Ultra-Low Voltage Nano-Scale Embedded RAMs

K. Itoh, M. Horiguchi*, and T. Kawahara
Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo 185-8601, Japan,
k-itoh@crl.hitachi.co.jp
*Renesas Technology Corp., Kodaira, Tokyo 187-8588, Japan

***Abstract*—Ultra-low voltage nano-scale embedded RAMs are described, focusing on RAM cells and peripheral circuits. First, challenges and trends of low-voltage RAM cells are discussed in terms of signal charge, signal voltage, and noise. ECC to cope with the ever-increasing soft-error rate, power-supply controls to widen the voltage margin of cells, and a fully-depleted SOI to reduce V_T-variation are also investigated. Then peripheral circuits are explained in terms of leakage reduction and compensation for speed variations. Based on this, it is concluded that ultra-low voltage RAMs cannot be achieved without reducing speed variations caused by variations in V_T, thus resulting in a further need for compensation circuits and new devices with reduced V_T variation.**

I. Introduction

Ultra-low voltage nano-scale embedded (e-) RAMs are becoming increasingly important because they play critical roles in reducing power dissipation and chip size of MPUs/MCUs/SoCs. Thus, sub-1-V RAMs have been actively researched and developed [1-4], including a 0.6-V 16-Mb e-DRAM [5], a 1.2- to 1-V 16-Mb e-DRAM [6, 7], and a 24-MB SRAM cache in a 0.8- to 1-V MPU [8]. To create such e-RAMs, however, many challenges [1-4] remain with RAM cells and peripheral circuits. In addition to being the smallest cells possible, they are high signal-to-noise-ratio (S/N) designs for stable and reliable RAM cells, and reductions in the ever increasing leakage and speed variation of RAM cells and/or peripheral circuits as devices and V_{DD} are scaled down.

This paper describes circuit designs for ultra-low voltage e-RAMs using the one-transistor one-capacitor (1-T) DRAM cell and the six-transistor (6-T) SRAM cell. First, a high S/N design for RAM cells is discussed in terms of signal charge, signal voltage, and noise. Then peripheral circuits are investigated in terms of leakage and speed variation. Finally, future prospects are given. In this paper, leakage currents denote subthreshold currents.

II. RAM Cells

A. Signal Charge

The ever decreasing signal charge (Q_S) of non-selected cells (Fig. 1) [1-4] restricts low-voltage operations with increased soft-error rate (SER), because Q_S is almost equal to the soft-error critical charge. The Q_S of DRAM cells is given by C_S $V_{DD}/2$, while that of SRAM cells is approximately given by $(C_1 + 2C_2)$ V_{DD} [9]. Here, C_S is the cell-node capacitance of DRAM cells, and C_1 and C_2 are the parasitic cell-node capacitances of SRAM cells. Due to $(C_1 + 2C_2) \ll C_S/2$, the SER of SRAM cells is always much larger than that of DRAM cells, and it rapidly increases with device scaling [2] because of the rapid decrease in C_1 and C_2, although spatial scaling reduces the collected charge. In contrast, the SER of 1-T DRAM cells gradually decreases because C_S needs to be gradually decreased to maintain a large signal voltage. One remedy for SRAM cells is to add a large capacitance to $(C_1 + 2C_2)$ [10], even though this requires more complicated processes. A triple-well structure shielding the cell array, as a soft-error barrier, is also effective in reducing the SER of RAM cells. The most effective way is to use ECC, as will be explained later.

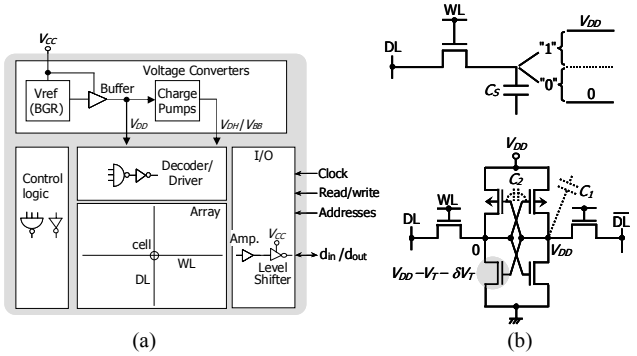


Fig. 1 RAM chip (a) and RAM cells (b) with DRAM cell for the upper and SRAM cell for the lower. δV_T:V_T-mismatch between paired MOSTs.

