Video Analytics Using Beyond CMOS Devices

Vijaykrishnan Narayanan[1,*], Suman Datta[2,*], Gert Cauwenberghs[3,~], Don Chiarulli[4,+], Steve Levitan[5,+], Philip Wong[6,^]
[*]The Pennsylvania State University, [~]University of California at San Diego, [+]University of Pittsburgh, [^]Stanford University
[1]vijay@cse.psu.edu, [2]sdatta@engr.psu.edu, [3]gert@ucsd.edu, [4]don@cs.pitt.edu, [5]levitan@pitt.edu, [6]hspwong@stanford.edu,

***Abstract*— The human vision system understands and interprets complex scenes for a variety of visual tasks in real-time while consuming less than 20 Watts of power. The holistic design of artificial vision systems that will approach and eventually exceed the capabilities of human vision systems is a grand challenge. The design of such a system needs advances in multiple disciplines. This paper focuses on advances needed in the computational fabric and provides an overview of a new-genre of architectures inspired by advances in both the understanding of the visual cortex and the emergence of devices with new mechanisms for state computations.**

Keywords—Emerging Device; Coupled Oscillator

I. INTRODUCTION

Cameras are already ubiquitous in our lives - in phones, game consoles, computers, cars, shopping malls, and airports [1, 2]. Transforming these cameras from passive recording devices with simple image processing capabilities to smarter systems that can understand and interpret complex scenes can have a multi-faceted impact on society, including assistance to blind or visually-impaired people, enhanced driver safety, enhanced experience for retail shopping or a vacation visit, and enhanced safety for critical public infrastructure (Figure 1). While tremendous progress has been achieved in past decades, the best engineered systems today still fall short of the robustness, flexibility, capability, and power usage of biological vision. While humans can understand scenes (at least roughly) within 150msec [3], no computer vision system can understand complex scenes. While computer vision systems can perform face detection and text detection in real time on smart phones, more complex and robust algorithms do not meet real-time needs. On a different dimension, the brain performs these complex vision tasks using 20W of power, at least three orders of magnitude more energy efficient than customized state-of-the-art artificial vision systems [4, 5]. Through concurrent advancements in algorithms and hardware, we anticipate realization of systems that can understand complex scenes, enhance the energy-efficiency by 100-1000X and accelerate performance to enable these algorithms to be deployed in end-applications.

Recent breakthroughs in neuroscience-inspired vision algorithms have shown that performance and power budgets which approach human levels are achievable in restricted domains such as visual attention, computing a visual scene's

