An On-chip Memory-Path Architecture on Merged DRAM/Logic LSIs for High-Performance/Log-Energy Consumption

Inoue, Koji
Department of Computer Science and Communication Engineering, Kyusyu University

Kai, Koji
Institute of Systems and Information Technologies/Kyushu

Murakami, Kazuaki
Department of Computer Science and Communication Engineering, Kyusyu University

出版情報：Proceedings of International Symposium on Low-Power and High-Speed Chips (COOL Chips III), pp.283-283, 2000-04
バージョン：
権利関係：

An On-chip Memory-Path Architecture on Merged DRAM/Logic LSIs for High-Performance/Log-Energy Consumption

Koji Inoue†, Koji Kai‡, and Kazuaki Murakami†

† Dept. of Computer Science and Comm. Eng.
Kyushu University

‡ Institute of Systems & Information
Technologies/KYUSHU

1 What is the problem?

Integrating a main memory (DRAM) and processors into a single chip, or a merged DRAM/logic LSI, makes available high on-chip memory bandwidth provided by widening on-chip bus and on-chip DRAM array. This approach is well known as a good solution to break the memory wall problem. For merged DRAM/logic LSIs having cache memory, we can exploit the high on-chip memory bandwidth by replacing a whole cache line at a time. This approach tends to increase the cache-line size if we attempt to exploit the attainable high on-chip memory bandwidth.

A large cache-line size can give a benefit of prefetching effect if programs have rich spatial locality. However, it will bring the following disadvantages with poor spatial locality:

1. a number of conflict misses will take place due to frequent evictions,
2. as the result, a lot of energy at the on-chip DRAM (main-memory) will be wasted by a number of DRAM accesses, and
3. activating the wide on-chip bus and the wide DRAM array will also dissipate a lot of energy.

Although increasing cache associativity can achieve high cache hit rate, it makes cache access time longer. In addition, it can not resolve the third disadvantage.

2 Solution

In order to resolve all of the disadvantages, we propose an on-chip memory-path architecture employing *dynamically variable line-size cache (D-VLS cache)*. The D-VLS cache can exploit the high memory bandwidth by means of larger cache lines. At the same time, it can alleviate the negative effects of larger cache-lines by partitioning the large cache line into multiple small cache lines (sublines). Activating only the DRAM subarray corresponding to the sublines to be replaced can reduce energy consumption at the on-chip DRAM. The appropriate cache-line sizes, or the number of sublines to be involved in cache replacements, will be determined by special hardware assists on run time.

Figure 1 shows the construction of a direct-mapped D-VLS cache having three cache-line sizes. If programs have rich spatial locality, the D-VLS cache chooses the largest cache-line size for cache replacements, as shown in figure 1 (c), in order to obtain prefetching effects aggressively. In this case, all of the DRAM subarrays and on-chip buses are activated. Otherwise, the cache attempts to reduce unnecessary evictions caused by the large cache lines by partitioning the large cache line into multiple small cache lines (sublines). As the result, only a few number of sublines are replaced, as shown in figure 1 (a) or (b). Activating the DRAM subarrays and the on-chip buses corresponding to the cache sublines to be replaced can reduce the energy consumption for accessing to the main memory.

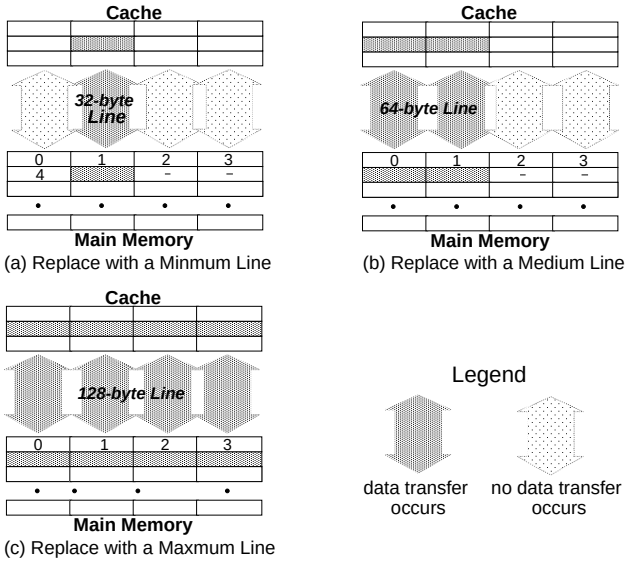

図 1: Three Cache-Line Sizes on Cache Replacements

表 1: Normalized Cache Miss Rate

Cache	Miss Rate	$AMAT'$	E_{DRAM}	$E_{DRAM} \times AMAT'$
Fix128	1.000	1.000	1.000	1.000
Fix128W4	0.418	1.132	0.418	0.511
Fix128double	0.617	0.823	0.617	0.527
D-VLS128-32	0.754	0.825	0.317	0.279

3 Evaluations

In our evaluation, our proposed on-chip memory-path architecture including a direct-mapped D-VLS cache have achieved about 20performance improvement, while it have produced about 70energy reduction, compared to a on-chip memory-path architecture with a conventional direct-mapped cache.