Low-Voltage Embedded-RAM Technology: Present and Future

Kiyoo Itoh, and Hiroyuki Mizuno

Central Research Laboratory, Hitachi, Ltd. Kokubunji, Tokyo 185-8601, Japan

Abstract: First, key issues for low-voltage ($<1V$) embedded RAMs are summarized in terms of stable operation, suppression of leakage (gate-tunneling/subthreshold) currents, and speed variation of memory cells and peripheral logic circuits. Next, DRAM and SRAM cells to cope with the above issues, the circuit design focusing on subthreshold-current issue, and suppression of or compensation for design-parameter variations to reduce the speed variations are discussed. Voltage converters and power management for low-power and low-voltage operation are also explained. Finally, based on the above discussions, a perspective is given with emphasis on needs for simple/high signal-to-noise ratio memory cells (such as gain cells) with a pure logic compatible process, high-speed subthreshold-current reduction focusing on active mode, and memory-rich SoC architectures.

Key words: subthreshold current, gate tunneling current, DRAM, SRAM, peripheral circuits, gate-source backbias, dynamic V_T , multi-static V_T , power management, I_{DDQ} test, gain cells, memory-rich architectures.

1. INTRODUCTION

Low-voltage embedded RAM (eRAM) is vital for low-power system-on-a-chip (SoC) technology [1]. To take advantage of device miniaturization, however, the power-supply (V_{DD}) of eRAMs must keep up with the rapid lowering of the V_{DD} of MPUs, as discussed later. As V_{DD} makes the transition to sub- V levels there are three major challenges [1]: stable operation of memory cells against reduced signal charges; suppression of the ever-increasing leakage (subthreshold and tunneling) currents of the MOSTs

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35597-9_40](https://doi.org/10.1007/978-0-387-35597-9_40)

M. Robert et al. (eds.), *SoC Design Methodologies*

© IFIP International Federation for Information Processing 2002

in the peripheral logic circuits and memory cells; and compensation for the increased variations of speed that are seen at lower levels of V_{DD} .

In this paper, low-voltage eRAM circuits are discussed with emphasis on subthreshold-leakage current issue: First, low-voltage trends are discussed, and then recent developments to do with the above issues for DRAM and SRAM cells, and peripheral logic circuits are investigated. Other design issues for low- V_{DD} operation, such as on-chip supply-voltage converters, power management, and testing methodology to cope with the subthreshold current are also described. A position is then taken that emphasizes needs for gain cells, subthreshold-current reduction circuits for use in active mode, and memory-rich SoC architectures.

2. CURRENT GENERATION OF LOW-VOLTAGE eRAM TECHNOLOGY

2.1 Trends in Gate-Oxide Thickness

Figure 1. Trends in gate-oxide thickness for MPUs and standard DRAMs. L, M; logic core and memory array.

Device miniaturization for the core logic and embedded SRAM cache (i.e., eSRAM) of high-end MPUs has recently been accelerated by the use of an MOST structure with a thinner gate oxide (t_{ox}), and thus a lower V_{DD} and a lower threshold voltage (V_T) for lower power and higher speeds (Fig. 1) [2]. Here, the dual- V_{DD} and dual t_{ox} -device approach has been maintained with an MOST that has a thicker- t_{ox} and higher- V_T for the higher- V_{DD} I/O circuitry. As a result, 1-V V_{DD} and 2-3-nm t_{ox} have become popular for use in the core and eSRAM.

For standard (stand-alone) DRAMs, operation with a single external V_{DD} and an on-chip voltage-down converter (i.e., series regulator) has been used to realize power-supply standardization despite the internally dual- V_{DD} operation. In addition, the single thick t_{ox} (necessary for word-bootstrapping of the cells) has been used

Figure 2. Two kinds of leakage currents; subthreshold current (I_s) and gate-tunneling current (I_g), and a power switch for the reductions.

throughout the chip for low cost. Recently, however, a dual- V_{DD} and dual t_{ox} -device approach, similar to that taken with MPUs, has been adopted to achieve higher speeds for eDRAMs. One example is a recent 8-Mb eDRAM with 3.7-ns access (Fig. 1) [3].

