0.5-V Operation Variation-Aware Word-Enhancing Cache Architecture Using 7T/14T hybrid SRAM

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

ABSTRACT

This paper presents a novel cache architecture using 7T/14T hybrid SRAM, which can dynamically improve its reliability with control lines. Our proposed 14T word-enhancing scheme can enhance its operating margin in word granularity by combining two words in a low-voltage mode. The proposed scheme is suitable for dynamic voltage and frequency scaling (DVFS). In a 65-nm process, it can reduce the minimum operation voltage (V_{min}) to 0.5 V, which is 42% and 21% lower, respectively, than the conventional 6T SRAM and the cache word-disable scheme. The respective power reductions are 90% and 65%.

Categories and Subject Descriptors

B.3.2 [**Memory Structures**]: Design Styles – *Cache memories*

General Terms

Design

Keywords

Cache memory, microarchitecture, variation, low voltage, low power, fine-grain control.

1. INTRODUCTION

As process technology advances, its minimum feature size decreases, which enables manufacturing with higher density and lower costs. 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 due to power-supply noise, IR drops, and/or soft errors. Because the dynamic power is proportional to the square of an operating voltage (V_{dd}), V_{min} is an important parameter in power dissipation. 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 a 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 transistor sizes are smaller. To make matters worse, the capacity of SRAM bitcells on a processor is huge. Consequently, large SRAM blocks such as an L1 data/instruction caches and last level cache (LLC) determine the V_{min} on the processor.

ISLPED'10, August 18–20, 2010, Austin, Texas, USA.

The V_{th} variation in each SRAM bitcell is distributed randomly in the whole SRAM block, which is called 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. For this reason, to reduce V_{min}, fine-grain control must be applied for the SRAM block that adaptively addresses the V_{th} variations.

In this paper, we present a word-level enhancing scheme for a large-capacity cache, using 7T/14T SRAM. 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.

2. RELATED WORKS

Wilkerson *et al.*, proposed the cache word-disable scheme (hereinafter "the word-disable scheme") 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 among eight words. A defect word map (one-bit information per word) that represents 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 the literature.

The bit-fix scheme exploits one of the ways 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 area overhead is smaller than the word-disable scheme. However, the bit-fix scheme takes a three-cycle penalty, whereas that in the word-disable scheme is a one-cycle penalty. In low-voltage operation, the reliability in the redundant way is lowered as well 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 a failure rate is increased rapidly in the redundancy way. Even ECC cannot fix it.

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

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 a 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. 7T/14T HYBRID SRAM

3.1 Failures in SRAM

Failures in SRAM are categorizable as a read margin failure, write margin failure, and access time violation.

- **Read margin failure:** read operation is signified by a read static noise margin (read SNM) [9]. 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:** write operation is explainable by a write-trip point (WTP) as a metric (= write margin) [10]. The WTP indicates the maximum voltage that can write "0" to a bitcell and can then flip an internal datum.
- **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 considered in this paper because it is dependent on the clock frequency. The read SNM and the write margin are dominant at a low frequency.

3.2 7T/14T hybrid SRAM

Figure 1 depicts the 7T bitcell (14T for two bitcells) [5]. Two pMOSes are added to internal nodes ("N00 and N10", "N01 and N11") in a pair of the conventional 6T bitcells shown in Figure 2. The area overhead in the 7T bitcell is 11% greater than that of the conventional 6T bitcell.

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

- 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 write operation, both WL0 and WL1 are driven, but in 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 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 [5]. In addition, the 14T dependable mode has better soft-error tolerance than the 7T normal mode because its internal node has more capacitance.

Table I: 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")

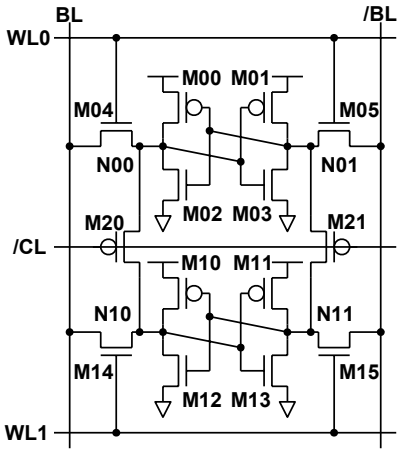

Figure 1. A 7T/14T bitcell pair.

