Department of Electrical and Computer Engineering
University of Rochester, Rochester, NY
Ph.D. Public Defense

Thursday January 7, 2016
2:00 PM
Computer Studies Building, Room 426

Energy Efficient Data Movement with Sparse Representation

Yanwei Song

Supervised by
Professor Engin Ipek

Abstract

Energy efficiency is one of the most significant requirements in the study of computer systems, from mobile devices to large-scale data centers. Data movement is responsible for a significant portion of the total energy in today's computer systems. Therefore, this research focuses on improving the energy efficiency of data movement. Sparse representation is a general data representation with more data bits compared to the binary representation; however, it exhibits advantageous traits that can be exploited to improve energy efficiency at the cost of greater bandwidth demand. This dissertation works on the application of sparse representation to the data movement problem by striking a balance between the energy benefit and the bandwidth cost of sparse representation.

This dissertation first characterizes the interconnects used for on-chip data movement, and surveys data communication and energy-efficient codes to provide background knowledge.

The dissertation then studies the data movement over long on-chip interconnects, a major contributor to on-chip system energy. It presents novel signaling and encoding techniques that together improve the energy efficiency of data communication between the processor cores and the last level cache. The proposed techniques make the interconnect energy proportional to the number of ones in the transferred data block (*i.e.*, the block's hamming weight), regardless of the previous state of the interconnect. The hamming weight of each cache block is kept low through a sparse data encoding approach to minimize the interconnect energy.

Finally, this dissertation focuses on the data movement over the long and highly capacitive o↵-chip interconnects. DDRx, the most broadly adopted family of DRAM interfaces, contributes significantly to the overall system energy in a wide range of computer systems. To reduce the energy cost of data transfers, DDR4 adopts a pseudo open-drain IO circuit that consumes power only when transmitting or receiving a **0**, which makes the IO energy proportional to the number of **0**s transferred over the data bus. A data bus invert (DBI) coding technique is therefore supported by the DDR4 standard to encode each byte with a fewer number of **0**s. Although sparse coding techniques that are more advanced than DBI can reduce the IO power further, the relatively high bandwidth overhead of these codes has heretofore prevented their application to the DDRx bus. This dissertation examines MiL (More is Less), a novel data communication framework built on top of DDR4, which exploits the data bus under-utilization caused by DRAM timing constraints to selectively apply sparse codes, thereby reducing the IO energy without compromising system performance.