IPSJ Transactions on System LSI Design Methodology Vol.5 32–43 (Feb. 2012)

[DOI: 10.2197/ipsjtsldm.5.32]

Regular Paper

0.5-V 4-MB Variation-Aware Cache Architecture Using 7T/14T SRAM and Its Testing Scheme

Yohei Nakata[1,a)] Shunsuke Okumura[1] Hiroshi Kawaguchi[1]
Masahiko Yoshimoto[1,2]

Received: May 27, 2011, Revised: September 2, 2011,
Accepted: October 19, 2011, Released: February 21, 2012

Abstract: This paper presents a novel cache architecture using 7T/14T SRAM, which can improve its reliability with control lines dynamically. Our proposed 14T word-enhancing scheme can enhance its operating margin in word granularity by combining two words in a low-voltage mode. Furthermore, we propose a new testing method that maximizes the efficiency of the 14T word-enhancing scheme. In a 65-nm process, it can reduce the minimum operation voltage (V_{min}) to 0.5 V to a level that is 42% and 21% lower, respectively, than those of a conventional 6T SRAM and a cache word-disable scheme. Measurement results show that the 14T word-enhancing scheme can reduce V_{min} of the 6T SRAM and 14T dependable modes by 25% and 19%, respectively. The respective dynamic power reductions are 89.2% and 73.9%. The respective total power reductions are 44.8% and 20.9%.

Keywords: cache memory, low voltage, low power, process variation, fine-grain control

1. Introduction

As process technology advances, the minimum feature size decreases, which enables manufacture of components with higher density at lower cost. However, technology scaling increases the threshold-voltage (V_{th}) variation of MOS transistors mainly because of the random dopant fluctuation. A minimum operating voltage (V_{min}) becomes higher as the V_{th} variation increases with technology scaling. The increase of V_{min} degrades the device reliability because of power supply noise, IR drops, and/or soft errors. The dynamic power is proportional to the square of the operating voltage (V_{dd}). Therefore, V_{min} is an important parameter in power dissipation because, when using larger V_{min}, dynamic voltage and frequency scaling (DVFS) cannot be exploited. Therefore, the range of power scaling is restricted.

The V_{min} on an entire processor including logic blocks and memory components is determined by the circuit that has the highest value of V_{min} [1]. The SRAM has a larger standard deviation of threshold voltage than logic blocks because its transistors are smaller. To make matters worse, the capacity of SRAM bitcells on a processor is huge. Consequently, large SRAM blocks such as L1 data/instruction caches and last level cache (LLC) determine the V_{min} on the processor.

The V_{th} variation in each SRAM bitcell is distributed randomly throughout the whole SRAM block, which is known as random variation or local variation. Therefore, failures in the whole SRAM block or in the entire processor are distributed. Coarse-grain control on an SRAM block level basis or a cache way level basis cannot prevent these failures efficiently. Therefore, to reduce V_{min}, fine-grain control that adaptively addresses the V_{th} variations must be applied to the SRAM block.

As described in this paper, we present a word-level enhancing scheme using 7T/14T SRAM for a large-capacity cache. The proposed 14T word-enhancing scheme is implemented with leveraging the word cut-off and with combining a 7T less-marginal bitcell to an adjacent 7T bitcell. The 14T word-enhancing scheme can reduce V_{min} lower than the cache word-disable scheme proposed by [2] because it can enhance the operating margin of the defective bitcell by making use of the 14T structure.

In the next section, we describe works related to the cache for low-voltage operation or yield enhancement. We then introduce the 7T/14T SRAM bitcell and its operating modes, and compare bit error rates (BERs) of 7T/14T SRAM with other conventional schemes in Section 3. Section 4 presents a description of the proposed 14T word-enhancing scheme and the proposed incremental testing scheme. Then, the simulated and measured improvements of V_{min} compared with the conventional scheme are reported. Detailed descriptions of the physical implementation of the 14T word-enhancing scheme are also presented. In Section 5, we describe a comparison of performance, energy, and power between the conventional scheme and the proposed scheme. Finally, Section 6 concludes the paper.

A previous work [13] did not include a description of the effects of V_{min} reduction in measurements of the fabricated silicon chip. Furthermore, it did not provide a detailed evaluation of energy and power consumptions. Those discussions are presented in this paper, respectively, in Sections 4.6 and 5.2.

[1] Graduate School of System Informatics, Kobe University, Kobe, Hyogo 657–8501, Japan
[2] JST CREST, Chiyoda, Tokyo 102–0076, Japan
[a)] nkt@cs28.cs.kobe-u.ac.jp

© 2012 Information Processing Society of Japan

2. Related Work

Wilkerson et al. proposed the cache word-disable scheme ('the word-disable scheme' hereinafter) and the cache bit-fix scheme ('bit-fix scheme') enabling low-voltage operation [2]. The word-disable scheme disables defective words and selects four workable words from eight words. A defect word map (one-bit information per word), which shows which words are defective and valid, is stored in a cache tag. The word-disable scheme purges the remaining four words. Therefore, the cache size and associativity must be halved. The number of ways is reduced to four from eight in studies described in the literature.

The bit-fix scheme exploits one strategy for redundancy: it stores locations of defective bits in the remaining three ways along with patch bits for them. Then, the defective bits are replaced with the patch bits. The number of ways results in six from eight, which means that the area overhead is smaller than the word-disable scheme. However, the bit-fix scheme suffers a three-cycle penalty, whereas that in the word-disable scheme suffers only a one-cycle penalty. In low-voltage operation, the reliability in the redundant way is lowered as much as the other three ways, where slow error correction coding (ECC) must be implemented. The bit-fix scheme cannot operate at a lower voltage than the word-disable scheme because the failure rate is increased rapidly in the redundancy way. Even ECC cannot fix it.

That earlier study applied a word-disable scheme and the bit-fix scheme to L1 caches and L2 cache, respectively, achieving V_{min} reduction to 0.5 V. Nevertheless, detailed conditions of the failure rate in their 6T SRAM were not described clearly. The failure rate for the redundancy way was not considered in their report.

Ozdemir et al. proposed a yield-aware cache architecture and specifically addressed cache access latency and leakage power [3]. They developed four schemes: The first one disables cache ways that have timing failures or excess leakage to improve the cache yield. The second also disables horizontal regions in the cache. The third one changes cache access latency in each cache way. The fourth is a hybrid scheme of the first, second, and/or third schemes. They reduced the yield losses by 81.1% using the fourth hybrid scheme. However, they evaluated the yield only with access latencies and leakage power, although margin analysis in SRAM is fundamental to the yield evaluation at a low voltage.

3. 7T14T SRAM

3.1 Failures in SRAM

Failures in SRAM are categorizable as read margin failures, write margin failures, and access time violations.

- **Read margin failure**: a read operation is signified by a read static noise margin (read SNM) [8]. If the read SNM becomes zero by a low V_{dd}, a noise source, or destructive readout, then the stored datum flips.
- **Write margin failure**: a write operation is explainable by a write-trip point (WTP) as a metric (= write margin) [9]. The WTP represents the maximum voltage that can write '0' to a bitcell and can then flip an internal datum.

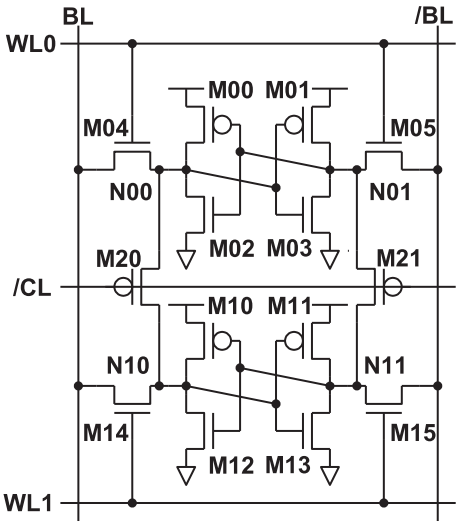

Fig. 1 A 7T/14T bitcell pair.