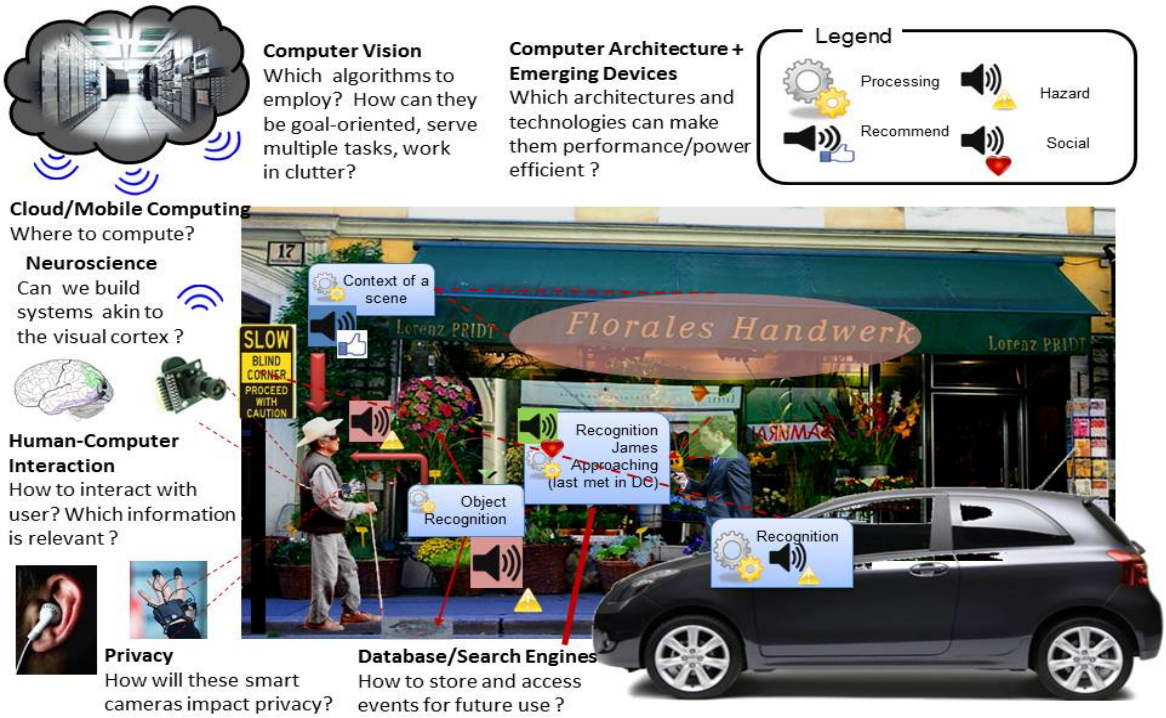

Figure 1: Illustrative scene of people and cars with smart cameras

978-3-9815370-2-4/DATE14/©2014 EDAA

coarse "gist" or scene category, and recognizing objects. Innovations in device technology and hardware design are enabling new computational paradigms beyond the confines of Von-Neumann architectures that hold promise for more efficient processing of more complex algorithms. New architectural and device paradigms offer an opportunity for radically different forms of algorithmic design and implementation. This paper outlines two complementary approaches to designing smart camera systems: the neuronal spike architecture models individual spiking neurons in visual cortex; the coupled oscillator model takes the view of local field potentials reflecting and inducing synchrony across populations of spiking neurons. From technology viewpoint, the two approaches are based on analog spike based architectures that leverage emerging analog memories and associative memory architectures that leverage nano-oscillators for pattern matching respectively.

II. NEUROMORPHIC HARDWARE: BRAIN-LIKE FABRICS FOR COMPUTER VISION

The human brain consists of $\sim 10^{11}$ neurons and an extremely large number of synapses, $\sim 10^{15}$, which act as a highly complex interconnection scheme among neurons. Synapses dominate the architecture of the brain and are responsible for massive parallelism, structural plasticity, and robustness of the brain. They are also crucial to biological computations that underlie perception and learning. Therefore, a compact nano-electronic device emulating the functions and plasticity of biological synapses will be the most important building block of brain-inspired computational systems. We provide insight on how recent technology innovations in dense analog memory are enabling efficient realization of similar spiking neuron architectures on silicon. Cauwenberghs' initial work has resulted in event-driven spiking neural arrays with dynamically reconfigurable synaptic connections for large scale emulation of neocortical vision (Figure 2) [6, 8]. The synaptic connections as well as the weights of the synaptic connection are achieved through an external memory look up in this architecture. Hierarchical address-event routing (HiAER) offers scalable long-range neural event communication tailored to locally dense and globally sparse synaptic connectivity [9], while integrate-and-fire array transceiver (IFAT) CMOS neural arrays offer low-power implementation of continuous-time analog membrane dynamics. The current 65k-neuron IFAT chip in 130nm double-stack 3-D integrated CMOS operates at 48pJ/spike energy [7]. This architecture is amenable to mapping various vision processing algorithms based on spiked neural networks. However, the use of external DRAM memory lookup for synaptic connection greatly impacts the efficiency of this architecture. Our objective is to replace the core of the external DRAM memory lookup with nanoscale analog memory implementing synapse arrays vertically interfacing with neuron arrays, and further optimize integrated IFAT-HiAER circuits, towards sub-pJ/spike overall energy efficiency in neocortical neural and synaptic computation, communication, and learning.

