Research Statement

Richard Dorrance
Electrical Engineering Department
University of California, Los Angles
rdorrance@ucla.edu

November 15, 2014

My research interests cover a broad swath of engineering disciplines, including post-CMOS device development and integration; high-performance, energy-efficient computing architectures; neuromorphic engineering; and high-performance computing for bioinformatics. I frequently collaborate with others, having worked closely with other students and professors at UCLA, UC Irvine, the University of Minnesota, Stanford, and KU Leuven during my time as a graduate student at UCLA.

As a master's student, my early research focused on the design and integration of post-CMOS devices into VLSI computing architectures (most notably spintronic memories). Gradually, this research morphed into spin-based, non-volatile logic, a completely new paradigm in the field of computer architecture. At the same time, my exposure to neuroengineering and biomedical research, through various collaborative projects, broadened the scope of my research.

As a Ph.D. student, I began to collaborate with my colleagues in the physics department at UCLA to develop a spin-based, artificial neuron for neuromorphic computing applications [1]. At the same time, my dissertation work focuses on the development of a scalable VLSI architecture for improving the performance and energy-efficiency of sparse linear algebra—a computational kernel that arises in a wide variety of computational disciplines, including medical imaging, compressive sensing, neural networks, applied mathematics, and bioinformatics.

In the following sections, I will describe the substance and impact of my past and present research, including my master's thesis work and the pursuits of my Ph.D. dissertation. I conclude with an overview of my future research agenda, to which I will bring the same enthusiasm and curiosity that has driven me to pursue my Ph.D. degree.

Circuit Design and Optimization of STT-MRAM

My early graduate work at UCLA focused on the development of spin-transfer torque magnetic random-access memory (STT-MRAM) as a "universal memory" replacement for the current trio of electronic memory technologies—SRAM for speed, DRAM for density, and Flash for non-volatility. Each of these memory technologies face significant challenges in scaling below the 22-nm node as CMOS begins to hit the atomic limits of 2D device scaling. STT-MRAM uses the properties of electron spin to store digital information in a magnetic tunnel junction (MTJ) device. Fabricated in the higher metal layers, STT-MRAM is well suited for 3D device integration, pushing the memory density limits beyond that of traditional CMOS.

The goal of my master's thesis was to tackle the two major issues preventing the widespread adoption and commercialization of STT-MRAM: the poor reading performance and poor yield of the MTJ bit-cell. As such, our research efforts focused on developing a comprehensive circuit-device co-design methodology. Extensive device modeling allowed us to map the design space, explore the feasibility of deploying

multiple MTJs per access transistor for improved memory density, and conduct several benchmarking studies to inform the future direction of device development and materials optimization [2]. At the same time, we developed and patented a reading circuit that can effectively and reliably enable the ultra-high-speed reading of STT-MRAM at GHz frequencies using a short pulse of sensing current [3]. Coupled with a systematic calibration methodology for optimizing the reference voltage of local sensing circuits in the presence of device variations, the body-voltage sensing circuit was able to successfully demonstrate low power operation with minimal bit error rate. Funding for the work was provided by the DARPA STT-RAM Program (HR0011-09-C-0114).

Non-volatile Logic for Ultra-Low-Power Electronics

Building upon our previous success with spintronic memories, the goal of the DARPA Non-Volatile Logic (NVL) Program (HR001-10-C-0153) was to address the issues plaguing the continued scaling of CMOS for high-performance and low-power applications: unsustainable power densities, poor device matching, lithography limitations, and production costs. Our proposal aimed to solve the fundamental issues of scaling by moving to three dimensions. This was done in two stages.

The first stage of the project focused on augmenting CMOS with MTJs to create hybrid spintronic-CMOS logic devices with the recently discovered voltage-controlled magnetic anisotropy (VCMA) effect. We pursued a 2.5D logic-in-memory architecture using our previously developed circuit-device co-design methodology: stacking MTJ memory layers over CMOS logic layers. Using VCMA-based MTJ devices, we developed and patented several novel memory circuits, including a read-disturbance-free non-volatile content addressable memory (CAM) [4] and a high-speed magneto-electric random access memory (MeRAM) [5].

This work led to the discovery of magnetic devices based on the giant Spin-Hall Effect (SHE) that can operate with power supplies as low as 100 mV—and showed excellent promise towards the development of a single electron transistor. This formed the foundation of the second phase of the project: the creation of spin-wave logic devices to encode digital information into the phase of a resting spin-wave. The basic building blocks—voltage-to-spin wave and wave-to-voltage converters, spin waveguides, spin modulators, and the magnetoelectric cell—enable reconfigurable and 3D stackable digital logic without the use of CMOS transistors.