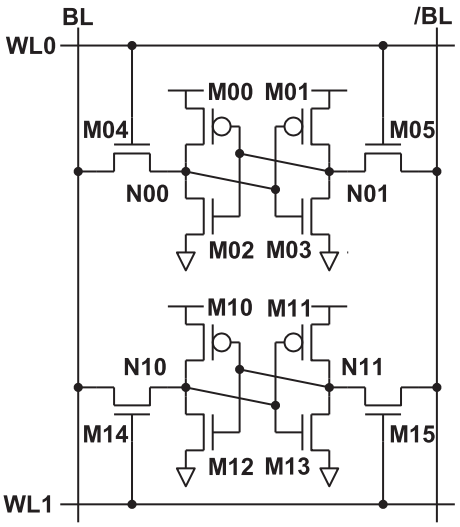

Fig. 2 Conventional 6T bitcells.

- **Access time violation** occurs when a differential voltage between bitlines is small and a sense amplifier cannot sense it in a predetermined acceptable time. The access time violation is dependent on the clock frequency and a timing guard band. This failure type is not incorporated into the discussion presented in this paper because it is dependent on the clock frequency. The read SNM and the write margin are dominant at low frequencies.

3.2 7T/14T SRAM

Figure 1 depicts the 7T bitcell (14T for two bitcells) [4]. Two pMOSes are added to internal nodes ('N00 and N10', 'N01 and N11') in a pair of the conventional 6T bitcells presented in **Fig. 2**. The area overhead in the 7T bitcell is 11% greater than that of the conventional 6T bitcell.

Table 1 shows that the 7T/14T bitcells have two modes.

- Normal mode (7T): The additional transistors are turned off (CL = 'H'); the 7T cell acts as a conventional 6T cell.
- Dependable mode (14T): The additional transistors are turned on (CL = 'L'); the internal nodes are shared by the bitcell pair. In a write operation, both WL0 and WL1 are

© 2012 Information Processing Society of Japan

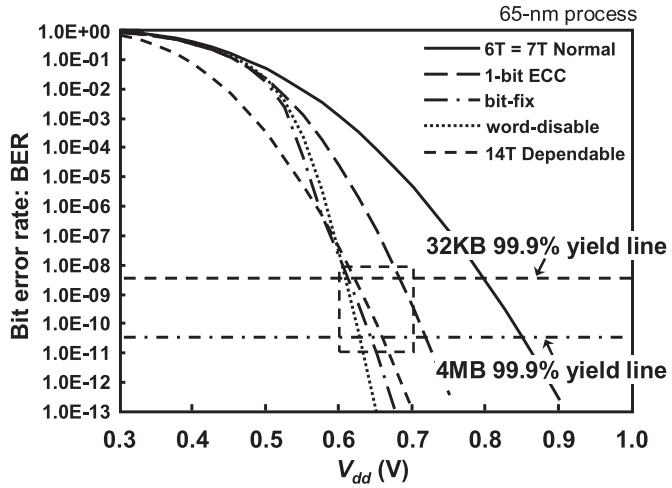

Fig. 3 BERs for 32-bit cache: "6T", "1-bit ECC", "bit-fix" and "word-disable" use conventional 6T bitcell schemes; "7T normal" and "14T dependable" use 7T/14T bitcells.

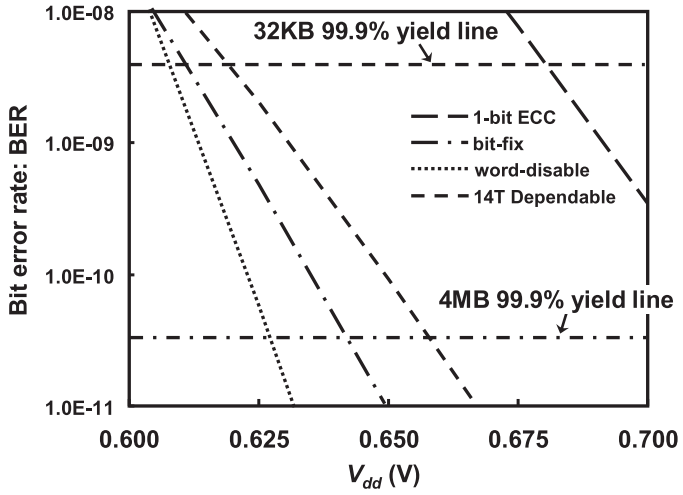

Fig. 4 BERs: magnifying the area bounded by the dashed line in Fig. 3.

Table 1 Two modes in 7T/14T bitcell.

	# of bitcells comprising 1 bit	# of WL drives	CL
Normal	1 (7T/bit)	1	Off ("H")
Dependable (write)	2 (14T/bit)	2	On ("L")
Dependable (read)	2 (14T/bit)	1	On ("L")

driven, but in a read operation, either WL0 or WL1 is asserted, which ensures stable operations.

In the normal mode, a one-bit datum is stored in one bitcell, which means that it is more area-efficient. In the dependable mode, a one-bit datum is stored in two bitcells, although the reliability of the information differs from that of the normal mode. The 'more dependable with less failure rate' information is obtainable by combining two bitcells [4]. In addition, the 14T dependable mode has better soft-error tolerance than the 7T normal mode because its internal node has more capacitance.

3.3 Bit Error Rates (BERs)

Figure 3 presents the bit error rates (BERs) simulated in a commercial 65-nm process. As described herein, the BER is referred as a metric in terms of the failure rate. The BERs in the 7T normal bitcell and the 14T dependable bitcell were obtained through Monte Carlo circuit simulation. The BERs in other scheme were obtained by probabilistic calculations using the above BERs in the 7T and 14T bitcells. Detailed descriptions of the probabilistic calculations are presented in the Appendix section. We also consider the worst-case parameters: temperature and a process corner.

Figure 4 portrays a magnified view of the area bounded by the dashed line in Fig. 3. Assuming 99.9% yield in 32-KB caches (999 good 32-KB caches out of 1,000), the respective V_{min} in the conventional 6T bitcell, one-bit ECC for a 32-bit word (= 32 bits + 6 correction bits) using 6T bitcells, the word-disable scheme, the bit-fix scheme, and the 14T dependable mode are 0.8 V, 0.685 V, 0.61 V, 0.615 V, and 0.620 V. Furthermore, assuming 99.9% yield in 4-MB cache, their V_{min} values respectively become 0.855 V, 0.72 V, 0.63 V, 0.645 V, and 0.66 V. The BER curve in the 7T normal mode is the same as that of the conventional 6T bitcells. The word-disable scheme can operate at lower V_{min} than the other schemes at both 32 KB and 4 MB sizes. In this simulation, the 14T dependable mode is applied uniformly to

© 2012 Information Processing Society of Japan

the entire cache (see Fig. 9 (a)); its BER slope is gentler than that of the word-disable scheme and the bit-fix scheme that exploits the word-grain control and the bit-grain control. Fine-grain control such as the word-grain control or the bit-grain control is more efficient than uniform control for a low BER at a low voltage because it can choose superior bitcells selectively and can abandon less-margin bitcells in the fine-grain region. However, the uniform control of the 14T dependable mode in this simulation uses all pairs of bitcells. Therefore, we apply fine-grain control to the 14T dependable mode in the next section.