Using phase change materials (PCM), Wong's group has developed nanoscale analog programmable devices with 100-step grey scale conductance modulation capable of emulating

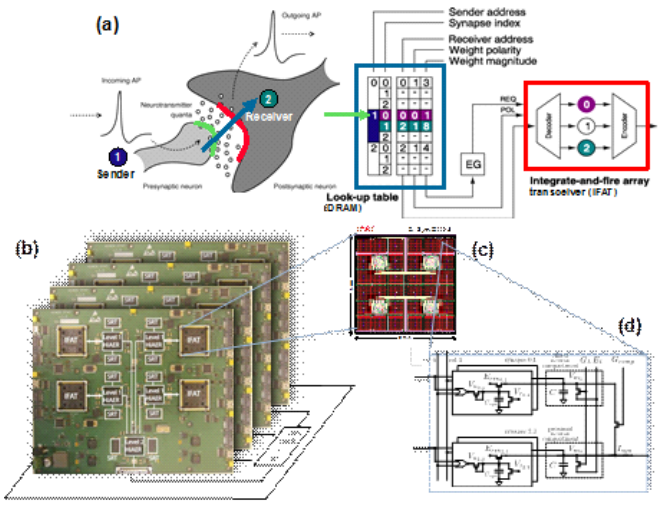

Figure 2: Hierarchical Address-Event Routing (HiAER) Integrate-and-Fire Array Transceiver (IFAT) for scalable and reconfigurable neuromorphic neocortical processing [6, 7].
(a) Dynamic reconfigurable synaptic connectivity across IFAT arrays of addressable neurons is implemented by routing neural spike events through DRAM synaptic routing tables (SRT).
(b) Full-size HiAER-IFAT network with 4 boards, each with 4 IFAT modules, serving 1M neurons and 1G synapses, and spanning 4 levels in connection hierarchy.
(c) Each IFAT chip module comprises a 65k-neuron Tezzaron 130nm CMOS IFAT microchip, Xilinx Spartan-6 FPGA (Level 1 HiAER), and two 2Gb DDR3 SDRAM SRTs serving 65M synapses.
(d) Each neural cell models conductance based membrane dynamics in proximal and distal compartments for synaptic input with programmable axonal delay, conductance, and reversal potential. IFAT chip measured energy consumption is 48 pJ per spike event, several orders of magnitude more efficient than emulation on CPU/GPU platforms.

the plasticity of the bio-logical synapse [10] (Figure 3). Device density obtained is 1.7×10^{10}cm^{-2} (75nm bottom electrode diameter) and is scalable to 6×10^{12}cm^{-2} (4nm×4nm per programmable element). Programming energy (to bring the device to full crystalline high-conductance state) is demonstrated at 50pJ and is scalable to 2pJ. Reset energy (to bring the device to full amorphous low-conductance state) reaches as low as 500fJ with novel methods [11].

PCM synapse devices can be fabricated on top of CMOS neural circuits using E-beam lithography [12]. As in HiAER-IFAT, sparse long-range synaptic connectivity will be implemented over the HiAER network by address-event routing (AER) of spike-event synaptic inputs and neural outputs. In contrast, high-density and high-efficiency local synaptic connectivity will be obtained by connecting each neuron in the CMOS IFAT neural array to one dedicated output line of the PCM crossbar array. This spatial interleaving of vertical interconnects avoids the need to pitch match synapse and neuron cells in the two arrays, and allows layout-optimal near-square cell geometries, with the neuron cell area greater than the synapse cell area by a factor equal to the synaptic fan-in/out. Hence large synaptic fan-in/out and relatively large neuron cell sizes are supported in fully scalable architecture.

