Abstracts for Posters Presented at the 4th IEEE Rebooting Computing Summit

---

Adapting to Thrive in a New Economy of Memory Abundance

Kirk Bresniker and R. Stanley Williams
Hewlett Packard Labs

Processing technology has eclipsed memory technology for the past six decades, but processor-centric architectures are reaching their terminal efficiency. We can reboot computing on the basis of abundant memory enabled by emerging device physics, which will make computation, communication, and memory more efficient. This approach also provides a unique opportunity to address novel security threats with modern, systemic solutions.

Neural Computing at Sandia National Laboratories

Craig M. Vineyard, Erik Debenedictis, James B. Aimone, Michael L. Bernard, Kristofor D. Carlson, Frances S. Chance, James C. Forsythe, Conrad D. James, Fred Rothganger, William M. Severa, Ann E. Speed, Stephen J. Verzi, Christina E. Warrender, and John S. Wagner
Sandia National Laboratories

An understanding of how the brain operates in performing computations has for decades served as an inspiration to build and design computers. Physical limits in computer hardware, advances in neuroscience, and the success of artificial neural network software have led to an emergence in neural inspired computing approaches. Over the course of a decade, leveraging neural principles for computational benefit at Sandia National Laboratories has resulted in a variety of research focusing upon neural theory, modeling and simulation, and neural inspired application development.

At the intersection of computability and neuroscience understanding, neural theory research strives to yield a formal understanding of various aspects of the brains operation for computational benefit. An analysis of the tradeoffs associated with computations performed and space requirements yields insights into the low power operating regime of brains. An information theoretic analysis of neural ensembles yields insights into representation and encoding paradigms. And neural ensemble research has investigated adaptive encodings.

In addition to developing theories regarding neural computing, we have also led several modeling and simulation efforts. Leveraging the high performance computing expertise at Sandia National Laboratories we have developed and analyzed several large scale neural models, advocated uncertainty quantification and sensitivity analysis in neural models, and developed a language and tool for describing large-scale dynamical systems.

And finally, the insights gained by an increased understanding of neural theory and building modeling and simulation capabilities has allowed us to develop a variety of neural applications. These include the development of neural inspired machine learning algorithms, a neural modeling approach to decision making, neural circuit approaches for information encoding and retrial, and the development of neural inspired computer architectures.

Scaling Up of Neuromorphic Computing Systems using 3D Wafer Scale Integration

Arvind Kumar[1], Zhe Wan[2,3], Subramanian Iyer[3], and Winfried Wilcke[4]
[1]IBM TJ Watson Research Center, [2]IBM Albany Nanotech, [3]UCLA, [4]IBM Almaden Research Center

The cognitive era is just beginning, with hopes of computing machines that can solve "unstructured" computational problems. such as sensing, learning, and inferring; detecting patterns and anomalies; and predicting and discovering. These type of data-centric applications require a fundamental shift from the von Neumann architecture which has defined computing systems since the 1940s. Inspired by the brain, we propose a radically different architecture consisting of a large number of highly interconnected simple processors intertwined with very large amounts of low-latency memory. This memory-centric architecture can be realized using 3D wafer scale integration, which provides massive interconnectivity through very high bandwidth directly between processors. Combined with mature CMOS technologies, it provides a path toward early realization of a highly scaled-up neuromorphic computer. The natural fault tolerance and lower power requirements of neuromorphic processing make 3D wafer stacking particularly attractive.

Acceleration of Neural Algorithms using Nanoelectronic Resistive Memory Crossbars

Matthew J. Marinella, Sapan Agarwal, David Hughart, Steve Plimpton, Ojas Parekh, Tu-Thach Quach , Erik Debenedictis, Ron Goeke, Pat Finnegan, Derek Wilke, Denis Mamaluy, Harry Hjalmarson, Brian Tierney, Dave Henry, Alex Hsia, Brad Aimone, and Conrad James
Sandia National Laboratories

The size and "depth" of deep neural algorithms are currently limited by available hardware. It is typically not practical to run simulations that require more than one week to run, and hence the neural field is limited to problems that can be run in this length of time with a modern supercomputer (typically 50k-3M CPU/GPU cores). Although impressive results training deep networks using modern GPU clusters have recently been reported, training much larger deep networks and datasets is highly desirable. Numerous groups are making progress in the short term toward this goal, though the development of highly efficient GPU, FPGA, and ASIC cluster architectures, which will likely increase the size of these networks by as much as two orders of magnitude in the short term. For the longer term, we are exploring the use of emerging nanoelectronic resistive memory technologies, which could provide as much as eight orders of magnitude improvement over implementing the same algorithm on a modern CPU. We will report an overview and share recent results from our effort at Sandia to create a neural algorithm accelerator, which includes multidisciplinary work ranging from basic materials science, device fabrication and characterization, through the architecting, theoretical modeling, and simulation of this accelerator.

