SANDIA REPORT
SAND2010 -6312
Unlimited Release
Printed September 2010

Toward Exascale Computing through Neuromorphic Approaches

Amber McKenzie, Darren W. Branch, Chris Forsythe, and Conrad D. James

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory operated and managed
by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for
the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2010-6312

Unlimited Release

Printed September 2010

Toward Exascale Computing through Neuromorphic Approaches

Conrad D. James

Biosensors and Nanomaterials Department
Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185-1425

ABSTRACT

While individual neurons function at relatively low firing rates, naturally-occurring nervous systems not only surpass manmade systems in computing power, but accomplish this feat using relatively little energy. It is asserted that the next major breakthrough in computing power will be achieved through application of neuromorphic approaches that mimic the mechanisms by which neural systems integrate and store massive quantities of data for real-time decision making. The proposed LDRD provides a conceptual foundation for SNL to make unique advances toward exascale computing. First, a team consisting of experts from the HPC, MESA, cognitive and biological sciences and nanotechnology domains will be coordinated to conduct an exercise with the outcome being a concept for applying neuromorphic computing to achieve exascale computing. It is anticipated that this concept will involve innovative extension and integration of SNL capabilities in MicroFab, material sciences, high-performance computing, and modeling and simulation of neural processes/systems.

TABLE OF CONTENTS

I. INTRODUCTION

The human brain with a volume of ~1200cm^3 consumes 20W to perform >10^{16} operations/s. Current supercomputer technology has reached 10^{15} operations/s, yet requires 1500m^3 and 3MW, giving the brain a 10^{12} advantage in operations/s/W/cm^3. Thus, exascale computing (10^{18}) presents a significant challenge if held to reasonable power and volume limitations. One path forward is to support a paradigm shift towards "neuromorphic" computing where hardware circuits are given architectures that mimic biological neural tissue in the hopes of achieving the computational capabilities of such systems with similar volume and power metrics. During the last several decades, neuromorphic computing strategies have emerged as a means to mimic the architecture of biological networks in silicon hardware [1]. However, these efforts have focused solely on mimicking neural tissue *structure* as opposed to mimicking neural tissue *function*. Thus, these circuits mimic the arrangement of neurons and neuron-neuron connections, but fail to mimic the molecular mechanisms by which actual computation occurs in neural tissue. The purpose of the effort here is to examine the attributes of biological neural networks and to determine paths forward to implement these attributes into silicon-based hardware for next generation computing systems with reduced volume and power consumption. Our tasks for this project are to 1) identify these attributes, 2) condense these attributes into larger categories to simplify the evolution of computing architectures, and 3) explore a path forward towards implementing one or more of those attributes into actual hardware. As a final effort for this project, we chose memristor technology as the neuromorphic-centric technology to further explore and to develop a conceptual design of a device that will demonstrate a biologically-inspired hardware architecture.

II. RESULTS AND DISCUSSION

A. Computational Attributes of Biological and Silicon-based Systems

A series of brainstorming sessions was organized in order to elicit ideas and thoughts about neural and computing systems that could be used to inspire development of next-

generation computing strategies. The session included SNL staff members from various organizations with relevant expertise in the areas of robotics, microfabrication, high-performance computing, cognitive science, computer science, biological science, and computational modeling. Participants included: B Rohrer (6473), F. Rothganger (1434), R. Abbott (1432), C. Warrender (1434), B. Carson (8622), M. Okandan (1749), R. Schiek (1437), J. Wagner (1432).

The first brainstorming session addressed two questions:

1. What are attributes or characteristics of naturally-occurring neural systems that make them superior to current computing technology?
2. What are attributes or characteristics of naturally-occurring neural systems that limit them or make them inferior to current computing technology?

The second brainstorming session addressed the following related questions:

3. What are attributes or characteristics of current computing technologies that make them superior to naturally-occurring neural systems?
4. What are attributes or characteristics of current computing technologies that limit them or make them inferior to naturally-occurring neural systems?