In any event, the above three issues arise with a rapid lowering of V_{DD} . In particular, leakage-current reduction is essential for all the LSIs of the future. There are two major leakage currents; the subthreshold current and the gate-oxide tunneling current of the MOSTs (Fig. 2). A tunneling-current reduction circuit with a power switch, in which an MOST with a thick t_{ox} and high V_T is used, has recently been proposed (Fig. 2) [4]. However, continued reduction in t_{ox} will eventually come to depend on the developments of an MOST with a gate insulator of some high- k material, because even the circuit mentioned above would not be able to manage the exponential increases in current that are a result of lower and lower values for t_{ox} .

2.2 Memory Cells and Relevant Circuits

DRAM: For the low-voltage one-transistor one-capacitor DRAM (1-T) cell, using vertical capacitors and high-permittivity thin-films to maintain sufficient signal charge, and adjusting the potential profile of the storage node to suppress the pn -leakage current are critical ways of maintaining the cell's signal voltage, soft-error immunity and retention times even as V_{DD} is lowered.

Low-voltage circuits are also important. The negative word-line (NWL) scheme (Fig. 3) [5] cuts the subthreshold current from the storage node to the data line despite the low-actual V_T with a gate-source back-bias of δ during the non-selected period. This makes it possible to reduce the word-line voltage in a full- V_{DD} write operation by δ , which allows the use of an MOST with a thinner t_{ox} for a given device reliability. The resultant small subthreshold swing (i.e. S -factor) enables low-voltage operation. The Multi-divided data-line scheme, in which data-line capacitance (C_D) and line delay

Figure3. Conventional DRAM cell and negative word-line scheme.

Figure4. 1-T SRAM™ using 1-T DRAM cells and operations of multi banks.

are reduced, is another way of obtaining attractive eDRAMs. If the storage capacitance (C_S) is sufficient, this allows ultra-low voltage operation by maintaining cell-signal voltages ($\sim C_S V_{DD} / 2 C_D$) with the help of a small C_D . Moreover, if V_{DD} is raised even a small C_S using a simple fabrication process that is essential for eDRAMs is acceptable. A good example is the so-called 1-T SRAM™ (Fig. 4) [6], in which a 1-T DRAM cell with a C_S as small as less than 10 fF is realized by a single poly-Si planar capacitor, and an extensive multi-bank scheme with 128 banks (32 Kb in each) that are simultaneously operable is used. A greater than 300-MHz row-access frequency was achieved for a 0.18- μ m 1.8-V 2-Mb eDRAM. A DRAM cache with the same capacity as a single bank is used to hide the refresh operation: Even when there is

a conflict between a normal access and a refresh operation at the same bank, a cache hit occurs while permitting a refresh cycle for that bank since all of the data in that bank have been copied to the cache.

Low-voltage high-speed sensing is also essential for the well-known mid-point sensing by which the data-line power is halved without using a dummy cell, since using the half- V_{DD} data-line voltage makes the sense-amplifier (SA) operation quite slow. In the overdrive sensing scheme [7,8,9] this problem is solved by applying a higher voltage solely to the SA-inputs with isolating the data line from SA or with capacitive coupling. The use of additional capacitors may be acceptable in eDRAMs for which area is not a concern.

Figure 5. 6-T SRAM cell and the voltage margin (a), and improved cell (b).

Figure 6. Static noise margin, SNM, of the 6-T SRAM cell. V_T s of all MOSTs are varied at the same time (a) while V_T of transfer MOST is varied and V_T of cross-coupled MOSTs is 0.8V (b).