The Dot-product engine: programming memristor crossbar arrays for efficient vector-matrix multiplication

John Paul Strachan, Miao Hu, J. Joshua Yang, Emmanuelle Merced-Grafals, Noraica Davila, Catherine Graves, Eric Montgomery, R. Stanley Williams
Hewlett Packard Labs

Vector-matrix multiplication dominates the computation time and energy for many workloads, particularly neural network algorithms and linear transforms (e.g, the Discrete Fourier Transform). We developed the Dot-product Engine (DPE), an enhanced memory array that exploits the fundamental relationship between row voltage and column current, to realize an analog multiply-accumulate unit with high power efficiency and throughput. We first invented a conversion algorithm to map arbitrary matrix values appropriately to memristor conductances in a realistic crossbar array, accounting for device physics and circuit issues to reduce computational errors. Accurate device resistance programming in large arrays is enabled by closed-loop pulse tuning and access transistors. To validate our approach, we simulated and benchmarked one of the state-of-the-art neural networks for pattern recognition on the DPEs. The result shows no accuracy degradation compared to software approach (99% pattern recognition accuracy for MNIST data set) with

only 4 Bit DAC/ADC requirement, while the DPE can achieve a speed-efficiency product of 1,000x to 10,000x compared to a comparable digital ASIC.

Cortical Processing

Paul Franzon
North Carolina State University

Cortical Processing refers to the execution of emerging algorithms relying on probabilistic spatial and temporal recognition. In this work we are building processors customized towards execution of these algorithms. Examples of these algorithms include Cogent Confabulation, Hierarhical Temporal Memory, and Long Short Term Memory. Customization features include Processor in Memory and functional accelerators. Improvements in performance/power of up to 10^5 have been demonstrated over GPUs.

Low-Power Image Recognition Challenge (LPIRC)

Yung-Hsiang Lu
Purdue University

Many mobile systems (smartphones, electronic glass, autonomous robots) can capture images. These systems use batteries and energy conservation is essential. This challenge aims to discover the best technology in both image recognition and energy conservation. Winners will be evaluated based on both high recognition accuracy and low power usage. Image recognition involves many tasks. This challenge focuses on object detection, a basic routine in many recognition approaches. The first LPIRC was held in June 2015 and the top two winners presented their solutions in the International Conference on Computer Aided Design in November 2015. The second LPIRC is planned for June 2016 in Austin Texas.

Cryogenic Computing Complexity (C3) Program

Marc Manheimer
IARPA

The ultimate goal of the Intelligence Advanced Research Projects Activity (IARPA)'s Cryogenic Computing Complexity (C3) program is to demonstrate a complete superconducting computer including processing units and cryogenic memory. IARPA expects that the C3 program will be a five-year two-phase program. Phase one, which encompasses the first three years, primarily serves to develop the technologies that are required to separately demonstrate a small superconducting processor and memory units. Phase two, which is for the final two years, will integrate those new technologies into a small-scale working model of a superconducting computer. Program goals are presented, and the approaches of the phase-one teams are reviewed.

Superconducting Computing in Large-Scale Hybrid Systems

Alan M. Kadin[1], D. Scott Holmes[2], and Mark W. Johnson[3]
[1]Consultant for Hypres, Inc.; [2]Booz-Allen Hamilton; [3]D-Wave Systems, Inc.

The past, present, and future of superconducting computing will be discussed, based on the feature article in the December issue of IEEE Computer Magazine. Specific systems addressed will include processors for digital radio receivers, quantum annealing, neural simulators, and ultra-low-power adiabatic computing.

