IEICE TRANS. ELECTRON., VOL.E93–C, NO.3 MARCH 2010

INVITED PAPER *Special Section on Circuits and Design Techniques for Advanced Large Scale Integration*

Adaptive Circuits for the 0.5-V Nanoscale CMOS Era

Kiyoo ITOH[†a)], ***Fellow, Honorary Member***, **Masanao YAMAOKA**[†], ***Nonmember***, ***and*** **Takashi OSHIMA**[†], ***Member***

SUMMARY The minimum operating voltage, V_{min}, of nanoscale CMOS LSIs is investigated to breach the 1-V wall that we are facing in the 65-nm device generation, and open the door to the below 0.5-V era. A new method using speed variation is proposed to evaluate V_{min}. It shows that V_{min} is very sensitive to the lowest necessary threshold voltage, V_{t0}, of MOSFETs and to threshold-voltage variations, ΔV_t, which become more significant with device scaling. There is thus a need for low-V_{t0} circuits and ΔV_t-immune MOSFETs to reduce V_{min}. For memory-rich LSIs, the SRAM block is particularly problematic because it has the highest V_{min}. Various techniques are thus proposed to reduce the V_{min}: using RAM repair, shortening the data line, up-sizing, and using more relaxed MOSFET scaling. To effectively reduce V_{min} of other circuit blocks, dual-V_{t0} and dual-V_{DD} circuits using gate-source reverse biasing, temporary activation, and series connection of another small low-V_{t0} MOSFET are proposed. They are dynamic logic circuits enabling the power-delay product of the conventional static CMOS inverter to be reduced to 0.09 at a 0.2-V supply, and a DRAM dynamic sense amplifier and power switches operable at below 0.5 V. In addition, a fully-depleted structure (FD-SOI) and fin-type structure (FinFET) for ΔV_t-immune MOSFETs are discussed in terms of their low-voltage potential and challenges. As a result, the height up-scalable FinFETs turns out to be quite effective to reduce V_{min} to less than 0.5 V, if combined with the low-V_{t0} circuits. For mixed-signal LSIs, investigation of low-voltage potential of analog circuits, especially for comparators and operational amplifiers, reveals that simple inverter op-amps, in which the low gain and nonlinearity are compensated for by digitally assisted analog designs, are crucial to 0.5-V operations. Finally, it is emphasized that the development of relevant devices and fabrication processes is the key to the achievement of 0.5-V nanoscale LSIs.

key words: *minimum operating voltage, SRAM, DRAM, FD-SOI, FinFET*

1. Introduction

Low-voltage scaling limitations of memory-rich CMOS LSIs are one of the major problems in the nanoscale era [1]–[4] because they cause the evermore-serious power crises with device scaling. The problems stem from two unscalable device parameters: The first is the high value of the lowest necessary threshold voltage V_t (that is,V_{t0}) of MOSFETs needed to keep the subthreshold leakage low. Although many intensive attempts to reduce V_{t0} through reducing leakage have been made since the late 1980s [4]–[6], V_{t0} is still not low enough to reduce the operating voltage, V_{DD}, to the sub-1 V region. The second is the variation in V_t (that is, ΔV_t), that becomes more prominent in the nanoscale era [1]–[4]. The ΔV_t caused by the intrinsic random dopant fluctuation (RDF) is the major source of various ΔV_t components. It increases with device scaling and thus intensifies various detrimental effects such as variations in speed (and delay) and/or the voltage margins of circuits, and it significantly increases the soft-error rates in RAM cells and logic gates. To offset such effects, V_{DD} must be increased with device scaling, which causes an increase in the power dissipation, as well as degrades the device reliability due to increased stress voltage. Due to such inherent features of V_{t0} and ΔV_t, V_{DD} is facing a 1-V wall in the 65-nm generation, and is expected to rapidly increase with further scaling of poly-Si bulk MOSFETs [1]–[4], as shown in Fig. 1. To reduce V_{DD}, the minimum operating power supply V_{DD} (that is, V_{min}), as determined by V_{t0} and ΔV_t, must be reduced, while the power supply integrity is ensured. This is because V_{DD} is the sum of V_{min}, ΔV_{ps}, and ΔV, where ΔV_{ps} is usually much higher than ΔV in the nanoscale era. Here, ΔV_{ps} is the power-supply droop and noise in the power supply lines and substrate. The ΔV is the sum of the voltage needed to compensate for the extrinsic ΔV_t due to short-channel effects and line-edge roughness and of the voltage needed to meet the speed target. Thus, ΔV depends on the quality and maturity of the fabrication process and on the design target, which cannot be specified here. An associated problem in the nanoscale era is the ever-higher resistance of interconnects [7]–[9]. This is closely related to the voltage-limitation problem at the chip and subsystem levels, since it not only degrades the speed of ever-larger chips, but also affects power supply integrity by increasing ΔV_{ps}. As well, integrity depends on the chip packaging such as 3D integration [10]. Mixed-signal LSIs present a similar problem, and special attention must be paid to the analog block on the chip because it consists of unique circuit configurations and elements, which differ from those of memory-rich LSIs (Fig. 2(a)). Differential and other circuits need an inherently higher V_{DD} to achieve a high gain and/or small offset. Moreover, some circuits require larger capacitors and high-Q inductors. In any event, for the LSI industry in order to flourish and proliferate, the 1-V wall must be breached in the nanoscale era. This requires a multidisciplinary approach since the problem covers different fields, including devices, circuits (digital and analog), and subsystems.

Concerns relating to adaptive circuits and relevant technologies to reduce V_{min} are addressed in this paper. The focus will mainly be on memory-rich LSIs, since such LSIs have usually driven the frontend of scaled devices development. Mixed-signal and other types of LSIs will sooner or later encounter similar problems. The V_{min} issue for

Manuscript received September 7, 2009.
Manuscript revised November 2, 2009.

[†]The authors are with Central Research Laboratory, Hitachi, Ltd., Kokubunji-shi, 185-8601 Japan.
a) E-mail: kiyoo.itoh.pt@hitachi.com

DOI: 10.1587/transele.E93.C.216

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers

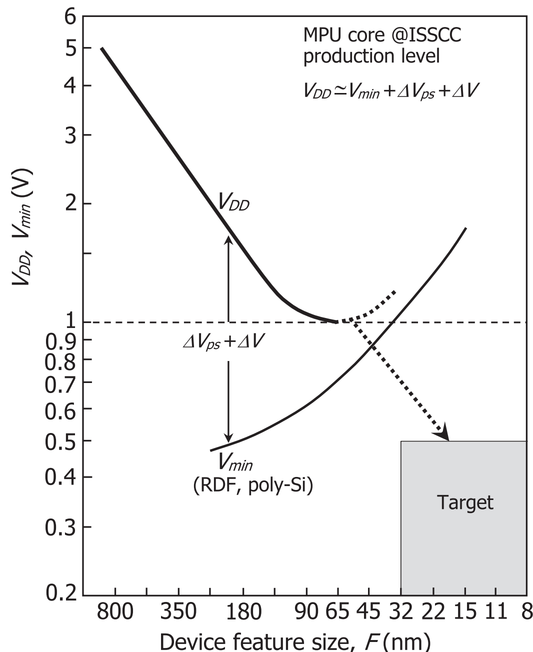
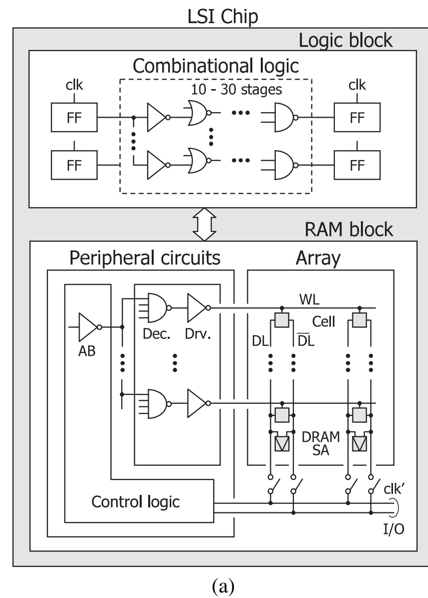


Fig. 1 Trends in V_{DD} and V_{min} of high-performance MPUs [3].