B. Signal Voltage of DRAM Cells

The small signal voltage and the ever slower sense-amplifier operation of a half-V_{DD} sensing prevents a low-voltage operation. The floating signal voltage v_S developed on the data line (DL) with capacitance C_D is given by $v_S \cong (V_{DD}/2)\ C_S/C_D$ for $C_D \gg C_S$. It is successfully sensed if $v_S > \delta V_T + v_N$, where δV_T is the offset voltage (i.e., V_T-mismatch between paired MOSTs) of sense amps, and v_N is the noise at sensing. To lower V_{DD} with a C_S small enough to accept a planar capacitor, which is a key to e-DRAMs, reducing C_D, δV_T, and v_N is essential. The C_D is reduced adequately enough by having a short DL, as exemplified by a 32-cell-connected DL and 5-fF C_S for a 1.2-V 322-MHz 16-Mb e-DRAM [6]. The δV_T can be reduced considerably if the largest MOST possible is used for sense amps despite an area penalty. Reducing v_N is extremely important not only for successful sensing, but also for fast sensing despite a half-V_{DD} data-line precharge (i.e., mid-point sensing). The mid-

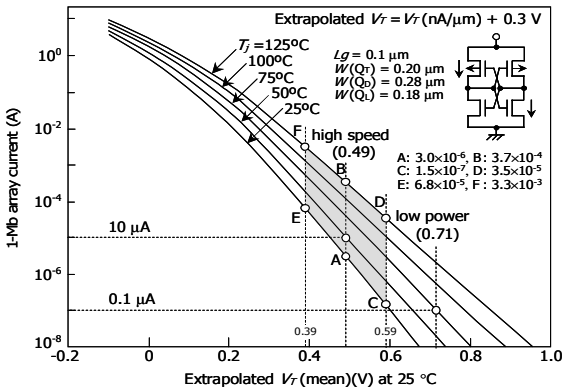
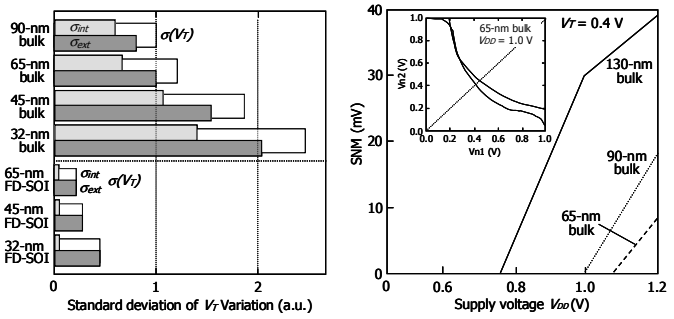


Fig. 2 1-Mb array current vs. V_T of cross-coupled MOSTs [3].



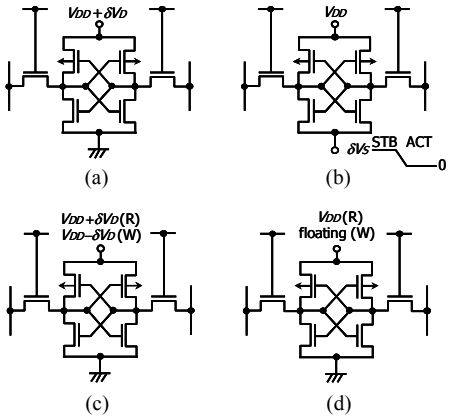
(a) (b)

Fig. 3 Standard deviations of V_T variation, $\sigma(V_T)$, and intrinsic (σ_{int}) and extrinsic (σ_{ext}) V_T variations (a) [3, 11, 20, 21], and SNM of the 6-T bulk-CMOS SRAM cell taking 6σ of V_T variation into consideration (b) [11].

