256-KB Associativity-Reconfigurable Cache with 7T/14T SRAM for Aggressive DVS Down to 0.57 V

Jinwook Jung[1], Yohei Nakata[1], Shunsuke Okumura[1], Hiroshi Kawaguchi[1], and Masahiko Yoshimoto[1, 2]
[1]Graduate School of System Informatics, Kobe University, Japan [2]JST, CREST
jung@cs28.cs.kobe-u.ac.jp

***Abstract*—This paper presents a dependable cache memory for which associativity can be reconfigured dynamically. The proposed associativity-reconfigurable cache consists of pairs of cache ways. Each pair has two modes: the normal mode and the dependable mode. The proposed cache can dynamically enhance its reliability in the dependable mode, thereby trading off its performance. The reliability of the proposed cache can be scaled by reconfiguring its associativity. Moreover, the configuration can be chosen based upon current operating conditions. Our chip measurement results show that the proposed dependable cache possesses the scalable characteristic of reliability. Moreover, it can decrease the minimum operating voltage by 115 mV. The cycle accurate simulation shows that designing the L1, L2 caches using the proposed scheme results in 4.93% IPC loss on average. Area estimation results show that the proposed cache adds area overhead of 1.91% and 5.57% in 32-KB and 256-KB caches, respectively.**

I. Introduction

Feature sizes in transistors continue to shrink along with the advance of process technology, achieving higher density and lower cost. Technology scaling, however, increases variations in different device parameters and creates a considerable spread in a transistor threshold voltage, mainly because of random dopant fluctuation (RDF), which has deviation that is inversely proportional to the square of a channel area [1]. Such variations strongly impact functionality and reliability in deep sub-micro process technology [2].

This situation yields severe problems, particularly in SRAM, because minimum-sized transistors are used in its design. Device mismatching between transistors in SRAM caused by the process variations make the memory cell unreliable, which results in cell failures (read failure, write failure, and access time failure) [3]. In addition, the minimum operating voltage ($V_{\min}$) tends to increase according to technology scaling [1]: a low $V_{\min}$ decreases read/write margins in SRAM and reliability deterioration arises. To make matters worse, SRAMs occupy a substantial fraction of the total die area and a transistor count in processors [4]. Consequently, a large SRAM block such as an L1 (Level-1) cache or a last level cache (LLC) determines the $V_{\min}$ of the whole processor.

Several studies of low voltage cache technique have been reported [5-6]. However, the low-voltage cache proposed in [5] has only two operating modes: a high-voltage mode and a low-voltage mode. The architecture proposed in [6] necessitates additional circuitry or large spare SRAM blocks.

As described in this paper, we propose an associativity-reconfigurable cache using 7T/14T SRAM for low-voltage operation. Our proposed cache can trade off its associativity (the number of cache ways) with low-voltage reliability. The *N*-way set associative cache can be reconfigured dynamically; its associativity can thereby be halved (from *N*/2 to *N*) with the 7T/14T SRAM. This reconfigurability provides scalable reliability and makes it possible to leverage optimal cache configuration in dynamic voltage scaling (DVS).

II. 7T/14T Dependable SRAM

We have proposed the 7-Transistor/14-Transistor (7T/14T) dependable SRAM [7]. Figure 1 shows a structure of 7T/14T dependable SRAM: two pMOS transistors (M20 and M21) are added to internal nodes (N00 and N10, N01 and N11) in a pair of the conventional 6-Transistor (6T) memory cells. The structure thereby achieves a dependable mode that features margin enhancements by combining two memory cells, especially in a low-voltage region.

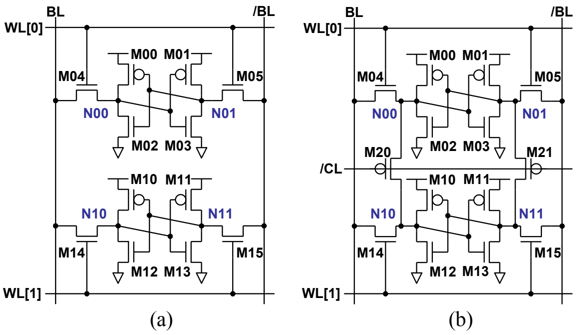

Figure 1. Schemetic of SRAM cell pairs.
(a) Conventional 6T SRAM and (b) 7T/14T Dependable SRAM.