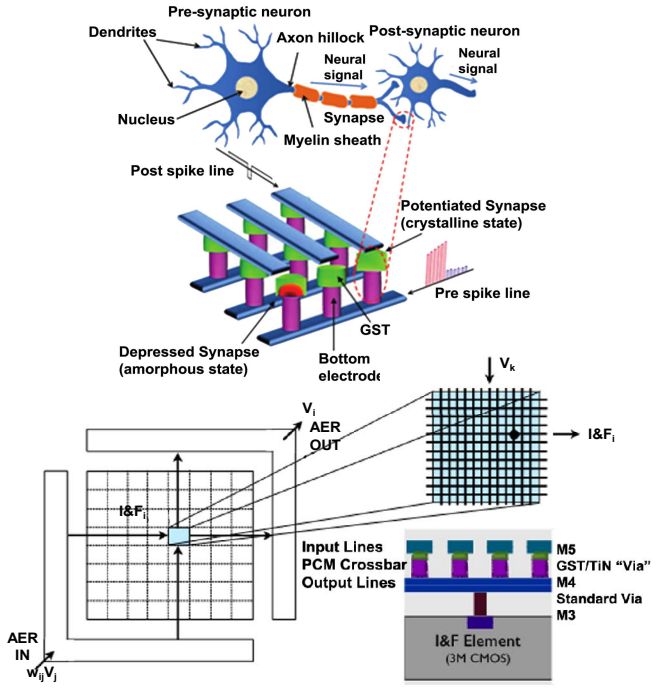

Figure 3: Hybridization and nanoscale integration of CMOS neural arrays with phase change memory (PCM) synapse crossbar arrays.
(top) Nanoelectronic PCM synapse with spike-timing dependent plasticity (STDP) [10, 19].
(bottom) CMOS IFAT array vertically interfacing with nanoscale PCM synapse crossbar array by interleaving via contacts to crossbar rows. The integration of IFAT neural and PCM synapse arrays externally interfacing with HiAER neural event communication combines the advantages of highly flexible and reconfigurable HiAER-IFAT neural computation and long-range connectivity with highly efficient local synaptic transmission

The proposed concept is similar to CMOL CrossNets [13] without a need for relative lattice rotation between neural and synapse arrays, and is compatible with state-of-the-art deep-submicron CMOS fabrication by deposition of phase change material (GST, W, and TiN) "vias" between two adjacent higher-level metal layers in the CMOS process. With currently available 20nm metal line pitch for 20nm pitch PCM cell arrays, and with currently implemented CMOS neuron cell arrays of 2µm pitch, a 10,000 synaptic fan-in/fan-out is currently achievable with 20M neurons and 200B synapses per square cm.

The neurons in the proposed architecture implements stochastic integrate-and-fire (I&F) neurons for Bayesian/Boltzmann inference and the PCM arrays provide synaptic plasticity for learning. Voltage clamping of each crossbar output line with a sense amplifier bypasses the need for selector elements and avoids cross-talk (conductance sneak paths) across synapses. The currents thus acquired from the array are linear in the voltage inputs with weights given by synaptic conductance.

Integrating sense amplifiers with on/off pulsing asynchronous level-crossing comparator with additive Bernoulli noise source and comparator feedback at the input produce a stochastic I&F response. These responses are constant amplitude voltage pulses of positive and negative polarity that encode discrete "on" and "off" events at the output. Such events could represent, for instance, rising and falling intensities in a pixel, or positive or negative gradients in intensities across pixels. While not directly biophysical, encoding on and off cells in a single unit offers the advantage of a compact implementation. The differential signal encoding in this representation offers robustness to noise and common-mode errors. Noise of varying amplitude at the input produces a sigmoidal response with varying degree of saturating nonlinearity at the output. The saturating nonlinearities along with linear synaptic summing and additive noise of the implemented I&F model, each with digitally controlled parameters, map directly to the functional form of the Dynamic Bayesian Network (DBN) and Restricted Boltzmann Machine (RBM) algorithms studied in the machine vision models for low-power embedded applications. An integrate-and-fire neuron implementation, interfacing with a PCM array, and operating on pulsed input voltage waveforms produces stochastic pulsed output waveforms. With the same reference potential used as the baseline for the pulse waveforms, the static current and power consumption of the array are close to zero, and the power is essentially proportional to pulse activity. With nanosecond pulses and 100fF feedback capacitance, we anticipate better than pJ/pulse energy levels with 100kΩ range of resistance of the PCM nano array elements.

