INTEGRATION, the VLSI journal 54 (2016) 24–36

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Cross-matching caches: Dynamic timing calibration and bit-level timing-failure mask caches to reduce timing discrepancies with low voltage processors

Po-Hao Wang [a,*], Shang-Jen Tsai [a], Rizal Tanjung [a], Tay-Jyi Lin [c], Jinn-Shyan Wang [b], Tien-Fu Chen [a]

[a] Department of CSIE, National Chiao Tung University, Hsinchu, Taiwan, ROC
[b] Department of EE, National Chung Cheng University, Chiayi, Taiwan, ROC
[c] Department of CSIE, National Chung Cheng University, Chiayi, Taiwan, ROC

ARTICLE INFO

Article history:
Received 2 October 2015
Received in revised form
14 January 2016
Accepted 18 January 2016
Available online 6 February 2016

Keywords:
Cache memory
Low voltage
Timing discrepancy
Timing-failure tolerance

ABSTRACT

Voltage scaling is an effective technique to reduce power consumption in processor systems. Unfortunately, timing discrepancies between L1 caches and cores occur with the scaling down of voltage. These discrepancies are primarily caused by the severe process variations of a few slow SRAM cells. Most previous designs tolerated slow cells by adjusting access latency based on a coarse-grained track of cache blocks. However, these methods become insufficient when the amount of slow cells increases. This paper addresses the issue for an 8T SRAM cache and proposes a cross-matching cache that includes dynamic timing calibration and actual bit-level timing-failure toleration.

1. Introduction

Voltage scaling is an effective method for saving energy in modern processor systems. In the past, caches limited the minimum operating voltage of systems because of the poor reliability and long-latency of static random-access memory (SRAM) in low-voltage operations. To increase the cache reliability, numerous fault-tolerance caches, such as disabling [1], redundancy [2,3], error correction code (ECC) designs [4,5] and robust SRAM cell designs [6,7], have been proposed. Unfortunately, most fault-tolerance designs necessarily sacrifice cache latency to increase reliability. Therefore, these designs are not suitable for latency-sensitive level 1 (L1) caches. To provide reliable access and dual-port access (better performance), robust 8T SRAM is widely used in modern L1 caches [8] without any fault-tolerance mechanism. The reliability issue of modern L1 caches has been solved by using 8T SRAM; however, low-voltage environments cause L1 caches to require long-latency for access. This overly long access latency causes a timing discrepancy between a core and a cache that restricts the performance of the entire system, particularly in sub-threshold voltage operations.

Aggressive voltage scaling worsens timing discrepancy problems. Assuming the access cycle of an L1 cache is 2 cycles at normal voltage, for 0.5 V, the worst case of cache latency can be up to 4 cycles [9]. The gray and black lines in Fig. 1 represent the increasing latency of the core and cache, respectively, as the voltage is scaled down. When the voltage is decreased to a certain level, the cache is not able to be accessed correctly within the access cycle of normal voltage operation (2 cycles). Thus, the core needs to decrease its operating frequency or extend the access cycles of the cache. However, both of these methods impact the performance of the entire system.

The severe increase in timing discrepancy between a core and a cache is primarily caused by the severe process variations of slow SRAM cells. These slow cells increase the overall SRAM access latency. The three dots in the upper right part of Fig. 1 represent the best-, average- and worst-case latencies of an SRAM cell. In the average case, the cache can be accessed correctly within the access cycle, which can catch up with the core's speed. Thus, only a few cells with long-latency compromise the performance of the entire system. Fig. 2 shows the delay distribution of SRAM cells at normal voltage and low voltage. Only a small fraction of the SRAM cells are slow. Nevertheless, the number of slow cells is increased by aggressive voltage decreases and technology node advancement.

* Corresponding author. Tel.: +886 35712121; fax.: +886 35721490.
E-mail address: pohaow@cs.nctu.edu.tw (P.-H. Wang).

http://dx.doi.org/10.1016/j.vlsi.2016.01.001
0167-9260/

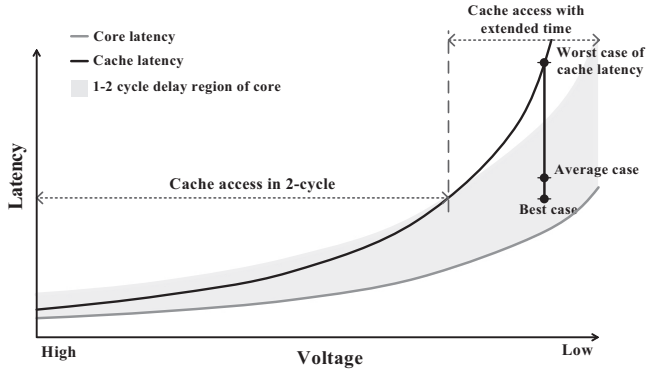

Fig. 1. Timing discrepancy between a core and a cache.