At each of these two brainstorming sessions, participants were asked to put individual thoughts on separate post-it notes. These notes were collected, organized, and compiled into a list. Table 1 shows attributes of both biological and silicon networks. Then researchers translated the concepts into a graphical structure, attempting to represent the underlying intent of the ideas and how they related to each other with regards to neural and computing systems. This analysis produced two separate graphs, which were then condensed into a list of six desired attributes:

1. Connectivity, parallel, distributed
2. Tunability, adaptability, plasticity
3. Self-preservation, healing, robust, recoverable
4. Compact, high-density, 3D geometries
5. Standardizable, predictable, specifiable
6. Flexibly sustained

The next series of brainstorming sessions involved proposing metrics and/or test cases that might serve to target development of a specific attribute.

Table 1: Attributes of biological and silicon networks

Biological networks	**Silicon Hardware**
Multi-scale interconnectivity	Standardizable
Analog and digital	Digital information storage and processing
Self-aware	Durable and robust
Graceful degradation	Easily copied
Memory and information processing	Programmable
Compact/High density	High accuracy/precision
Self-healing	Hardware and software
Adaptable	Interconnectable

B. Current research efforts

Attributes 1 and 4: Large interconnectivity and 3D High density

Issues involving connectivity, parallelism, and distributed computing are becoming extremely important for the computing community in general. Having come close to the threshold of making processors faster, system designers are turning more to system design innovations to try keep up with demand for faster, more efficient computing systems. Much recent work has been conducted with consideration for attributes of natural neural systems. A key advantage held by biological systems is that they are truly 3D in the sense that there is no structural limitation in any dimension. This enables biological systems to be more compact and reduces transport and communication delay times. Current computer technology is only 2D, mostly due to limitations in the fabrication processes used to manufacture chips. Current efforts have reached ~2.5D, which means that only 1-3 layers of active devices have been stacked in the z-dimension [2].

One area of interest is in connectivity, in which researchers strive to develop the most efficient ways to connect nodes and/or components. Vainbrand and Ginosar are working on innovations in flexible connectivity by transforming the typical system on chips architecture into a network on chip architecture for neural networks [3]. With this research, they hope to overcome limitations inherent in rigid hardware connections and introduce flexibility more like that of natural neural systems.

Both the Blue Brain and the SpiNNaker projects aim to address many of these issues surrounding connectivity and parallelism by mimicking the workings of the brain. The Blue Brain project uses the computing power of IBM's Blue Gene supercomputer to power a myriad of nodes representing neurons in an effort to model an actual mammalian brain [4]. "The SpiNNaker system is a biologically-inspired massively parallel architecture of bespoke multi-core System-on-Chips designed with the aim of simulating up to a billion spiking neurons in (biological) real-time" [5].

Parallelism, in particular, is the focus of research efforts at many universities. In 2008, researchers at the University of Illinois at Urbana-Champaign produced a document detailing their work and research avenues in parallel computing [6]. Their three main areas of focus are applications and patterns, disciplined programming models, and development and execution environments. Likewise, Berkley researchers make recommendations framing the parallel landscape in the future, including programming model and hardware architecture recommendations [7].

Size and energy requirements have been significant limitations for the advancements in these aspects of computing. For this reason, researchers are trying to find alternative ways to design hardware that will cut down on these requirements. Some of the most recent work involves a nanoparticle-protein hybrid implementation of a Set-Reset machine [8].