Energy Recovery and Recycling in Computation: Reversible Adiabatic Logic

Gregory L. Snider[1], Ismo K. Hänninen[1], César O. Campos-Aguillón[1], Rene Celis-Cordova[2], Alexei Orlov, and Craig S. Lent
[1]*University of Notre Dame,* [2]*Tecnológico de Monterrey, Mexico*

Energy use in computation has become the dominant challenge in the design of computational logic and systems. Here energy is dissipated to heat in two ways, by passive dissipation due to leakage, and by active dissipation caused by the processing of information. As supply voltages are lowered to reduce the active dissipation, the passive dissipation increases, so resent research has concentrated on reducing the passive dissipation. Even if passive dissipation is eliminated, active dissipation in conventional computation will still set a lower limit on total dissipation, limiting future progress.

Recent experiments testing the Landauer principle have shown that, as predicted, there is a minimum limit of dissipation, $k_BT \ln(2)$, if information is destroyed, and that dissipation can be less than $k_BT \ln(2)$, with no lower limit, if information is not destroyed. Since these experiments have shown that ultra-low energy dissipation is possible, the question becomes how to extend these results to real computing systems. One approach is to use reversible logic implemented with adiabatic circuits to avoid information destruction, so that energy can be recovered and loss is minimized in state transitions. In such a system the energy needed to process information is sent to the logic by power clocks, and then returned from the logic when the computation is complete. To achieve overall energy savings the energy returned must be recycled and reused in the next computation, rather that dissipated to heat in the clock generator.

This poster presents reversible adiabatic circuits designed using adiabatic CMOS as a test bed. As an existing technology, adiabatic CMOS can be used to evaluate the performance and active power dissipation of circuits. Simple test circuits and a simple reversible adiabatic microprocessor will be presented. To recycle the energy used in computation, MEMS resonators are proposed as clock circuits. Molecular quantum-dot cellular automata (QCA) is presented as a beyond-CMOS paradigm that maps well onto reversible adiabatic computational systems.

Improving Energy Efficiency via Nonlinear Dynamics and Chaos

Erik P. DeBenedictis[1], Neal G. Anderson[2], Michael P. Frank[1], Natesh Ganesh[2], R. Stanley Williams[3]
[1]*Sandia National Laboratories,* [2]*University of Massachusetts Amherst,* [3]*Hewlett Packard Labs*

The Boolean logic abstraction offers intellectual elegance and reduces design effort, but may also limit energy efficiency. This poster gives one example where a new circuit based on a new MeRAM device theoretically improves energy efficiency by severl orders of magnitude over accepted projections of Boolean logic gates. A route to improved energy efficiency was demonstrated for a simple "learning machine," but generalization to other problems is beyond the scope of this poster.

Revealing Fundamental Efficiency Limits for Complex Computing Structures: The FUELCOST Methodology

Neal G. Anderson, Ilke Ercan*, and Natesh Ganesh
University of Massachusetts Amherst

The energy efficiency of computation doubled every ~1.57 years from the dawn of digital computation to around 2010 (Koomey's Law), after which progress has slowed. Restoration of exponential efficiency scaling over the long term will likely be achievable only through the development of new computing technologies based on unconventional computing strategies and paradigms. Given the major investment that will be required to develop any new computing paradigm, and the critical importance of energy

efficiency, evaluation of alternative computing paradigms should include limiting efficiency as an integral component.

In this poster, we describe an evolving physical-information-theoretic methodology—the FUELCOST methodology—that enables determination of the FUndamental Efficiency Limits of complex COmputing STructures. This methodology is based on a fundamental physical description of the *dynamics of information* as it is processed by computing hardware, as opposed to a physics-based description of the *dynamics of computing hardware* as it processes information (e.g. as in standard models and simulations). This enables isolation of fundamental sources of inefficiency that are deeply rooted in physical law and incurred at different levels of abstraction in complex computing systems. We discuss the underlying theoretical approach; previous studies of various computing structures (finite-state automata, simple processor architecture), logic blocks and functions (ALUs, decoders, adders), and nanocircuit implementations (both transistor- and non-transistor-based); progress toward full synthesis, integration, and automation of the multi-level evaluation methodology; and exploratory application directions (digital/discrete-analog neuromorphic, approximate, and Brownian approaches).

* Present Address: Boğaziçi University