4. Implementation of the 14T Word-Enhancing Scheme

In this section, we describe the proposed 14T word-enhancing scheme that enhances the operating margins of bitcells on the word-grain level. Then we will introduce incremental testing to improve the yield further. That is to say, the degree to which V_{min} is reduced using the proposed schemes will be demonstrated through comparison with the conventional word-disable scheme.

4.1 Conventional Word-disable Scheme

As described in Section 2, the word-disable scheme was proposed in an earlier report in the literature [2]. The word-disable scheme purges defective words, combines two cache lines in two consecutive ways, and thereby produces one logical cache line. Consequently, this scheme halves the cache size and associativity with cutting out of the defective words. Each way's tag has a defect word map as one-bit information that signifies a defective word (1) or a valid word (0). In a single 64-B cache line, it includes 16 sets of 32-bit words, which means that each cache line has the additional 16-bit defect word map in its tag.

Figure 5 portrays a comprehensive view of the cache word-disable scheme. A 16-word cache line is halved (Word0–Word7 and Word8–Word15). In every stage, a word shifter removes a defective word (or weak word). That is, four defective words are removed in all through the four stages. Four defect-free words (strong words) remain in each path. Eventually, 8 defect-free words are obtainable out of 16 by merging the two sets of 4 defect-free words.

Figure 6 presents a block diagram of a word shifter that removes defective words, and presents an example in which the second word is defective and removed. First, a defect vector ('01000') is extracted from the defect word map. The converting logic, similarly to a decoder, converts the 1-hot defect vector into a multiplexer control vector (0111) that controls four 32-bit 2:1 multiplexers to shift out the defective word.

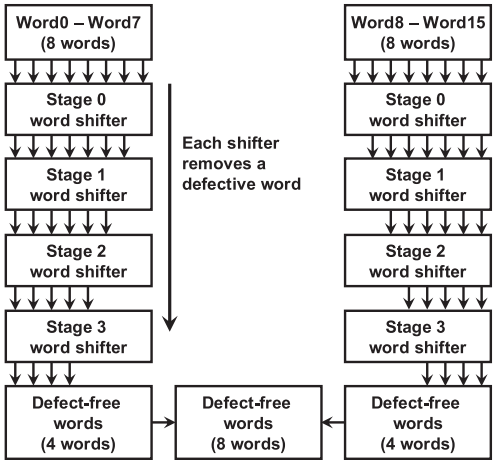

Fig. 5 Comprehensive view of the cache word-disable scheme.

4.2 Proposed 14T Word-enhancing Scheme with Divided Control Line

The proposed 14T word-enhancing scheme is a method to use of the 14T dependable mode for word-grain control. We assert a control line using a divided control line (DCL) scheme to select either the 7T normal mode or the 14T dependable mode on the word-grain level. The circuit function of the DCL scheme resembles the divided word-line (DWL) scheme [10]. The DCL scheme divides a global control line (GCL) into local control lines (LCLs) dedicated to each word. **Figure 7** depicts a schematic of the 7T/14T SRAM with the DCL scheme. A GCL and a control line selection (CLS) signal control an LCL on row-by-row and column-by-column bases. In addition, a global word line (GWL) is divided into local word lines (LWL), one of which is asserted by the GWL and a word line selection (WLS) signal in the same way. Dedicated decoders, which are controlled by a defect vector from the defect word map, assert a CLS and WLS signals.

4.3 Incremental Testing for the 14T Word-enhancing Scheme

Figure 8 portrays BERs including a word-level BER of the 14T word-enhancing scheme. The BER of the bit-fix scheme is removed. It is not included in the following comparison because the word-disable scheme is superior to the bit-fix scheme in terms of low-voltage operation and the cycle penalty.

On the 32-KB and 99.9% yield line, V_{min} of the 14T word-enhancing scheme is 0.605 V. On the 4-MB and 99.9% yield line, V_{min} is 0.62 V. As this figure shows, the 14T word-enhancing scheme yields only a small benefit compared to the conventional word-disable scheme because the BER of the 14T word-enhancing scheme is extracted from conventional testing without consideration of its features. Conventional testing means testing by lowering voltage, with subsequent checking to determine whether each bitcell fails or not.

The conventional scheme, which performs control on a whole block level, applies the 14T dependable mode uniformly to all word pairs, as portrayed in **Fig. 9** (a), whereas the 14T word-

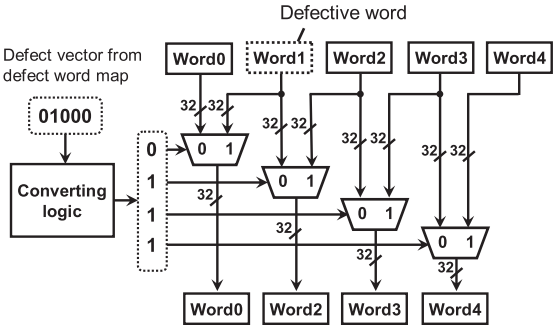

Fig. 6 Block diagram of a word shifter.

© 2012 Information Processing Society of Japan

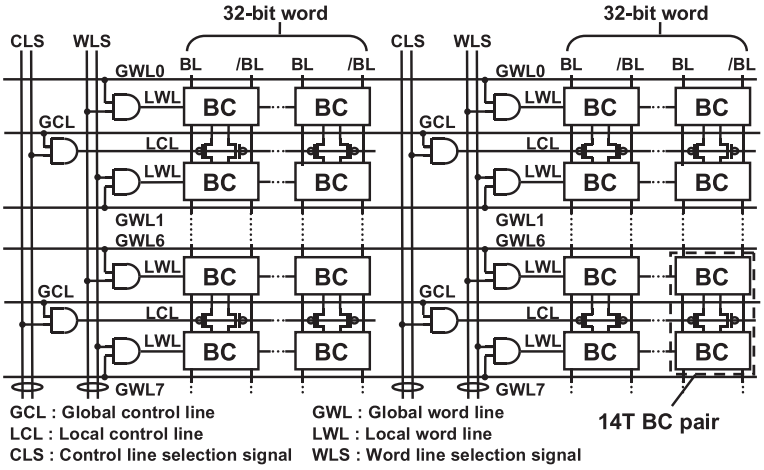

Fig. 7 7T/14T SRAM bitcell (BC) array with the divided control line (DCL) scheme.

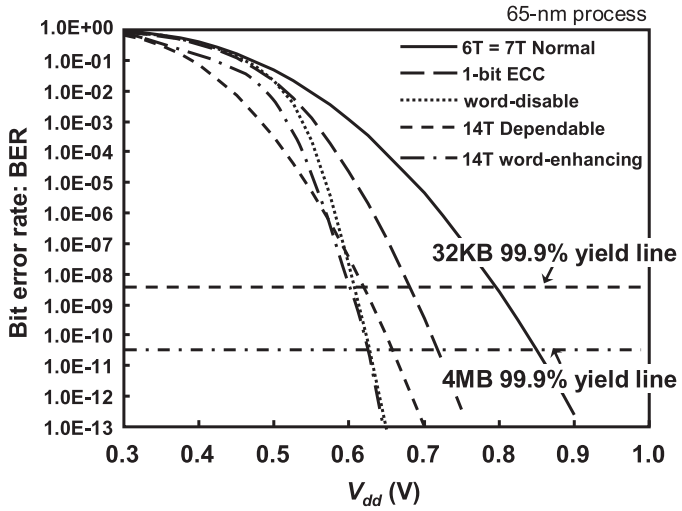

Fig. 8 BERs: including the 14T word-enhancing scheme with conventional testing.

enhancing scheme reinforces a defective word using another half of a pair connected to the word in a testing phase. In low-voltage testing, however, if both words in a 14T pair are recognized as defective words simultaneously at a certain voltage, then such a word pair cannot be applied to the 14T dependable mode, as shown in Fig. 9 (b). In fact, the 14T word-enhancing scheme can reduce its V_{min} efficiently in the case in which the 14T dependable mode is applied to all word pairs, as presented in Fig. 9 (c). To do so, we propose incremental testing that exploits the salient feature of the 14T dependable mode.

