A Buffered Dual-Access-Mode Scheme Designed for Low-Power Highly-Associative Caches

Yul Chu, University of Texas Pan American, Edinburg, TX, USA
Marven Calagos, University of Texas Pan American, Edinburg, TX, USA

ABSTRACT

This paper proposes a buffered dual-access-mode cache to reduce power consumption for highly-associative caches in modern embedded systems. The proposed scheme consists of a MRU (most recently used) buffer table and a single cache structure to implement two accessing modes, phased mode and way-prediction mode. The proposed scheme shows better access time and lower power consumption than two popular low-power caches, phased cache and way-prediction cache. The authors used Cacti and SimpleScalar simulators to evaluate the proposed cache scheme by using SPEC benchmark programs. The experimental results show that the proposed cache scheme improves the EDP (energy delay product) up to 40% for instruction cache and up to 42% for data cache compared to way-prediction cache, which performs better than phased cache.

Keywords: Embedded System, Energy Delay Product (EDP), Low-Power, Phased Cache, Way-Prediction Cache

INTRODUCTION

Fabrication technology has been developed at a rapid pace, resulting in large transistor budgets for chips and processors (Hennessy & Patterson, 2012). This, in turn, enables processor designs with large caches, i.e., more than 32KB for level-one (L1) cache memory with highly-associative cache structures. Even though those caches lead to higher performance, they might consume a large amount of power; hence, low-power consumption has been a critical issue for mobile or hand-held devices, which use batteries as a power source. Those large cache structures occupy more than 60% of modern microprocessor’s die area (Montanaro et al., 1996) and cause more than 50% of total power dissipation (Flynn & Hung, 2005).

Highly-associative caches like 16- or 32-way have an inherent property that provides better performance by reducing the conflict misses (Hennessy & Patterson, 2012). However, they significantly increase power consumption because of simultaneous accesses to all the banks in parallel, i.e., the n-way cache has n banks to access at the same time. Therefore, some
on-chip caches in mobile devices have been designed with less than 32-way set-associative. For example, ARM Cortex A9, which is one of popular mobile microprocessors, uses 4-way or 8-way cache architectures (ARM, 2011).

This paper aims to reduce the power consumption by predicting one bank out of \( n \) banks (e.g., \( n \)-way) and by using two cache access modes, phased or way-prediction mode, according to a tag match via MRU buffer table (refer to ‘Proposed scheme’ section). In this paper, banks (or \( n \) banks) and ways (or \( n \)-way) will be used interchangeably with the similar concept.

The rest of the paper is organized as follows: First, we discuss some related works for popular low-power cache schemes and present the proposed cache scheme. Then, we explain experimental methodology and performance metrics. Finally, we provide the evaluation results, analysis, and conclusion.

**RELATED WORKS**

On-chip power consumption in memories becomes a challenge with the use of deep sub-micron technologies. Alipour et al. (2011) explores a design space for memory architecture. This helps chip designers find cache sizes for optimum power consumption and performance for embedded processors.

Set-associative caches are used to improve cache hit rate but have higher energy consumption than direct-mapped cache due to some wasted energy dissipation (Powel et al., 2001) since, regardless of the number of banks in a set, only one bank has the desired data for a cache hit.

To resolve the energy issue, Hasegawa et al. (1995) proposed a low-power set-associative cache scheme, now commonly referred to as phased cache. In phased cache, all the tags are accessed in the first phase, and if one tag matches to a reference address, only one data block in a bank is accessed as the second phase. The basic idea is to avoid unnecessary data access to reduce power consumption. The disadvantage of a phased cache is a poor performance caused by using more clock cycles to access desired data, compared to other conventional caches.

Taking advantage of power savings over phased cache, Inoue et al. (1999) proposed a low-power set-associative cache scheme, called way-prediction cache that improves the latency of phased cache. The MRU (most recently used) algorithm is popular to predict one of the \( n \) banks to access. If the prediction is correct, the tag and data block are accessed in one cycle with the speed and power savings like direct-mapped cache. If the prediction is wrong, the rest of the banks are accessed during the next cycle in parallel. The performance and power efficiency of way-prediction cache is highly dependent on the accuracy of a way-prediction algorithm used.

Chung et al. (2008) proposes a pipeline change for way determination. An early tag lookup stage, between branch prediction and fetch stage, is used to determine the next way to be accessed. In this method, prediction accuracy and hit rate of the original way-prediction cache are maintained while reducing power consumption. Chung et al. (2008) did not evaluate this scheme for data caches since the early tag lookup stage was proposed for instruction cache only.

Data caches in set-associative schemes are inherently inefficient since data used to be in a dead block or no locality causing more cache misses and power consumption. Liu et al. (2008) uses cache bursts and prefetching to hide the irregularity of individual references. It can also identify dead blocks with 96% accuracy leading to an average L1 improvement of 9% and L2 improvement of 10%. Tseng et al. (2009) proposed a method to improve the problems of spatial and temporal locality by using a 2-bit counter to store the most recently used (MRU) information and Modified Pseudo LRU replacement algorithm (MPLRU): 1) To implement MRU, each cache set needs to have 2 bits of index to save the MRU status; and 2) To conduct MPLRU replacement policy, 3 bits are required to hold history information per
Related Content

Cross-Layer Adaptive Packet Scheduling over Fading Channel: A Decision Theoretic Approach
[www.irma-international.org/chapter/cross-layer-adaptive-packet-scheduling/65663/](www.irma-international.org/chapter/cross-layer-adaptive-packet-scheduling/65663/)

A Novel Secret Key Generation Method in OFDM System for Physical Layer Security
[www.irma-international.org/article/a-novel-secret-key-generation-method-in-ofdm-system-for-physical-layer-security/159141/](www.irma-international.org/article/a-novel-secret-key-generation-method-in-ofdm-system-for-physical-layer-security/159141/)

Application of Parametric Cost Estimation Model to Telecommunication Networks
[www.irma-international.org/article/application-parametric-cost-estimation-model/37204/](www.irma-international.org/article/application-parametric-cost-estimation-model/37204/)

Remote Management of a Province-Wide Youth Employment Program Using Internet Technologies
[www.irma-international.org/chapter/remote-management-province-wide-youth/6458/](www.irma-international.org/chapter/remote-management-province-wide-youth/6458/)
MPSoC Architecture for Macro Blocks Line Partitioning of H.264/AVC Encoder
www.irma-international.org/article/mpsoc-architecture-macro-blocks-line/121728/