(a)

	Logic block	SRAM block		DRAM block		
		Periphery	Cells	Periphery	Sense amps	Cells
LW (av.)	4-12F^2	4-12F^2	1.5-2.5F^2	4-12F^2	10-20F^2	1F^2
V_t (av.)	0.2-0.4 V	0.2-0.4 V	0.2-0.7 V	0.2-0.4 V	0.2-0.4 V	0.7-1.3 V
t_{OX}	Thin	Thin	Usually thick	Usually thin	Thin	Thick
ΔV_t	Small	Small	Large	Small	Small	Large
Circuit count	Large	Small	Large	Small	Large	Large
Repair	No	No	Yes	No	Yes	Yes
Fan out	Small	Large*	–	Large*	–	–
Logical depth	Deep	Shallow	–	Shallow	–	–
Power off	Yes	Yes	No	Yes	No	No

*Iterative-circuit sub-blocks

(b)

Fig. 2 (a) LSI composed of logic block and RAM block; (b) features of blocks [1]. RAM block denotes SRAM block or DRAM block. AB: Address buffer.

memory-rich LSIs is described in the first part of the paper. First, V_{min}, as a methodology to evaluate the low-voltage potential of MOSFETs, is proposed in terms of a tolerable speed variation, and the general features are described. Then, the V_{min}s of logic gates, SRAMs, and DRAMs are compared, and state-of-the-art SRAM circuits to tackle the highest V_{min} problem of SRAMs are reviewed. After that, circuits and devices to reduce V_{min} to the sub-1 V region are described. Finally, the V_{min} issue for analog circuits in mixed-signal LSIs is briefly discussed.

2. Low-Voltage Scaling Limitations

If a MOSFET has an average V_t ($\cong V_{t0}$) with a maximum deviation (ΔV_{tmax}) in V_t from V_{t0}, the speed variation ($\Delta\tau$), that is, the ratio of the slowest speed at the highest V_t to the average speed at the average V_t ($\cong V_{t0}$) is approximately given as [1]

$$\Delta\tau = \{1 - \Delta V_{tmax}/(V_{DD} - V_{t0})\}^{-1.2}. \quad (1)$$

Fortunately, for conventional MOSFETs, $\Delta\tau$ was negligible up till about the 130-nm-device generation because V_{DD} is much higher than V_{t0} and ΔV_{tmax} is sufficiently small. In the nanoscale era below the 130-nm-device generation, however, $\Delta\tau$ rapidly increases with device scaling due to the ever-increasing ΔV_{tmax}, as shown in Fig. 3. To offset the increase, V_{DD} must be increased, but this results in a continually increasing V_{DD} with device scaling. If V_{DD} is reduced under such circumstances, the increase in $\Delta\tau$ becomes catastrophic, as seen in Eq. (1).

2.1 Definition of V_{min}

In practice, the increase in $\Delta\tau$ must be within a tolerable value ($\Delta\tau_0$) for reliable operation. The minimum operating voltage (V_{min}) is the V_{DD} necessary for achieving a tolerable $\Delta\tau_0$. Thus, V_{min} increases with device scaling, as shown in Fig. 3. V_{min} is obtained by solving Eq. (1) for V_{DD}:

$$V_{min} = V_{t0} + (1+\gamma)\Delta V_{tmax}, \; \gamma = 1/(\Delta\tau_0^{1/1.2} - 1),$$
$$\Delta V_{tmax} = m\sigma(V_t), \quad (2)$$
$$\sigma(V_t) = A_{vt}(LW)^{-0.5}, \text{ and } A_{vt} \propto t_{ox}. \quad (3)$$

For a conventional bulk MOSFET, $\sigma(V_t) = B_{vt}[t_{ox}(V_{t0} - V_{FB} - \Phi_S)/LW]^{0.5} \propto t_{ox}N_A^{0.25}(LW)^{-0.5}$, where m depends on the circuit count in the block, $\sigma(V_t)$ is the standard deviation of V_t distribution, A_{vt} and B_{vt} are the Pelgrom and Takeuchi constants [11], [12], respectively, t_{ox} is the inversion electrical gate-oxide thickness, V_{FB} is the flat-band voltage, Φ_S is the surface potential, N_A is the impurity concentration of the channel, and LW is the MOSFET size. The $\Delta\tau_0$ can take two values, $\Delta\tau_0(+)$ and $\Delta\tau_0(-)$, corresponding to plus and minus values of ΔV_t. Here, $\Delta\tau_0(+)$ will be used after this, simply expressed as $\Delta\tau_0$.

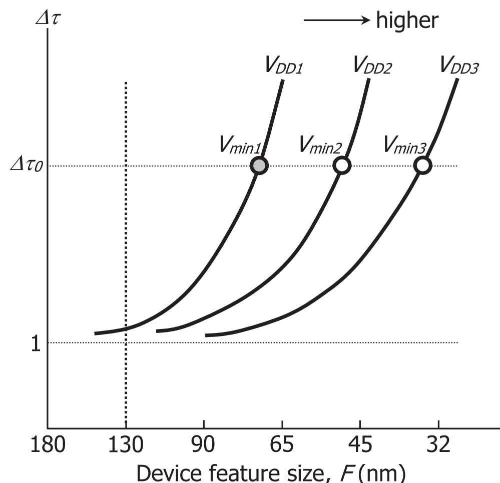


Fig. 3 Speed variation τ vs. device feature size, F.

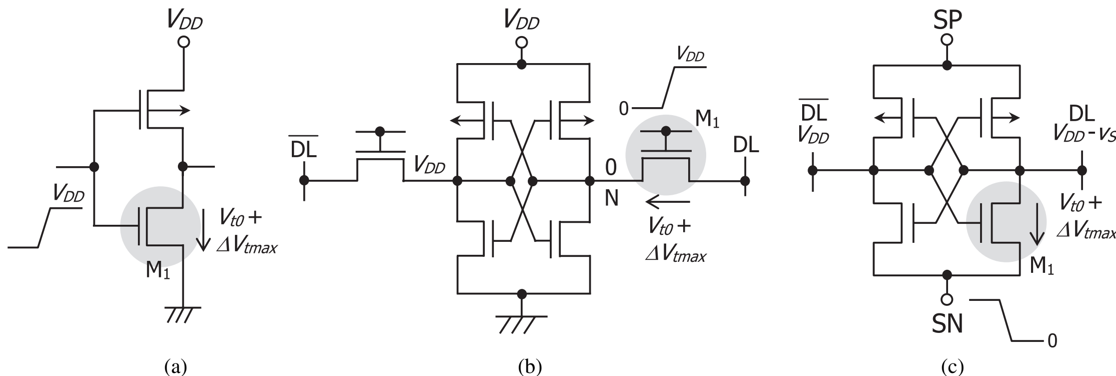
2.2 General Features of V_{min}

MOSFETs Governing V_{min}: The V_{min} of a chip is equal to the highest of V_{min}s of the three blocks (logic, SRAM, and DRAM) in the chip. The V_{min} of each block is governed by the circuit having the highest V_{min} in the block. Furthermore, the V_{min} of each circuit is governed by the MOSFET having the highest V_{min} in the circuit. Therefore, the V_{min} of each block is eventually determined by the MOSFET having the highest V_{min} in the block. Here, the MOSFET must be in a major core circuit impacting on power dissipation and speed of the block that are our major concerns in this paper. Note that the smaller the MOSFET, the higher its V_{min} with a larger $\sigma(V_t)$. If a specific MOSFET is used more often in the block, causing an iterative circuit block, the V_{min} of the MOSFET is statistically higher with a larger ΔV_{tmax}. Furthermore, the larger the C_L/W (C_L: the load capacitance), the higher the V_{min} with a larger γ. This is because $\Delta\tau_0$ must be smaller as the C_L/W is large, so it becomes less influential in the block speed. For RAMs, the V_{min} is also influenced by operation modes, that is, the non-destructive read-out (NDRO) and thus ratio operations for SRAMs, and the destructive read-out (DRO) and refresh operations for DRAMs. Taking these general features into account, the MOSFET can be specified as M_1 in each circuit in Fig. 4. Note that the DRAM sense amplifier (SA) operates simpler than the SRAM cell for lack of any transfer MOSFET despite the same cross-coupled circuit configuration. The details are in what follows.