Incremental testing is based on the idea of applying the 14T dependable mode incrementally to the word pairs to maximize the number of word pairs. Incremental testing adopts one word pair on even and odd lines for the 14T dependable mode within a single execution of testing.

Figure 10 portrays a flow chart showing the incremental testing process. We take a step of an incremental V_{dd} as 50 mV [6]. First, the testing V_{dd} is set to a nominal voltage. Next, testing is executed to evaluate whether defective words are detected or not. If detected, then the 14T dependable mode is applied to the defective words: one word in a pair at most. Then testing is executed again for the updated 14T pair. If defective words are not detected, then the testing V_{dd} is decreased by 50 mV and testing continues. Before every testing execution, the number of disable words is checked to determine whether it equals or exceeds eight words (= half of the whole words in a cache line) or not. The incremental testing finishes if it is equal. If it is greater, then the number of disable words is limited to half for the cache line function, so that the 14T dependable mode is not applied to the excess words.

4.4 Improved BER in the 14T Word-enhancing Scheme

Figure 11 shows the BER of the 14T word-enhancing scheme with incremental testing. On the 32-KB and 99.9% yield line, V_{min} in the 14T word-enhancing scheme is improved further to 0.49 V. On the 4-MB and 99.9% yield line, it is 0.5 V, which is 42% and 21% lower, respectively, than the conventional 6T SRAM and the word-disable scheme. The figure shows that the 14T word-enhancing scheme with the incremental testing can reduce V_{min} effectively and that incremental testing is necessary for the 14T word-enhancing scheme.

4.5 Implementation

Figure 12 shows a layout plot of a 4-MB cache implemented with the 14T word-enhancing scheme using the 65-nm design rule.

© 2012 Information Processing Society of Japan

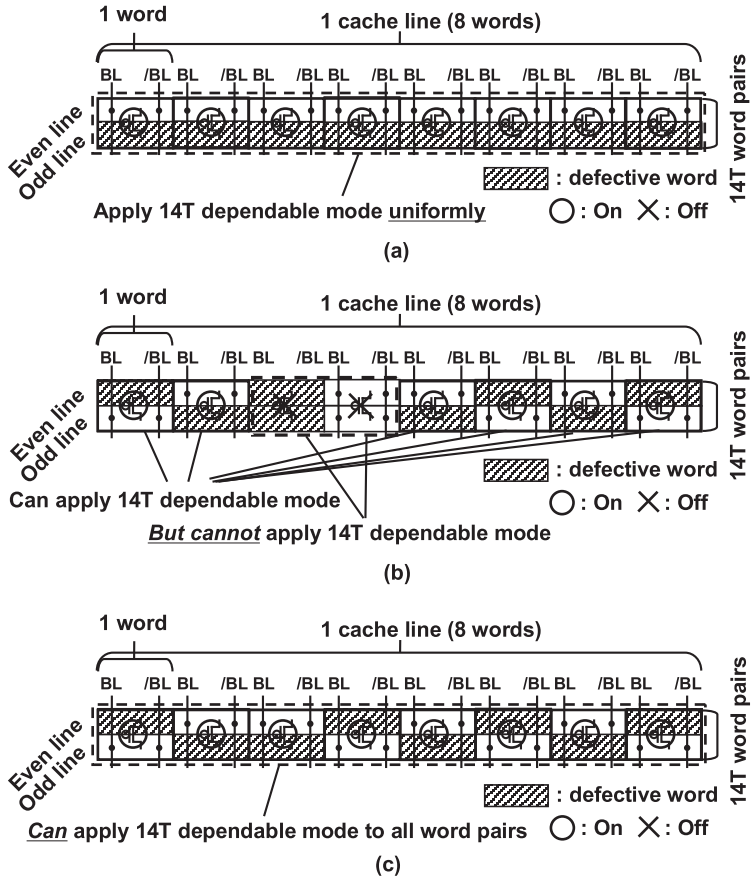

Fig. 9 Applying the 14T dependable mode in testing. These examples use eight-word cache lines for simplicity. Only asserted bitlines are shown. (a) Dependable mode is applied uniformly to all word pairs. (b) Conventional testing by which the 14T dependable mode is not applied to all word pairs. (c) Incremental testing, where the 14T dependable mode is applied to all word pairs.

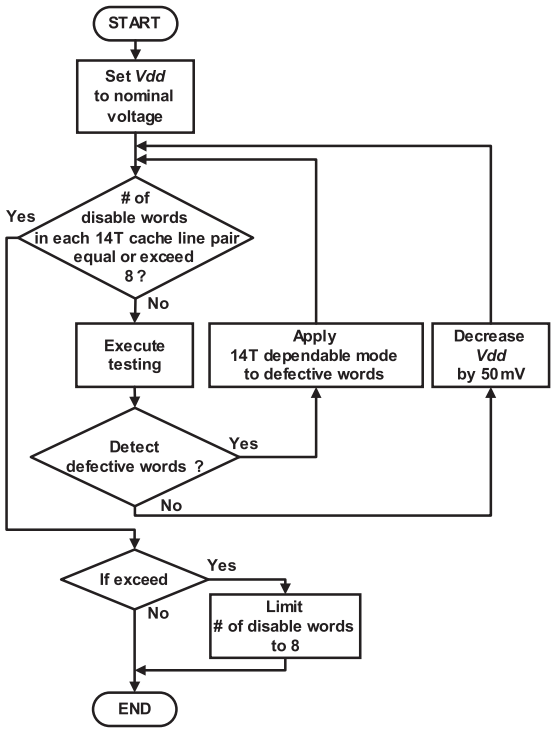

Fig. 10 Flow chart of incremental testing (this figure shows the case of an eight-word cache line).

The tags must also operate under 0.5 V. The word-disable scheme guarantees low-voltage operation capability in the tags by application of 10T sub-threshold (ST) bitcells [5]. The ST 10T bitcells, however, constitute a large area overhead. Instead, we implement a tag with large 6T bitcells that can suppress random (local) variation. The 6T bitcells for the tags are 1.3 times larger than normal 6T cells, which is 35% smaller than the ST 10T bitcell. The large 6T bitcell can operate reliably at 0.5 V.

The respective area overhead values attributable to the tags and DCL with the dedicated decoders are 4% and 8.9% of those in the conventional 6T SRAM. The total area overhead including the tags, the DCL with the dedicated decoders, and the 7T/14T SRAM, is 24% and 8% of the respective overhead values of the conventional 6T SRAM and the word-disable scheme.

4.6 Measurement Result

To show the voltage reduction in our scheme, we fabricated a 512-kb SRAM macro with the proposed 14T word-enhancing scheme in a 65-nm process. **Figure 13** shows a chip micrograph of the 512-kb SRAM macro with the proposed 14T word-enhancing scheme.