SRAM: Of the many proposed designs [1], the 6-T full-CMOS cell is most suitable despite the fact that it is large. This is because of the simple process and the ease of design provided by the cell's wide-voltage margin. Even for this cell, subthreshold-currents must be reduced and voltage margin must be widened as V_T is lowered by lowering V_{DD} . Note that the subthreshold current for $V_T = 0$ V would lead to a retention current of as much as 10 A in a 1-Mb SRAM array [1]. This places a strict limit on the reduction of V_T . The worsening of voltage (noise) margins as V_{DD} and V_T are lowered (Figs. 5 and 6) creates a demand for a decrease in the ratio of transconductance of the transfer MOST to that of the driver MOST. The voltage margin is also worsened by variations in V_T and mismatches of V_T between paired MOSTs, both of which increase by device miniaturization [1]. In addition, for a low- V_T transfer MOST, the margin is further worsened by the data pattern of non-selected cells in the column, as shown in Fig. 7. The total of the

Figure 7. Leakage currents of low- V_T transfer cell.

Figure 8. SRAM cell using a gate-source offset driving.

subthreshold leakage currents from the non-selected cells may be greater than the selected cell's current. This causes a read failure. A hierarchical data-line scheme provides a partial solution to this problem by limiting the number of memory cells connected to the data line [10,11]. Using a data equalizer for current compensation [12] is also effective, despite the penalty in terms of area. Applying a gate-source offset driving scheme (Fig. 8) to the memory cell allows the use of a low- V_T transfer MOST for high speed and a negligible data-line current [13], although eleven transistors are required for each cell.

The loadless CMOS 4-T SRAM (Fig. 9) [14] is attractive because the cell size is only 56% of that of a conventional 6-T cell. Transfer MOSTs of the non-selected cells at V_{DD} -word line (WL) voltage work as load elements, and the load current to keep the storage node high is supplied from a data line that has been precharged to V_{DD} . The WL voltage, however, must be precisely controlled, as shown in the figure, so as to keep the load current (i.e., the off-current of the transfer MOST) to more than a hundred times that of the off current of the driver MOST. The data-pattern problem outlined above also arises in this approach.

Figure 9. 4-T SRAM cell and control of the nonselected WL level.

Almost all of the above problems may be solved by using high- V_T cross-coupled MOSTs combined with a boosted power supply, as shown in Fig. 5(b) [15]. High V_T for the paired MOSTs is necessary to obtain a low subthreshold current. The raised power-supply voltage increases the transconductance of the driver MOST and compensates for variations in

V_T and mismatches of V_T . Even a low- V_T transfer MOST for high-speed is then acceptable, due to the increased transconductance of the driver MOST. Negative word-line biasing reduces the leakage currents from cells in the column. Using a high- V_T transfer MOST with word bootstrapping is an alternative approach that achieves the same result.

2.3 Peripheral Logic Circuits

Subthreshold Current: Many attempts [1] have been made to reduce subthreshold currents in standby mode. Gate-source back-biasing [1] is effective most for memory applications. For example, applying a back-bias of as little as 0.25 V to a word-driver block reduces the standby subthreshold-current by 2-3 decades without inflicting penalties in terms of speed and area. In combination with a multiply divided power-line scheme [1], this approach is able to confine subthreshold currents to a single sub-power line. Applying dual- V_T [16] to the critical paths of logic circuits is quite effective, reducing the standby subthreshold current to one-fifth of its value for a single low V_T . In another form of multiple- V_T design called random modulation [17], four kinds of V_{TS} are applied to further reduce the current in sub-1V circuitry despite additional cost and complexity of design. Applying the well-known dynamic substrate back-bias (V_{BB}) reduces the current by 1-2 decades. Even so, the required V_{BB} swing is as large as 1-3 V and this approach becomes less effective with device scaling [18]. The smaller body constant, enhanced short-channel effects, and increases in other leakage currents such as the gate-induced drain-lowering (GIDL) current [18] are responsible for the drawback. Even if the subthreshold current is still a small fraction of the active current because V_T is still high, the subthreshold current in the active mode makes the operation of dynamic circuits such as dynamic NAND for the decoder unstable, discharging the floating nodes. The level keeper proposed for logic circuits (Fig. 10) [19] would also be effective in memory design, although its effects would be fatal if subthreshold current is increased to the extent that the keeper cannot manage.

Figure 10. A level keeper in a dynamic NOR against precharge-level degradation by leakage current.