point sensing always involves the slowest speed in a chip despite distinctive features of a quiet array and halved data-line charging power [1]. This is because the gate-over-drive of turned-on MOST in cross-coupled MOSTs in sense amps is the lowest in a chip, which is given by $V_{DD}/2 - (V_T + v_N)$. Thus, to lower V_{DD}, the value of $(V_T + v_N)$ must be minimized for fast sensing, which calls for reducing v_N and finding the lowest V_T possible while reducing the subthreshold currents involved. The noise v_N consists of many components [1]: an inaccurate $V_{DD}/2$-level setting caused by a DC level fluctuation of the $V_{DD}/2$ generator and a capacitive coupling to the DL from the precharge circuit and equalizer; a capacitive imbalance between a pair of DLs; the word line to DL coupling; and the adjacent DL coupling. Recently, an accurate $V_{DD}/2$-level setting with fuse trimming, a differential driving of the precharger and equalizer, and DL transpositions without area penalty have been reported [6, 7]. The V_T of the sense-amp MOST was reportedly lowered to 0.2 V, coupled with power switches [5], enabling 0.6-V V_{DD} sensing.

C. Signal Voltage of SRAM Cells

A large necessary V_T and a larger V_T variation of cross-coupled MOSTs in the 6-T SRAM cell are major obstacles to low-voltage operation. The V_T of cross-coupled MOSTs must be quite high to reduce the leakage that rapidly increases as V_T decreases, as shown in Fig. 2 [3]. Here, the V_T is the average of the cross-coupled MOSTs in a chip, because it is the average V_T that determines chip leakage. For example, for a low-power 1-Mb e-SRAM that allows a leakage of 0.1 μA at T_{jmax} = 75°C, the V_T at 25°C might be higher than 0.71 V. For a high-speed 1-Mb e-SRAM that can tolerate a leakage of 10 μA at T_{jmax} = 50°C, the V_T can be as low as 0.49 V. In addition to such high V_Ts, the intra-die or inter-die V_T variation (ΔV_T) that increases as devices get smaller (Fig. 3(a)) [3, 4, 11] reduces the signal voltage on the data line because of a reduction in the drive current of on-MOST in cross-coupled MOSTs, which is proportional to the gate-over-drive ($V_{DD} - V_T - \delta V_T$). Here, the V_T is also the average V_T in a chip, and δV_T ($= 2\sqrt{\Delta V_T}$) is again the V_T-mismatch of cross-coupled n-MOSTs. Hence, even for a fixed V_{DD}–V_T, each cell can have a different drive current, depending on its δV_T. For



(a) (b) (c) (d)

Fig. 4 Power controls of SRAM cells [4].

V_T = 0.71 V and δV_T = 0.1 V, the minimum V_{DD} (V_{DDmin}) for a successful operation is as high as 0.81 V. In practice, the V_{DDmin} must be higher than this value to suppress the variation of access time of cells that is prominent at around $V_{DD} \cong V_T + \delta V_T$ [12]. The continually increasing V_T variation also degrades the static noise margin (SNM): The V_{DDmin}, defined as the V_{DD} for an SNM of 0, becomes higher with device scaling, as shown in Fig. 3(b) [11].

To solve the problems many power-supply controls for the cells (Fig. 4[4]) have been proposed. Although they are effective only for high V_{DD} over 1 V, or they prevent MOSTs from being scaled down, they nevertheless reduce leakage, widen the voltage margin, or compensate for the V_T variation. Type (a) features the raised supply ($V_{DD} = V_{DD} + \delta V_D$) and dual-$V_T$ scheme [3, 13, 14]. The raised supply offsets the high V_T and the δV_T of cross-coupled MOSTs, though MOSTs are unscalable due to their need for a high stress-voltage. Moreover, the well known negative word line scheme [1-4] applied to low-V_T transfer MOSTs cuts leakage during non-selected periods, while increasing the cell read current. Type (b) features the source offset driving [15]. In the active to standby mode transition, it lowers the data-line voltage from 1.5 V to 1 V to relax the electric field of all MOSTs, and raises the ground line to 0.5 V to increase the V_T of off-MOST. However, reducing the supply voltage by δV_S in the standby mode restricts the low-voltage operation with

