FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 15, No. 2, April 2002, 1-12

Trends in Low-Voltage Embedded-RAM Technology

Invited Paper

Kiyoo Itoh

Abstract: First, trends in the gate-oxide thickness of MOSFET for DRAM and MPU are discussed to clarify the strong need for low-voltage operation of embedded RAMs. Then, modern peripheral logic circuits for reducing leakage currents, and DRAM/SRAM cells to cope with the ever-decreasing signal charges are described. Finally, needs for developments of subthreshold-current reduction circuits for use in active mode, memory-rich SoC architectures, and gain cells and non-volatile cells are emphasized.

Keywords: Subthreshold current, gate tunneling current, DRAM, SRAM, peripheral circuits, gate-source backbias, dynamic V_T, multi-static V_T, signal charge, soft errors, memory-rich architectures, non-volatile RAM cells.

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 major challenges [1] to peripheral logic circuits and RAM cells. As for peripheral circuits (e.g. address buffers, decoders, row/column drivers, and other control logic circuits), they are the ever-increasing leakage (subthreshold and tunneling) currents of the MOSFETs,

Manuscript received March 16, 2002. An earlier version of this paper was presented at the 23rd International Conference on Microelectronics, MIEL 2002, May 12-15, 2002, Niš, Serbia.

The author is with Central Research Laboratory, Hitachi, Ltd. Kokubunji, Tokyo 185-8601, Japan (e-mail:k-itoh@crl.hitachi.co.jp).

and the chip-to-chip compensation for the increased variations of speed that are seen at lower levels of V_{DD}. For RAM cells, in addition to leakage current suppressions, stable operation to cope with reduction of signal charge due to low V_{DD} is vital. Other design issues for low-V_{DD} operation [2], such as on-chip supply-voltage converters, power management, testing methodology, and DA tools to manage the subthreshold current are also crucial.

In this paper, low-voltage eRAM circuits are discussed with emphasis on the subthreshold-leakage current and stable operation issues: First, trends in the gate-oxide thickness, which is closely related to V_{DD}, of MOSFET in DRAM and MPU are briefly discussed. Second, general features of leakage currents are clarified. Third, recent developments to do with the above issues for peripheral logic circuits, and DRAM and SRAM cells are investigated. Finally, a position is taken that emphasizes needs for gain cells, subthreshold-current reduction circuits for use in active mode, and memory-rich SoC architectures. Importance of non-volatile RAMs whose operations are not based on charges, such as MRAM and OUM, is also emphasized.

2 Current Generation of Low-Voltage eRAM Circuits

2.1 Trends in gate-oxide thickness

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 a MOSFET 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 a MOSFET 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 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]. Even a 3.3-ns cycle 6.6-ns access 16Mb macro with a dual-V_{DD} (1.5/2.5V) and triple t_{ox} (1.7/2.2/5.2 nm) approach was presented [4]. Therefore, urgent

issues for eRAMs of the near future are to develop a thinner-t_{ox} MOSFET without tunneling current, and low-voltage devices/circuits to keep up with the rapid lowering of V_{DD} in MPU.

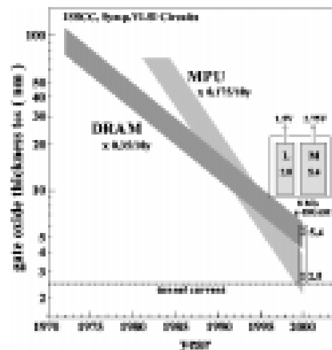
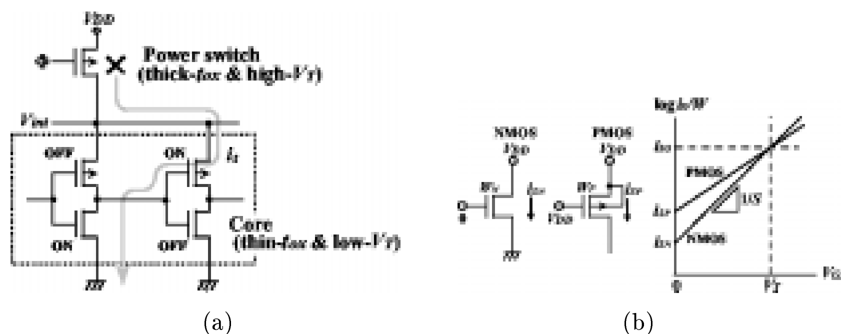