For the logic block, the statistical expression for ΔV_{tmax} in Eq. (2) has some ambiguity, unlike RAM blocks. Each gate does not work independently and randomly, and some gates form logical configurations with considerable logical depth and small fan out (see Fig. 2(b)), enabling the $\sigma(V_t)$ to be reduced due to the averaging effect of random variations. The V_{t0} differs for some gates. For example, the well-known dual-V_{t0} logic block combines a low-V_{t0} MOSFET for the critical path and a high-V_{t0} MOSFET for the non-critical paths. The critical path tends to reduce the $\sigma(V_t)$ due to the low V_{t0} and large MOSFETs necessary to attain high speed. In addition, the small total MOS width of the path (typically about 10% of the total for the whole logic block) effectively reduces the m, so the non-critical paths inevitably determine the V_{min} of the whole block. Furthermore, the actual MOS size is different, ranging from 4 to $12F^2$(F is feature size). To validate the equation even for such a logic block, however, it is assumed that the logic block consists of many identical CMOS inverters, in which n/p MOSFETs have the same V_{t0} and size (that is, $LW = 8F^2$ on average). The V_{min} of the logic block calculated under these assumptions and using Eq. (2) may be higher than the actual V_{min} including at least the averaging effect. This is also the case for peripheral logic circuits in RAM blocks because the circuit configurations are almost the same as those of the logic block. For array-relevant circuits in RAM blocks, however, the expression for ΔV_{tmax} is valid since each of the cores comprises MOSFETs with the same V_{t0} and the same size, and operates independently and randomly.

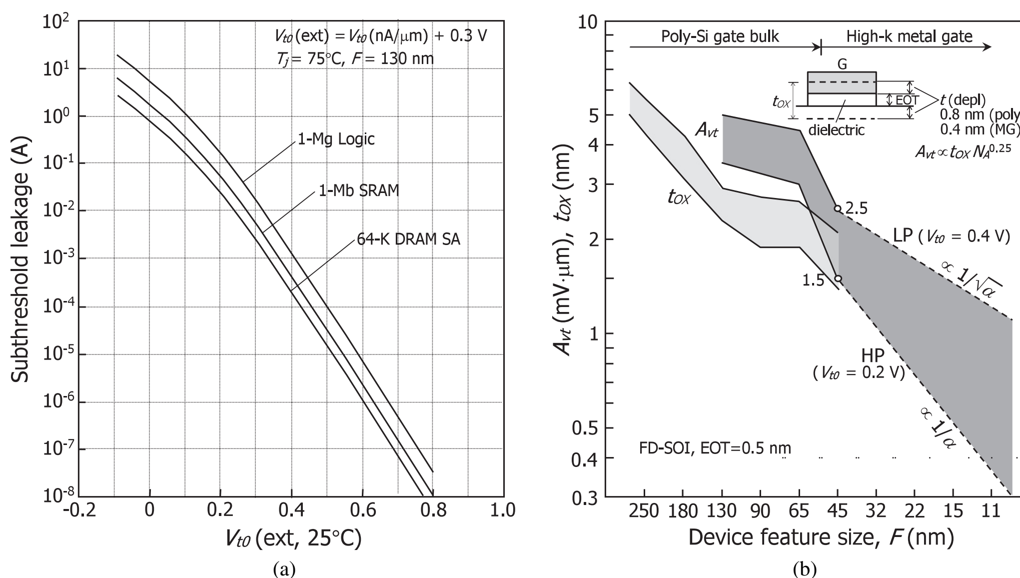
For SRAMs using the six-transistor (6-T) cell, the V_{min} is equal to the highest of the three V_{min} values determined by cell stabilities at write and read, and tolerable speed variation at read. The V_{min} for write stability can be reduced sufficiently by power control of pMOSFET loads [26], [27] for a wider write margin, as explained in Sect. 2.4. The V_{min} for read stability can also be lowered by reducing the word-line voltage from V_{DD} [60], as will be mentioned later. Hence, the V_{min} of SRAMs is determined by the speed variation of the transfer MOSFET. Unfortunately, the MOSFET always involves a slow and wide speed variation. The drawbacks come from the smallest MOSFET and the source voltage raised from ground (V_{SS}) level, caused by a ratio operation of the transfer and driver MOSFETs, and the largest V_{t0} variation due to the largest MOSFET count. The V_{min} can be calculated with Eq. (2) on the assumption that the source stays at 0 V during read operations, although this assumption makes the V_{min} lower than the actual V_{min} taking the raised source voltage into account. Here, the size of the transfer MOSFET is assumed to be $1.5F^2$. The V_{t0} is also assumed to be the same as those of cross-coupled MOSFETs shown in Fig. 5(a), since their leakage currents must be comparable in conventional designs.

For DRAMs, the DRO calls for restoring of the cell [4] by utilizing the amplified signal by SA. It takes a long time because a small read signal must be amplified to a full V_{DD} on the heavily capacitive data line, requiring a small $\Delta\tau_0$ and thus high V_{min} for confining the array speed within a tolerable value. Moreover, the refresh operation calls for simultaneous restoring of many cells along the selected word line. This involves charging and discharging of many heavily capacitive data-lines and operations of many SAs at a high voltage, causing high power. If the full-V_{DD} sensing (i.e., full-V_{DD} data-line precharging) is used, and activation of cross-coupled nMOSFETs in an SA precedes that of cross-coupled pMOSFETs [1], [4], Eq. (2) is appli-



(a) (b) (c)

Fig. 4 (a) Inverter, (b) 6-T SRAM cell, and (c) DRAM sense amplifier.



(a) (b)

Fig. 5 (a) Leakage vs. V_{t0} for various blocks; (b) trends in t_{ox} and A_{vt} [3].

cable to M_1. The challenge of the full-V_{DD} sensing, however, is to generate a stable and reliable reference voltage for signal discrimination [4]. If it is difficult to accomplish, the conventional mid-point sensing [4] (i.e. half-V_{DD} data-line precharging) must be used instead, although the sensing doubles the V_{min} of full-V_{DD} sensing. A remedy for the doubled V_{min} problem will be discussed in Sect. 3.1. Note that others are not core circuits. DRAM cells adopt the well-known word bootstrapping to perform a full-V_{DD} write of the cell [4], in which the word-line voltage is higher than the sum of the highest data-line voltage and the V_t of the cell transfer MOSFET. Therefore, to be exact, the word driver can have the highest V_{min} in the block. However, the driver quickly drives the word-line with a large MOSFET, and less contributes to power dissipation of the block because only one word diver is activated, unlike SAs. Moreover, in the past, DRAMs have solved the high voltage problem by using high-voltage tolerant word drivers [4]. Furthermore, although the transfer MOSFET has the largest ΔV_{tmax} in the block due to the largest size/count, it never dominates the block speed. In fact, the developing speed of cell signal on the data line is quickly and insensitive to the V_t-variation thanks to a small voltage swing needed on the data-line and the word-bootstrapping. In any event, the full-V_{DD} sensing is assumed in the following, and the size of the nMOSFET is also assumed to be $15F^2$.

Lowest necessary $V_t(V_{t0})$: The lowest necessary V_t for the above-described MOSFETs depends on subthreshold-leakage specifications. Figure 5(a) plots the leakage versus V_{t0} and was prepared using previously reported SRAM data [13] for a device feature size of 130 nm. Note that V_{t0} is an extrapolated value that is familiar to circuit designers [5]. It is the sum of constant-current V_t (nA/μm) that is familiar to device designers and 0.3 V [5]. Moreover, the total MOS size, which governs subthreshold leakage, is assumed to be $16 \times 10^6 F^2$ for 1-Mgate logic if a gate generates leakage from two MOSFETs with an average size of $8F^2$; $3.5 \times 10^6 F^2$ for 1-Mb SRAM if a 6-T cell generates leakage from the two MOSFETs (total $LW = 3.5\ F^2$) of four cross-coupled MOSFETs in the cell; and $2 \times 10^6\ F^2$ for 64-k DRAM SAs, which contribute to leakage in the active standby mode if each SA generates leakage from two MOS-

FETs with an average LW of 15 F^2 for each. Obviously, the V_{t0} depends on the leakage. If the tolerated leakage is about 1 to 100 mA for a 1-Mgate logic block, 0.5 to 70 mA for a 1-Mb SRAM, and 0.15 to 20 mA for 64-k DRAM SAs, the V_{t0} is between 0.2 V (for high-speed designs) and 0.4 V (for low-power designs). However, the leakage of the chip increases as logic gate and memory integration in the chip increases. Many reduction circuits have been developed for offsetting the increase, as exemplified by power gating with power switches [1]–[4]. Further reduction in V_{t0}, however, requires the development of innovative low-leakage circuits.

