

Emulation of Neural Networks at Nanoscale¹

Mary M. Eshaghian-Wilner,* Aaron Friesz,** Alex Khitun,* Shiva Navab,*
Alice C. Parker,** Kang L.Wang, * and Chongwu Zhou**

*Department of Electrical Engineering
University of California at Los Angeles
Los Angeles, CA 90095

**Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089

Abstract. In this paper, we propose using a nanoscale spin-wave-based architecture for implementing neural networks. We show that this architecture can efficiently realize highly interconnected neural network models such as the Hopfield model. In our proposed architecture, no point-to-point interconnection is required, so unlike standard VLSI design, no fan-in/fan-out constraint limits the interconnectivity. Using spin-waves, each neuron could broadcast to all other neurons simultaneously and similarly a neuron could concurrently receive and process multiple data. Therefore in this architecture, the total weighted sum to each neuron can be computed by the sum of the values from all the incoming waves to that neuron. In addition, using the superposition property of waves, this computation can be done in $O(1)$ time, and neurons can update their states quite rapidly.

Keywords: neural networks, nanoscale architectures, fully interconnected networks, spin waves.

1. Introduction

During the past several decades, researchers have developed electronic models of neurons designed to emulate neural behavior with electrical signals that mimic in some ways the measured potentials of biological neurons. However, due to certain inherent limits in VLSI technology, emulation of complex neural network models has not been feasible.

There are several challenges in realizing a neural network model. One of the most difficult of the challenges is the massive interconnectivity. Neurons in the human cortex possess on the average of 10,000 post synaptic terminals, amounting to massive fan in and fan out [1]. Also, multiple synapses can converge (fan in) to a single postsynaptic terminal, either from a single oversized presynaptic terminal or from multiple presynaptic terminals. The structure of the dendritic tree can be quite complex as well, affecting the processing of the neuron [2]. In this paper, we are focusing on solutions to the interconnection challenges. Fortunately many neural connections are local, and our solution exploits this property.

A suitable topological candidate for emulating neural circuits would be a fully interconnected network, where each of the computing nodes could send (or receive) data to (or from) all the other nodes. However, implementing such a network on a CMOS VLSI chip is not practical due to inherent interconnection limits in this technology. For instance, the VLSI area of a fully interconnected

¹ Authors are listed alphabetically

network of multiprocessors on a chip, in which each of the computing nodes can send (or receive) data to (or from) all the other nodes directly and in constant time, is on the order of $O(N^4)$ [3], which would be unreasonably large to implement. Furthermore, the implementation of such an organization would require having nodes with $O(N)$ fan-in and fan-out, which also is not practical in VLSI. If constant degree nodes were to be used instead, then there would be an $\Omega(\log N)$ intercommunication delay lower-bound in realizing such a network in VLSI using electrical interconnects[3]. Recent publications [4] discuss the feasibility of modeling neural networks using future nanoscale CMOS technology. The replacement of electrical wires with free space optical interconnects could significantly improve the VLSI interconnection area and fan-in fan-out limitations [3]. However electro-optical designs have their own challenges [5].

Another alternative for reducing the VLSI area requirement is to use nanoscale architectures. Several methods have been proposed during the past few years for implementation of such digital circuits. Among them are designs having single electron transistors, molecular switches, quantum-dots, carbon Nanotubes, and spins [6]. In traditional spin-based devices, information is transmitted via the spin of a carrier with the orientation of the spin in one of two directions (along with or opposite to the external magnetic field) representing the bit to be transmitted. The spins, attached to carriers, transfer information from one spin-based device to another through a conducting wire. We propose to use spin-wave-based devices in which the wave transmits the information without any charge transfer. We present a nanoscale spin-wave-based fully interconnected architecture for implementing neural networks. We show that this architecture can efficiently realize highly interconnected local neural-network models such as the Hopfield model. In our proposed architecture, no point-to-point interconnection is required, so, unlike standard VLSI design, no fan-in/fan-out constraint limits the interconnectivity. Using spin waves, each neuron can broadcast to all other neurons simultaneously and similarly a neuron can concurrently receive and process multiple data. Therefore in this architecture, the total weighted sum to each neuron can be computed by the sum of the values from all the incoming waves to that neuron. In addition, using the superposition property of waves, this computation can be done in $O(1)$ time, and neurons can update their states quite rapidly.

