Extremely Lightweight Root-cause Memory Bottleneck Analysis in Modern Parallel Architectures

Modern architectures employ deep memory hierarchies to bridge the speed gap between the CPU and memory. Usually, frequent data movement in the memory hierarchy not only incurs high latency but also increases energy consumption. There is a significant amount of work focusing on optimizing memory performance for programs. However, most of the work involves heavyweight memory simulation, preventing the analysis from being applied (1) in efficient online tuning systems, (2) to real applications running with real data sets and real scale on real architectures, and (3) to understand time-sensitive memory contention between threads or processes in the same or different programs.

Intellectual Merit: This project will develop new techniques for memory system performance analysis suitable for online introspection of program executions. These techniques will enable users to establish an upper limit on their execution overhead, e.g., 10%, and require only modest space. The lightweight techniques that we envision for online introspection of memory system performance will enable (1) establishment of efficient online tuning systems for full-scale executions, (2) measurement of production runs of data-intensive applications, and (3) understanding of time-sensitive memory contention when multiple threads or applications co-exist in the system. The proposed research consists of three parts:

- Lightweight Memory Analysis to Accerlerate Application Execution: For many programs, one can reduce average memory latency by staging data into caches and accessing it thoroughly before it is evicted. Access patterns that do so are said to have excellent data locality. The lightweight memory analysis not only pinpoints poor locality in a program's code but also identifies the *root causes* of the poor locality. This information can be directly used by JIT compilers or auto-tuning systems to dynamically generate efficient code.

- Lightweight Memory Analysis to Improve System Throughput: Memory hierarchies in today's systems are shared between cores. Threads from one or more applications can easily contend for memory hierarchies, leading to poor system throughput, e.g., system-wide floating point operations per second (FLOPS). The lightweight memory analysis will identify the *root causes* of such contention and provide deep insights to avoid the contention with efficient online optimizations.

- Lightweight Memory Analysis Support in the Future System: The PI will study the lightweight memory analysis support in different architectures and investigate necessary but missing features for more powerful analysis. Moreover, the PI will study the accuracy and overhead of the lightweight memory analysis and understand the desired support in future compilers, operating systems, and architectures.

Broader Impacts: This research will provide a fundamental understanding of lightweight root-cause analysis of memory bottlenecks in modern parallel architectures, such as its functionality, accuracy, overhead, and necessary but missing features from the system and hardware support. The direct result, performance improvement, not only boosts productivity but also saves machine energy costs. Moreover, this research will yield benefits for the fields of program debugging, performance modeling, and hardware design. The PI will release open-source tools developed as part of the proposed research to benefit the community. An integral view of the proposed research is its integration with teaching undergraduate and graduate courses as well as student mentoring.