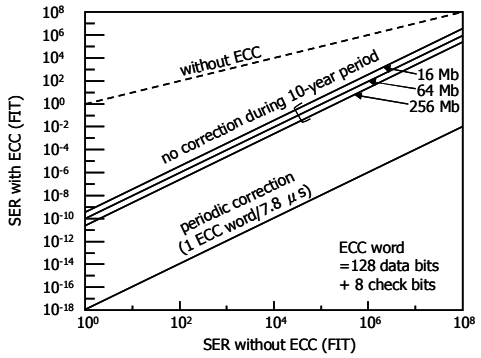


Fig. 5 SER reduction through on-chip ECC.

increased SER. Types (c) and (d) feature the dynamic control of cell-power supply. Type (c) switches the power supply [16] to a lower level in the write mode and a higher level in the read mode to widen the write and read margins. The voltage difference between read and write has been reported to be 200 mV at V_{DD} = 1.1 V. Type (d) leaves the supply line at a floating level during the write mode [17]. Raising both the power supply and word line in the read mode has also been proposed for a TFT load cell [18].

D. On-chip ECC

On-chip error checking and correcting (ECC) is the key to RAMs in the future, especially to SRAMs with their inherent small signal charge, as explained previously. The SER reduction of RAM by using a single-error correcting code is expressed by the equation $E = (N^2T/2N_0^2W)\,E_0^2$, [23], where E is SER with ECC, E_0 is SER without ECC, N is the number of bits of an ECC word including check bits, N_0 is the number of data bits of an ECC word, T is the correction period, and W is the number of ECC words in a RAM (i. e., $W = M/N_0$, where M is the memory capacity). Figure 5 shows the SER reduction using a code of N = 136 and N_0 = 128. Even if SER without ECC is as high as 10^6 FIT (one upset per 1,000 hours) and the errors are not corrected at all during a ten-year period (T = 10 y), the SER is improved by four to five orders of magnitude through ECC. If periodic error correction (one ECC word every 7.8 μs) is performed like a DRAM refresh operation (T = 7.8 μs × W), the resulting SER becomes as low as 10^{-6} FIT. ECC is also effective for hard errors, especially for random single-cell faults, such as the V_T mismatch described above. In addition, combining with redundancy produces a synergistic effect [24], which results in a drastic increase in fault tolerance.

III. Peripheral Logic Circuits

Obstacles to low-voltage operation of peripheral circuits are the ever-increasing subthreshold leakage and speed variations. The subthreshold leakage can be sufficiently reduced as far as RAMs are concerned. The speed-variation issue that is common to all nano-meter LSIs necessitates compensation circuits for, and new devices against the variation.

A. Leakage

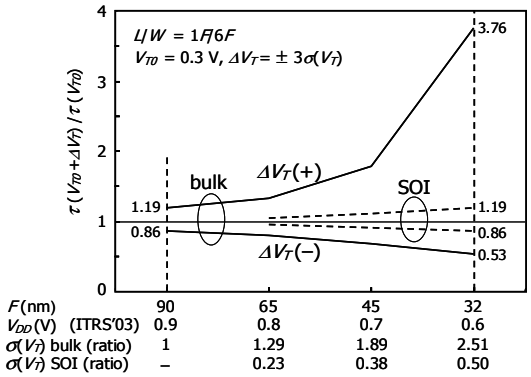
Reducing leakage in the active mode is especially important, although this is more difficult than in the standby mode because leakage needs to be controlled much faster. Fortunately, leakage currents can be quickly, simply, and drastically reduced by utilizing RAM's features [1-4]. The basic reduction concept is to use a high-V_T MOST that is achieved with a high actual V_T or effectively with a low actual V_T MOST. Of many proposals, the gate-source offset driving, the gate-source self-back-biasing, power-switches with a level holder, and multi-static V_T [1-4] are practical for RAMs. In fact, they sufficiently reduced the standby and/or active leakage of a 0.6-V 16-Mb DRAM [5], 256-Mb DRAMs [1], a hypothetical 1-V 16-Gb DRAM [1], and a 1.2-V SRAM [19].