Figure 14 shows the measured BERs of the 6T normal, 14T dependable, and 14T word-enhancing schemes. The function of the incremental testing is conducted off the chip in this evaluation environment. The respective first failure bits of the 6T normal, 14T dependable, and 14T word-enhancing schemes come

© 2012 Information Processing Society of Japan

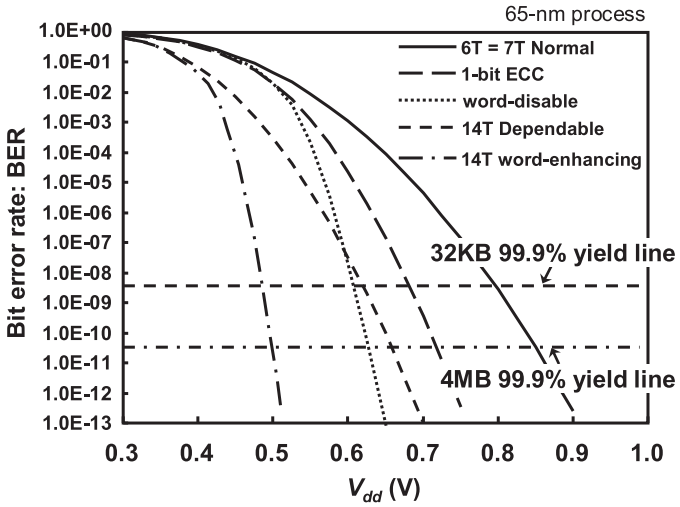

Fig. 11 Bit error rates (BERs): applying 14T word-enhancing with incremental testing.

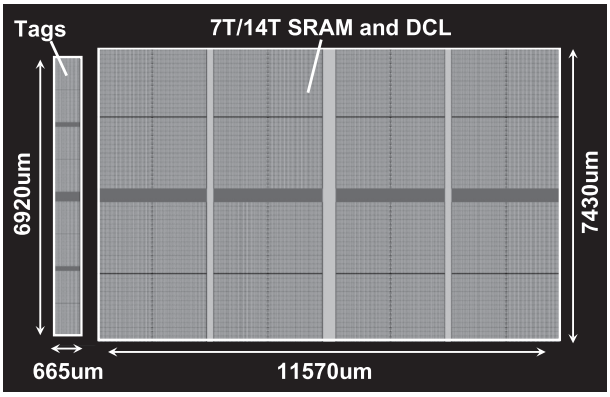

Fig. 12 Layout plot of a proposed 4-MB cache implemented with a 65-nm process.

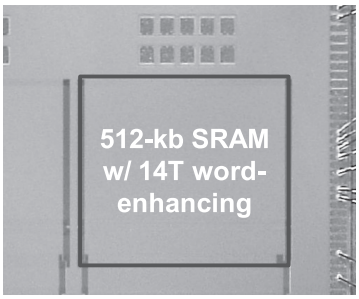

Fig. 13 Chip micrograph of the 512-kb SRAM macro with 14T word-enhancing scheme.

out at 0.53 V, 0.49 V, and 0.3975 V (i.e., the respective V_{min} are 0.5325 V, 0.4925 V, and 0.4 V). From this measurement, it is apparent that the 14T word-enhancing scheme can function effectively in a low-voltage region and reduce V_{min} under the variation of the fabricated 65-nm chip.

5. Performance, Energy, and Power Comparison

5.1 Performance Evaluation

In this section, we will make a performance comparison between the conventional scheme and the proposed scheme. The performance degradation derived from the additional latencies and the cache capacity reduction must be evaluated quantitatively.

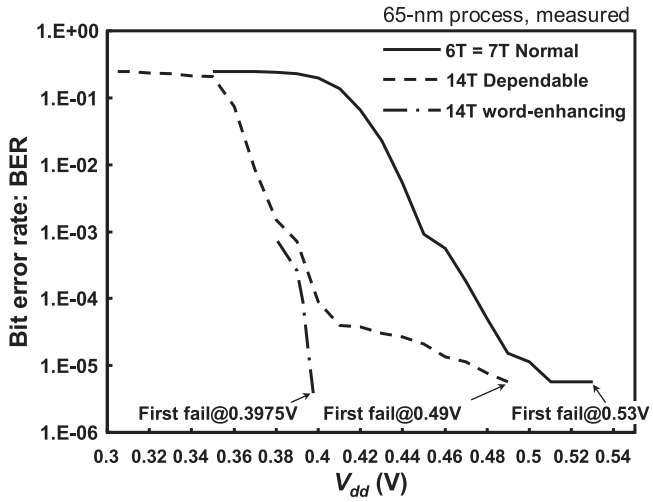

Fig. 14 Measured BERs of 6T, 14T dependable, and 14T word-enhancing scheme in 512-kb SRAM macro.

We used the SESC [7] cycle-accurate simulator. **Table 2** presents the architectural configuration parameters dependent on V_{dd} and the energies consumed on the cache in a single operation. The cache energies presented in Table 2 will be explained in Section 5.2.

We assumed a 20 FO4 gate delay for a single pipeline stage and obtained the operating frequencies in these 65-nm SPICE simulations. **Table 3** presents the architectural configuration parameters that are independent of V_{dd}. The 14T word-enhancing and the word-disable have access time overhead derived respectively from the dedicated decoder assertion of the CLS and WLS signals and the word-disable circuitry. Consequently, the 14T word-enhancing scheme and the word-disable scheme have a one-cycle penalty each for all cache accesses over the 6T normal mode.

We conducted SPEC2000 CINT (gzip, vpr, gcc, mcf, crafty, parser, gap, vortex, twolf)/CFP (wupwise, swim, mesa, ammp, equake) benchmarks and SPLASH2 benchmark [12] (fft, fmm, ocean, lu, radix, barnes, raytrace) as a performance evaluation. **Figure 15** presents normalized IPCs in the conventional scheme and the proposed scheme. The IPC reductions in the word-disable and 14T word-enhance schemes are, respectively, 3.8% and 3.7%,

© 2012 Information Processing Society of Japan

Table 2 Cache architecture configuration parameters dependent on V_{DD}: Energies of single operation are derived from 65-nm SPICE simulations and CACTI.

	6T normal		Word-disable		14T word-enhancing	
	High-voltage operation	Low-voltage operation	High-voltage operation	Low-voltage operation	High-voltage operation	Low-voltage operation
Vdd (supply voltage)	1.2 V	0.855 V	1.2 V	0.63 V	1.2 V	0.5 V
Frequency	2.6 GHz	1.7 GHz	2.6 GHz	900 MHz	2.6 GHz	500 MHz
DRAM access latency	260 cycles	170 cycles	260 cycles	90 cycles	260 cycles	50 cycles
L1$ read op. energy	0.187 nJ	0.095 nJ	0.267 nJ	0.072 nJ	0.188 nJ	0.033 nJ
L1$ write op. energy	0.181 nJ	0.092 nJ	0.256 nJ	0.071 nJ	0.183 nJ	0.032 nJ
L2$ read op. energy	0.984 nJ	0.500 nJ	1.059 nJ	0.299 nJ	0.992 nJ	0.172 nJ
L2$ write op. energy	0.892 nJ	0.453 nJ	0.969 nJ	0.298 nJ	0.895 nJ	0.155 nJ

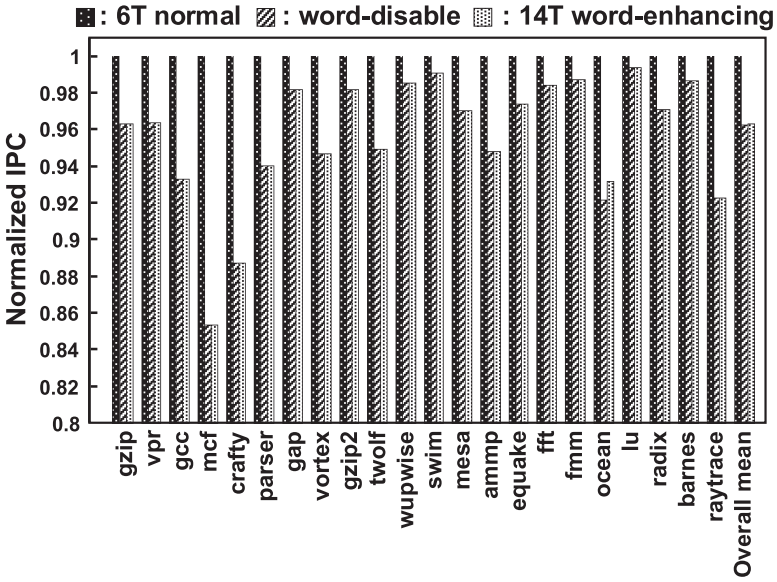

Fig. 15 Normalized IPCs in SPEC CPU2000 and SPLASH2 benchmarks.