The rest of the paper is organized as follows. In Section 2, we present a brief introduction to spin waves as well as the Hopfield neural network model. We then describe our proposed spin-wave-based fully-interconnected architecture and the emulation of the Hopfield neural network model using this architecture. Our concluding remarks and future research are discussed in Section 4.

2. Background

In this section, we first present an introduction to spin waves. We then give a brief description of the Hopfield neural network model that we emulate on our spin-wave-based architecture.

2.1. Spin Waves

The use of spin waves for computation is introduced in [7]. In traditional spin-based architectures, information is encoded into the electron orientation and is carried by the electron through conducting wires. In spin-wave based architectures however, information is encoded into the phase of spin waves and is transferred through the ferromagnetic buses, without any charge transmission [7].

Spin wave is an elementary excitation of a spin system [8]. In other words, spin wave is a collection of precession of electron's magnetic moment about a magnetic field, as shown in Figure 1. Attenuation of a spin wave is around 50 Microns, which makes it a suitable candidate for "nano" scale communication.

; ;

Figure 1- Spin Wave: Collection of Precession of Electron Magnetic Moment about a Magnetic Field

2.2. Hopfield Neural Network Model

The Hopfield model, an associative or content-addressable memory model, was proposed by John Hopfield. The publication of his work significantly contributed to the renewed interest in research in artificial neural networks [9]. In this model, each processing device (neuron) i has two states: $\alpha_i = 0$ ("not firing") and $\alpha_i = 1$ ("firing at maximum rate"). When neuron i has a connection made to it from neuron j , the strength of connection or so called weight of this connection is defined as w_{ij} . ($w_{ij}=0$ for non-connected neurons). The input coming from neuron j to neuron i is shown as I_{ij} . The total input to neuron i , I_i , is computed as the weighted sum of all its inputs. For each neuron i there is a fixed threshold θ_i . Each neuron randomly and asynchronously evaluates whether its total input is above or below threshold and readjusts accordingly [10,11].

$$\alpha_i = \begin{cases} 1, & I_i \geq \theta_i \\ 0, & \text{otherwise} \end{cases}$$

3. The Spin-Wave Architecture

In this section, we present the proposed layout of a spin-wave based nanoscale cluster of neurons on a semiconductor chip. In this architecture it may be possible for each neuron to concurrently broadcast to all other neurons and similarly each neuron may be able to receive and process multiple data simultaneously, depending on the physical implementation.

Figure 2 shows the top and cross-section view of the layout of our fully interconnected architecture proposed previously [12] in which the N computing nodes (neurons) are placed around a circle on a magnetic film. Each neuron communicates via a single asymmetric coplanar strip (ACPS) transmission line, used as either a sender or receiver but not both simultaneously.