The 7T/14T memory cells have two modes, as shown in Table I.

- Normal mode (7T): Additional transistors are turned off (CL = "H"); the 7T cell acts as a conventional 6T cell.
- Dependable mode (14T): Additional transistors are turned on (CL = "L"); the internal nodes are shared by the memory cell pair. During write operation, both WL0 and WL1 are driven. In the read operation, on the other hand, either WL0 or WL1 is asserted, which ensures stable operations.

In the normal mode, a one-bit datum is stored in one memory cell, which is more area-efficient. In the dependable read mode, only one wordline is asserted to gain a large β ratio (a ratio of two driver transistors' total size to one access transistor size). A memory cell with no static noise margin [8] is recovered by the other memory cell through the two

978-1-4577-1846-5/11/$26.00 ©2011 IEEE

connecting pMOS transistors. In the dependable write mode, a datum is written into a pair of memory cells by asserting both wordlines, which averages and mitigates the write margin degradation. In addition, the dependable mode has better soft-error tolerance because its internal node has more capacitance [9].

TABLE I. TWO MODES IN THE 7T/14T DEPENDABLE SRAM

Mode	# of memory cells comprising 1 bit	# of WL drivers	CL
Normal	1 (7T/bit)	1	Off ("H")
Dependable (write)	2 (14T/bit)	2	On ("L")
Dependable (read)	2 (14T/bit)	1	On ("L")

The dependable mode and normal mode can be switched according to the operating voltage, power limit, and required reliability in an application. If the 7T/14T SRAM has sufficient operation margins, for instance, then the recovery feature of 7T/14T SRAM can be disabled when it operates at high voltage by negating the two connecting pMOS transistors. In the dependable mode, a one-bit datum is stored in two memory cells, although the quality of the information differs from that of the normal mode. We designate this concept as 'quality of a bit (QoB)', in which the operating voltage, power, and bit error rate (BER) are controlled as attributes of one-bit information [7]. The information quality is scalable in the 7T/14T memory cell.

III. PROPOSED CACHE ARCHITECTURE

In this section, we describe the proposed associativity-reconfigurable scheme. The *N*-way set associative cache using the proposed 7T/14T SRAM can change its associativity dynamically between *N*/2 and *N*. Figures 2 and 3 depict the organization of the proposed cache; adjacent two ways compose a pair of odd-even ways. These pairs of two ways enable application of the 7T/14T SRAM-dependent feature to the cache design. If there is no need to leverage the enhanced reliability of the dependable mode (for instance, when the operating voltage margin is sufficient to operate properly or application software requires large cache capacity for high performance rather than improved reliability), two ways in a pair operate separately in the normal mode, as shown in Figure 2. Figure 3 portrays the case of the dependable mode; odd-even ways in a pair are logically bound together.

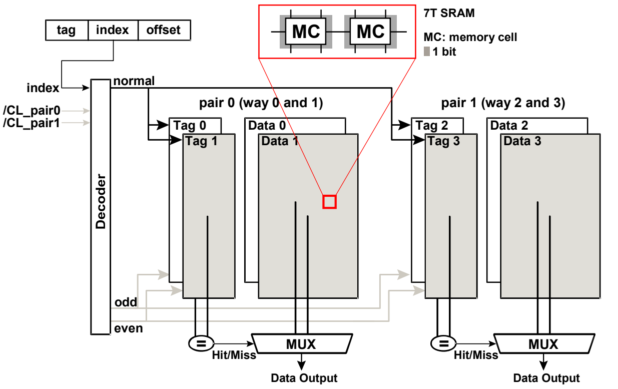

Figure 2. Organization of the associativity-reconfigurable cache. All pairs of odd-even ways operate in the normal mode using 7T SRAM.