Parameter γ: This parameter strongly depends on the tolerable speed variation, $\Delta\tau_0$. In general, the logic block needs a small $\Delta\tau_0$ (that is, large γ) because the timing control must be quickly and stringently managed so as to meet the targeted speed from one flip flop (FF) to the other at every combinational logic stage (Fig. 2(a)). The speed is usually one clock latency when measured in terms of the necessary number of clocks. In contrast, for RAM blocks, such a quick and stringent timing control is extremely difficult because of a large physical memory array, which inherently contains large delay components throughout the array. This difficulty occurs to the SRAM cell and the DRAM SA, each of which dominates the block speed with a large C_L/W. For example, a small transfer MOSFET in an SRAM cell must discharge a heavily capacitive data (bit) line, which takes a long time. Unfortunately, the discharge time varies greatly due to a wide variation in the V_t of the MOSFET and the ratio operation. The discharging signal must be aligned with a column clock (clk' in Fig. 2(a)), waiting for the signal from the slowest cell, so that the signal transferred to I/O is discriminated correctly. Such an operation unavoidably tolerates a two-clock latency, as typically seen in actual designs, as a result of giving up one-clock latency that requires an extremely high V_{min} to offset the speed variation. This is also the case for DRAM SAs. Therefore, γ = 3.09 and $\Delta\tau_0$ = 1.4 for the logic block and γ = 2.09 and $\Delta\tau_0$ = 1.6 for the SRAM and DRAM blocks are used here, with practical designs taken into account.

Maximum deviation, ΔV_{tmax}: The number m ranges from 4.9 to 6.0 for 0.6- to 320-Mgate logic blocks, from 5.2 to 6.3 for 4-Mb to 2-Gb SRAMs, and from 4.8 to 5.9 for the 16-Mb to 8-Gb DRAMs connecting 64 cells to an SA [4]. It also depends on the repairable percentage, r, for RAMs. For the upper limit of r (that is, 0.1% for SRAMs and 0.4% for DRAMs), attained by a combination of error correcting code (ECC) and redundancy, m is reduced to about 3.29 for SRAMs and to about 2.88 for DRAMs [1], [2]. Note that, for a conventional bulk MOSFET, $\sigma(V_t)$ also depends on V_{t0}, as mentioned above. For $V_{FB} = -0.9$ V and $\Phi_S = 0.8$ V, $\sigma(V_t)$ is reduced to 0.45 of $\sigma(V_t = 0.4\,\text{V})$, when V_{t0} is reduced from 0.4 to 0 V. Furthermore, $\sigma(V_t)$ depends on A_{vt} and F^2, as expected from Eq. (3).

The expected trends in t_{ox} and A_{vt} are plotted in Fig. 5(b). For 130-nm poly-Si gate bulk nMOSFETs [14], [15], A_{vt} is about 4.2 mV·μm when V_{t0} and t_{OX} are 0.30 to 0.45 V and 2.1 to 2.4 nm, respectively. The most advanced planar MOSFETs in the 45-nm generation have a low A_{vt} (1.0 to 2.5 mV·μm) [16]–[18] with high-k metal-gate materials for a thinner t_{OX} and/or a fully-depleted silicon-on-insulator (FD-SOI) structure for a smaller N_A. Figure 6 plots trends in the $\sigma(V_t)$ for three values of A_{vt} [3]. Obviously, the $\sigma(V_t)$ of each block rapidly decreases with A_{vt}.

2.3 Comparison of V_{min} for Logic Block, SRAMs, and DRAMs

Figure 7 compares the V_{min} for the logic block and repaired RAMs for three values of A_{vt} [3], showing the strong dependence of V_{min} on A_{vt}. For A_{vt} = 4.2 mV·μm, the V_{min}s of the logic and SRAM blocks were almost the same but still high, reaching an intolerable level of about 1.5 V in the 32-nm generation. For A_{vt} = 1.5 mV·μm, however, they were reduced to less than 1 V even in the 22-nm generation. Obviously, the V_{min} of DRAMs is the lowest due to the smallest $\sigma(V_t)$ and fewer SAs. The prime concern is the SRAM because its V_{min} is actually the highest when repair techniques are not used and the raised cell-node voltage is taken into consideration.

2.4 State-of-the-Art SRAM Circuits

Recent research on high-speed 6-T SRAMs has focused on widening the voltage margin at a fixed operating voltage of around 1 V rather than reducing V_{min} and thus V_{DD}. Managing the power of the cell is an effective way of tackling the rapidly degrading voltage margin caused by an ever increasing $\sigma(V_t)$, despite a lithographically symmetric cell layout being used [4]. Figure 8 illustrates three practical 6-T cells using power management and an 8-T cell. The one shown in (a) has a cell supply voltage higher than the data-line voltage, V_{DL}. The combination of a low-V_t(V_{tL}) transfer MOSFET and a negative word-line scheme [19] results in a read margin wider than that of a high-V_t (V_{tH}) transfer MOSFET and boosted word-line scheme [20] combination. This is because the low V_t reduces the $\sigma(V_t)$ for conventional MOSFETs. In this scheme, as the data (bit)-line voltage can be scaled in accordance with MOSFET scaling in the peripheral circuits, high density and low power are achieved for data-line-relevant circuits. A reduced word-line voltage scheme in accordance with the V_{t0} of the transfer MOSFET [60] has also been proposed to widen the read margin. Dynamic power control of the driver nMOSFET [13], [21]–[23] (Fig. 8(b)) or load pMOSFET [24], [25] reduces the V_t of the MOSFETs in active mode (ACT) while reducing leakage in standby mode (STB) with increased V_t (= δV_t) due to the body bias effects. Power control of pMOSFET loads (Fig. 8(c)) [26], [27] to increase load impedance during write periods improves the write margin. It has been reported that 8-T SRAM cells (Fig. 8(d)) [28] widen the read and write margins due to separation of the read and write functions in a cell. This is true for the selected cell. However, the half-selection problem is always involved for the non-selected cells along the selected word line. A read op-

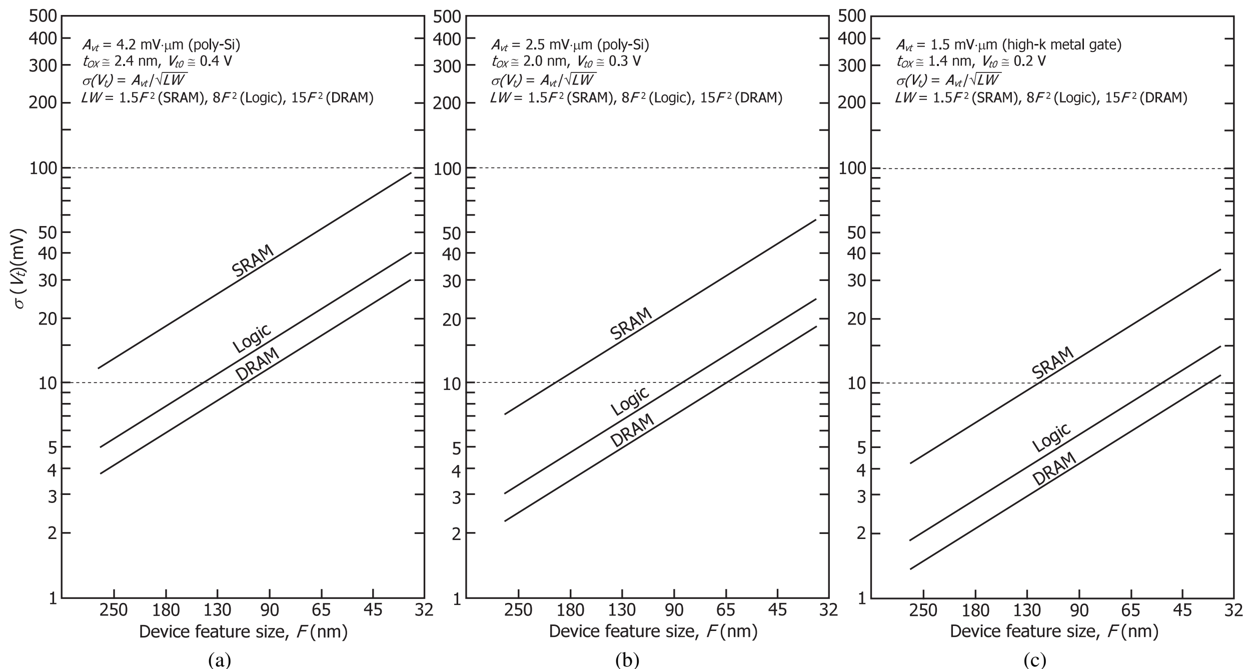


Fig. 6 Trends in $\sigma(V_t)$ for (a) $A_{vt} = 4.2\,\text{mV}{\cdot}\mu\text{m}$, (b) $A_{vt} = 2.5\,\text{mV}{\cdot}\mu\text{m}$, and (c) $A_{vt} = 1.5\,\text{mV}{\cdot}\mu\text{m}$ [3].