As mentioned previously, one of the downsides of neuromorphic computing efforts has been the focus on mimicking neural structure as opposed to neural function. For instance, synaptic connections between biological neurons are thought to be the basic unit of computation, yet many early efforts in neuromorphic computing focused on hardware CMOS transistor "neurons" connected by software "synapses." Other efforts

[9] have involved the fabrication of large multi-transistor hardware "synapses" that are 10^4 too large to be scaled to biological synapse densities (10^{10} connections/cm^2). However recent work has demonstrated the potential of mimicking molecular mechanisms in neurons. Recent work in the nanoscience arena has focused on memristors, the 4th theoretically-predicted in 1971 but only recently experimentally-verified circuit element [11]. These structures have a resistance value that changes based on the historical record of electrical charge that has passed through them. Thus, they are capable of storing information, and these components scale in sizes similar to that of biological synapses [12]. Jo et al. demonstrated spike-timing-dependent plasticity (STDP) in memristor structures [10], a phenomenon observed decades ago in biological synapses in which a synapse connection grows stronger when the pre-synaptic neuron fires before the post-synaptic neuron. Figure 1 shows an example comparing the demonstration of STDP in a memristor structure and in a biological neural synapse. Arrays of memristors were recently combined with CMOS to create logic gates and flip-flops [13]. Thus, these structures hold great promise for the next generation of neuromorphic computing hardware devices.

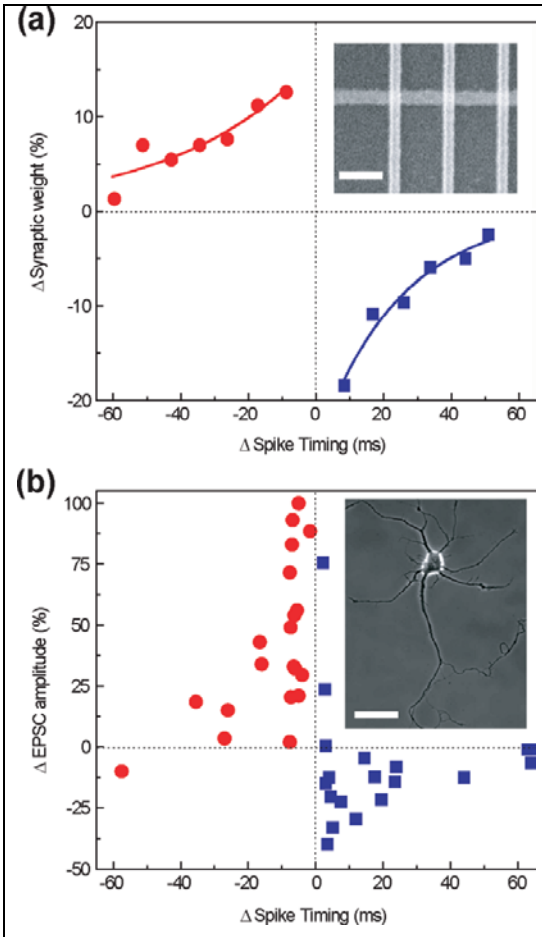

Figure 1: From Jo et al. [10]. Demonstration of spike-timing-dependent plasticity in a a) memristor structure and b) a biological neural synapse.

Attributes 2 and 5: Adaptability and standardization

Biological networks are incredibly adaptable and plastic – meaning that they have the ability to handle data in different forms and when neural tissue is damaged, nearby tissue takes over some of the duties and functions of lost tissue. Next generation computers should be able to handle damage in similar ways.

Another aspect of silicon networks is that they are standardizable. There are defined rules and procedures for fabricating and utilizing silicon hardware and this enables such devices to be volume manufacturable and easily exchanged from one device to another. One of the advantages of biological systems is that the components of living organisms are standardizable: items such as DNA and cells are confined by certain rules that are common to all organisms. What is not standardizable in living systems are the individual expressions of genes and the interconnections between cells. Thus biological systems have standardization, yet there is a degree of freedom and chaos that allows biological systems to adapt to changes in the environment or to changes in the organisms itself.

One approach to affording this blend of standardization and non-standardization to next generation computing hardware is to maintain standardized components such as transistors while allowing interconnections between transistors and integrated circuits to be malleable and modified based on use.