Associative learning is central to neural computation and formation of episodic memory in brain. Our architecture can support biophysically-based spike timing-dependent plasticity (STDP) in PCM synaptic conductance as previously demonstrated [10]. Based on measured PCM full-swing programming and reset energies, we anticipate typical energies lower than 1pJ per increment and decrement learning update.

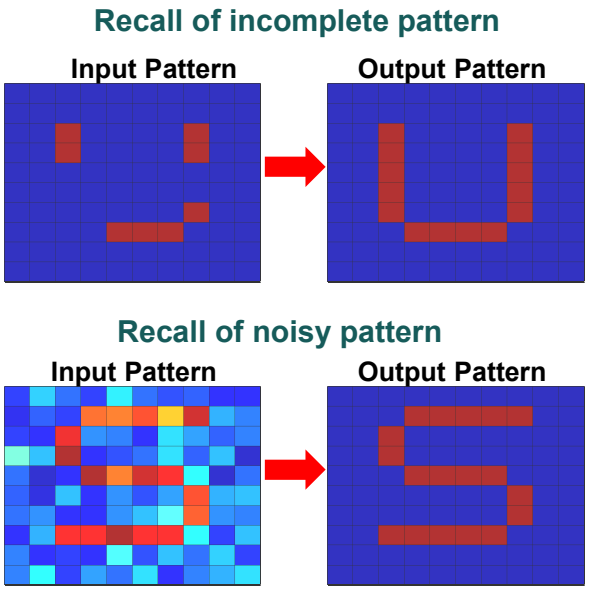

Figure 4: Character Recognition using learning.

The learning schemes have been used to recognize noisy characters as shown in Figure 4. If an incomplete pattern is presented, the potentiated synapses can recruit the missing neurons in order to recall original pattern. A recurrent network of 100 neurons and 10000 synapses with asymmetric STDP is constructed. The network can recruit missing neurons and recall the original pattern, when an incomplete pattern with up to 50% missing neuron spikes is presented. After training, the network is stimulated with 50% incomplete "S" or "U" patterns and is shown to recover the full pattern except the case where missing parts are more than 70% and the stimulated parts have strong overlaps with both patterns.

III. Coupled Oscillator Based Associative Architecture

The proposed architecture couples the oscillatory neurons dynamically through a time dependent input rather than direct connections between them. This complements the approach in neuromorphic architectures that aim to perform neuromorphic computing through flexible adjustment of synaptic connection strength between individual neurons. Hoppensteadt and Izhikevich [14] suggested a mathematical model for an associative processing unit using coupled oscillator arrays. This dynamic model was proved to be able to form attractor basins at the minima of Lyapunov energy function by adjusting a coupling matrix though the use of a Hopfield rule. Consequently, vision problems that require the computation of distance metrics in N-dimensional spaces can be directly mapped to that of observing the phase and frequency synchronization behavior of such coupled oscillator systems. The relative phase relationship among a group of loosely coupled non-linear oscillators can be used as a representation of state. Using phase as the basis of state, our primitive computational operations are based on direct interactions between the frequency and phase relationships of clusters of oscillators. For associative memory comparison operations, synchronization of the oscillators based on the degree of similarity between stored and input vectors, encoded in frequency and phase, becomes our primitive operation.

Many emerging post-CMOS devices exhibit oscillatory behavior that can be leveraged for such designs. The coupling between these oscillator systems can be achieved by connecting the outputs of the oscillator. Spin torque oscillators are a type of frequency-coherent spin devices, based on the interaction of a spin-polarized current with a thin magnetic layer. This device was discovered to have an oscillatory behavior based on exchange coupling, which is similar to the weakly coupled neural network model.