Speed Variation: The amount of speed variation for a given variation in design-parameters is increased by lowering V_{DD} . Control of V_{BB} and internal V_{DD} [20,21] in accordance with the parameter variations reduces the speed variation. Controlling a forward V_{BB} is more effective in reducing speed variations [22,23] than controlling a reverse V_{BB} . This is because the V_T - V_{BB} characteristics is more sensitive to V_{BB}

[1]. Forward V_{BB} in [23], for example, reduced the variation of logic circuits and improved operation speed by 10%. If a forward V_{BB} is used, however, requirements for noise suppression become more stringent, i.e., there is a call for the uniform distribution of the forward V_{BB} throughout the chip. However, having a quasi-floating V_{BB} at each MOST induces instabilities such as history effects, as in SOI, resulting in intra-die fluctuations in V_T [24]. This is especially so for high-speed LSIs. When the clock cycle is smaller than the noise decay time in the substrate, which may be more than 10 ns for a substrate capacitance of 1 nF, substrate noise may accumulate and cause a variation in V_T . In particular, eDRAMs may couple spike noises to the floating V_{BB} when many data lines are simultaneously charged and discharged. This is so unless a triple-well substrate is used. Additional current consumption, in the form of bipolar current induced by the forward V_{BB} , is another matter that must be considered.

2.4 Other Design Issues

On-Chip Supply-Voltage Converters: On-chip voltage conversion is essential for the stable operation of memory cells, reduction of the subthreshold current, and suppression of variations in speed with fewer external supply voltages. A high efficiency of voltage conversion, high degree of accuracy in the output voltage and low cost of implementation are the key issues. Using voltage-up converter with a charge pump suffers from the poor current-driving ability of this approach, which means that subthreshold currents at the load must be strictly limited to maintain the boosted voltage level. Voltage-down converters have two types; the series regulator and the switching regulator. The series regulator, widely used in modern commercial DRAMs, provides a small output current and a poor efficiency (η) of around 50% when the difference between input and output voltages is large. A regulator of this type, however, offers a highly accurate output voltage and it is possible to implement all of the required components on a single chip. The switching regulator has a good efficiency, even in the high output-current region. However, a regulator of this type inherently suffers from a large ripple in the output voltage and a poor transient response for large changes in the current. In general, it also requires a large NMOST and PMOST, and external parts such as an inductor and capacitor. There have been many attempts to solve these problems. A combination of a charge pump and series regulator achieves a high efficiency and precise output voltages [25], and is applicable to the NWL scheme described above. In the hybrid converter [26], consisting of both a series regulator and switching regulator, the switching regulator is turned on and off according

Figure 11. I_{DDQ} testing by reducing subthreshold current.

the load current to maintain a high efficiency across a wide range of load currents.

Power Management: Here are some view points of memory designers on the methods of power management for SoC technology that have been proposed by logic designers. Although it incurs a long recovery time at heavily capacitive internal power-lines, power switching with a high- V_T MOST completely cuts the leakage currents of the internal core circuits. Dynamic voltage scaling (DVS) [27,28], in which the clock frequency and V_{DD} are dynamically varied in response to the computational load, provides a reduced energy consumption per process during the periods of little computational, while still providing peak performance when this is required. However, this approach becomes less effective in the low- V_{DD} era since the range across which it is possible to vary V_{DD} becomes narrow. In addition, successful operation over a wide range of V_{DD} requires the accurate tracking of all circuit delays. Furthermore, applying DVS makes dynamic circuits (e.g. eDRAMs) unstable. This is due to the trapping of charge at floating nodes when voltage bumps are applied. DVS does not reduce subthreshold currents. However, in elastic- V_T CMOS [29], where the clock frequency, V_{DD} , and V_{BB} are all dynamically varied, these currents are reduced. The cost is the added complexity of design. Recently, low-power designs for which it is possible to specify the maximum level of power have been proposed. In ChipOS [30], for example, each circuit-block's power is managed by controlling the gated clock and power switch to achieve a given power budget. The confinement of junction temperature and load current that this allows may permit the use of temperature-sensitive eRAM, or the suppression of fluctuations in the voltage output of on-chip voltage converters.