Table 3 Architecture configuration parameters independent of V_{dd}.

# of cores	2
Technology	65-nm CMOS
L1 Instruction cache	32KB, 8-way, 2-cycle latency
L1 Data cache	32KB, 8-way, 2-cycle latency
Shared L2 cache	4MB, 8-way, 14-cycle latency
Cache line size	64B
Fetch / Issue / Retire	4/4/4
INT / FP registers	128/128

on average. They are almost identical.

5.2 Energy and Power Comparison

In the 14T dependable mode, internal nodes of the bitcell have almost double the capacitance of the 7T normal mode. However, the read energy in the 14T dependable mode does not increase from the 7T normal mode because the bitline current is the same as that of the 7T normal mode because the number of asserted wordlines is the same. Nevertheless, the write energy increases because charging and discharging the capacitance associated with the internal node increases. The energy consumed on the wordline is also increased because the number of asserted wordlines is doubled.

CACTI [11] was used to estimate energy overheads in the 14T dependable mode, word-disable and 14T word-enhancing schemes for the entire cache. Before the evaluation of cache energy, we first evaluated the write energies in the 7T normal mode and 14T dependable mode for a single 7T/14T bitcell by 65-nm SPICE simulations. The write energies per bitcell in the 7T normal mode and 14T dependable mode were, respectively, 5.5214 fJ at 1.2 V and 11.208 fJ at 1.2 V. Furthermore, we evaluated additional peripheral circuitry including the word shifter and additional dedicated decoders in the word-disable scheme, plus driving circuits of GCL, LCL, GWL, and LWL and additional dedicated decoders in the 14T word-enhancing scheme. By feeding back the energies in the bitcells and additional peripheral circuits to CACTI, the read and write operation energy per cache access is obtainable.

In Table 2, we assumed L1I, L1D, L2 caches in the 65-nm technology (LSTP for cell array and HP for peripheral circuitry) for the cache energy evaluation. The read and write energies of each cache are shown in Table 2. During high-voltage operation, compared with the read energy overhead of the 6T normal mode, those of the word-disable scheme and 14T word-enhancing scheme are, respectively, 40.05% and 0.32% for L1 caches, and

© 2012 Information Processing Society of Japan

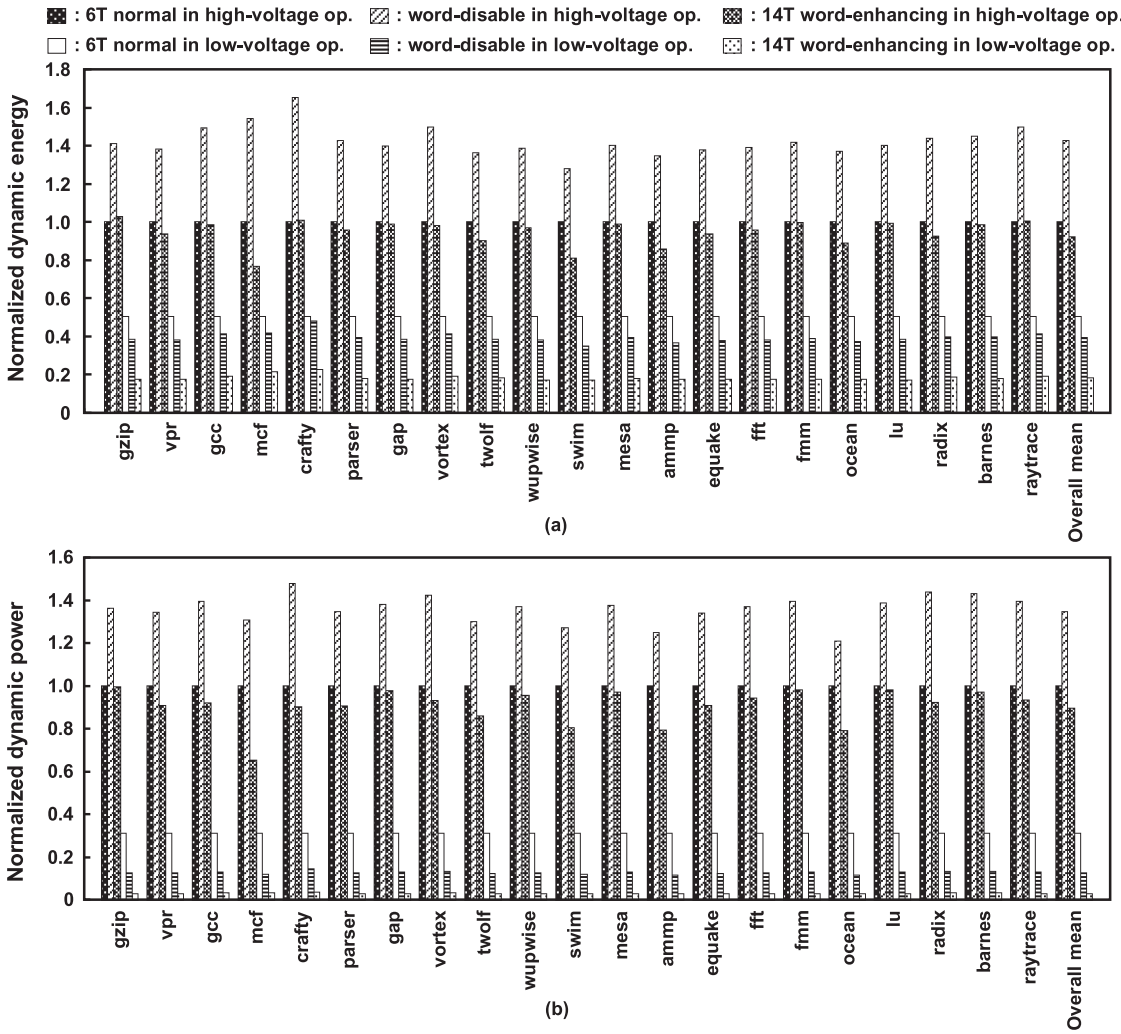

Fig. 16 Dynamic energy and dynamic power in high-voltage operation and low-voltage operation in SPEC2000 and SPLASH2 benchmarks. Each figure is normalized by 6T normal in high voltage operation and sums up the energies and powers on the L1D cache, L1I cache, and L2 cache: (a) normalized total energy in each benchmark, (b) normalized total power in each benchmark.

7.62% and 0.83% for L2 cache. The write energy overheads are, respectively, 41.48% and 1.22% for L1 caches, and 8.72% and 0.32% for L2 caches. In the word-disable scheme, the word shifter consumes great amounts of energy: 75.4 pJ per cache operation for each cache. Consequently, the word-shifter is a major contributor to the large energy overhead of the word-disable scheme. In contrast, the 14T word-enhancing scheme has a reasonable energy overhead even if the 14T dependable mode's write operation and the additional peripheral circuitry are considered.