Fig. 1. Trends in gate-oxide thickness for DRAMs and MPU [2]. L, M: logic core and memory array .

2.2 General features of leakage currents

In addition to stable operations of RAM cells, leakage-current reduction is essential for all the low-voltage LSIs of the future. There are two major leakage currents; the gate-oxide tunneling current and the subthreshold current of the MOSFETs (Fig. 2). The tunneling current is proportional to $Wt_{ox}^{-2}\exp(-\beta t_{ox})$ for a given V_{DD}, where W and β are the channel width of the MOSFET and a constant, causing a larger standby current of chip with making t_{ox} thinner. A thick t_{ox} and high-V_T power switch (Fig. 2) [5]



(a) (b)

Fig. 2. Power switch for reducing gate tunnel current i_t (a) and subthreshold current i_s (b).

reduces the current of the internal thin-t_{ox} and low-V_T core with turning off

the switch during standby periods. A problem, however, is an unmanageably large current at active mode when further reducing t_{ox}. Thus, as the final solution, a new MOSFET with a high-k gate insulator material must be developed, which is up to device/process designers. The subthreshold current is proportional to $W10^{-V_T/S}$, Where S is the subthreshold swing, also causing a larger standby current of chip with lowering V_T. Note that more contribution of PMOSFETs to the current because of a larger total channel width and a larger S-factor [1].

2.3 Peripheral logic circuits

Many attempts [1] have been made to reduce the subthreshold current although they are almost limited to the standby mode. Gate-source backbiasing approach is effective and practical most for memories that feature to use many iterative circuit blocks. For example, a low-V_T PMOS switch (Qs in Fig. 3) shared with n word drivers allows for the common power

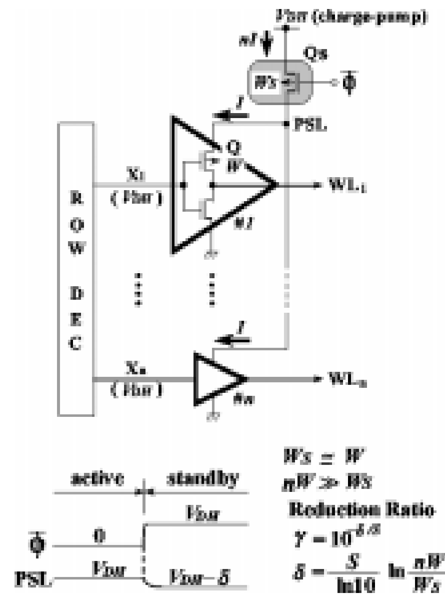


Fig. 3. Gate-source backbiasing approach applied to DRAM word drivers [1].

line (PSL) to drop by δ as a result of the total subthreshold current flow of nI, when the switch is off in standby mode. It gives a backbias of δ to each PMOS driver (Q) so that the subthreshold current (I) is eventually re-

duced. In active mode the selected word line is driven after connecting PSL to a supply voltage (V_{DH}) by turning on Qs. Here, the Qs channel width can be reduced to the extent comparable to that of the Q channel width without speed penalty because only one of n drivers is turned on. For a 256 Mb chip, a δ as little as 0.25 V reduces the standby subthreshold-current by 2-3 decades without inflicting penalties in terms of speed and area. The multi-static V_T approach (Fig. 4) is well known. In this approach a high-V_T power switch completely cuts off the standby current of the low V_T core