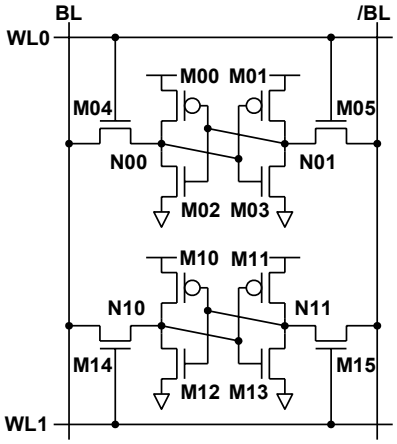

Figure 2. Conventional 6T bitcells.

3.3 Bit Error Rates (BERs)

Figure 3 shows 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 by Monte Carlo circuit simulation. The BERs in other scheme were obtained by probabilistic calculations using the above BERs in the 7T and 14T bitcells. 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 Figure 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 the conventional 6T bitcells. The word-disable scheme can operate at lower V_{min} than the other schemes in both of 32 KB and 4 MB. In this simulation, the 14T dependable mode is applied to the whole cache uniformly (see Figure 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 the uniform control for a low BER at a low voltage because it can pick superior bitcells selectively and abandon less-margin bitcells in the fine-grain region. On the other hand, the uniform control of the 14T dependable mode in this simulation uses all pairs of bitcells. Therefore, we will apply the fine-grain control to the 14T dependable mode in the next section.

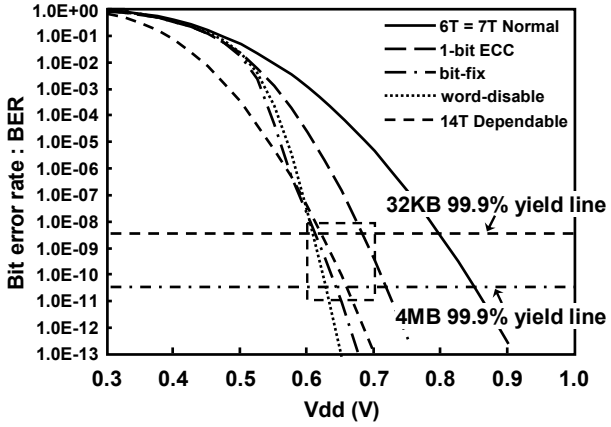

Figure 3. BERs for 32-b cache: "6T", "1-bit ECC", "bit-fix", and "word-disable" use conventional 6T bitcell schemes; "7T normal" and "14T dependable" use 7T/14T bitcells.

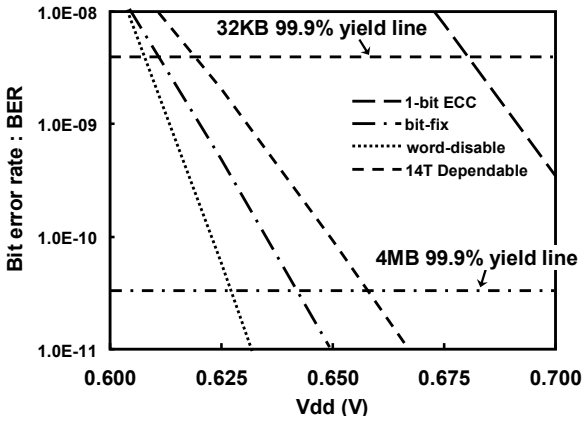

Figure 4. BERs: magnifying the area bounded by the dashed line in Figure 3.

4. IMPLEMENTATION OF THE 14T WORD-ENHANCING SCHEME

In this section, we will 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: namely how much the V_{min} is reduced using the proposed schemes will be exhibited by comparison with the conventional word-disable scheme.

4.1 Conventional word-disable scheme

As described in Section 2, the word-disable scheme was proposed by [2]. The word-disable scheme purges defective words, combines two cache lines in two consecutive ways and makes one logical cache line. Consequently, this scheme halves the cache size and associativity with cutting out 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 shows the overall view of the cache word-disable scheme. A 16-word cache line is divided into two halves (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 four defect-free words.

