CSE530 Project Proposal
Cache Hierarchy Design for Bandwidth-Bounded Applications

Cong Xu and Jishen Zhao
{czx102, juz138}@psu.edu

1 Project Description

One of the challenges for modern computing systems is the growing bandwidth gap between processor cores and off-chip main memory. Emergy memory technologies bring in potential opportunities to fill this gap. The cache hierarchy design, however, need to be carefully designed in order to bridge the this gap with low performance and power overhead.

In this project, we would like to explore the optimal cache hierarchy design with various memory technologies specifically for bandwidth-bounded applications. We will first estimate the performance of an application on a baseline system, in terms of the bandwidth demanded by the application for a range of cache capacities and the bandwidth that can be provided by the system with such cache capacities. We would like to find out whether the bandwidth provided by the cache hierarchy can satisfy the demand. If it is not satisfied, we can modify the cache hierarchy, for example, by changing the capacity of existing caches, or by adding extra levels of caches. Based on our exploration, we would like to find a cache hierarchy design solution optimized for bandwidth, including the number of cache levels, as well as the memory technology, the bandwidth, and the capacity of each cache level. In addition, we will enhance the cache hierarchy design from the energy-efficiency point of view by constraining the total power consumption of the caches.

2 Motivation

Many of modern computing systems are designed to perform well on various applications to achieve high performance by exploiting their inherent parallelism.

Such systems support large number of threads and single instruction multiple data (SIMD) execution, which puts a lot of pressure on the memory system. Memory latency is typically not a bottleneck, since the latency can be hidden via multithreading or hardware prefetching. However, bandwidth becomes a potential bottleneck. A high rate of computing often brings in a high rate of data transitions. In some cases, the working set of an application fits in the on-die caches, which can typically provide sufficient bandwidth to keep up with the processing cores. However, if the working set does not fit in the on-die caches, the main memory needs to provide much of the data. Since the bandwidth of off-chip main memory is quite limited, applications with such working sets are potentially bandwidth-bound. Therefore, it is crucial to design the cache hierarchy to overcome the bandwidth limitation.

It is known that performance improvement of a computing system can be achieved via multiple cache levels. Adding extra levels of caches, however, can also help alleviate the bandwidth bottleneck of off-chip memory. In addition, emerging memory technologies such as Magnetoresistive random access memory (MRAM), Phase-change memory (PCM), memristor, etc., have shown potential to to be used as on-chip caches [5]. Therefore, we would like to explore how to enhance the cache hierarchy from the bandwidth point of view. Specifically, we would like to examine (1) the number of levels in the optimal cache hierarchy, (2) the appropriate memory technology of each level, and (3) the capacity and bandwidth of each level. We will also explore the energy-efficiency of the cache hierarchy design constrain the cache hierarchy design with energy efficiency with a fixed power budge in our design.

3 Experimental Setup

We will use our modified CACTI [1] to estimate the capacitance and bandwidth provided by each cache configuration with different memory technologies. We will use Simics [2] as the simulator to evaluate our design method. The benchmarks will be selected from bandwidth-bound applications, which possibly include benchmarks from SPEC CPU2006 [4] and NPB [3].

References

[1] HP Labs. CACTI, http://www.hpl.hp.com/research/cacti/. 2010.

[2] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner. Simics: a full system simulation platform. *IEEE Transactions on Computer*, 35(2):50–58, 2002.

[3] NASA. NPB, http://www.nas.nasa.gov/resources/software/npb.html.

[4] SPEC. SPEC CPU2006, http://www.spec.org/cpu2006/.

[5] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In *Proceedings of the International Conference on High-Performance Computer Architecture*, pages 239–249, 2009.