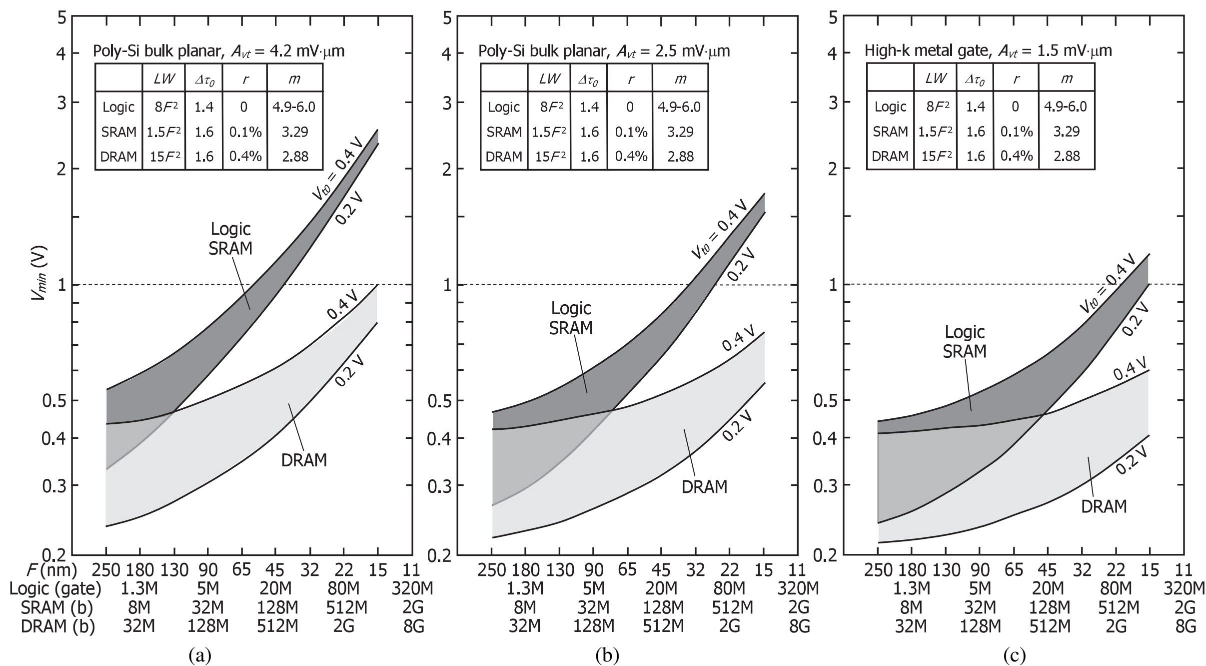


Fig. 7 V_{min}s for the logic block and repaired RAMs for various MOSFETs having (a) $A_{vt} = 4.2\,\text{mV}{\cdot}\mu\text{m}$, (b) $A_{vt} = 2.5\,\text{mV}{\cdot}\mu\text{m}$, and (c) $A_{vt} = 1.5\,\text{mV}{\cdot}\mu\text{m}$ [3].

eration is performed using read path M_4–M_5 without ratio operation of M_1 and M_2, unlike for the 6-T cell. M_1 can thus be enlarged for a wider write margin of the selected cell while M_2 is kept the same, resulting in a tolerable increase in cell size. However, half-selected cells are all read, necessitating the ratio operation of M_1 and M_2, while the selected cell is written. The reduced ratio of M_1 and M_2 for the non-selected cells, however, tends to cause destructive read operations due to the reduced margin. Therefore, application of the 8-T cell is strictly limited to wide bit configurations, in which all cells along the same word line are simultaneously written.

Shortening the data line [29] reduces the V_{min} of the 6-T cell because a large speed variation $\Delta\tau_0$ is allowed. For example, if the data-line length is halved to increase $\Delta\tau_0$ from 1.6 to 3.6, V_{min} is reduced, as shown in Fig. 9(a). Us-

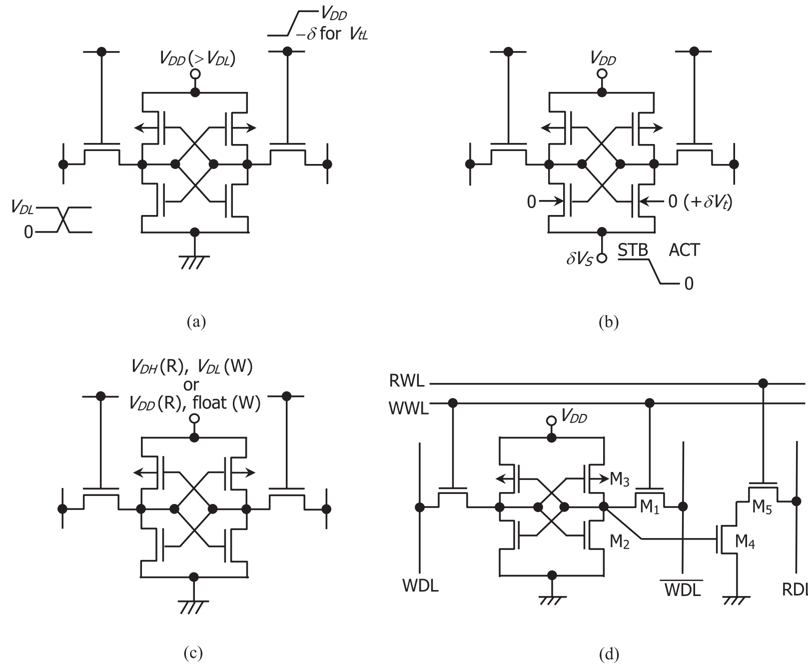


Fig. 8 Practical schemes to maintain voltage margin of SRAM cells: (a)–(c) for 6-T cell, and (d) for 8-T cell.

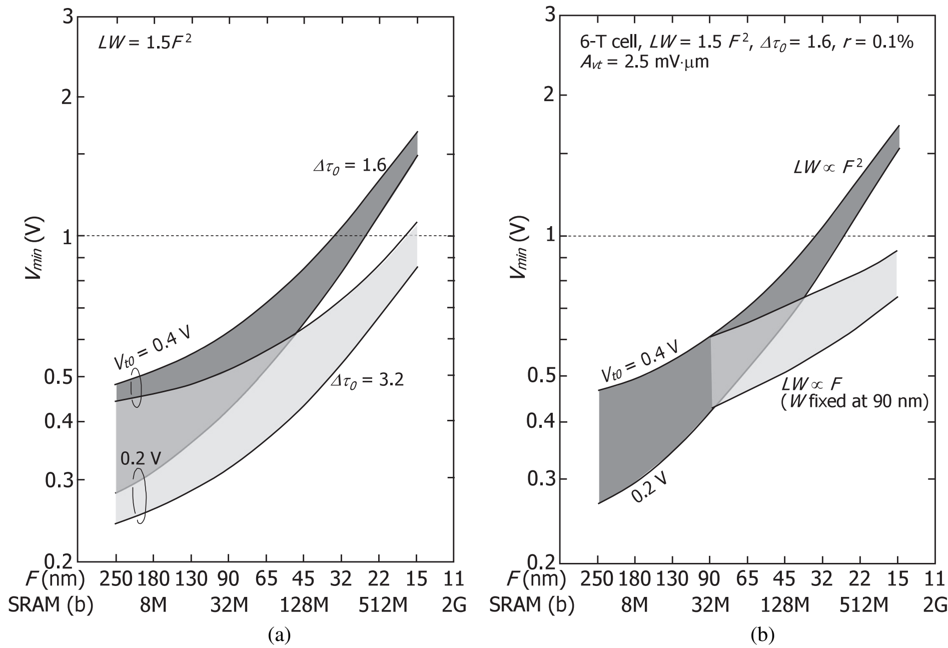


Fig. 9 (a) V_{min} of 6-T cell: (a) shortening data line and (b) up-sizing.