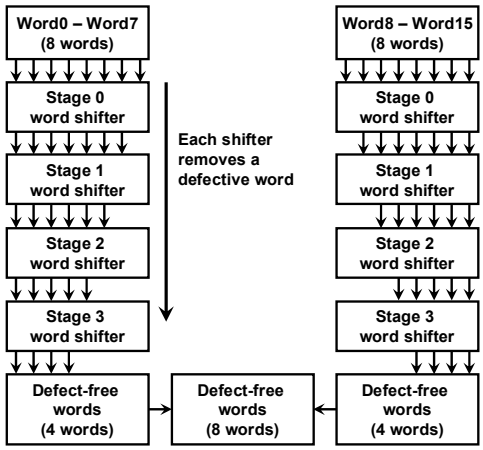

Figure 5. Overall view of the cache word-disable scheme.

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

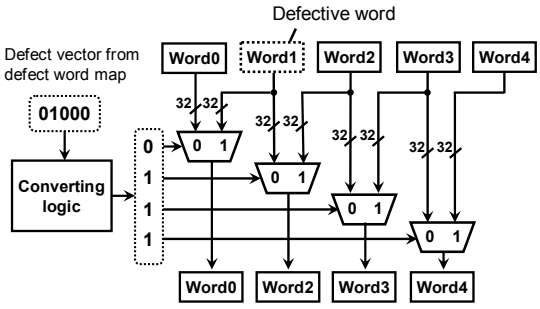

Figure 6. Block diagram of a word shifter.

4.2 Proposed 14T word-enhancing scheme with divided control line

The proposed 14T word-enhancing scheme combines the 7T/14T SRAM with the word-disable scheme to make 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 of 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 [11]. The DCL scheme divides a global control line (GCL) into local control lines (LCLs) dedicated to each word. Figure 7 shows 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. Also, 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. A CLS and WLS signals are asserted by dedicated decoders that are controlled by a defect vector from the defect word map.

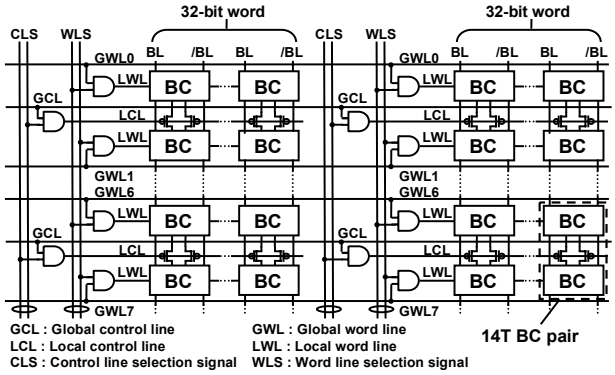

Figure 7. 7T/14T SRAM bitcell (BC) array with the divided control line (DCL) scheme.

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 and 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 size and 99.9% yield line, V_{min} of the 14T word-enhancing scheme is 0.605 V. On the 4-MB size and 99.9% yield line, V_{min} is 0.62 V. This figure shows that 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 considering its features. The conventional testing means testing that decreasingly lowers voltage and then checks if each bitcell fails or not.

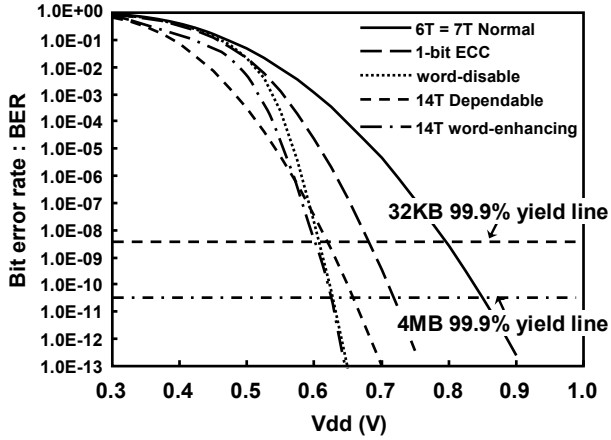

Figure 8. BERs: including the 14T word-enhancing scheme with conventional testing.