Figure 16 (a) and (b) portray dynamic energy and dynamic power of 6T normal mode, word-disable, and 14T word-enhancing schemes in the high-voltage operation and low-voltage operation. The SPEC2000 and SPLASH2 benchmarks are used. Each figure is normalized by the 6T normal mode in the high-voltage operation and sums up energies and powers of the L1D, L1I, and L2 caches.

5.2.1 Overheads in high-voltage operation

The word-disable scheme in the high-voltage operation has 42.43% energy overhead and 34.33% power overhead on average, against the 6T normal mode. The word-disable consumes large amounts of dynamic energy and dynamic power because of its word shifter for the variation-aware low-voltage operation. In stark contrast, the 14T word-enhancing scheme in the high-voltage operation consumes 7.8% less energy and 10.54% less power on average, compared with the 6T normal mode. This difference results from the increase in cache access latency, which reduces the energy and power used in the 14T word-enhancing scheme.

5.2.2 Energy and power reduction in low-voltage operation

During low-voltage operation, the word-disable and 14T word-enhancing schemes respectively reduce dynamic energy usage by 22.23% and 63.1% compared with the 6T normal mode. The dynamic power reductions are, respectively, 58.66% and 89.22%. Each scheme in the low-voltage operation has a different frequency.

5.2.3 Leakage power

Leakage power is also calculated using CACTI, augmented with the data obtained in SPICE simulations. We also assumed a 65-nm LSTP process for a cell array and a 65-nm HP process for peripheral circuitry, as in the energy calculation. During high-voltage operation, the word-disable and 14T word-enhancing schemes consume 14.9% and 25.0% more leakage power than

© 2012 Information Processing Society of Japan

Table 4 Performance comparison: V_{min}, area, frequency, IPC, and power during low-voltage operation.

	6T cell	Word-disable	14T word-enhancing
V_{min} (mV)	855	630	500
Normalized area	1	1.15	1.24
Frequency (MHz)	1700	900	500
IPC	1.357	1.310	1.309
Normalized dynamic power	1	0.413	0.108
Normalized total power w/ HP peripheral	1	0.698	0.552
Normalized total power w/ LSTP peripheral	1	0.415	0.111

6T normal consumes. The respective increase in leakage power of the word-disable and 14T word-enhancing schemes is caused mainly by the increase in the number of transistors and area. During low-voltage operation, the respective leakage power reductions of the word-disable and the 14T word-enhancing schemes are 27.1% and 40.0%.

5.2.4 Total power

Total power includes the dynamic power and leakage power. During high-voltage operation, the total power used by the word-disable is higher by 17.9% and that used by 14T word-enhancing schemes is higher by 19.6%. During low-voltage operation, however, they are lower, respectively, by 30.2% and 44.8%.

Additionally, we estimate the total power considering the 65-nm LSTP process for both the cell array and peripheral circuitry assuming a low-power mobile processor. During high-voltage operation, the total power of the word-disable scheme is higher by 34.3%, and the total power of the 14T word-enhancing scheme is lower by 10.2%. During low-voltage operation, they are reduced, respectively, by 58.5% and 88.9%. The average ratio of the dynamic power to the leakage power is 1:8.48 for the LSTP process for the cell array in the HP process for the peripheral circuitry. The average ratio is 3400:1 for the LSTP process for both the cell array and peripheral circuitry. The leakage power is dominant in the former case. The dynamic power is dominant in the latter case.

Table 4 presents a comparison of the performance of the conventional schemes and the proposed 14T word-enhancing scheme during low-voltage operation. Our proposed scheme can reduce the minimum dynamic power significantly, by 89.2% and 73.9%, respectively, compared to the conventional 6T cell and the word-disable scheme. It can also reduce the total power consumption of the LSTP cell array and the HP peripheral by 44.8% and 20.9%, respectively, and reduce the total power consumption of the LSTP cell array and the LSTP peripheral by 88.9% and 73.2%, respectively.

Wider-range power scaling is possible when using the proposed scheme, which is suitable for low-power mobile devices that have a low-power operation mode with DVFS.

6. Conclusion

We proposed a 14T word-enhancing scheme that lowers V_{min}. It uses a 7T/14T SRAM with divided control lines. The proposed incremental testing expands the efficiency of the 14T word-enhancing scheme, and it can further reduce V_{min}. The proposed architecture achieves V_{min} reduction of 42% and 21%, respectively, for a 4-MB cache compared to the conventional 6T SRAM and the word-disable scheme. Measurement of a 512-kb macro implemented with the 14T word-enhancing scheme revealed 25% and 19% lower V_{min}, respectively, than in the 6T normal mode and 14T dependable mode. The minimum dynamic power was 89.2% and 73.9% lower, and the minimum total power was lower by 44.8% and 20.9%.

Acknowledgments This work was supported by VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Cadence Design Systems, Mentor Graphics and Synopsys, Inc. The authors appreciate Prof. Itsuro Kakiuchi with Kobe University for valuable technical discussions.

Reference

[1] Itoh, K.: Low-voltage scaling limitations for nanoscale CMOS LSIs, *International Conference on Ultimate Integration of Silicon* (*ULIS*), pp.3–6 (Mar. 2008).
[2] Wilkerson, C., Gao, H., Alameldeen, A.R., Chishti, Z., Khellah, M. and Lu, S.-L.: Trading off Cache Capacity for Reliability to Enable Low Voltage Operation, *International Symposium on Computer Architecture* (*ISCA*), pp.203–214 (June 2008).
[3] Ozdemir, S., Sinha, D., Memik, G., Adams, J. and Zhou, H.: Yield-Aware Cache Architectures, *IEEE/ACM International Symposium on Microarchitecture* (*MICRO*), pp.15–25 (Dec. 2006).
[4] Fujiwara, H., Okumura, S., Iguchi, Y., Noguchi, H., Kawaguchi, H. and Yoshimoto, M.: A Dependable SRAM with 7T/14T Memory Cells, *IEICE Trans. Electronics*, Vol.E92-C, No.4, pp.423–432 (Apr. 2009).
[5] Kulkarni, J.P., Kim, K. and Roy, K.: A 160 mV Robust Schmitt Trigger Based Subthreshold SRAM, *IEEE Journal of Solid-State Circuits*, Vol.42, No.10, pp.2303–2313 (Oct. 2007).
[6] Stackhouse, B., Bhimji, S., Bostak, C., Bradley, D., Cherkauer, B., Desai, J., Francom, E., Gowan, M., Gronowski, P., Krueger, D., Morganti, C. and Troyer, S.: A 65 nm 2-Billion Transistor Quad-Core Itanium Processor, *IEEE Journal of Solid-State Circuits*, Vol.44, No.1, pp.18–31 (Jan. 2009).
[7] Renau, J., Fraguela, B., Tuck, J., Liu, W., Prvulovic, M., Ceze, L., Strauss, K., Sarangi, S., Sack, P. and Montesinos, P.: SESC Simulator (Jan. 2005), available from ⟨http://sesc.sourceforge.net⟩.
[8] Seevinck, E., List, F.J. and Lohstroh, J.: Static-noise margin analysis of MOS SRAM cells, *IEEE Journal of Solid-State Circuits*, Vol.22, No.5, pp.748–754 (Oct. 1987).
[9] Heald, R. and Wang, P.: Variability in sub-100 nm SRAM designs, *IEEE/ACM International Conference on Computer Aided Design* (*ICCAD*), pp.347–352 (Nov. 2004).
[10] Yoshimoto, M., Anami, K., Shinohara, H., Yoshihara, T., Takagi, H., Nagao, S., Kayano, S. and Nakano, T.: A divided word-line structure in the static RAM and its application to a 64 K full CMOS RAM, *IEEE Journal of Solid-State Circuits*, Vol.18, No.5, pp.479–485 (Oct. 1983).
[11] Thoziyoor, S., Muralimanohar, N., Ahn, J.H. and Jouppi, N.: CACTI 5.1, Technical Report HPL-2008-20, Hewlett Packard Labs (Apr. 2008).
[12] Woo, S.C., Ohara, M., Torrie, E., Singh, J.P. and Gupta, A.: The SPLASH-2 programs: characterization and methodological considerations, *International Symposium on Computer Architecture* (*ISCA*), pp.24–36 (June 1995).
[13] Nakata, Y., Okumura, S., Kawaguchi, H. and Yoshimoto, M.: 0.5-V operation variation-aware word-enhancing cache architecture using 7T/14T hybrid SRAM, *ACM/IEEE International Symposium on Low-Power Electronics and Design* (*ISLPED*), pp.219–224 (Aug. 2010).