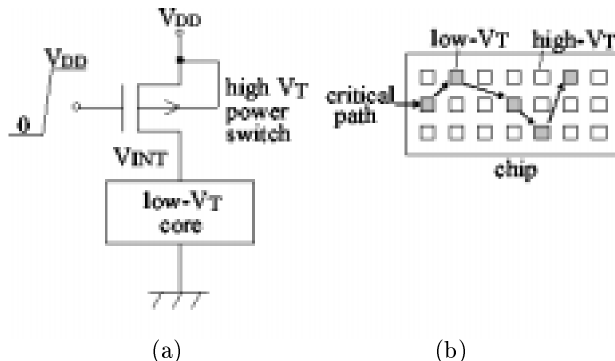
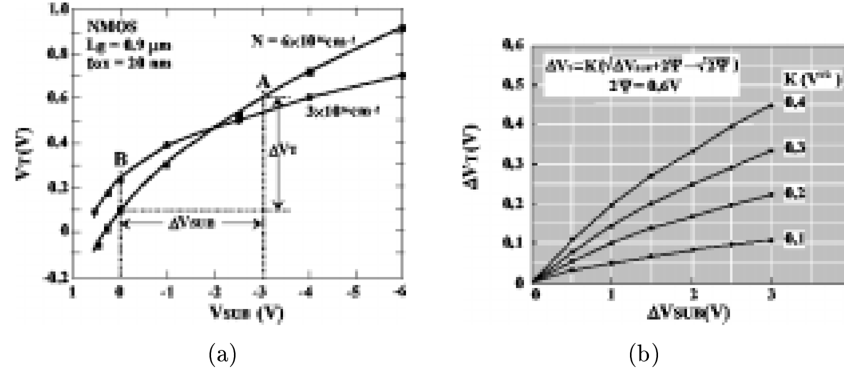


Fig. 4. Multi-static V_T applied to power switch (a) and signal paths [1]. A high-V_T MOSFET is available with ionimplatation or static substrate biasing.

with turning off the switch. The drawbacks, however, are the high power and slow recovery, which come from a large voltage swing at the internal power line, and a large switch. Application of a high V_T to the non-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. Applying the well-known dynamic substrate back-bias (V_{SUB}), as shown in Fig. 5, reduces the current by 1-2 decades. The substrate forward biasing recently proposed [6] achieves a larger V_T-change for a given V_{SUB}-swing, even though the forward bias is strictly limited to less than 0.4 V due to a rapid increase in the substrate current. The required V_{SUB} swing, however, is as large as 1-3 V (Fig. 5(b)) [2] and this approach becomes less effective with device scaling [7]. This is because of the smaller body constant, enhanced short-channel effects, and increases in other leakage currents such as the gate-induced drain-lowering (GIDL) current [7]. Even if the subthreshold current is still a small fraction of the active current because V_T is still high, the subthreshold current in



(a) (b)

Fig. 5. Dynamic-V_T approach with dynamic substrate-bias control (a), and V_T changes (b) as a parameter of body constant (K) [1]. K is usually 0.2-0.3.

the active mode makes the operation of dynamic circuits such as dynamic NAND for the decoder unstable, since the floating nodes are discharged. The level keeper proposed for logic circuits [8] 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.

2.4 Memory cells and relevant circuits

Soft Errors of RAM Cells: Soft error issue is increasingly important with lowering V_{DD} and device scaling. There are two kinds of soft errors (SEs); alpha-particle induced SEs, and cosmic-ray induced SEs [1]. Recently, it has been found that the neutron-induced SEs are more serious [9] because they generate about ten times as many charges as alpha- particles, as shown in Fig. 6. SE rate of DRAM cell decreases with device scaling due to an intentionally increased cell capacitance and spatial scaling that causes less collection of charges. On the contrary, SE rate of SRAM cell increases due to decreased parasitic capacitance of cell node despite spatial scaling. Solutions are to increase the signal charge, and to use a triple well structure and ECC.