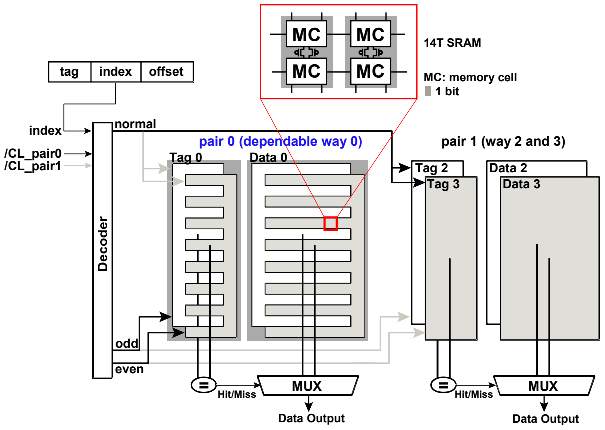

Figure 3. Organization of the associativity-reconfigurable cache. Pair 0 operates in the dependable mode using 14T SRAM.

Figure 4 illustrates detailed views f the dependable cache ways interleaving odd-even cache lines. Because a one-bit datum in the dependable way is made up of a pair of memory cells, the capacity is halved and the associativity is decreased by one in a pair of ways, but improved reliability is obtainable. If the operation margins are sufficient or if high performance is a critical issue, each pair of ways might be detached and be operated respectively as described above.

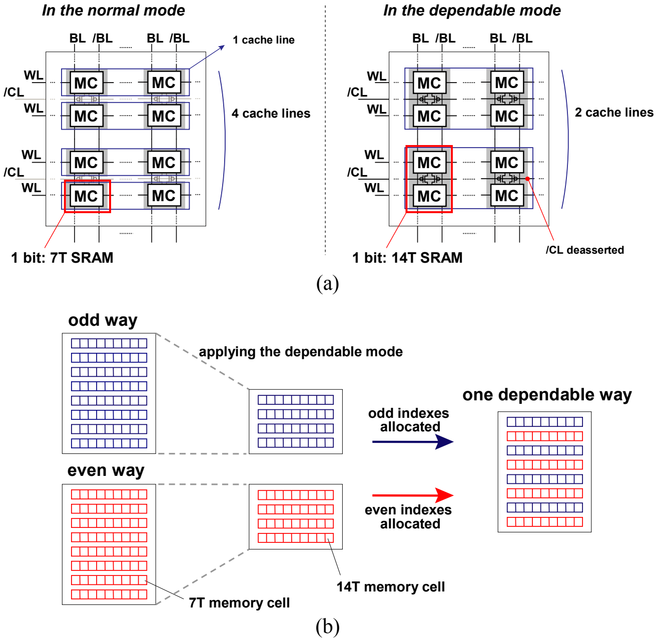

Figure 4. Composition of a dependable cache way in the dependable mode: (a) physical and (b) logical allocations.

Configuration of the proposed cache can be determined arbitrarily by enabling or negating corresponding control lines (CLs in Figure 1); which pair of two ways and how many pairs to be applied can be parameterized. It is necessary to write back dirty cache lines to the next level cache or main memory before entering the dependable mode: dirty odd lines in the even way of the pair, and dirty even lines in the odd way. In this way, the proposed cache scheme can dynamically change its associativity, and obtain adaptive reliability. Appropriate associativity can be chosen by applying the dependable mode to some pairs of ways selectively, and therefore optimal reliability is obtainable.