B. Speed Variation

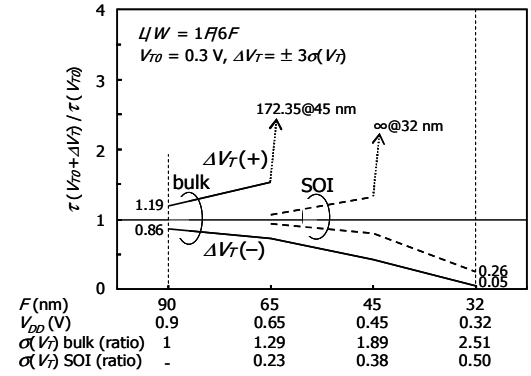
Inter-die and intra-die V_T variations increase not only variations in leakage, but also variations in speed. For example, for the high-speed SRAM design with V_T = 0.49 V, the leakage varies as much as four orders of magnitude for a V_T variation of ± 0.1 V and a temperature variation of 100°C, as shown in Fig. 2 [3]. Such is the case for peripheral circuits. For any ΔV_T, the degree of speed variation, ΔV_T / ($V_{DD} - V_T$), increases with lower V_{DD}. It is enhanced by device scaling involving the ever larger ΔV_T (Fig. 3). Figure 6(a) shows delay versus feature size, F, for a low-power design with V_{DD}-scaling based on ITRS 2003[22]. Delay times for ± 3σ(V_T) are normalized by that for the average V_T (i.e., V_{T0} = 0.3 V) for each generation. For the bulk CMOS, the speed spread is from 1.19 to 0.86 in the 90-nm generation. However, it rapidly increases with device scaling, reaching as large as 3.76 to 0.53 in the 32-nm generation. This is an unacceptable increase. A 2.5-time increase in V_T variation and a decrease in V_{DD} from 0.9 V to 0.6 V are responsible for the increase. If V_{DD} is scaled down as for F, the speed spread increases to an unacceptable level, as in Fig. 6(b), although such V_{DD} scaling is ideal in terms of low power and ease of device development.

For an excessive intra-die variation, there may be no solution without new devices with less V_T variation. A double-depleted (FD)-SOI [11, 20, 21] is promising in the nano-meter era if the expected features are fully verified. This is because the ultra-thin and lightly-doped channel of the SOI structure suppresses the V_T variation (Fig. 3). As a result, even in the 32-nm generation the speed spread remains in the same range as that for the 90-nm bulk CMOS, as seen in Fig. 6(a). This implies that the SOI would extend the low-voltage limitation of bulk CMOS by at least three generations. Moreover, an ultra-thin BOX (buried oxide) layer allows the V_T to be widely controlled by positive and negative back-bias controls, making it possible to create new low-voltage circuits such as a dynamic-V_T MOS circuit [20, 21]. This is true for a relatively high V_{DD}. If V_{DD} is scaled down, as in Fig. 6(b), however, not even the SOI will be able to manage the speed variation, calling for new techniques.

For the inter-die variation, compensation by controlling the substrate voltage is unavoidable. It has been reported that



(a)



(b)

Fig. 6 Speed variations of an inverter for the V_{DD} projected by ITRS 2003 [22](a), and the V_{DD} scaled down as for F(b). The delay time is assumed to be proportional to $V_{DD}/(V_{DD}-V_T)^{1.25}$ [25].

positive body bias improved the speed by 63 % in slow process conditions, and negative body bias reduced leakage by 75 % in fast process conditions [5].

