Abstract

Cache memories have been usually implemented with Static Random-Access Memory (SRAM) technology since it is the fastest electronic memory technology. However, this technology consumes a high amount of leakage currents, which is a major design concern because leakage energy consumption increases as the transistor size shrinks. Alternative technologies are being considered to reduce this consumption. Among them, embedded Dynamic RAM (eDRAM) technology provides minimal area and leakage by design but reads are destructive and it is not as fast as SRAM.

In this thesis, both SRAM and eDRAM technologies are mingled to take the advantatges that each of them offers. First, they are combined at cell level to implement an n-bit *macrocell* consisting of one SRAM cell and *n-1* eDRAM cells. The macrocell is used to build n-way set-associative hybrid first-level (L1) data caches having one SRAM way and *n-1* eDRAM ways. A single SRAM way is enough to achieve good performance given the high data locality of L1 caches. Architectural mechanisms such as way-prediction, swaps, and scrub operations are considered to avoid unnecessary eDRAM reads, to maintain the Most Recently Used (MRU) data in the *fast* SRAM way, and to completely avoid refresh logic. Experimental results show that, compared to a conventional SRAM cache, leakage and area are largely reduced with a scarce impact on performance.

The study of the benefits of hybrid caches has been also carried out in second-level (L2) caches acting as Last-Level Caches (LLCs). In this case, the technologies are combined at bank level and the optimal ratio of SRAM and eDRAM banks that achieves the best trade-off among performance, energy, and area is identified. Like in L1 caches, the MRU blocks are kept in the SRAM banks and they are accessed first to avoid unnecessary destructive reads. Nevertheless, refresh logic is not removed since data locality widely differs in this cache level. Experimental results show that a hybrid LLC with an eighth of its banks built with SRAM technology is enough to achieve the best target trade-off.

This dissertation also deals with performance of replacement policies in heterogeneous LLCs mainly focusing on the energy overhead incurred by refresh operations. In this thesis it is defined a new concept, namely MRU-Tour (MRUT), that helps estimate reuse

information of cache blocks. Based on this concept, it is proposed a family of MRUT-based replacement algorithms that randomly select the victim block among those having a single MRUT. These policies are enhanced to leverage recency of information for a few blocks and to adapt to changes in the working set of the benchmarks. Results show that the proposed MRUT policies, with simpler hardware complexity, outperform the Least Recently Used (LRU) policy and a set of the most representative state-of-the-art replacement policies for LLCs.

Refresh operations represent an important fraction of the overall dynamic energy consumption of eDRAM LLCs. This fraction increases with the cache capacity, since more blocks have to be refreshed for a given period of time. Prior works have attacked the refresh energy taking into account inter-cell feature variations. Unlike these works, this thesis proposes a selective refresh policy based on the MRUT concept. The devised policy takes into account the number of MRUTs of a block to select whether the block is refreshed. In this way, many refreshes done in a typical distributed refresh policy are skipped (i.e., in those blocks having a single MRUT). This refresh mechanism is applied in the hybrid LLC memory. Results show that refresh energy consumption is largely reduced with respect to a conventional eDRAM cache, while the performance degradation is minimal with respect to a conventional SRAM cache.