The conventional scheme which controls on a whole block level applies the 14T dependable mode to all word pairs uniformly, as illustrated in Figure 9 (a), while the 14T word-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 illustrated in Figure 9 (b). In fact, the 14T word-enhancing scheme can reduce its V_{min} efficiently in the case where the 14T dependable mode is applied to all word pairs, as presented in Figure 9 (c). For doing so, we will propose incremental testing that exploits the salient feature of the 14T dependable mode.

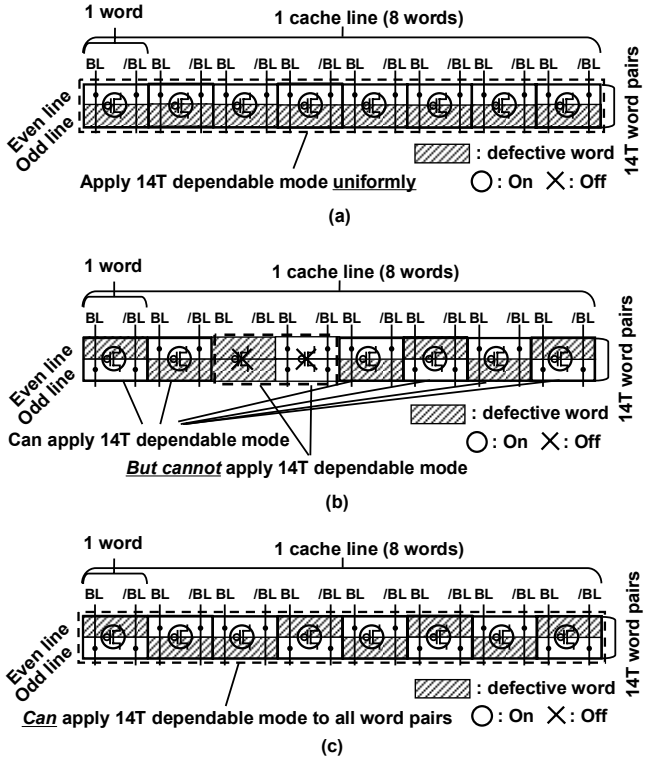

Figure 9. Applying the 14T dependable mode in testing (these examples use eight-word cache lines for simplicity and only asserted bitlines are shown): (a) dependable mode is applied to all word pairs uniformly, (b) conventional testing where the 14T dependable mode is not applied to all word pairs, and (c) incremental testing where the 14T dependable mode is applicable to all word pairs.

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

Figure 10 shows a flow chart of the incremental testing. We take a step of an incremental V_{dd} as 50 mV [7]. 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 at most one word in a pair; then testing is executed for the updated 14T pair again. If not detecting defective words, then the testing V_{dd} is decreased by the 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 a half to function the cache line, so that the 14T dependable mode is not applied to the exceeded words.

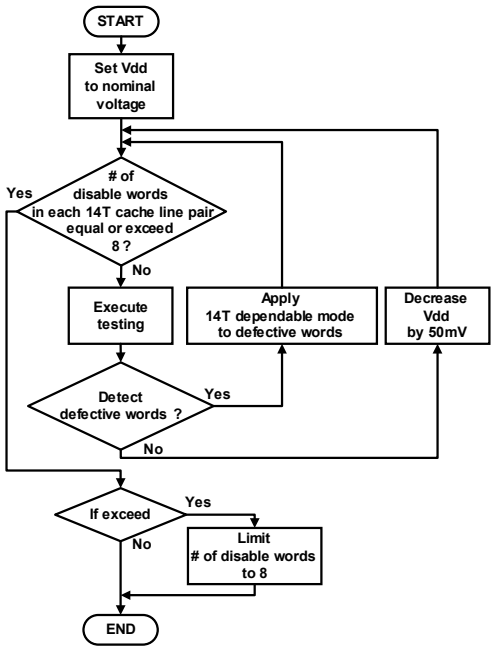

Figure 10. Flow chart of incremental testing (this figure shows in the case of eight-word cache line).