Attributes 3 and 6: Self-healing and flexibly sustained

Motivated by the increasing complexity and speed of new computing systems, research into self-healing and self-preserving systems has attempted to target many different facets of computing. "The essence of autonomic computing systems is self-management, the intent of which is to free system administrators from the details of system operation and maintenance and to provide users with a machine that runs at peak performance 24/7" [14]. Self-healing software has been developed to automate processes such as fault and error detection and the decision making behind determining solutions to problems. These involve both centralized control, in which a centralized entity or program decides a course of action to resolve an issue or adapt to a system or objective change, and distributed control, in which individual nodes or programs collaborate for a

solution [15]. In terms of hardware, two avenues of self-healing have involved both redundancy in hardware components, which would allow transferring of functionality to a non-failing component, and self-healing materials, which would allow for automated hardware maintenance. Challenges facing self-healing technology researchers include finding techniques that can be deployed on different, and often times heterogeneous, platforms, as well as techniques that are both consistent and efficient.

With regards to self-healing and redundant hardware, the majority of research involves basic structures of reconfigurable logic components. Much recent research has targeted field programmable gate arrays (FPGAs) as a way to produce self-healing, adaptive architectures necessary for mission-critical systems. Sreeramareddy et al. "present an attractive option for creating reliable platforms that adapt to changes in user objectives over time and respond to hardware/software anomalies automatically with self-healing action" [16]. The first implementation of their technique involves a two-level hierarchical self-testing, at the component and network levels, that strives to identify faults at both fine and coarse granularity, with the intent of reducing the amount of logic required [17]. They later move to addressing the node and network levels [18]. Most recently, they strive to address limitations of their previous method by implementing a hardware-based partial bitstream relocation technique they call Accelerated Relocation Circuit [16]. This technique allows flexible module relocation, less communication overhead, and faster performance for relocation over more traditional techniques.

Other avenues of self-healing exploration for computing involve investigating materials that would facilitate self-repair at the hardware level. This broad field of research has focused on creating materials capable of self-repairing cracks and small physical damage to the structure of the material. Of the two main approaches to developing such materials, a layered surface versus the incorporation of nanocontainers or hollow fibers, the second seems to be the less problematic and more promising option [19]. Research of this nature spans the materials world with work involved with polymers, metallic and non-metallic coatings, nanocomposites, concretes, and ionomers [20]. Most recent research focuses on refining techniques and creating domain-specific solutions.

With regard to energy harvesting, one of the difficult aspects of modern technology has been that while disk capacity and processer speed has increased supra-linearly, battery technology has been relatively slow in progress. Thus, new computing architectures would significantly benefit from the varied energy harvesting technologies under investigation [21]. Several of these technologies include harvesting motion using piezoelectric materials or scavenging RF energy from the surrounding environment.

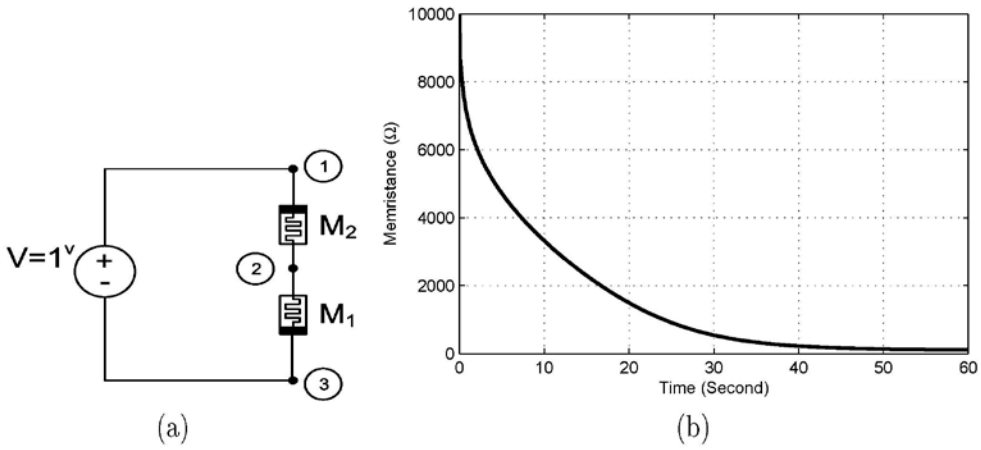

Figure 2: a) Schematic of two memristors in series tied to a voltage source. b) Memristance as a function of time for the condition when a set of memristors has a similar output curve to a biological synapse.