A Spin-Based Artificial Neuron and Neural Spike Sorting

In early 2013, during the development of the 3-terminal SHE-based MTJs, we realized that thresholding and probabilistic switching behavior of the devices very closely resembled the stochastic spiking resonance of a biological neuron. The core of the spin-based, artificial neuron consists of a highly resistive MTJ stacked on top of a metallic wire with large SHE. Switching of the MTJ device is accomplished by applying current though the wire over a certain magnitude, with the direction of the current determining the final state of the device. This perfectly emulates how biological neurons pass messages between themselves via electrical signals from the axon to excitatory and inhibitory dendrites. The spin-based, artificial neuron has been patented [1] and was also recently presented at the Qualcomm Research Center in Bridgewater, New Jersey, where Hochul Lee and I were finalists for the 2014 Qualcomm Innovation Fellowship.

Prior to this work, several members of my lab had been working on a series of projects involving the recording and analysis of neural spikes. They developed a series of implantable ASICs for the real-time detection, alignment, and clustering of neural spikes which have been tested in vivo in rats. However, one of the major problems my colleagues have faced is the deluge of data, with a single neural recording

generating 1.2 terabytes of data a week for 64 recording channels. These recorded neural spikes are usually very sparse in the time domain, but require a tremendous amount of memory bandwidth to be processed in real-time. My own dissertation work focuses on improving the memory utilization and energy efficiency of algorithms that rely on sparse linear algebra. This is why I have recently become involved in a project, in collaboration with Stanford, to develop an FPGA platform to accelerate the detection and analysis of retinal neural recordings with several hundred simultaneous channels. In the future, we hope to enable real-time detection of retinal neural spikes to create a closed-loop system able to provide therapeutic stimulation to help in the treatment of degenerative retinal diseases.

Scalable VLSI Architectures for Bioinformatics

As a continuation of my prior research, the focus of my dissertation work has also been low-power, high-performance computing in the post-CMOS era—in particular, the integration of spintronic memories to create new architectures for high-performance data processing to address the issue of poor computational throughput of sparse linear algebra. Why sparse linear algebra? Sparse linear algebra arises in a wide variety of computational disciplines beyond the aforementioned problem of neural spike sorting. These fields include, and are not limited to: medical imaging, 3D graphics, compressive sensing, neural networks, applied mathematics, various optimization problems, and bioinformatics. As a memory-bound mathematical kernel, the performance of the sparse basic linear algebra subroutines (sparse-BLAS) has always significantly lagged behind that of their dense counterparts, by 2 to 3 orders of magnitude, due to a fundamental mismatch between the compression formats required to efficiently store sparse matrices and traditional Von Neumann computing architectures.

The Von Neumann architecture is a byproduct of 1970s-era when hardware for digital circuit design, particularly memory, was very expensive. Even now—despite the availability of high-capacity, dense, power-efficient memories—each processor core in a modern CPU or GPU is very compact, usually containing only a handful of general purpose registers and a single floating-point unit (FPU). Under this kind of an architecture, the only viable option for computing sparse-BLAS is to use row-major operations due to the limited number of registers. For a sparse matrix-vector multiplication, the CPU/GPU is forced to load each element of the compressed matrix from main memory one at a time, uncompress it, and then load the matching vector element from main memory before the multiplication in the FPU can occur. Memory accesses are very irregular and incur large unmaskable latencies due to cache misses, leading to poor computational performance and energy efficiency.

Our solution to this problem has been to reorder the sparse-BLAS operations on an algorithmic level to support very large, vectorized register banks to enable continuous data streaming. Instead of performing row-major operations, these vectorized register banks allow us to perform column-major operations. Fundamentally, the column-major operations can be thought of as simultaneously performing multiple row-wise dot products with out-of-order execution to optimize the memory access profile via sequential memory accesses. This also allows for a very simple processing element, with each containing a single floating-point adder and multiplier, as well as a simple dual-port RAM. Potential data hazards due to the non-zero latency of the floating-point adder are dealt with by a simple circuit to reorder the incoming data stream. Our sparse-BLAS coprocessor is able to achieve up to a 300 times improvement in computational efficiency over state-of-the-art CPUs and GPUs [6]. Additionally, an FPGA implementation of the coprocessor shows a 50 times improvement in energy efficiency for various bioinformatics algorithms in a high-performance computing environment [7].

We are currently in the process of taping-out the sparse-BLAS coprocessor in a 40nm CMOS technology. Preliminary estimates for a single processing element show that it would consume only 2mW of power at 200MHz for 0.09mm^2 of area. This is 60 times more energy efficient than our FPGA demonstration of the

sparse linear algebra kernel, and almost 3,000 times more energy efficient than state-of-the-art CPUs and GPUs. In the future, we also plan to replace the dual-port SRAMs that make up the vectorized register banks with non-volatile, high-speed spintronic memories to eliminate leakage power.

