Department of Electrical and Computer Engineering

University of Rochester, Rochester, NY

Ph.D. Public Defense

Thursday, June 16, 2016

2:00 PM

Computer Studies Building, Room 523

Memory System Optimizations for Energy and Bandwidth Efficient Data Movement

Mahdi Nazm Bojnordi

Supervised by
Professor Engin Ipek

Abstract

Since the early 2000s, power dissipation and memory bandwidth have been two of the most critical challenges that limit the performance of computer systems, from data centers to smartphones and wearable devices. Data movement between the processor cores and the storage elements of the memory hierarchy (including the register file cache levels, and main memory) is the primary contributor to power dissipation in modern microprocessors. As a result, energy and bandwidth efficiency of the memory hierarchy is of paramount importance to designing high performance and energy-efficient computer systems.

This research explores a new class of energy-efficient computer architectures that aim at minimizing data movement, and improving memory bandwidth efficiency. We investigate the design of domain specific ISAs and hardware/software interfaces, develop physical structures and microarchitectures for energy efficient memory arrays, and explore novel architectural techniques for leveraging emerging memory technologies (*e.g.*, Resistive RAM) in energy efficient memory-centric accelerators.

This dissertation first presents a novel, energy-efficient data exchange mechanism using synchronized counters. The key idea is to represent information by the delay between two consecutive pulses on a set of wires connecting the data arrays to the cache controller. This time-based data representation makes the number of state transitions on the interconnect independent of the bit patterns, and significantly lowers the activity factor on the interconnect. Unlike the case of conventional parallel or serial data communication, however, the transmission time of the proposed technique grows exponentially with the number of bits in each transmitted value. This problem is addressed by limiting the data blocks to a small number of bits to avoid a significant performance loss. A viable hardware implementation of the proposed mechanism is presented that incurs negligible area and delay overheads.

The dissertation then examines the first fully programmable DDRx controller that enables application specific optimizations for energy and bandwidth efficient data movement between the processor and main memory. DRAM controllers employ sophisticated address mapping, command scheduling, and power management optimizations to alleviate the adverse effects of DRAM timing and resource constraints on system performance. These optimizations must satisfy different system requirements, which complicates memory controller design. A promising way of improving the versatility and energy efficiency of these controllers is to make them programmable—a proven technique that has seen wide use in other control tasks ranging from DMA scheduling to NAND Flash and directory control. Unfortunately, the stringent latency and throughput requirements of modern DDRx devices have rendered such programmability largely impractical, confining DDRx controllers to fixed-function hardware. The proposed programmable controller employs domain specific ISAs with associative search instructions, and carefully partitions tasks between specialized hardware and firmware to meet all the requirements for high performance DRAM management.

Finally, this dissertation presents the memristive Boltzmann machine, a novel hardware accelerator that leverages *in situ* computation with RRAM technology to eliminate unnecessary data movement on combinatorial optimization and deep learning workloads. The Boltzmann machine is a massively parallel computational model capable of solving a broad class of combinatorial optimization problems and training deep machine learning models on massive datasets. Regrettably, the required all-to-all communication among the processing units limits the performance of the Boltzmann machine on conventional memory architectures. The proposed accelerator exploits the electrical properties of RRAM to realize *in situ*, fine-grained parallel computation within the memory arrays, thereby eliminating the need for exchanging data between the memory cells and the computational units. Two classical optimization problems, graph partitioning and boolean satisfiability, and a deep belief network application are mapped onto the proposed hardware.