ing the largest MOSFET possible (i.e., up-sizing) [13], [23] in the 6-T cell also reduces V_{min} with reduced $\sigma(V_t)$. For example, if the channel lengths of all MOSFETs are scaled down while keeping the channel widths fixed, such as in the 90-nm generation (where $LW \propto F$ with W fixed at 90 nm; Fig. 9(b)), the increase in V_{min} can be suppressed. In contrast, with conventional scaling (that is, $LW \propto F^2$), V_{min} rapidly increases as F decreases. The cell size (Fig. 10) of the W-fixed approach, however, is gradually reduced since all Ws in the cell are fixed at each generation. Thus, the size becomes equal to that of an 8-T cell having a size of 156 to $185F^2$ in the 45-nm generation, while conventional scaling reduces cell size more rapidly (that is, to $120F^2$). In practice, the sizes of MOSFETs in a 6-T cell can be adjusted between the two approaches, so the V_{min} is between about 0.6 and 1 V in the 32-nm generation for $A_{vt} = 2.5\,\text{mV}{\cdot}\mu\text{m}$,

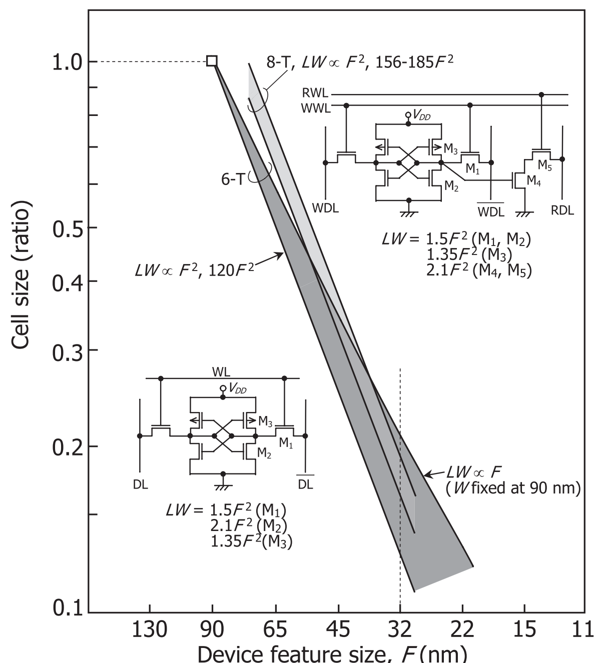


Fig. 10 Cell size of 6-T and 8-T cells [3].

as seen in Fig. 9(b). The investigation suggests that multiple cell sizes and types combined with multi-V_{DD} operation on a chip are feasible, depending on the length of the data line and the required memory chip capacity. For example, for a small-capacity SRAM, in which overhead due to the use of ECC is intolerable but a larger cell size is tolerable, up-sizing of MOSFETs in the cell enables low-V_{DD} operation. For a large-capacity SRAM necessitating a small cell size, repair techniques and/or a dedicated high-voltage supply are a viable solution. However, even if V_{DD} can be managed so that it remains at about 1 V even in 45- to 32-nm generations, it will still continue increasing, especially for conventional scaling aiming at higher density, as long as conventional MOSFETs are used.

3. Challenges to Low-Voltage Circuits and Devices

If the V_{min} of each block needs to be lowered by a factor of at least $\alpha^{-0.5}$ (α: scaling factor > 1) by device scaling, and given the past trends (Fig. 1), both V_{t0} and ΔV_{tmax} must be scaled down by the same factor, as predicted by Eq. (2). Thus, repair techniques for RAMs, shortening the RAM data lines, and relaxed size scaling and up-sizing of MOSFETs, as described above, are crucial. In addition, a real challenge is to develop low-V_{t0} circuits. To minimize V_{min}, V_{t0} must be made much lower than that in Fig. 5(a), which means that leakage must be drastically reduced. Another challenge is to develop new MOSFETs suitable for low-voltage operations, such as small-A_{vt} MOSFETs for small $\sigma(V_t)$ and/or $\sigma(V_t)$-scalable MOSFETs. Indeed, conventional poly-Si gate MOSFETs having an A_{vt} as large as 2.5 to 4.2 mV·μm are of no use in reducing V_{min}, as mentioned above.

3.1 Dual-V_{DD}, Dual-V_{t0} Circuits

Low-V_{t0} MOSFETs in a circuit reduce the V_{min} of the circuit, thus enabling the use of low V_{DD}, as mentioned above. Their major challenge is to reduce the resultant leakage. Three examples of such circuits will be discussed here. They are logic circuits utilizing gate-source reverse biasing, a low-V_{t0} temporarily activated DRAM pre-amplifier, and power switches using series-connected small low-V_{t0} MOSFETs. Here, V_{t0} is defined as the sum of V_{t0} (nA/μm) and 0.3 V, as explained previously.

Logic Circuits: One way to reduce the resultant leakage is to make the V_{t0} effectively high. Obtaining an effectively high V_{t0}, despite a low actual V_{t0}, can be achieved by using the gate-source reverse biasing with the help of a high V_{DD} provided by a high-V_{DD}, high-V_{t0} circuit. This necessitates the use of a dual-V_{DD}, dual-V_{t0} circuit and a dynamic circuit configuration. The V_{min} of the whole circuit is higher because it is equal to that of the high-V_{DD}, high-V_{t0} circuit. This dual circuit, however, is extremely vital to reduce power dissipation if low-V_{DD}, low-V_{t0} MOSFETs are used in the outputs of inherently high-power circuits, such as the buffers, for driving heavy capacitive loads. For a given power dissipation, the use of such a circuit effectively reduces the V_{min} of the whole circuit. Figure 11(a) shows the concept of dual-V_{DD} (V_{DD}; $V_{DL} < V_{DD}$) and dual-V_{t0} (V_{tH}; $V_{tL} < V_{tH}$) dynamic circuits using gate-source (G-S) reverse biasing [30]. It works with a large difference in V_{t0}, as exemplified by a high $V_t(V_{tH})$ of 0.4 V and a low $V_t(V_{tL})$ of zero. For example, reverse biasing is applied to a V_{tL}-pMOSFET during inactive periods with the help of a higher power supply. As a result, a sufficiently high $V_{t0}(V_{teff})$, despite a low-actual V_{tL}, is obtained, thereby reducing leakage during inactive periods. Even so, the gate-over-drive voltage (effective gate voltage, V_{geff}) is maintained at a high level during active periods. Thus, V_{t0} can be scaled by adjusting the gate-source bias. Note that even depletion (normally on) MOSFETs (i.e., D-MOSFETs) can be used, as long as the MOSFET is cut by using a sufficiently high V_{DD}.

Figure 11(b) illustrates application of this concept to a self-resetting inverter [31]–[34]. When M_1 is on, the output goes to low, and the gate of M_2 becomes low, so M_2 drives the output to V_{DL}. Subsequently, M_2 is kept off because a high-V_{DD}, high-V_{t0} CMOS inverter chain drives the gate to V_{DD}, so the gate-source is reverse biased by $V_{DD} - V_{DL}$. Although a leakage flows at the first inverter in the chain when output is at V_{DL}, it is small due to the small W. Figure 12(a) shows another application — to a dynamic inverter (D-INV) with a V_{DD}-clock, CK, [1], [3], [35], in which a low V_{t0}, V_{tL}, is assigned only for input detector M_1 and output driver M_3. Node N and output OUT are at V_{DD} and zero, respectively, during inactive periods, while M_3 is off owing to reverse biasing by $V_{DD} - V_{DL}$. Once the CK enables the set of differentially driven V_{DL}-inputs (IN, /IN), N is discharged in case of 0-V IN, causing M_3 to drive the output to V_{DL} while cutting off M_4. Here, the substrate of M_3 is connected to

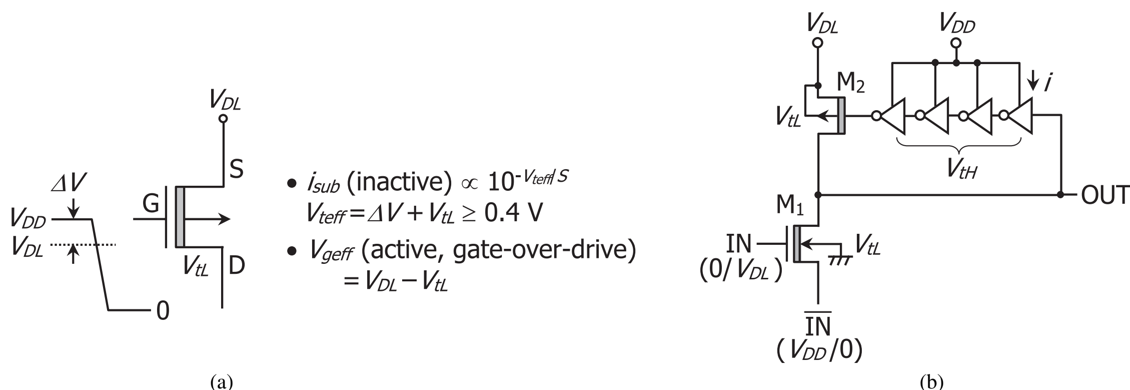


Fig. 11 Dual-V_{DD}, dual-V_{t0} circuits using gate-source offset driving: (a) concept behind gate-source offset driving [3] and (b) self-resetting inverter [3].