Testing: A large subthreshold current makes the discrimination of defective and non-defective I_{DDQ} currents difficult, and thereby poses a problem for the I_{DDQ} testing of low-voltage CMOS circuits. I_{DDQ} testing with

the application of a reverse V_{BB} (Fig. 11) [31] is effective when low-temperature measurement and multi- V_T design are combined.

Figure 12. Effective cell area of various RAMs vs. power-supply voltage.

3. FUTURE TRENDS

The 1-T DRAM cell is not suitable for low-voltage operation because the reduction of its signal voltage is not acceptable. More division of the data-line, aiming at a smaller C_D , to offset the reduction in cell-signal levels at a lower V_{DD} sharply increases the effective cell area, as shown in Fig. 12. This area is defined as the sum of the actual cell area and the total area of the additional cell-related circuits (such as sense amp) in each division. Using gain cells would be one solution for the problem, because they generate a high enough signal voltage without requiring any data-line division even at low values of V_{DD} , and thus provide a fixed effective cell area that is independent of V_{DD} . For eSRAM, the advanced 6-T cell (Fig. 5(b)) may continue to be used.

The issue of leakage current presents a real challenge in the design and testing of a low-voltage high-speed SoC. The developments of innovative MOSTs that have new gate materials and subthreshold-current reduction circuits are the keys to tackling this challenge. In particular, with a further reduction in V_T , subthreshold-current reduction circuits for use in high-speed active mode will be indispensable, even for static logic circuits. This is because the subthreshold current comes to exceed the capacitive current and eventually dominates the total active current of the chip [1]. In this case, the low-power advantage of CMOS circuits is lost. If this problem is not solved,

we can envision a scenario in which even CMOS SoC would suffer from huge levels of dc power dissipation caused by subthreshold currents, as was the case in the bipolar and BiCMOS LSI eras of the recent past. Thus, low-power techniques for bipolar and BiCMOS circuits might be revived to help reduce the power dissipation of the SoC. Reducing the number of random logic gates is one vital way of reducing the active-mode subthreshold current, since control of subthreshold currents from random logic gates at sufficiently high speed may remain impossible. Hence, new SoC architectures such as a memory-rich SoC, by which the subthreshold currents are effectively reduced, will be needed.

4. REFERENCES

- [1] K. Itoh, *VLSI Memory Chip Design*, Springer-Verlag, March 2001.
- [2] K. Itoh et al., "Reviews and Prospects of High-Density RAM Technology," CAS 2000, Sinaia (Romania), Oct. 2000.
- [3] O. Takahashi, et al., "1GHz Fully Pipelined 3.7ns Address Access Time 8kx1024 Embedded DRAM Macro," ISSCC 2000, pp. 396-397.
- [4] T. Inukai, et al., "Suppression of Stand-by Tunnel Current in Ultra-Thin Gate Oxide MOSFETs," SSDM 1999, pp. 264-265.
- [5] S. Miyano and M. Takahashi, "Embedded DRAM SOC's and Its Application for MPEG4 Codec LSI's," Proceedings of VLSI CIRCUITS SHORT COURSE, pp. 101-121, June 2001.
- [6] W. Leung, et al., "The Ideal SoC Memory: 1T-SRAM," 13th Annual IEEE Int. ASIC/SOC Conference, Sept. 2000.
- [7] M. Asakura, et al., "An Experimental 256-Mb DRAM with Boosted Sense-Ground Scheme," IEEE JSSC, vol. 29, no. 11, pp. 1303-1309, Nov. 1994.
- [8] T. Kawahara, et al., "A Circuit Technology for Sub-10ns ECL 4Mb BiCMOS DRAMs," 1991 Symp. on VLSI Circuits, pp. 131-132.
- [9] H. Mizuno, et al., "CMOS-Logic-Circuit-Compatible DRAM Circuit Designs for Wide-Voltage and Wide-Temperature-Range Applications," 2000 Symp. on VLSI Circuits, pp. 120-121.
- [10] B. Baas, et al., "An Energy-Efficient Single-Chip FFT Processor," 1995 Symp. on VLSI Circuits, pp. 164-165.
- [11] K. Zhang, et al., "The Scaling of Data Sensing Schemes for High Speed Cache Design in Sub-0.18 μ m," 2000 Symp. on VLSI Circuits, pp. 226-227.
- [12] K. Agawa, et al., "A Bit-Line Leakage Compensation Scheme for Low-Voltage SRAM's," 2000 Symp. on VLSI Circuits, pp. 70-71.
- [13] R. Krishnamurthy, et al., "A 0.13 μ m 6GHz 256x32b Leakage-tolerant Register File," 2001 Symp. on VLSI Circuits, pp. 25-26.
- [14] K. Takada, et al., "A 16Mb 400MHz Loadless CMOS Four-Transistor SRAM Macro," ISSCC 2000, pp. 264-265.
- [15] K. Itoh, et al., "A Deep Sub-V_t Single Power-Supply SRAM Cell with Multi-V_t Boosted Storage Node and Dynamic Load," 1996 Symp. on VLSI Circuits, pp. 132-133.
- [16] N. Rohrer, et al., "A 480 MHz RISC Microprocessor in a 0.12 μ m Leff CMOS Technology with Copper Interconnects," ISSCC 1998, pp. 240-241.