4.4 Improved BER in the 14T word-enhancing scheme

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

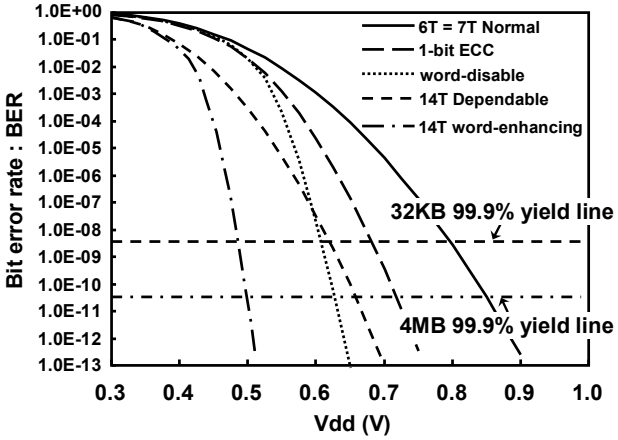

Figure 11. Bit error rates (BERs): applying 14T word enhancement with incremental testing.

4.5 Implementation

Figure 12 shows a layout plot of a 4-MB cache implemented with the 14T word-enhancing scheme including word shifters using the 65-nm design rule. The word shifters have 147,200 transistors; the additional latency (delay penalty) for the cache access is less than 300 ps, which is comparable to a delay of 20 fanout-4 (FO4) gates.

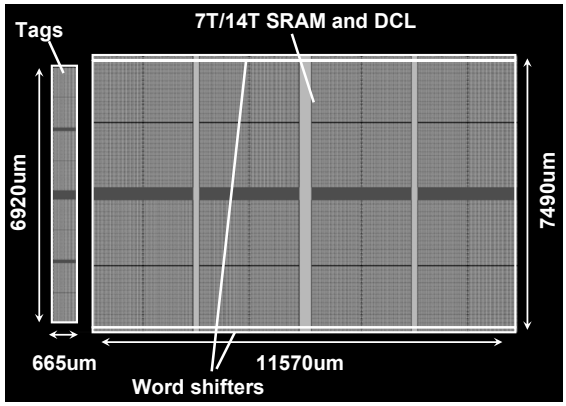

Figure 12. Layout plot of a proposed 4-MB cache implemented with a 65-nm process.

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 [6]. The ST 10T bitcells, however, incur a large area overhead. Instead, we implement the tag with large 6T bitcells that can suppress random (local) variation. The 6T bitcells for the tags are 1.3 times larger than a normal 6T cells, which is 35% smaller than the ST 10T bitcell. The large 6T bitcell can operate reliably under 0.5 V.

The respective area overhead attributable to the tags and DCL are 4% and 9.2% compared with the conventional 6T SRAM. Regarding the word shifters, no area overhead exists compared with the word-disable scheme. The total area overhead including the tags, the DCL, and the 7T/14T SRAM, is 24% and 8%, compared with the conventional 6T SRAM and the word-disable scheme.

5. 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. We used the SESC [8] cycle-accurate simulator.

Table II shows the architectural configuration parameters dependent on V_{dd}. We assumed a 20 FO4 gate delay for a single pipeline stage and obtained the operating frequencies in these 65-nm SPICE simulations. Table III presents the architectural configuration parameters that are not dependent on V_{dd}. The 14T word-enhancing scheme and the word-disable scheme have a one-cycle penalty each for all caches' accesses over the baseline.

We conducted SPEC2000 CINT/CFP benchmarks and SPLASH2 benchmark as a performance evaluation. Figure 13 shows normalized IPCs in the conventional scheme and the proposed scheme. The IPC reductions in the word-disable and 14T word-enhance schemes are 3.8% and 3.7%, respectively, on average. They are almost identical.

Table II. Architecture configuration parameters dependent on V_{dd}: the baseline signifies the normal 6T SRAM in high-voltage operation

	High-voltage operation (baseline)	Low-voltage operation w/ word-disable	Low-voltage operation w/ 14T word-enhancing
Vdd (supply voltage)	1.2V	0.63V	0.5V
Frequency	2.6GHz	900MHz	500MHz
Memory access latency	260 cycles	90 cycles	50 cycles

Table III. Architecture configuration parameters not dependent on 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

Table IV. Performance comparison: V_{min}, area, frequency, power, and IPC.