DRAM: For the low-voltage one-transistor one-capacitor DRAM (1-T) cell (Fig. 7), using vertical capacitors and high-permmittivity thin-films to maintain sufficient signal charge, and adjusting the potential profile of the storage node to suppress the *pn*-leakage current [1] 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. 7) [10] cuts the subthreshold current from the storage node

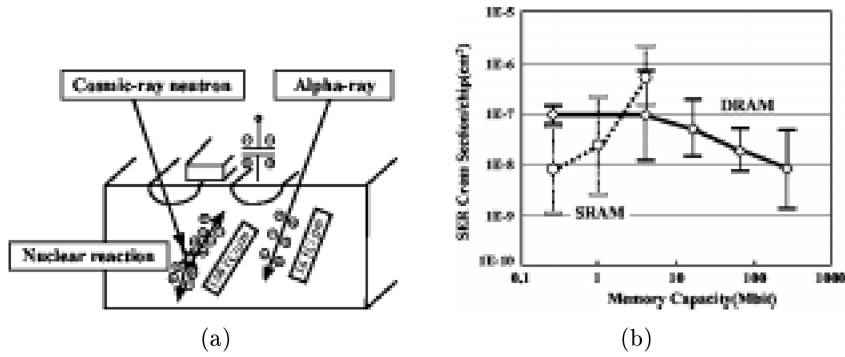


Fig. 6. Soft-error mechanisms (a) and soft-error cross section (b) in RAM cells [1, 9].

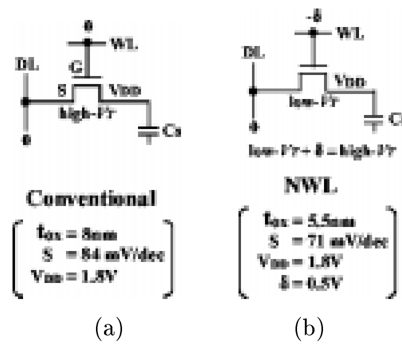


Fig. 7. Conventional high-V_T DRAM cell (a), and negative word-line (NWL) DRAM cell (b) for reducing subthreshold current despite a low-V_T MOSFET [1, 10].

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 a MOSFET with a thinner t_{ox} for a given device reliability. The resultant small S-factor enables low-voltage operation. The multi-divided data-line scheme, in which data-line capacitance (C_D) and line delay 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 ($\simeq C_S V_{DD}/2C_D$) 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$^{\text{TM}}$ [11], 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. Low-voltage high-speed sensing is also essential [2] 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 operation quite slow. Thus, many attempts [2] have been made.

SRAM: Of the many proposed designs [1], the 6-T full-CMOS cell (Fig. 8) 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