C. Conceptual Design and Fabrication of a Neuromorphic Memristor Device

A final task for this project was to develop a conceptual design of a neuromorphic device. We chose to investigate an array of memristor structures as these devices have the necessary attributes to provide a neuromorphic platform for silicon hardware development. Recent work in memristor technology has showed that a series of opposite polarity memristors will provide a characteristic biological synapse time-dependent output curve [22-23]. In the case of a single memristor, the output curve has the opposite curvature, indicating that the first set of voltage pulses applied to the memristor reduce the resistance by a small amount, and only after the accumulation of many pulses does the resistance drop significantly. This is the opposite situation of a biological synapse where the first set of pulses causes a large drop in the synaptic weight which then

plateaus and saturates after many pulses. Figure 2 shows the output of a series of memristors, and in this case the output curve looks similar to that of a biological synapse (large drop initially and then a plateau). This is achieved by having the condition of $\Delta M_1 / \Delta M_2 << 1$ [22].

Given this theoretical framework, our next objective is to identify a candidate fabrication process to construct memristor arrays with neuromorphic characteristics. A recent article details a method using titanium dioxide doped with oxygen in order to garner memristor characteristics [24]. Previous work by Hewlett-Packard utilized titanium dioxide with oxygen vacancies made by heating the substrate to drive oxygen out of the material. With a multilayered stacked device, such a thermal process would drive oxygen out of all the titanium dioxide layers and prevent the fabrication of more than one junction between oxygen-rich and oxygen-depleted titanium dioxide. The scheme demonstrated by Prodromakis et al. utilizes a process of native titanium dioxide and oxygen-enriched titanium dioxide using an oxygen feed into the inert atmosphere during the deposition process [24]. Thus, this method allows for multiple vertically-stacked memristors to be fabricated since the doping process is confined to each layer

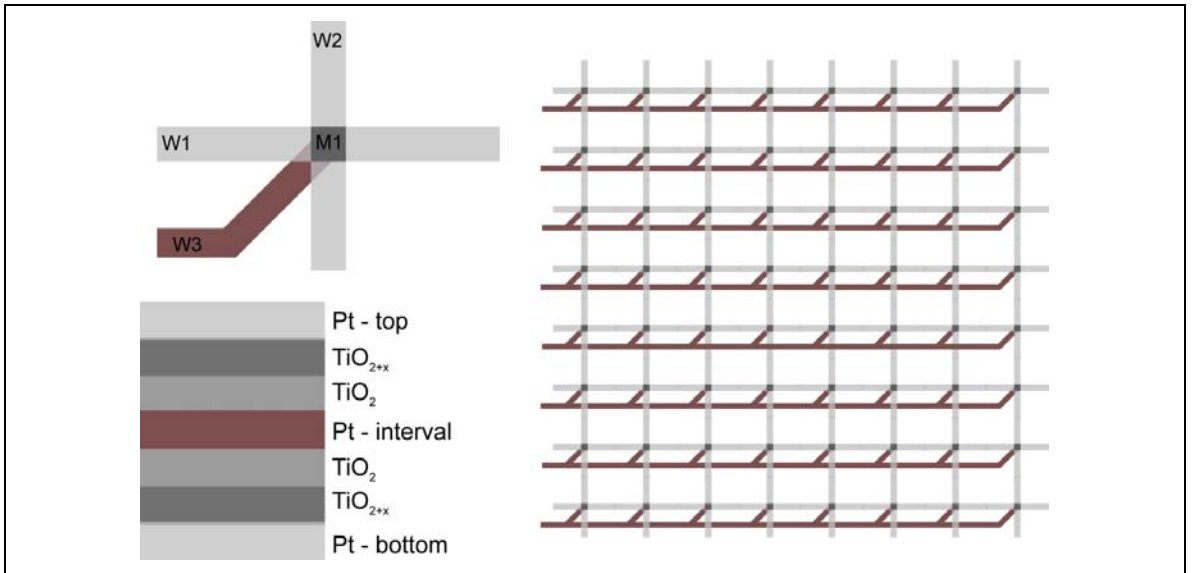

Figure 3: Top-down (left, top) and cross-section (left, bottom) of a candidate dual memristor in series device fabricated using the process described in [24]. An array of these structures (right) is also pictured.