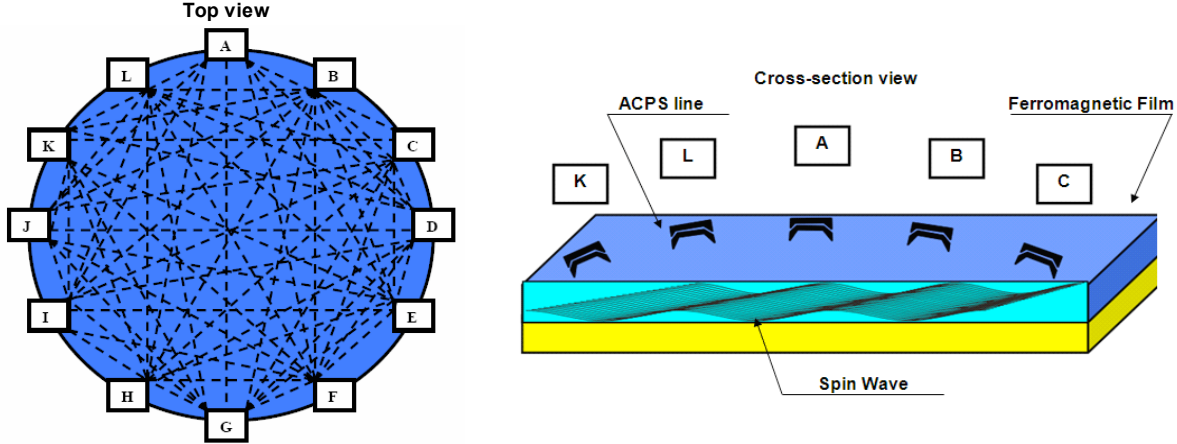


Figure 2- The top and Cross-section view of the architecture with full spin-wave interconnectivity

The area requirement of the above layout of the spin-wave cluster architecture is $O(N^2)$ as opposed to the $O(N^4)$ area requirement if electrical interconnects were to be used. We should also note that in this architecture all the distances are nanoscale dimensions. Also unlike an electrical interconnection network, in which only one transmission can be done at a time, here multiple simultaneous communications are possible by transmitting the spin waves over different frequencies. The information is coded into the phase of the spin waves in the sender and is detected by the receivers. In addition, within each frequency, data can be sent to one or more other neurons from each neuron. In the following, we present a brief discussion on the placement of neurons and the communication mechanisms among them.

3.1. Placement of the Neurons

Normally in architectures where the phases of the waves are the means of information transmission, the exact location of the nodes with respect to the size of topology is an important design issue. The distance between the sender and receiver has to be at a length that is a multiple of the wave's wavelength; otherwise the receiver might receive the wave with a π radian phase-shift, which is a "0" instead of a "1" or vice versa. However, in our design, this is not an issue since the wavelengths of spin waves are considerably larger than the distances between the neurons. The speed of spin waves is around 10^5 m/s. Assuming the input frequency range of 1-10 GHz (as in our experiment), the wavelength will be in the order of 10^{-4} to 10^{-5} m, while the distances are nanoscale or 10^{-9} m. In other words, the wavelengths of the spin waves are some orders of magnitude greater than the distances between the neurons. Therefore, all the neurons receive the same phase regardless of their location, and there is no need to place the neurons in specific distance relative to the other ones.

3.2. Communication among the Neurons

To distinguish the data being transmitted to different neurons, transmissions are done at distinct frequencies, using frequency division multiplexing. This is similar to having various radio stations broadcasting at different frequencies. To listen to a specific station, one dynamically tunes to the corresponding frequency. Here similarly each neuron can broadcast or receive at a specific frequency. For instance, neuron A can broadcast to all the other neurons. This requires that all the neurons' receiving frequencies to be tuned to the same frequency as neuron A's transmitting frequency. Similarly, a neuron can receive and process multiple data simultaneously. For instance, neuron G can receive multiple data simultaneously from other neurons. In this case, the requirement is that all the neurons should transmit at the same frequency at which G's receiver is tuned. Furthermore at a given frequency, each neuron can listen to multiple waves simultaneously. Using the superposition property of waves, it can compute the sum of all waves sent to it.

3.3. Data Detection at the Neurons

The input information is coded into the polarity of the voltage pulse applied to the edge ACPS lines (for example, $V_{input} = +1V$ corresponds to the logic state 1, and $V_{input} = -1V$ corresponds to the logic state 0). In order to detect the output signal V_{ind} we use the time-resolved inductive voltage measurement.

In analog detection mode, the ACPS line detects the inductive voltage produced by the superposition of multiple waves. For example, if ten waves are sending a "1", then their analog sum through their cumulative amplitude is computed instantly as 10. Also this property can be used to compute logical functions as described previously. In digital detection mode, this value is digitized to just a "1," and then the computations are continued digitally. We can take advantage of this property to implement a simple Hopfield neural network. Each neuron can simultaneously sum the weight of all incoming weights broadcast over the waves and see if it is above a certain threshold or not and update its state accordingly. We propose to study this further for implementation of more sophisticated neural models such as those that are updated based on the frequencies of the input signals to neurons.