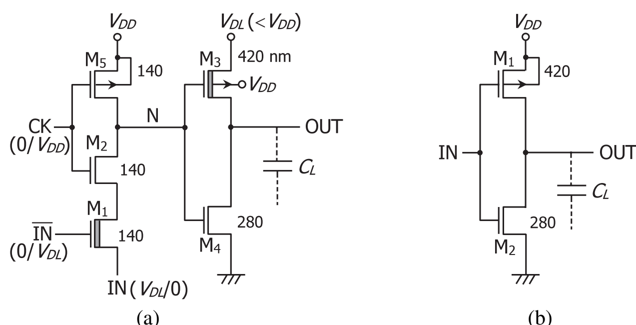


Fig. 12 (a) Dual-V_{DD} and dual-V_{t0} dynamic inverter (D-INV) [35]; (b) conventional high-V_{t0} static inverter (S-INV).

V_{DD} to eliminate an additional well isolation. The concept of D-INV is widely applicable to NAND, NOR, and other logic circuits [1]. When compared with the conventional static inverter (S-INV, Fig. 12(b)) with the assumption of a fixed total width of 700 nm for both output inverters, the D-INV reduces delay to 0.49 (i.e., 0.78 to 0.38 ns) for V_{DL} = 0.2 V, $V_{t0}(M_3)$ = −0.2 V (i.e., depleted), and $V_{t0}(M_1)$ = 0.1 V, while reducing power dissipation to 0.18 (i.e., 138 to 25 nW at a 20-ns cycle time), as shown in Fig. 13. Thus, the power-delay product is reduced to about 0.09. The D-INV further improves performance when driving a larger load capacitance.

DRAM Sense Amplifier: Second way to reduce the leakage is to temporarily activate the low-V_{t0} circuit while leaving the subsequent low-leakage operations to a high-V_{DD}, high-V_{t0} circuit. This concept is vital to reduce the V_{min} of DRAMs using the mid-point sensing. The mid-point sensing (i.e., half-V_{DD} data-line precharging) has widely been used for DRAM products due to advantages of generation of a stable reference level, low power and low noise [5]. Unfortunately, however, the sensing doubles the V_{min} of the full-V_{DD} sensing because the V_{DD} in Eq. (1) is regarded as $V_{DD}/2$, making the V_{min} equal to $2(V_{t0} + (1{+}\gamma)\Delta V_{tmax})$. In principle, even for the mid-point sensing, the V_{min} is maintained to the same, if V_{t0} and ΔV_{tmax} are halved.

Figure 14 compares two mid-point sensing schemes; the conventional sensing and a new sensing using the above-described concept [36], [37]. The conventional sensing (Fig. 14(a)) necessitates using a dual-V_{DD} (a half-V_{DD} and V_{DD} in this case) and a high-V_{t0} SA for performing two functions, sensing and data holding. In this sensing, a small signal, v_s, is read out on the floating data line after the data line is precharged to a half-V_{DD} level and then amplified by activating the SN. After that, the amplified signal is latched and held in the SA by activating the SP. Therefore, for the nMOS M in order to turn on when the SN is activated, $V_{DD}/2$ must be higher than the V_t. Since the V_t must be higher than 0.35 V for the succeeding low-leakage data holding, V_{DD} must be higher than 0.7 V. The circuit shown in Fig. 14(b) features separation of sensing and data holding with two SAs [36]: a low-V_{t0} temporarily activated pre-amplifier (PA) for low-voltage, low-leakage sensing, and a conventional high-V_{t0} SA for low-leakage data holding. After amplifying the signal by applying a short pulse, P, the low-V_{t0} PA is turned off to cut the leakage path. The high-V_{t0} SA is then activated to latch and hold the amplified signal. In this manner, low-voltage sensing and low-leakage data holding are simultaneously performed. To be more precise, the PA stops amplification when DL drops to $V_{t0}(M_P)$, enabling the signal to be finally amplified to $V_{t0}(M_P)$. For the high-V_{t0} SA in order to successfully latch the amplified signal, $V_{t0}(M_P)$ must be higher than the offset voltage of the high-V_{t0} SA. This voltage is usually less than 0.2 V. Therefore, $V_{t0}(M_P)$ must be higher than 0.2 V, and the PA thus turns on when the half-V_{DD} is higher than 0.2 V. This implies that V_{DD} can be reduced to 0.4 V, which is almost half the voltage of a conventional SA. In addition, a low-offset-voltage PA can be achieved owing to a low-V_{t0} M_P. The area penalty is 2.2% for a 128-Mb DRAM [36].

Power Switches: Third way to reduce the leakage is to confine the leakage with series-connected small leaky (that is, low-V_{t0}) MOSFETs. The applications to power switches are particularly vital, considering key roles of power switches in the nanoscale era. The details are in what follows. Small cores and chips, new architectures such as multi-core MPUs, and 3-D thermally conscious small-chip integration with high-density through silicon vias (TSVs) [10] will enable the development of compact subsystems, which, with their

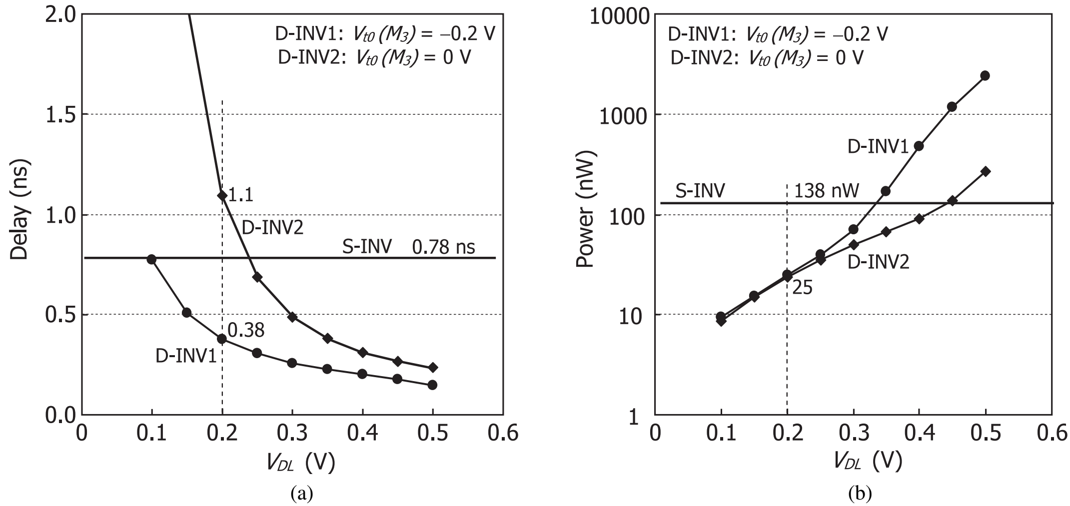


Fig. 13 (a) Delay and (b) power dissipation of dynamic inverter (D-INV) compared with those of the conventional static inverter (S-INV): For D-INV, V_{t0} $(M_1) = 0.1$ V, V_{t0} $(M_3) = 0$ or -0.2 V, and V_{t0s} of others are all 0.3 V. $V_{DD} = 0.5$ V, $C_L = 20$ fF + 4MOSFETs, and W of each MOSFET is given as numeral in nm in Fig. 12 for $L = 65$ nm. For S-INV, V_{DD} and V_{t0s} are fixed to be 0.5 V and 0.3 V, respectively.