without affecting previously deposited layers. Figure 3 shows a candidate design for a

simple array of dual-series memristor devices. The system relies on three layers of nanoelectrodes, with an interval Pt electrode layer that serves to connect the opposite polarity memristive layers in series. The polarities are opposite (doped-undoped, undoped-doped) in order to facilitate the neuromorphic characteristic of the curve shown in Figure 2. The interval Pt leads are all tied together on each row of the array, and addressing occurs through simultaneous voltages being applied at either the top/bottom Pt lead and the interval lead, similar to the scheme shown by Xia et al. [13].

D. Benchmark Application

In an effort to design and develop next generation computing strategies, a benchmark problem should be established to specifically target the desired attributes of the future computing systems. One such test case could be the next generation internet/cell mobile devices, which would conceivably address many unresolved difficulties faced by the current mobile devices that access the internet. Desirable attributes of such a system and measurements of those attributes in terms of performance might include:

- Adaptability, tunability, plasticity
    - Ability to find alternative routes
        - If some portion of the system is compromised, can redirect those functions to other portions of the system
        - Reroute traffic to minimize bandwidth usage and sluggish communication
    - Interpret signals in channels for which there is extensive crosstalk
    - Dynamic optimization of resources – e.g. power consumption, tactile sensitivity, voice analysis
- Self-preservation, self-healing, robust, recoverable
    - Able to replicate itself or portions of itself;
    - Can analyze data about itself to determine areas and means for self-improvement
    - Able to diagnose and provide self-protection from cyber attacks
        - Shut down part of the network or reconfigure the hardware connections for a device that is being infected by virus/worm
    - Recover from faults
        - Move functions to other components
        - Change objective/function of component with failure

    - Adapt to changes in user objectives
    - Self-repairing materials: repair physical damage to system; built-in materials to repair insulators and conductors
- Compact, high-density, 3D geometries
    - Be space conscious and fully utilize the 3rd dimension
    - Increased speed and bandwidth, with less crosstalk and wasted power.
- Standardizable, predictable, specifiable
    - Standardized and flexible media types
    - Coordination of wireless technologies
    - Standardized components with flexible interconnections
- Flexibly sustained
    - Can scavenge for parts or energy
        - vibrations, temperature gradient, ambient light and RF
    - Energy use during off-peak times

III. CONCLUSIONS

Current computing technology faces footprint and energy consumption barriers as hardware systems scale up to exascale computing levels. Neuromorphic computing provides a paradigm shift in the design and fabrication of computing hardware in order to reduce power consumption and footprint while maintaining computational capabilities. Several aspects of biological neural networks are important to incorporate into neuromorphic computing strategies including adaptability and high interconnectivity. Memristive technology is one example of a recently developed neuromorphic technology that has the potential for achieving biological synapse characteristics such as small footprint and reversible storage of information. An array of series memristor structures was schematically designed in order to demonstrate the fabrication and layout of a device with neuromorphic architecture. Future work will consist of fabricating such an array followed by characterization of the memristive properties of the system.

IV. REFERENCES

[1] L. Smith and A. Hamilton, *Neuromorphic systems: engineering silicon from neurobiology*: World Scientific Pub Co Inc, 1998.

[2] R. Wieland, *et al.*, "3D Integration of CMOS transistors with ICV-SLID technology," *Microelectronic Engineering*, vol. 82, pp. 529-533, Dec 2005.