IV. Future Prospects

For RAMs, low-voltage operations are eventually restricted by the retention characteristics and signal-to-noise-ratio of RAM cells, even if coupled with ECC, a fast and reliable sensing, and the leakage current and speed variation of peripheral circuits, as described previously. For existing RAM cells, a high and unscalable V_T, which requires a high V_{DD}, is necessary to ensure a small retention current and a long enough refresh time [1]. Thus, new RAM cells, which do not rely on the charge and are thus insensitive to leakage, such as nonvolatile RAMs, will be strongly needed for lower-voltage operation. Even if a dual V_{DD} scheme [4] with a high V_{DD} for the cell array and a low V_{DD} for the peripheral circuits is used, ever increasing speed variations with device scaling will eventually limit low-voltage operations of the whole chip. If V_T variations continue to get larger and larger, as in existing bulk CMOSTs, ultra-low voltage RAMs could not be achieved without reducing the resultant speed variations. Note that even if the inter-die variations can be compensated for by the improving the existing circuits, intra-die speed variations cannot be compensated for. In this case, the V_{DD} of peripheral circuits must get higher and higher to offset the speed variation. Alternatively, new MOSTs with small V_T variations, such as a fully depleted SOI [11] despite expensive wafers, will be necessary. Here, reducing the gate tunneling current is also vital: it is not discussed in this paper, though, because this kind of reduction is the intended responsibility of the process and device designers [3]. In any event, two approaches in the nanometer era can be envisioned based on the above discussion: One is high-V_{DD} bulk-CMOS e-RAMs for low-cost applications, and the other is low-V_{DD} FD-SOI e-RAMs for high-speed and low-power applications.

V. Conclucion

Ultra-low voltage nano-scale embedded RAMs were described, focusing on RAM cells and peripheral circuits. First, challenges and trends of low-voltage RAM cells were discussed in terms of the signal charge, signal voltage, and noise, clarifying the importance of ECC in coping with the ever increasing soft-error rate and a fully depleted SOI to reduce variations in V_T. Then, peripheral circuits were discussed in terms of leakage reduction and compensation for speed variations, and it was concluded that ultra-low voltage RAMs would not be achieved without reducing speed variations caused by the V_T variation, thus prompting a need for compensation circuits and new devices with reduced V_T variation.

Acknowledgment

The authors are grateful to Dr. Osada, Mr. Yamaoka, and Dr. Tsuchiya for their discussion and support with valuable data.

References

[1] K. Itoh, *VLSI Memory Chip Design*, Springer-Verlag, 2001.
[2] Y. Nakagome, IBM J. R & D,47, no.5/6, pp. 525-552, 2003.
[3] K. Itoh, CICC Dig., pp. 339-344, 2004.
[4] K. Itoh, ICICDT Dig., pp. 235-242, 2005.
[5] K. Hardee, ISSCC Dig., pp. 494-495, 2004.
[6] M. Iida, ISSCC Dig., pp. 460-461, 2005.
[7] M. Shirahama, ISSCC Dig., pp. 462-463, 2005.
[8] S. Naffziger, ISSCC Dig., pp. 182-183, 2005.
[9] P. Carter, IEEE J. SSC, 22, No.3, pp. 430-436, 1987.
[10] S-M Jung, IEDM Dig., pp. 289-292, 2003.
[11] R. Tsuchiya, IEDM Dig., pp. 631-634, 2004.
[12] J. Wuu, ISSCC Dig., pp. 488-489, 2005.
[13] K. Itoh, Symp. VLSI Circuits Dig., pp. 132–133, 1996.
[14] K. Itoh, PATMOS Dig., pp. 3-15, 2004.
[15] K. Osada, ISSCC Dig., pp. 302–303, 2003.
[16] K. Zang, ISSCC Dig., pp. 474-475, 2005.
[17] M. Yamaoka, ISSCC Dig., pp. 480-481, 2005.
[18] H. An, Symp. VLSI Circuits Dig., pp. 282-283, 2004.
[19] M. Yamaoka, ISSCC Dig., pp. 494-495, 2004.
[20] M. Yamaoka, Symp. VLSI Circuits Dig., pp. 288-291, 2004.
[21] M. Yamaoka, Int'l SOI Conf. Dig., pp. 109-111, 2004.
[22] ITRS 2003 EXECUTIVE SUMMARY Table 6a, p. 57.
[23] M. Horiguchi et al., IEEE J. SSC, 23, p. 27, Feb. 1988.
[24] H. L. Kalter et al., IEEE J. SSC, 25, p. 1118, Oct. 1990.
[25] K. Chen et al., Trans. on Electron Devices, pp. 1951-1957, 1997.