In order to implement the proposed cache, it is necessary to use decoders as shown in Figure 5; these are one n-to-2^n decoder and one n−1-to-2^{n-1} decoder (where n is a bit width of the cache index). In the normal mode, the upper n-to-2^n decoder is activated and it drives each cache way independently. On the other hand, in the dependable mode, the n−1-to-2^{n-1} decoder is asserted in the dependable mode; a pair of odd-even ways comprises one dependable way. Operating as one dependable way, because the capacity is halved, it is necessary to decide either the odd way or the even way in the pair of ways is chosen. The LSB (least significant bit) in the cache index is used to do so; if the LSB is 0, n−1-to-2^{n-1} decoder's output connected to the even way in the pair of ways, otherwise (if the LSB is 1) n−1-to-2^{n-1} decoder's output connected to the odd way.

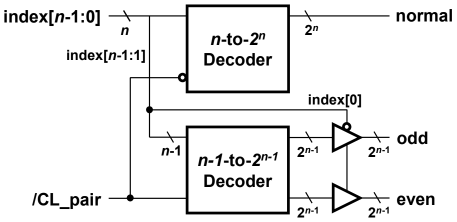

Figure 5. Implementation of decoders for the proposed cache.

IV. Evaluation

In this section, we describe evaluation of the proposed associativity-reconfigurable cache. The cache system configuration is presented in TABLE II. The minimum operating voltage ($V_{\min}$) improvement is evaluated based on measurement results, and the simulated performance overhead is analyzed. We also estimated the area overhead of the proposed cache.

TABLE II. Cache System Configuration

Instruction L1 (IL1) Cache, Data L1 (DL1) Cache	32 KB 8-way set associative cache (with 2.75 KB tag array)
Unified L2 Cache	256 KB 8-way set associative cache (with 19 KB tag array)

A. Measurement Results

To evaluate the $V_{\min}$ improvements in our proposed cache, we measured 512-Kbit 7T/14T SRAM macros manufactured in a 65-nm process. It can be regarded as two ways of eight-way 256-KB L2 cache: i.e., a pair of odd-even ways in the proposed associativity-reconfigurable cache. Therefore, $V_{\min}$ improvement in the eight-way 256-KB L2 cache is evaluated with measuring four 512-Kbit 7T/14T SRAM macros. A 512-Kbit 7T/14T SRAM can also be matched to a 32-KB IL1 cache and 32-KB DL1 cache. Therefore, we can evaluate $V_{\min}$ improvements in the eight-way 32-KB IL1 and DL1 caches by measuring one 512-Kbit SRAM macro. We measured one more 512-Kbit SRAM macro to evaluate the tag arrays in the L1 and L2 caches.

Figure 6 shows the measured $V_{\min}$s of each pair of ways in the 256-KB L2 cache tag and data arrays. In the normal mode, the measured $V_{\min}$s are 0.685 V, 0.645 V, 0.630 V, and 0.610 V, respectively, whereas in the dependable mode, the respective values are reduced to 0.570 V, 0.550 V, 0.540 V, and 0.530 V. As an eight-way L2 cache, the proposed cache can operate at 0.685 V. If Pair 0 changes its mode to the dependable mode, the proposed cache can operate at 0.645 V as 7-way cache. Likewise, the $V_{\min}$ can be scaled by trading off the associativity.

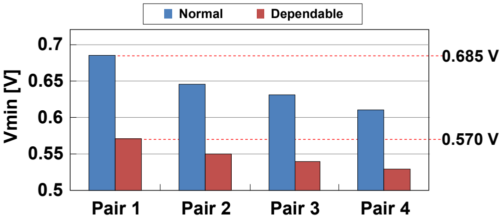

Figure 6. Measured $V_{\min}$s of each pair of ways in the L2cache.

Figure 7 summarizes the $V_{\min}$ scalability in the proposed associativity-reconfigurable cache. If all the four pair of ways enter the dependable mode, the proposed cache can operate as a four-way 128-KB cache at 0.570 V, achieving a 115-mV lower $V_{\min}$ than the eight-way 256-KB in the normal mode. Applying the dependable mode in one pair of ways reduces $V_{\min}$ by 30 mV ~ 40 mV. We also evaluated the $V_{\min}$ of 32-KB IL1 and DL1 cache in the same way. The IL1 cache and the DL1 cache can operate eight-way 32-KB in the normal mode at 0.610 V and 0.600 V, respectively. This value is larger than the four-way L2 cache's $V_{\min}$. Consequently, in the case in which the L2 cache operates as a four-way 128-KB, two L1 caches enter the dependable mode for achieving minimum $V_{\min}$. In this case, the IL1 and DL1 caches operate as five-way 20-KB caches when the L2 cache operates as a four-way 128-KB cache.