Appendix

A.1 Probabilistic BER calculations

Procedures used for probabilistic BER calculations for the one-bit ECC, the bit-fix scheme [2], the word-disable scheme, and the 14T word-enhancing scheme are explained below.

© 2012 Information Processing Society of Japan

First, the procedure for the one-bit ECC is introduced and explained. The one-bit ECC can fix a one-bit error in a single word. The BER of the one-bit ECC for an n-bit word can be expressed as a binomial expression

$$BER(1 - bit\_ECC(n)) = 1-((1-BER(6T))^n+n\times((1-BER(6T))^{n-1}\times BER(6T)))^{1/n}, \quad (A.1)$$

where $BER(6T)$ denotes the BER for a single 6T bitcell.

Second, in the probabilistic BER calculation of the bit-fix scheme, the bit-fix scheme in the literature has 10 sets of two patch bits per 512-bit cache line. Therefore the bit-fix scheme can repair 10 defects per 512-bit cache line by replacing the 10 defects with the 10 sets of two patch bits. In principle, one patch bit is sufficient to fix one defect, but the address pointing to the defect requires nine bits ($512 = 2^9$) in this case. Therefore, in the literature, the two patch bits are adopted, which can repair two consecutive bits with an eight-bit address ($512/2 = 2^8$). The 10 bits (two patch bits and eight address bits are further encoded to one-bit-correction ECC, in which four bits are added to the ten bits and the total bits becomes 14 per defect (a one-bit defect and a two-consecutive-bit defect can be corrected by the 14 bits). The BER of the bit-fix scheme is expressed as shown below.

$$BER(Bit-fix) = 1 - \left(\sum_{i=0}^{10}\binom{256}{i}\times(1-BER(6T))^{2\times(256-i)}\times BER(6T)^{2\times i} \times BER(1-bit\_ECC(14))^{2\times i}\right)^{1/512} \quad (A.2)$$

Next, the probabilistic BER calculation for the word-disable scheme is introduced. The word-disable scheme can remove eight defective words from 16 words in one way. In a 512-bit cache line in one way, 16 sets of 32-bit words exist. The 16-word cache line is divided into two halves. The word-disable scheme can then remove four defective words from eight words in the two halves. The BER for the word disable scheme is therefore expressed as follows.

$$BER(Word-disable) = 1 - \left(\sum_{i=0}^{4}\binom{8}{i}\times(1-BER(6T))^{32\times(8-i)}\times BER(6T)^{32\times i}\right)^{2/512} \quad (A.3)$$

Finally, the probabilistic BER calculation for the 14T word-enhancing scheme is introduced. Actually, the BER of the 14T word-enhancing scheme can be expressed similarly to that for the word-disable as

$$BER(14T\_word-enhancing) = 1 - \left(\sum_{i=0}^{4}\binom{8}{i}\times(1-BER(14T))^{32\times(8-i)} \times(BER(14T))^{32\times i}\right)^{2/512}, \quad (A.4)$$

where $BER(14T)$ denotes a BER for a single 14T bitcell.

Yohei Nakata received his B.E. and M.E. degrees in Computer and Systems Engineering from Kobe University, Hyogo, Japan in 2008 and 2010, respectively, where he is currently pursuing a Ph.D. degree in Engineering. His current research interests include low-power and dependable processor designs and multi-core processor architecture. He is a student member of IPSJ, IEICE, and IEEE.

Shunsuke Okumura received his B.E. and M.E. degrees in Computer and Systems Engineering from Kobe University, Hyogo, Japan in 2008 and 2010, respectively, where he is currently pursuing a Ph.D. degree in Engineering. His current research interests include high-dependability and low-power SRAM designs. He is a student member of IPSJ, IEICE, and IEEE.

Hiroshi Kawaguchi received his B.E. and M.E. degrees in electronic engineering from Chiba University, Chiba, Japan, in 1991 and 1993, respectively, and Ph.D. degree in engineering from the University of Tokyo, Tokyo, Japan, in 2006. He joined Konami Corporation, Kobe, Japan, in 1993, where he developed arcade entertainment systems. He moved to the Institute of Industrial Science, the University of Tokyo, as a Technical Associate in 1996, and was appointed a Research Associate in 2003. In 2005, he move to the Department of Computer and Systems Engineering, Kobe University, Kobe, Japan, as a Research Associate. Since 2007, he has been an Associate Professor with the Department of Computer Science and Systems Engineering, Kobe University. He is also a Collaborative Researcher with the Institute of Industrial Science, the University of Tokyo. His current research interests include low-power VLSI design, hardware design for wireless sensor network, and recognition processor. Dr. Kawaguchi was a recipient of the IEEE ISSCC 2004 Takuo Sugano Outstanding Paper Award and the IEEE Kansai Section 2006 Gold Award. He has served as a Program Committee Member for IEEE Symposium on Low-Power and High-Speed Chips (COOL Chips), and as a Guest Associate Editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. He is a member of IPSJ, IEEE and ACM.

© 2012 Information Processing Society of Japan

Masahiko Yoshimoto received his B.S. degree in electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1975, and M.S. degree in electronic engineering from Nagoya University, Nagoya, Japan, in 1977. He received a Ph.D. degree in Electrical Engineering from Nagoya University, Nagoya, Japan in 1998. He joined the LSI Laboratory, Mitsubishi Electric Corp., Itami, Japan, in April 1977. From 1978 to 1983 he was engaged in the design of NMOS and CMOS static RAM including a 64 K full CMOS RAM with the world's first divided-wordline structure. From 1984, he was involved in research and development of multimedia ULSI systems for digital broadcasting and digital communication systems based on MPEG2 and MPEG4 Codec LSI core technology. Since 2000, he has been a Professor of the Department of Electrical and Electronic Systems Engineering at Kanazawa University, Japan. Since 2004, he has been a Professor of the Department of Computer and Systems Engineering at Kobe University, Japan. His current activity is focused on research and development of multimedia and ubiquitous media VLSI systems including an ultra-low-power image compression processor and a low power wireless interface circuit. He holds 70 registered patents. He served on the Program Committee of the IEEE International Solid State Circuit Conference from 1991 to 1993. In addition, he has served as a Guest Editor for special issues on Low-Power System LSI, IP, and Related Technologies of IEICE Transactions in 2004. He received the R&D100 awards from R&D Magazine for development of the DISP and development of a real-time MPEG2 video encoder chipset in 1990 and 1996, respectively.

(Recommended by Associate Editor: *Takashi Sato*)

© 2012 Information Processing Society of Japan