CSE530 Project Mid-Point Progress Report
Bandwidth-Aware Memory Hierarchy Design with Hybrid Memory Technologies

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

1 Project Description

One of the challenges for modern chip-multiprocessor (CMP) design is the growing bandwidth gap between processor cores and off-chip main memory. Emerging memory technologies bring in potential opportunities to fill this gap. The memory hierarchy, 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 memory 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 memory capacities and the bandwidth that can be provided by the system with such memory capacities. We would like to find out whether the bandwidth provided by the memory hierarchy can satisfy the demand. If it is not satisfied, we can modify the memory hierarchy, for example, by changing the capacity of existing cache and memories, or by adding an extra level of cache or memory. Based on our exploration, we would like to find a memory hierarchy design solution optimized for bandwidth, including the number of cache and memory levels, as well as the memory technology, the bandwidth, and the capacity used for each level. In addition, we will enhance the memory hierarchy design from the energy-efficiency point of view by constraining the total power consumption of the memory system.

2 Motivation

Many of modern chip-multiprocessors 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 multi-threading or hardware prefetching. To this end, 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 memory hierarchy to overcome the bandwidth limitation.

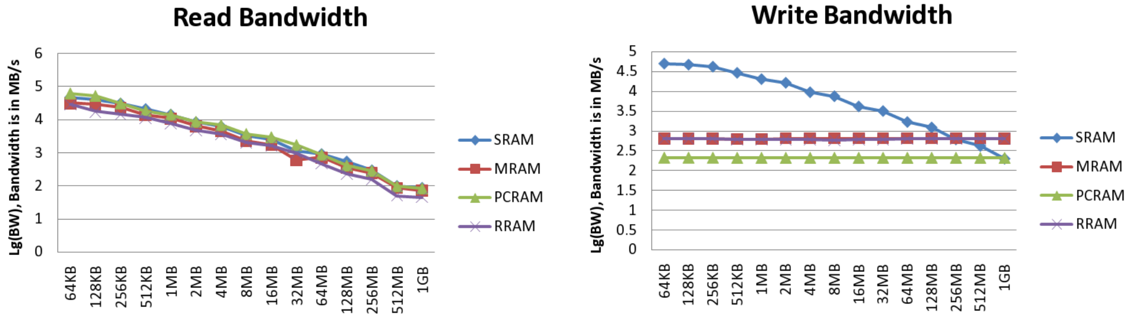
It is known that performance improvement of a computing system can be achieved via multiple memory levels. Adding an extra level of memory, however, can also help alleviate the bandwidth bottleneck of off-chip memory. In addition, emerging memory technologies such as Magnetoresistive Random Access Memory (MRAM), Phase-Change Memory (PCM), Resistive Random Access Memory (RRAM), etc., have shown potential to to be used as on-chip caches and the main memory [5]. Therefore, we would like to explore how to enhance the memory hierarchy from the bandwidth point of view. Specifically, we would like to examine (1) the number of levels in the optimal memory 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 memory hierarchy design constrain, and the memory hierarchy design within a fixed power budge.

3 Experimental Setup

We will use our modified CACTI [1] to estimate the bandwidths provided by different memory technologies. We will also 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].

4 Experiments

In this section, we describe the experimental results we have obtained so far using a modified version of CACTI [1].



(a) **(b)**

Figure 1: Read and write bandwidths provided by different memory technologies. (a) Read bandwidth provided by different memory technologes. (b) Write bandwidth provided by different memory technologies.

First of all, we estimate the read and write bandwidths that can be provided by different memory technologies. Figure 1 shows the results, with both x- and y-values in *log* scale. The figure illustrates both the provided read and write bandwidth as a function of memory capacity. Each of the memory technologies actually provide nearly the same read bandwidths, as is shown in figure 1(a). On the other hand, a straight forward observation from figure 1(b) is that the write bandwidth varies among different memory technologies. The shape of the SRAM write bandwidth curve is very similar to the read bandwidth curve. The write bandwidth curves of the other three memory technologies appear to be very different. The reason is that write latencies of the three non-volatile memories are much higher than read latencies. Another observation from figure 1(b) is that the curves cross to each other at different locations. Based on these two observations, we would like to design a bandwidth-aware hybrid memory hierarchy, which always provides the high memory bandwidth with the given capacity. For example, we probably would like to use 1) SRAM when the memory capacity is under 256MB, 2) MRAM or RRAM between 256MB and 1GB (if only bandwidth is considered), and 3) PCRAM when the memory cacity is over 1GB.

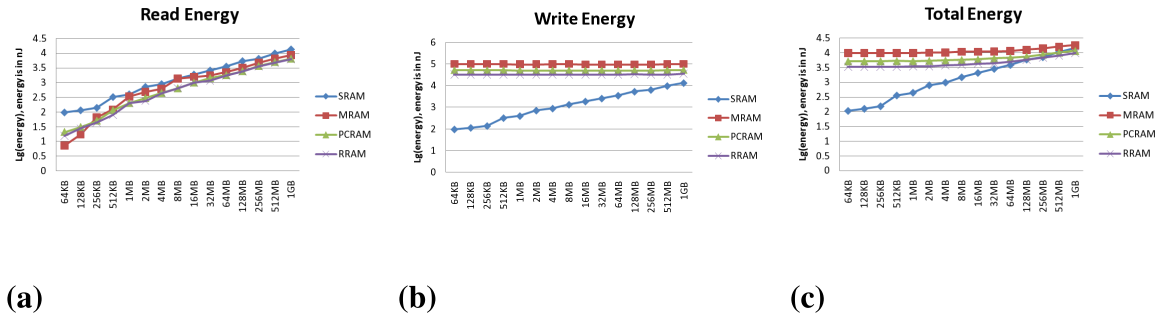


Figure 2: Dynamic energy consumption with the provided bandwidths of different memory technologies. (a) Read energy of different memory technologes. (b) Write energy of different memory technologies. (c) Total energy consumption (with 10% write) of different memory technologies.

In addition, we would like to put an upper bound to the total dynamic energy consumption of the memory hierarchy. Figure 2 shows the estimation of dynamic power consumptions of different memory technologies with the provided bandwidths. In figure 2(c), we estimate the total dynamic energy with 10% write activities.

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.