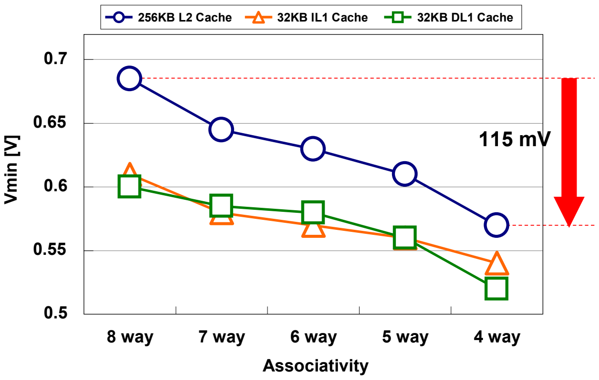

Figure 7. Measured $V_{\min}$s in the proposed caches when changing associativity.

B. Performance evaluation

We simulate the performance overhead of the proposed cache architecture. Since the proposed cache downsizes the associativity and capacity in the dependable mode, it affects a cache hit rate and thus a processor performance; it is necessary to evaluate the impact on the performance. The SESC [10] cycle-accurate simulator is used to measure the performance degradation of the proposed scheme. TABLE III shows the voltage-independent baseline processor configuration.

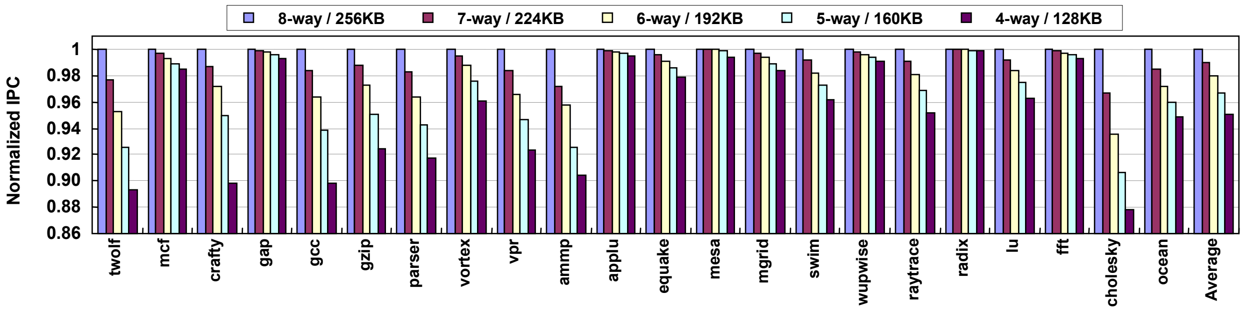

Figure 8. Normalized IPCs for SPEC CPU2000 and SPLASH2 benchmark suites.

TABLE III. VDD-INDEPENDENT BASELINE PROCESSOR PARAMETERS

Parameter	Value
Processor frequency	1 GHz
Number of cores	4
External DRAM latency	100 cycles
Fetch / Issue / Retire width	2/2/2
Cache line size	32 bytes
L1 Instruction Cache	32KB, 8-way, 3 cycles
L1 Data Cache	32KB, 8-way, 3cycles
Unified L2 Cache	256KB, 8-way, 16cycles
Replacement policy	Pseudo-LRU

Figure 8 plots a normalized IPC (instructions per cycle) for each benchmark in SPEC2000 and SPLASH2 with respect to the L2 cache's associativity. The IL1 and DL1 cache operate as 8-way cache, except in the case of four-way L2 cache: as described above, L1 caches operate as five-way caches to achieve minimum V_{min}. We simulated nine integer benchmarks and seven floating-point benchmarks from SPEC2000 and six benchmarks from SPLASH2 benchmark suite. The maximum IPC loss is 12.2% (a four-way case in cholesky). The IPC degradation is 1.24% on average when the dependable mode is applied to a single pair of ways. The average IPC degradation is 4.93% in the 128-KB four-way L2 cache (in the case in which all the pairs operate in the dependable mode).

