Name: Jiangjiang Liu
E-mail address: jliu3@cse.buffalo.edu

Dissertation Proposal Title:
INFORMATION PATTERN AWARE DESIGN STRATEGIES
FOR NANOMETER-SCALE ADDRESS BUSES

Committee:
Chair: Nihar Mahapatra
Members: Chunming Qiao
Shambhu Upadhyaya

ABSTRACT

The growing disparity in processor and memory performance has forced designers to allocate an increasing fraction of die area to *communication* (I/O buffers, pads, pins, on- and off-chip buses) and *storage* (registers, caches, main memory) *components* of the memory system to enable it to provide low-latency and high-bandwidth access to large amounts of *information* (addresses, instructions, and data). Consequently, the memory system has become critical to system performance, power consumption, and cost.

In this project, we consider the following three types of redundancies related to information communicated and stored in the memory system, with the main focus being on information communicated on nanometer-scale memory system buses. *Temporal redundancy* refers to the fact that there are time periods when memory system components carry no or non-performance-critical information (e.g., idle buses and invalid, stale, and "dead" blocks in caches). *Spatial redundancy* means redundancy in the number of bits used to represent information, which causes more resources (e.g., bus lines or memory cells) to be engaged than necessary—information compression techniques address this. Finally, *energy redundancy*

implies expending more than the required energy to communicate or store information; encoding schemes attempt to minimize this redundancy via energy-efficient information representations.

To take advantage of these redundancies, we are analyzing and designing *information pattern aware memory (IPAM)* systems that employ statistical and data mining techniques to exploit various patterns in information communicated and stored in a multi-level memory hierarchy to derive gains in performance, power consumption, and cost. We consider address, instruction, and data information, all types of communication and storage components at different memory levels, and various target system (embedded, desktop, server) and application (e.g., DSP, multimedia, integer-intensive, scientific) scenarios.

We have made significant strides in the analysis and design of IPAM systems in the context of real-world benchmark suites such as SPEC CPU2000 and using execution-driven simulators like Sun Microsystems' Shade and SimpleScalar. We have completed the most comprehensive analysis to date of the potential benefits that address, instruction, and data compression may yield at all levels of the memory system considering a wide variety of factors. We are in the process of developing a technique called *hardware-only compression (HOC)*, in which narrow bus widths are used for underutilized buses to reduce cost, novel encoding schemes are employed to reduce power consumption, and concatenation and other methods are applied to mitigate performance penalty.

In future research, we propose to exploit spatial and energy redundancy of information transmitted on memory system buses for performance, power, and cost improvements. In this

context, we are developing sophisticated address compression methods and associated hardware designs to take advantage of spatial redundancy by representing and transmitting address information using fewer bits on a normal-width bus, which we refer to as *information-only compression (IOC)*. This method exploits temporal locality, spatial locality, first order temporal context, and cluster patterns, both statically and dynamically, to transmit address information using significantly fewer bits. Using fewer bits to transmit address information potentially results in lower power consumption. By combining address compression with address packing, multiple compressed addresses are transmitted simultaneously on a normal-width bus, which improves performance. We are also considering *information-cum-hardware compression (IHC)* to obtain the benefits of both approaches.