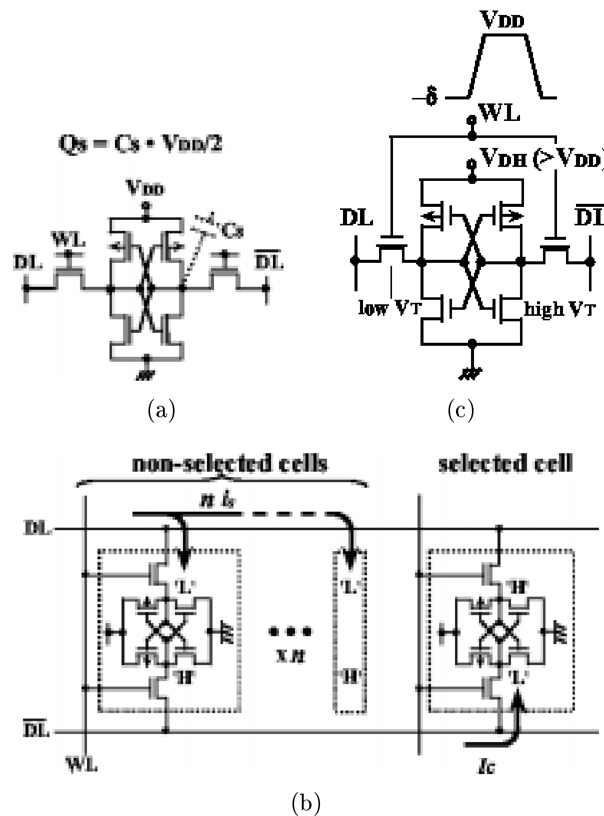


Fig. 8. Conventional SRAM cell (a), data pattern of cells along the data line (b), and low-voltage SRAM cell (c) [1]. Qs: signal charge.

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 [2] creates a demand for a decrease in the ratio of transconductance of the transfer MOSFET to that of the driver MOSFET. In addition, for a low-V_T transfer MOSFET, the margin is further worsened by the data pattern of non-selected cells in the column, as shown in Fig.8 (b). The total of the subthreshold leakage currents from the non-selected cells may be greater than the selected cell's current, causing 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 [2]. The voltage margin of the 6-T cell is also worsened by variations in V_T and mismatches of V_T between paired MOSFETs, both of which increase by device miniaturization [1].

The loadless CMOS 4-T SRAM [12] is attractive because the cell size is only 56% of that of a conventional 6-T cell. Transfer PMOSFETs 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 so as to keep the load current to more than a hundred times that of the off current of the driver MOSFET. The data-pattern problem outlined above also arises in this approach.

Almost all of the above problems may be solved by using high-V_T cross-coupled MOSFETs combined with a boosted power supply, as shown in Fig. 8(c) [13], in terms of a large signal charge, a low subthreshold current of cross-coupled MOSFETs, and V_T-imbalance immunity. The negative word-line scheme reduces the leakage currents from cells in the column. Using a high-V_T transfer MOSFET with word bootstrapping is an alternative approach that achieves the same result.

3 Future Trends

3.1 Peripheral logic circuits

The issue of leakage current presents a real challenge in the design and testing of a low-voltage high-speed SoC. For example, 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 eventually exceeds the capacitive current and dominates the total active current of the chip [1]. In this case, the low-power advantage of CMOS circuits is lost. Thus, the development of leakage-current reduction

devices/circuits is the key to tackling this challenge. If this problem is not solved, we can envision a scenario in which even CMOS SoCs would suffer from huge levels of dc power dissipation caused by leakage currents, as was the case in the bipolar and BiCMOS LSI eras of the recent past.

If a SoC consists of high-speed random logic gates and eRAMs, more serious problems will be caused due to the random logic gates. This is because control of subthreshold currents from the logic gates at sufficiently high speed may remain impossible. Fortunately, however, the subthreshold currents of eRAMs are effectively reduced by using iterative circuit blocks (see Fig. 3).

The following are the author's points of view on this issue. As for devices, developments of new high-k gate materials are essential, as discussed above. A fully-depleted SOI device with its inherently smaller S-factor will partly solve the subthreshold-current issue. For circuits, low-power techniques for bipolar and BiCMOS circuits might be revived to help reduce the power dissipation of SoCs. Such are the cases for old circuits such as current-mode logic, enhancement/depletion circuits, and gate-bootstrapping of high-V_T MOSFETs with capacitive coupling that were popular during the NMOS era in the 1970's [1]. With respect to architectures, reducing the number of random logic gates is one vital way of reducing the active-mode subthreshold current. Hence, new SoC architectures such as a memory-rich SoC [14], by which the subthreshold currents are effectively reduced, will be needed. Such architectures will enable a higher yield by using redundancy and ECC, and ease of test, if the soft-error issue for RAM cells is solved.