[3] D. Vainbrand and R. Ginosar, "Network-on-chip architectures for neural networks," presented at the 2010 Fourth ACM/IEEE International Symposium on Networks-on-Chip, 2010.

[4] H. Markram, "The Blue Brain project," *Nature Reviews Neuroscience*, vol. 7, pp. 153-160, 2006.

[5] J. Navaridas, *et al.*, "Understanding the interconnection network of SpiNNaker," presented at the International Conference on Supercomputing, Yorktown Heights, NY, 2009.

[6] S. V. Adve, *et al.*, "Parallel computing research at Illinois: The UPCRC agenda," University of Illinois at Urbana-Champaign2008.

[7] K. Asanovic, *et al.*, "The Landscape of Parallel Computing Research: A View from Berkeley," University of California, Berkeley2006.

[8] I. Medalsy, *et al.*, "Logic implementations using a single nanoparticle-protein hybrid," *Nature Nanotechnology*, vol. 5, pp. 451-457, 2010.

[9] G. Indiveri, *et al.*, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, pp. 211-221, 2006.

[10] S. Jo, *et al.*, "Nanoscale Memristor Device as Synapse in Neuromorphic Systems," *Nano Letters*, pp. 433-475.

[11] D. Strukov, *et al.*, "The missing memristor found," *Nature*, vol. 453, pp. 80-83, 2008.

[12] J. J. Yang, *et al.*, "Memristive switching mechanism for metal/oxide/metal nanodevices," *Nat Nanotechnol*, vol. 3, pp. 429-33, Jul 2008.

[13] Q. F. Xia, *et al.*, "Memristor-CMOS Hybrid Integrated Circuits for Reconfigurable Logic," *Nano Letters*, vol. 9, pp. 3640-3645, Oct 2009.

[14] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, pp. 41-50, 2003.

[15] L. D. Paulson, "Computer System, Heal Thyself," *Computer*, vol. 35, pp. 20-22, 2002.

[16] A. Sreeramareddy, *et al.*, "Self-configurable architecture for reusable systems with accelerated relocation circuit (SCARS-ARC)," presented at the IEEE International Symposium on Parallel & Distributed Processing, Atlanta, GA, 2010.

[17] S. K. Venishetti, *et al.*, "Hierarchical built-in self-testing and FPGA based healing methodology for system-on-a-chip," presented at the NASA/ESA Conference on Adaptive Hardware and Systems, Scotland, 2007.

[18] A. Sreeramareddy, *et al.*, "SCARS: Scalable self-configurable architecture for reusable space systems," presented at the NASA/ESA Conference on Adaptive Hardware and Systems, Noordwijk, 2008.

[19] D. G. Shchukin and H. Möhwald, "Self-repairing coatings containing active nanoreservoirs," *Small,* vol. 3, pp. 926-943, 2007.

[20] S. K. Ghosh, Ed., *Self-healing materials: fundamentals, design strategies, and applications*. Weinheim: Wiley-Vch, 2009, p.^pp. Pages.

[21] P. D. Mitcheson, *et al.*, "Energy harvesting from human and machine motion for wireless electronic devices," *Proceedings of the Ieee,* vol. 96, pp. 1457-1486, Sep 2008.

[22] F. Merrikh-Bayat, *et al.*, "Bottleneck of using single memristor as a synapse and its solution," *Arxiv preprint arXiv:1008.3450,* 2010.

[23] Y. Joglekar and S. Wolf, "The elusive memristor: properties of basic electrical circuits," *European Journal of Physics,* vol. 30, pp. 661-675, 2009.

[24] T. Prodromakis, *et al.*, "Fabrication and electrical characteristics of memristors with TiO< inf> 2</inf>/TiO< inf> 2+x</inf> active layers," 2010, pp. 1520-1522.

V. DISTRIBUTION LIST

1	MS 1425	Darren W. Branch
1	MS 1188	Chris Forsythe
1	MS 1425	Conrad D. James
1	MS 1425	Steve Casalnuovo
1	MS 0899	Technical Library (electronic), 9536