Future Agenda

In my future research, I plan to focus on two issues facing the future of computing. The first is the development of economical embedded sensing and computing systems for widespread healthcare services to exploit the ubiquitous nature of mobile devices in our everyday lives. The second is to innovate VLSI architectures to support energy-efficient and high-performance big data analysis in server systems for bioinformatics, personalized medicine, and related areas to tackle the exponential growth in data. To achieve these goals, we will need to look to new device technologies and architectures to overcome the challenges of CMOS scaling and the limitations of the antiquated Von Neumann style of thinking. I am very enthusiastic about pursuing these projects in an academic setting where opportunities for collaborations are abundant and essential.

The future of healthcare lies in our ability to continuously monitor the vitals of patients—both in clinical and everyday environments. To support this, we must develop a network of robust, low-power electronics that are also very economical. These data aggregators will likely be battery powered and part of a close-knit, wireless sensor network. This is where post-CMOS devices will be key. My prior work with spintronic memories and devices has shown that they have the potential to be several orders of magnitude more energy-efficient that CMOS, due to their non-volatility and improved memory density.

To support energy-efficient, high-performance computing, we need to first analyze the performance and execution characteristics of a variety of different classes of algorithms to identify the key structural weaknesses in existing architectures. For example, in the case of sparse-BLAS on CPUs and GPUs, the organization and size of the memory at various levels of hierarchy was the major bottleneck. Once identified, we can build new, specialized processor cores, or adapt the architecture of existing SoCs, to improve performance and energy-efficiency. My dissertation work has already done this for sparse algorithms—a class of algorithms which heavily rely upon sparse-BLAS. Furthermore, it is simply not enough to have efficient hardware: *we need efficient hardware that is also easy to program*. At UCLA, I have also been developing the soft IP-core of our sparse-BLAS kernel to act as a coprocessor for a Xilinix MicroBlaze processor. With the appropriate low-level C driver and compiler, a programmer will be able to use the standard programming interface for the sparse-BLAS routines and not have to be aware of the coprocessor to leverage it. The ability to support backwards compatibility and significantly accelerate existing code banks is the level of programmability I strive to meet.

Under the direction of my advisor, I have learned how to write a successful grant proposal that attracts interest from other colleagues, departments, and organization within my own university, as well as other academic and industry institutions. Having these collaborators makes a proposal more competitive and interesting to the grant selection committee. I am also very interested in assisting with on-going projects to help build relationships with future collaborators. Collaboration should be synergistic in nature to allow imaginative people to bring out the best in each other and together create something that none of them could achieve alone.

References

[1] R. Dorrance, H. Lee, S. Basir-Kazeruni, P. K. Amiri, D. Marković, K. L. Wang, "An Artificial Neuron based on the Spin-Hall Effect," *US Patent*, University of California Case No. 2014-925-1, Nov. 2014.

[2] R. Dorrance, F. Ren, Y. Toriyama, A.A. Hafez, C.-K.K. Yang, D. Marković, "Scalability and Design-Space Analysis of a 1T-1MTJ Memory Cell for STT-RAM," *IEEE Trans. Electron Devices (TED)*, vol. 59, pp. 878-887, no. 4, Apr. 2012.

[3] F. Ren, K. L. Wang, C.-K. Yang, D. Marković, "Body Voltage Sensing Based Short Pulse Reading Circuit," *International Patent*, WO2013043738 A1, Mar. 2013.

[4] R. Dorrance, P. K. Amiri, D. Marković, K. L. Wang, "Read-Disturbance-Free Nonvolatile Content Addressable Memory (CAM)," *US Patent*, US 20140071728 A1, Mar. 2014.

[5] R. Dorrance, P. K. Amiri, D. Marković, K. L. Wang, "Nonvolatile Magneto-Electric Random Access Memory Circuit with Burst Writing and Back-to-Back Reads," *US Patent*, US 20140071732 A1, Mar. 2014.

[6] R. Dorrance, F. Ren and D. Marković, "A Scalable Sparse Matrix-Vector Multiplication (SpMxV) Kernel for Sparse-BLAS on FPGAs," in *Proc. ACM/SIGDA Int. Symp. Field-programmable Gate Arrays (FPGA'14)*, pp. 161-170, Feb. 2014.

[7] R. Dorrance and D. Marković, "A Scalable Sparse Linear Algebra Kernel for Accelerating Bioinformatics on FPGAs," *IEEE/ACM Trans. Comput. Biol. Bioinf. (CBB)*, under review.