3.2 Memory cells

Existing RAM Cells: The 1-T DRAM cell is not suitable for low-voltage operation because its signal voltage reduction is not acceptable. Further 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 [2]. 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}. For eSRAM, the advanced 6-T cell (Fig. 8(c)) may continue to be used. In addition to the ever-increasing soft errors, a large 6-T cell size would be a serious concern in memory-rich architectures.

Non-Volatile RAM Cells: In the long run, high-speed high-density non-volatile RAMs would be attractive for eRAMs as well as stand-alone RAMs. In particular, non-destructive read-out RAM cells whose operations are not

based on charges, unlike DRAM cells, are important for achieving a fast cycle and soft-error-immune eRAMs. In this sense, MRAM (magnetic RAM) [15] and OUM (Ovonic Unified Memory) [16] are attractive. For MRAM, one of major challenges is to reduce the magnetic field necessary for switching the magnetization of the storage element, while for OUM it is to manage the proximity heating of the cell. At present, however, their scalabilities and stabilities for ensuring non-volatility still remain unknown as they are in early stages of development.

References

[1] K. Itoh: *VLSI Memory Chip Design*, Springer-Verlag, March 2001.
[2] K. Itoh and H.Mizuno: *Low-voltage embedded-RAM technology: present and future.* In: Proc. of the 11th IFIP Int. Conference on Very Large Scale Integration, pp. 393-398, Dec.2001.
[3] O. Takahashi et al.: *1GHz fully pipelined 3.7ns address access Time 8kx1024 embedded DRAM macro.*In: Proc. of the ISSCC2000, pp. 396-397.
[4] J. Barth et al., *A 300 MHz multi-banked eDRAM macro featuring GND Sense, bit-line twisting and direct reference cell write,.* In: Proc. of the ISSCC2002, pp. 156-157.
[5] T. Inukai et al.: *Suppression of stand-by tunnel current in ultra-thin gate oxide MOSFETs,.* In. Proc. of the SSDM1999, pp. 264-265.
[6] M. Miyazaki et al.: *A 175 mV multiply-accumulate unit using an adaptive supply voltage and body bias (ASB) architecture.* In: Proc. of the ISSCC2002, pp. 58-59.
[7] A. Keshavarzi et al.: *Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs.* In: Proc. of the ISLPED2001, pp. 207-212.
[8] A. Alvandpour et al.: *A conditional keeper technique for sub-0.13μm wide dynamic gates.* In: Proc. of the 2001 Symp. on VLSI Circuits, pp. 29-30.
[9] E.Ibe: *current and future trend on cosmic-ray-neutron induced single event upset at the ground down to 0.1-micron-devices,.* The Svedberg Laboratory Workshop on Applied physics, Uppsala, May 3, 2001.
[10] S. Miyano and M. Takahashi: *Embedded DRAM SOCs and its application for MPEG4 codec LSIs.* In: Proc. of VLSI Circuits Short Course, pp. 101-121, June 2001.
[11] W. Leung et al.: *The ideal SoC memory: 1T-SRAM.* In: Proc. of the 13th Annual IEEE Int. ASIC/SOC Conference, Sept.2000.
[12] K. Takada et al.: *A 16Mb 400MHz loadless CMOS four-transistor SRAM macro.* In: Proc. of the ISSCC2000, pp. 264-265.
[13] K. Itoh et al.: *A deep sub-V, single power-supply SRAM cell with multi-Vt, boosted storage node and dynamic load.* In: Proc. of the 1996 Symp. on VLSI Circuits, pp. 132-133.

[14] International Technology Roadmap for Semiconductors, 2001 Edition, System Driver (Fig.12)

[15] P.K. Naji et al.: *A 256kb 3.0V 1T1MTJ nonvolatile magnetoresistive RAM.* In: Proc. of the ISSCC2001, pp. 122-123.

[16] M.Gill et al.: *Ovonic unified memory - A high-performance nonvolatile memory technology for stand-alone memory and embedded applications.* In; Proc. of the ISSCC2002, pp. 202-203.