3.4. Limitations of the Proposed Implementation

The complexities of neural modeling challenge every possible engineering technology. In spite of the advantages of spin waves over other communication technologies, some limitations exist in the proposed model. The limitations arise from the fact that each neuron has a single transmitter/receiver. Specifically, the weighting of signals transmitted would be performed by the transmitter, not by the receiver, and therefore every recipient from a single transmitter would receive equally weighted signals. Also, assuming analog summation of multiple signals, simultaneous reception of different frequency signals at a single receiver would not be possible. This would limit sender/receiver

transmission combinations. Both of these limitations could be remedied by time-multiplexing the communication medium.

4. Conclusion

In this paper, we proposed to use a nanoscale spin-wave-based cluster architecture for implementing neural networks. We showed that this architecture can efficiently realize locally highly-interconnected neural network models such as the Hopfield model. In our proposed architecture, no point-to-point interconnection is required, so unlike standard VLSI design, no fan-in/fan-out constraint limits the interconnectivity. Using spin waves, each neuron can broadcast to all other neurons simultaneously and similarly a neuron can concurrently receive and process multiple data. Therefore in this architecture, the total weighted sum to each neuron can be computed by the sum of the values from all the incoming waves to that neuron. In addition, using the superposition property of waves, this computation can be done in $O(1)$ time, and neurons can update their states quite rapidly. Our future work includes the integration of such nanoscale spin-wave cluster of neurons into microscale chips with MEMS components for dynamically interconnecting the clusters. Furthermore, we are studying the incorporation of carbon nanotubes [13] for providing static connectivity among the neurons.

References

- [1] Shepherd, Gordon "Introduction to Synaptic Circuits," in *The Synaptic Organization of the Brain*, edited by Gordon Shepherd, 5th edition, Oxford University Press, 2004
- [2] Mel, B.W. & Schiller J. (2004). On the fight between excitation and inhibition: Location is everything. *Science STKE*, pe44.
- [3] M. M. Eshaghian, "Parallel Computing with Optical Interconnects," Ph.D. Dissertation, University of Southern California, December 1988.
M. M. Eshaghian, "Parallel Algorithms for Image Processing on OMC," *IEEE Transaction on Computers*, Vol. 40, No.7, 1991.
- [4] A. C. Parker, A. K. Friesz, and A. Pakdaman, "Towards a Nanoscale Artificial Cortex", The 2006 International Conference on Computing in Nanotechnology (CNAN'06): June 26-29, 2006, Las Vegas, USA
- [5] M. M. Eshaghian, and L. Hai, "A Glance at VLSI Optical Interconnects: From the Abstract Modelings of the 1980s to Today's MEMS Implementations," book chapter "Handbook on Innovative Computing," 2006.
- [6] M. M. Eshaghian-Wilner., A. H. Flood, A. Khitun, J. Fraser. Stoddart, and K. L. Wang, "Molecular and Nano-scale Computing and Technology," book chapter "Handbook of Innovative Computing," 2006.
- [7] A. Khitun, and Wang K. L., "Nano Scale Computational Architectures With Spin Wave Bus," *Superlattices & Microstructures*. 38(3),184-200, 2005.
- [8] C. Kittel, "Introduction to Solid State Physics," New York, Wiley, 1986.
- [9] <http://encyclopedia.thefreedictionary.com/Hopfield+network>
- [10] <http://www.comp.nus.edu.sg/~pris/AssociativeMemory/HopfieldModel.html>
- [11] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of National Academy of Sciences*, vol. 79 no. 8 pp. 2554–2558, 1982
- [12] M. M. Eshaghian-Wilner., A. Khitun, S. Navab, and K.L. Wang, "A Nano-scale Module with Full Spin-Wave Interconnectivity for Integrated Circuits", *NSTI Nanotech 2006*, Boston, May 2006.
- [13] F. Liu, M. Bao, H. Kim, K. L. Wang, C. Li, X. Liu, and C. Zhou, "Giant Random Telegraph Signals in the Carbon Nanotubes as a Single defect Probe," *Appl. Phys. Lett.* 86, 163102 "C 1-3 2005.