C. Area overhead

In this section, we present our evaluation of the area overhead of the proposed cache architecture. The 7T memory cell area is 11% greater than that of the conventional 6T memory cell. We assume that the dedicated decoders have a negligible impact on the overall area because they occupy a smaller proportion than SRAM arrays in the transistor number. Utilizing these facts, we estimated the area overhead of the proposed cache in the 32-KB L1 and 256-KB L2 cache with CACTI [11]. In 65-nm process technology, the proposed cache imposes only a 1.91% and 5.57% area overhead in 32-KB L1 and 256-KB L2 cache respectively. For a small cache, the proposed cache adds a small area overhead because the fraction of the peripheral circuitry in area is relatively large.

V. CONCLUSION

In this paper, we proposed a novel dependable cache architecture using 7T/14T SRAM. The dependable cache described in this paper has the reconfigurable associativity and thus the adaptive reliability is realizable. Measurement-based evaluation shows that the proposed dependable cache possesses the scalable characteristic of reliability and it can decrease the V_{min} by 115 mV. The cycle accurate simulation shows that applying the proposed cache architecture results in 4.91% average IPC loss. The area overhead of the proposed cache is only 1.91% and 5.57% in 32-KB and 256-KB cache respectively.

REFERENCES

[1] K. Itoh, "Adaptive Circuits for the 0.5-V Nanoscale CMOS Era," *IEEE International Solid-State Circuits Conference*, pp. 14-20, 2009.

[2] S. Borkar, T. Karnik, and V. De, "Design and realibility challenges in nanometer technologies," *ACM/IEEE Design Automation Conference*, pp. 75, 2004.

[3] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859-1880, 2005.

[4] E. J. Marinissen, B. Prince, and D. K.-S. Y. Zorian, "Challenges in embedded memory design and test," *Design, Automation and Test in Europe Conference and Exhibition*, pp. 722–727, 2005.

[5] C. Wilkerson, H. Gao, A. R. Alameldeen, Z. Chishti, M. Khellah, and S. -L. Lu"Trading Off Cache Capacity for Reliability to Enable Low Voltage Operation," *IEEE International Symposium Computer Architecture*, pp 203-214, 2008

[6] J. Kim, N. Hardavellas, K. Mai, B. Falasafi, and J. C. Hoe, "Multi-Bit Error Tolerant Caches Using Two-Dimensional Error Coding," *ACM/IEEE International Symposium MicroArchitecture*, pp. 197-209, 2007

[7] H. Fujiwara, S. Okumura, Y. Iguchi, H. Noguchi, H. kawaguchi and M. Yoshimoto, "A 7T/14T Dependable SRAM and Its Array Structure to Avoid Half Selection," *IEEE International Conference on VLSI Design*, pp. 295-300, 2009.

[8] E. Seevinick, F. J. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748-754, 1987.

[9] S. Yoshimoto, T. Amashita, S. Okumura, K. Yamaguchi, H. Kawaguchi, and M. Yoshimoto, "Bit Error and Soft Error Hardenable 7T / 14T SRAM with 150-nm FD-SOI Process," *IEEE International Reliability Phisics Symposium*(IRPS), pp. 876-881, 2011.

[10] J. Renau, B. Fraguela, J. Tuck, W. Liu, M. Prvulovic, L. Ceze, K. Strauss, S. Sarangi, P. Sack, and P. Montesinos, "SESC Simulator," 2005. http://sesc.sourceforge.net.

[11] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. Jouppi, "CACTI 5.1," Technical Report HPL-2008-20, Hewlett Packard Labs, 2008.