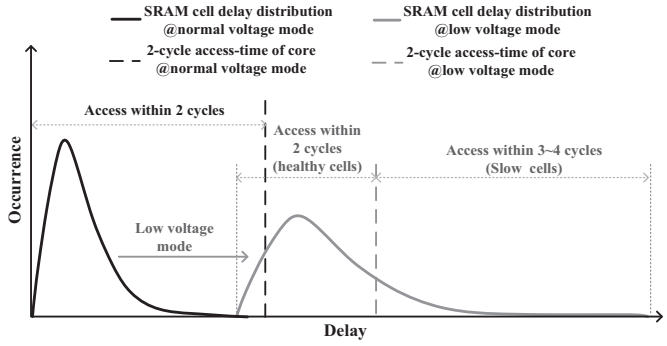

Fig. 2. Voltage scaling impact of the cell delay distribution.

Therefore, tolerating numerous slow cells to reduce the timing discrepancy will become a critical issue.

Mutyam et al. [10] proposed a variable-latency cache (VL-cache) to tolerate slow cells, and Wang et al. [11] proposed a variation-aware and adaptive-latency cache (VAL-cache) that used a timing table created by the manufacturer during the testing process to record the appropriate access cycles of each cache set/line. However, even if there is only one slow cell in the cache line/set, this cache line/set should be accessed with the worst-case access time. Thus, when the percentage of slow cells in SRAM increases to ~1%, these solutions can only slightly improve the performance because of the coarse-grained timing of the recording.

In this paper, we observe that the value stored in 8T SRAM significantly influences the read latency of the cache. Based on this observation, we propose a cross-matching cache (CM-cache) for modern L1 caches that includes a dynamic timing calibration for 8T SRAM. The CM-cache dynamically calibrates the read latency of each cache line with different stored data and then shortens the read latency of the cache line that can be read faster. Moreover, we propose three different cache management strategies for dealing with different usages:

Simple placement is the simplest way to tolerate slow cells by exploiting three cache characteristics to increase the probability of a faster read and to decrease the overhead.

Fixed mirror sacrifices one way to store the mirror data of an MRU line for providing the bit-level timing-failure mask. With failure masking, this strategy can tolerate significantly more slow cells.

Selective mirror adds a finite state machine to allow each cache line to switch between a simple placement mode and a fixed mirror mode.

Overall, our contributions are as follows:

1) We observe that value stored in 8T SRAM significantly influences the read latency of the cache.
2) We propose a dynamic timing calibration SRAM to dynamically adjust the read latency for each row with different stored data.
3) We build a cross-matching cache based on DTC-SRAM and perform three cache management strategies according to different execution scenarios for reducing the timing discrepancy between core and L1 caches with a slight overhead.
4) We propose a bit-level timing-failure mask to tolerate numerous slow cells against more aggressive voltage scaling and advanced process nodes in the future.

Section 2 discusses the impact of 8T SRAM caches in low voltage and details our observations. Section 3 shows the dynamic timing calibration SRAM in detail. Section 4 explains our CM-cache in three different strategies. Section 5 introduces the experiment and evaluates our design and the overhead estimation. Section 6 reviews related work and Section 7 concludes the paper.

2. Characteristics of 8T SRAM

In the L1 cache of a modern processor system, the 8T cell has gradually replaced the 6T cell for low-voltage applications and dual-port access. In this section, we present some observations on characteristics of 8T SRAM cells and discuss SRAM failure in low-voltage situations.

2.1. Wide delay distribution of SRAM cells in low voltage

In low-voltage mode, Fig. 2 shows a long tail distribution of an SRAM cell delay. Slow cells need more cycles to be accessed. An SRAM cell is more likely to be affected by process variation than a logic cell, and the most significant problem is access failure, which occurs when slow cells cannot complete their discharge in time due to variations. The logic part is not as vulnerable to slow cell problems, and the delay distribution is more balanced than with SRAM cells [12] because it is usually series connected by logic gates and works one after one. Therefore, the total access time will be balanced by the gates on the path. Although a SRAM cell is stored or loaded independently, it is more vulnerable to access-time failure. To access these slow cells successfully, they require extending access cycles to complete their discharge and to allow the sense amplifier to determine the correct value. If these slow cells can be tolerated and accessed with total cycles close to normal cells, their performance can be improved.

2.2. Effect of the stored value on the latency

Fig. 3 shows the cell structure of an 8T SRAM. To perform a read operation, the read word line (RWL) is activated and the read bitline (RBL) is pre-charged. When reading '0', the RBL is pulled down through the transistors M7 and M8. An access-time failure occurs when reading '0' if the RBL voltage drops too slowly for the sense amplifier to sense it in time. Contrarily, the datum '1' can be read via the RBL directly after pre-charging. Access-time failures will not occur because bitlines do not require any discharge time.

Fig. 4 shows the read operation waveforms of slow cells and healthy cells with different stored values on an 8T SRAM. There is no critical issue with either healthy cells or slow cells when reading the value '1'. Because the bitline does not need to be discharged and the bitline voltage is always greater than the sense amplifier sensitivity, the sense amplifier will always sense the correct value '1'. However, when the value '0' is read, the value sensed by the sense amplifier at a shortened fetch point (SFP) is different for healthy cells and slow cells. For a healthy cell, the read