neumann architectures.

5. Conclusions
In summary, we presented an overview of a basic brain inspired neuromorphic system architecture. In this work, we provided a survey of few essential challenges in terms of theoretical and experimental neuroscience. For the development of neuromorphic computing systems, inter-disciplinary efforts are required which involve contribution from neurologists/neuroscientists, computer engineers, technology engineers and researchers from material science. There is a strong need to train new generation of scientists and engineers with inter-disciplinary skill and encourage the area specialists on different domains to work together for the deep understanding of brain and to develop brain-inspired high performance system architectures in dark silicon era. As part of our future work, we will use GEM5 full system simulator to model a distributed memory multicore system and to simulate a real heterogeneous system with varying workloads under stringent power-performance constraints. With initial simulation results, we plan to present few fundamental problems that can be solve by the hardware/software research community.

Acknowledgement
We would like to extend our gratitude to the Senior Professors from the Department of Electrical and Electronics Engineering, BITS Pilani for their valuable suggestions.

References
[1] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner, Simics: A Full system Simulator Platform. Computer, 2002, vol. 35-2, pp. 50-58.
[2] C. Bienia, S. Kumar, J. P. Singh, and K. Li, The parsec benchmark suite: Characterization and architectural implications, in Procs. of the International Conference on Parallel Architectures and Compilation Techniques, 2008.
[3] C. Mead, Analog VLSI and Neural Systems. Reading, MA, USA: Addison-Wesley, 1989.
[4] E. Fragniere, A. van Schaik, and E. Vittoz, Design of an analogue VLSI model of an active cochlea, J. Analog Integr. Circuits Signal Process., vol. 13, no. 1/2, pp. 1935, May 1997.
[5] T. Horiuchi and C. Koch, Analog VLSI-based modeling of the primate oculomotor system, Neural Comput., vol. 11, no. 1, pp. 243265, Jan. 1999.
[6] K. Boahen, Neuromorphic microchips, Sci. Amer., vol. 292, no. 5, pp. 5663, 2005.
[7] R. Sarpeshkar, Brain powerVBorrowing from biology makes for low power computing VBionic ear, IEEE Spectrum, vol. 43, no. 5, pp. 2429, May 2006.
[8] S. H. Jo et al., Nanoscale memristor device as synapse in neuromorphic systems, Nano Lett., vol. 10, no. 4, pp. 12971301, 2010.
[9] K. Kim et al., A functional hybrid memristor crossbar-array/CMOS system for data storage and neuromorphic applications, Nano Lett., vol. 12, no. 1, pp. 389395, 2011.
[10] W. Maass, *Noise as a resource for computation and learning in networks of spiking neurons*, Proc. IEEE, vol. 102, no. 5, pp. 860–880, May 2014.

[11] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, and D. Modha, *A digital neurosynaptic core using embedded crossbar memory with 45 pJ per spike in 45 nm*, in IEEE Custom Integrated Circuits Conference, pp. 14, IEEE, Sept. 2011.

[12] T. Serre, L. Wolf, and T. Poggio, *Object Recognition with Features Inspired by Visual Cortex, in Conference on Computer Vision and Pattern Recognition*, pp. 994–1000, IEEE, 2005.

[13] Paul Merolla, John Arthur, Filipp Akopyan, Nabil Imam, Rajit Manohar, Dharmendra S. Modha, *A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45 pJ per Spike in 45 nm* in IEEE custom integrated circuits conference (CICC), pp.1, 2011.