- [17] K. Yano, "Logic Circuit Families for Year-2005 Multi-Giga-Hertz Processors," IEEE Workshop on Design for Multi-Giga-Hertz Processors, San Francisco, Feb. 2000.
- [18] A. Keshavarzi, et al., "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual V_t CMOS ICs," ISLPED 2001, pp. 207-212.
- [19] A. Alvandpour, et al., "A Conditional Keeper Technique for Sub-0.13 μ m Wide Dynamic Gates," 2001 Symp. on VLSI Circuits, pp. 29-30.
- [20] M. Kubo et al., "A Threshold Voltage Controlling Circuit for Short Channel MOS Integrated Circuits," ISSCC 1976, pp. 54-55.
- [21] T. Kuroda, et al., "A 0.9V, 150MHz, 10mW, 4mm², 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage (VT) Scheme," ISSCC 1996, pp. 166-167.
- [22] Y. Oowaki, et al., "A Sub-0.1 μ m Circuit Design with Substrate-over-Biasing," ISSCC 1998, pp. 88-89.
- [23] M. Miyazaki, et al., "1000-MIPS/W Microprocessor using Speed-Adaptive Threshold-Voltage CMOS with Forward Bias," ISSCC 2000, pp. 420-421.
- [24] H. Mizuno, et al., "18- μ A-Standby-Current 1.8-V 200-MHz Microprocessor with Self Substrate-Biased Data Retention Mode," ISSCC 1999, pp. 280-281.
- [25] H. Tanaka, et al., "A Precise On-Chip Voltage Generator for a Giga-Scale DRAM with a Negative Word-Line Scheme," 1998 Symp. on VLSI Circuits, pp. 230-231.
- [26] M. Hiraki, et al., "A 63 μ W-Standby-Power Microcontroller with On-Chip Hybrid Regulator Scheme," 2001 Symp. on VLSI Circuits, pp. 225-228.
- [27] D. R. Ditzel, "Transmeta's Crusoe: A Low-Power x86-Compatible Microprocessor Built with Software," COOL Chips III, pp. 1-30.
- [28] T. Burd, et al., "A Dynamic Voltage Scaled Microprocessor System," ISSCC 2000, pp. 294-295.
- [29] M. Mizuno, et al., "Elastic-V_t CMOS Circuits for Multiple On-Chip Power Control," ISSCC 1996, pp. 300-301.
- [30] H. Mizuno, et al., "ChipOS: Open Power Management Platform to Overcome the Power Crisis in Future LSIs," ISSCC 2001, pp. 344-345.
- [31] T. Miyake, et al., "Design Methodology of High Performance Microprocessor using Ultra-Low Threshold Voltage CMOS," CICC 2001, pp. 275-278.