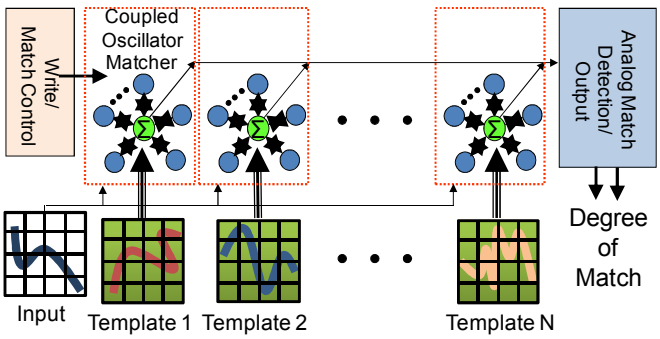

Figure 5: Associative Processor Architecture

Datta et al. have experimentally observed self-sustaining voltage oscillations due to the coexistence of correlated metal and insulating phases in epitaxial correlated oxide (e.g. vanadium oxide, VO2) thin films on rutile TiO2 substrate [15]. Since the origin of the oscillation is traced to atomic scale phenomenon, the oscillators can be scaled to very small dimensions requiring very little excitation signal. The periodicity and shape of the voltage oscillations can be modulated by an external resistor and capacitor in series. Beside the two nano devices mentioned above, there are many other new devices like resonant body transistors [16] and single-electron transistors that exhibit such behavior. These devices can provide opportunities for building novel architectures for cognitive tasks like pattern recognition or computer vision.

In this work we use the relative phase relationship among a group of loosely coupled non-linear oscillators as a representation of state. Based on this concept, information processing systems that perform image recognition tasks by using fast pattern recognition in high-dimension feature spaces can be designed.

Associative processing engine architecture: The associative processor architecture has three main sub-components (Figure 5). It has a template memory storing N-element vectors, multiple coupled oscillator modules each consisting of N oscillators and a post-processing analog module [17] that transforms the coupled oscillator module phase and frequency information to an output indicative of degree of match. The collective behavior of the coupled oscillators is determined by the difference between the template vector and the input vector. The difference in inputs manifests in synchronization behavior of the coupled system in frequency and/or phase. Based on device characteristics and underlying circuit efficiencies, the architecture of the system will need refinements. For example, the limitations on the maximum number of oscillators that can physically realized to couple with each other will require appropriate partitioning of our image vectors into sub-blocks. Similarly, limitation of the sensitivity of the analog circuits will require use of hierarchical matching circuits [18].

Directly using the correlated behavior of oscillators to measure the degree of match in contrast to current approaches that require multiple counters (for Hamming distance), or subtractor/multipliers (for Euclidean distances) results in ~100X device count reduction. Tapping the physics of the correlated oscillators to compute the degree of match and the scalability of these nano oscillator feature sizes are key advantages of this style of architecture.

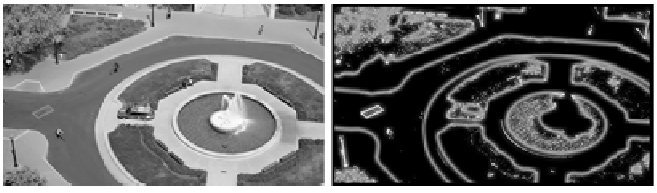

Figure 6: 19×19 Oscillator Saliency Responses
(a) Grayscale Image, (b) Raw Oscillator Response

We can also leverage analog memory such as PCM for efficient template memory design. Due to the large number of templates that are required in our architecture, using compact, low power memories such as phase change memory is very attractive. Further, these analog memories will obviate the need for expensive analog to digital conversion for the template data. Figure 6 shows the results from an oscillator array configured to model the visual attention model.

IV. CONCLUSION

Emerging devices offer a new opportunity for exploring new styles of architectures beyond traditional accelerator designs currently used for computationally intensive video analytics applications. This paper provided an overview of two emerging paradigms for efficient vision architectures. This work primarily focused on the functional design of such architectures. Our ongoing efforts are aimed at further quantification of performance and power benefits offered by these paradigms. This work was supported in part by the National Science Foundation Expeditions in Computing Award: 1317560, 1317407, 1317470, and 1317373.

REFERENCES

[1] E. Tingwall, "2013 Technology of the Year," *Automobile Magazine*, January 2013.