	6T cell	Word-disable	14T word-enhancing
Vmin (mV)	855	630	500
Normalized area	1	1.15	1.24
Frequency (MHz)	1700	900	500
Normalized power	1	0.29	0.1
IPC	1.36	1.31	1.31

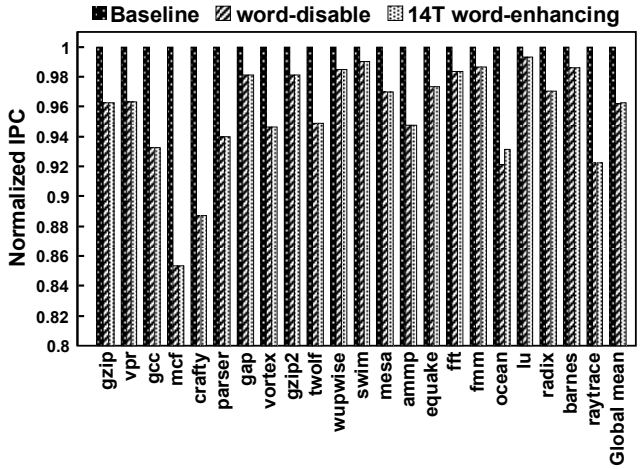

Figure 13. Normalized IPCs in SPEC CPU2000 and SPLASH2 benchmarks.

Table IV presents a comparison between the conventional schemes and the proposed 14T word-enhancing scheme. Our proposed scheme can significantly reduce the operating power consumption by 90% and 65% compared to the conventional 6T cell and the word-disable scheme. The wider-range power scaling can be enjoyed in the proposed scheme, which is suitable to low-power mobile devices that have a low-power operation mode with DVFS.

6. CONCLUSION

We proposed the 14T word-enhancing scheme that lowers V_{min}. It uses 7T/14T SRAM with the divided control lines. The proposed incremental testing expands the efficiency of the 14T word-enhancing scheme and can further reduce V_{min}. The proposed architecture achieves V_{min} reduction of 42% and 21% compared to the conventional 6T SRAM and the word-disable scheme, respectively. The respective power is reduced dramatically by 90% and 65%.

7. REFERENCES

[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.; Lu, S.-L., "Trading off Cache Capacity for Reliability to Enable Low Voltage Operation," *International Symposium on Computer Architecture (ISCA)*, pp. 203-214, Jun. 2008

[3] Ozdemir, S.; Sinha, D.; Memik, G.; Adams, J.; Zhou, H., "Yield-Aware Cache Architectures," *Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 15-25, Dec. 2006.

[4] Agarwal, A.; Paul, B.C.; Mahmoodi, H.; Datta, A.; Roy, K., "A process-tolerant cache architecture for improved yield in nanoscale technologies," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 1, pp. 27-38, Jan. 2005.

[5] Fujiwara, H.; Okumura S.;, Iguchi, Y.; Noguchi, H.; Kawaguchi, H.; and Yoshimoto, M., "A Dependable SRAM with 7T/14T Memory Cells," *IEICE Transaction. on Electronics*, vol. E92-C, no. 4, pp. 423-432, Apr. 2009.

[6] Kulkarni, J.P.; Kim, K.; 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.

[7] Stackhouse, B.; Bhimji, S.; Bostak, C.; Bradley, D.; Cherkauer, B.; Desai, J.; Francom, E.; Gowan, M.; Gronowski, P.; Krueger, D.; Morganti, C.; 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.

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

[9] Seevinck, E.; List, F.J.; 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.

[10] Heald, R.; Wang, P., "Variability in sub-100 nm SRAM designs," IEEE/ACM *International Conference on Computer Aided Design (ICCAD)*, pp. 347-352, Nov. 2004.

[11] Yoshimoto, M.; Anami, K.; Shinohara, H.; Yoshihara, T.; Takagi, H.; Nagao, S.; Kayano, S.; Nakano, T., "A divided word-line structure in the static RAM and its application to a 64K full CMOS RAM," *IEEE Journal of Solid-State Circuits*, vol.18, no.5, pp.479-485, Oct. 1983.