[2] TIME Magazine Staff, November 2012. [Online]. Available: http://techland.time.com/2012/11/01/best-inventions-of-the-year-2012/slide/google-glass/.

[3] S. Thorpe, D. Fize and C. Marlot, "Speed of Processing in the Human Visual System," *Nature*, vol. 381, no. 6582, pp. 520-522, 1996.

[4] IBM Corporation, [Online]. Available: http://www.conceivablytech.com/8953/science-research/i-ibm-a-chip-that-is-measured-in-neurons-and-synapses.

[5] DARPA, "Neovision2: Phase 1 Evaluation Results," 2011.

[6] J. Park, T. Yu, C. Maier, S. Joshi and G. Cauwenberghs, "Hierarchical Address-Event Routing Architecture for Reconfigurable Large Scale Neuromorphic Systems," *in IEEE Int. Symp. Circuits and Systems*, 2012.

[7] T. Yu, J. Park, S. Joshi, C. Maier and G. Cauwenberghs, "65k-Neuron Integrate-and-Fire Array Transceiver with Address-Every Reconfigurable Synaptic Routing," *in IEEE Biomedical Circuits and Systems Conference*, 2012.

[8] J. Park, S. Joshi, T. Yu, C. Maier, F. Broccard and G. Cauwenberghs, "Scalable Reconfigurable Emulation of Cortical Models on HiAER-IFAT," *in IEEE Biomedical Circuits and Systems Conference*, 2011.

[9] S. Joshi, S. Deiss, M. Arnold, J. Park, T. Yu and G. Cauwenberghs, "Scalable Event Routing in Hierarchical Neural Array Architecture with Global Synaptic Connectivity," *in IEEE Int. Workshop Cellular Nanoscale Networks and Their Applications*, 2010.

[10] D. Kuzum, R. Jeyasingh, S. Yu and H.-S. P. Wong, "Low energy, robust neuromorphic computation using synaptic devices," *IEEE Transaction on Electron Devices*, 2012.

[11] J. Liang, R. Jeyasingh, H.-Y. Chen and H.-S. Wong, "An Ultra-Low Reset Current Cross-Point Phase Change Memory With Carbon Nanotube Electrodes," *IEEE Transactions on Electron Devices*, pp. 1155-1163, 2012.

[12] B. Lee and H.-S. Wong, "NiO resistance change memory with a novel structure for 3D integration and improved confinement of conduction path," *in Symposium on VLSI Technology*, 2009.

[13] K. K. Likharev, "CrossNets: Neuromorphic Hybrid CMOS/Nanoelectronic Networks," *Science of Advanced Materials*, vol. 3, pp. 322-331, June 2011.

[14] F. C. Hoppensteadt and E. M. Izhikevich, "Oscillatory Neurocomputers with Dynamic Connectivity," *Physical Review Letters* , pp. 2983-2986, 1999.

[15] E. Freeman, A. Kar, N. Shukla, R. Misra, R. Engel-Herbert, D. Schlom, V. Gopalan, K. Rabe and S. Datta, "Characterization and modeling of metal-insulator transition (MIT) based tunnel junctions," *in 70th Annual Device Research Conference (DRC)*, 2012.

[16] D. Weinstein, *Nano Letters*, vol. 10, no. 4, p. 1234-1237, 2010.

[17] T. Shibata, R. Zhang, S. Levitan, D. Nikonov and G. Bourianoff, "CMOS supporting circuitries for nano-oscillator-based associative memories," *in International Workshop on Cellular Nanoscale Networks and Their Applications (CNNA)*, 2012.

[18] S. Levitan, Y. Fang, D. Dash, T. Shibata, D. Nikonov and G. Bourianoff, "Non-Boolean associative architectures based on nano-oscillator," *in International Workshop on Cellular Nanoscale Networks and Their Applications (CNNA)*, 2012.

[19] D. Kuzum, R. G. D. Jeyasingh, B. Lee and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano Letter*, p. 2179-2186, 2012.