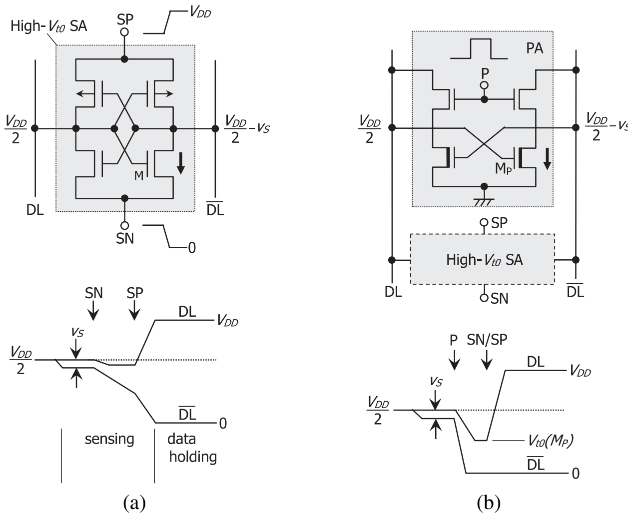


Fig. 14 (a) Conventional sense amplifier and (b) dual sense amplifier [5], [36], [37].

reduced wire-length distributions, are the key to overcoming the interconnect-delay problem in the nanoscale era. They will also ensure power-supply integrity throughout the subsystem, making low-V_{DD} operation possible with a reduced difference between V_{DD} and V_{min}. For such subsystems, drastically reducing the memory array area is particularly important since the array dominates the core or chip. Connecting small cores, each embedding a large-capacity DRAM, with low-resistive global interconnects and meshed power-supply lines, as found in the multi-divided array of modern DRAMs [5], will enable achievement of high-speed, multi-core LSIs [45], [46]. For example, a hypothetical 0.5-V 16 k-core LSI accommodating as many as 320-Mgate logic and 8-Gb DRAMs on a $10 \times 10\,\mathrm{mm}^2$ chip would be feasible in the 11-nm generation although the real challenge will be to find applications that can fully utilize such a powerful multi-core chip. Each homogeneous core, including 20-kgate logic and 512-Kb DRAM with a $5F^2$ cell [3], would be less than $56 \times 56\,\mu\mathrm{m}^2$. The main challenges are to develop redundant cores and a low-V_{t0} power switch. Note that the switch must sufficiently reduce the leakage of the core in the inactive mode. In addition, it must provide a large enough active current to the core with a channel width much less than the total width of the internal MOSFETs to minimize the area overhead. Furthermore, it must enable low-V_{DD}, high-speed, core-to-core hopping. The requirements impose uses of multi-V_{DD}, multi-V_{t0} circuits on the core.

Figure 15 compares three power switches designed for application to an internal low-V_{t0} core. To maximize the leakage reduction with the body bias effects [5], the substrates of the p- and nMOSFETs in the core are connected to V_{DD} and V_{SS}, respectively. The switch in (a) is a conventional high-V_{t0} nMOS (M_S) power switch (SW1), the gate of which is driven at V_{DD} swing. In the inactive mode (i.e., power shut down mode), it sufficiently cuts the core leakage. The channel width, $W(M_S)$, must be wide enough to provide a large active current to the core because of the reduced gate over-drive voltage ($=V_{DD} - V_{t0}$). In addition, a large V_{DD} swing in the heavy capacitive internal power line, N_S, results in long discharge and recovery times and high power dissipation during fast cycling of the switch. The noise coupled to other conductors at the transients may increase. Fast core-to-core hopping is thus prevented. Moreover, each node loses its logic state because it is completely discharged, meaning that a data latch is needed at the node in some cases.

The second switch (b) is a low-V_{t0} nMOS (M_S) power switch (SW2). In the inactive mode, the N_S voltage, V_{NS}, is adjusted so that the total leakage from all the n- and pMOSFETs in the core is reduced to the value of the current of the leaky M_S at $V_{GS} = 0$. This reduction stems from the

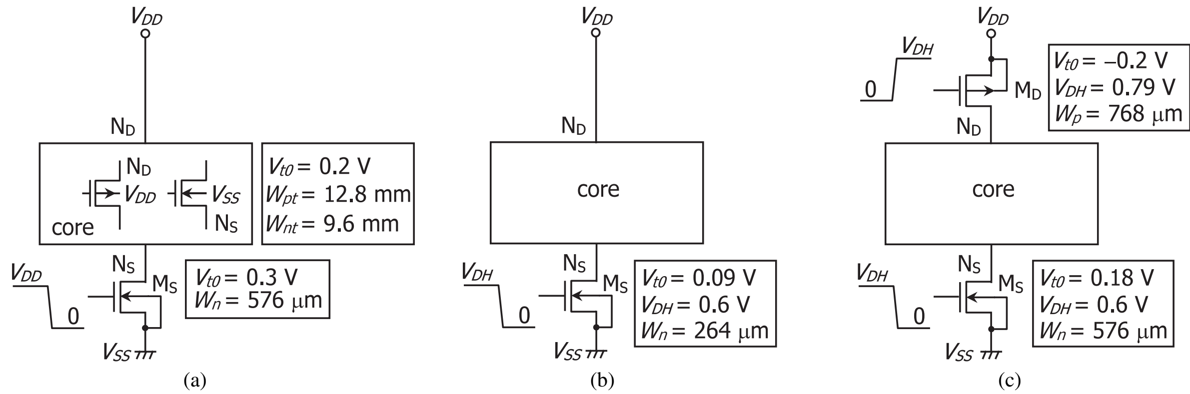


Fig. 15 (a) Conventional high-V_{t0} power switch (SW1), (b) low-V_{t0} power switch (SW2), and (c) differentially-driven power switch (SW3) [3]. V_{DD} = 0.5 V.

body bias effects and leakage characteristics of MOSFETs. For the nMOSFETs, the leakage is reduced as a result of increasing V_{t0} by raising V_{NS}. The supply voltage of the core is thus reduced to $V_{DD} - V_{NS}$. Note that the reduced supply voltage is simply the drain-source voltage, V_{DS}, of the switched-off pMOSFETs. The leakage of the pMOSFETs is thus reduced since it is reduced with the reduction in V_{DS} unless V_{DS} is sufficiently high [4]. In any event, M_S can supply more current to the core in the active mode due to the low V_{t0}, or $W(M_S)$ can be smaller for a given supply current. Moreover, discharge and recovery times and power dissipation are reduced owing to reduced voltage swing. If $V_{DD} - V_{NS} > V_{min}(h)$, the logic state is held, where $V_{min}(h)$ is the minimum supply voltage necessary to hold the logic state.

The third switch (c) is a differentially driven low-V_{t0} pMOS/nMOS (M_D, M_S) power switch (SW3). Leakage for both the n- and pMOSFETs is reduced due to the body bias effects more than for SW2. The differential operations of the internal N_D and N_S power lines at reduced swing results in short discharge and recovery times and low power dissipation. Differential operation occurs when the off-currents of M_S and M_D are equal. The internal logic state is preserved if the internal supply voltage (i.e., voltage difference between N_D and N_S) is higher than $V_{min}(h)$. However, $W(M_S)$ or $W(M_D)$ is always wider than that of the other two switches for a given internal supply, $V_{ND} - V_{NS}$, because the two switches use series connection. Note that a boosted gate voltage, V_{DH}, is applied to the gate of M_S to increase the supply current and to the gate of M_D to reduce the leakage. Since this means utilizing the gate-source back biasing, a multi-V_{DD}, multi-V_{t0} circuit is required.

Figure 16 compares the simulated internal waveforms for the three switches under the assumption of a 20-kgate core using 65-nm MOSFETs. The W, V_{t0}, and gate voltage of the switch MOSFETs were selected so as to provide almost the same and sufficient active current to the core. The V_{t0} and total channel width, W_t, of the core MOSFETs were 0.2 V, and 12.8 mm for the nMOSFETs and 0.2 V and 9.6 mm for the pMOSFETs. The V_{t0} and $W(M_S)$ were 0.3 V and 2.6% of W_t, 0.087 V and 1.2%, and 0.178 V and 2.6%

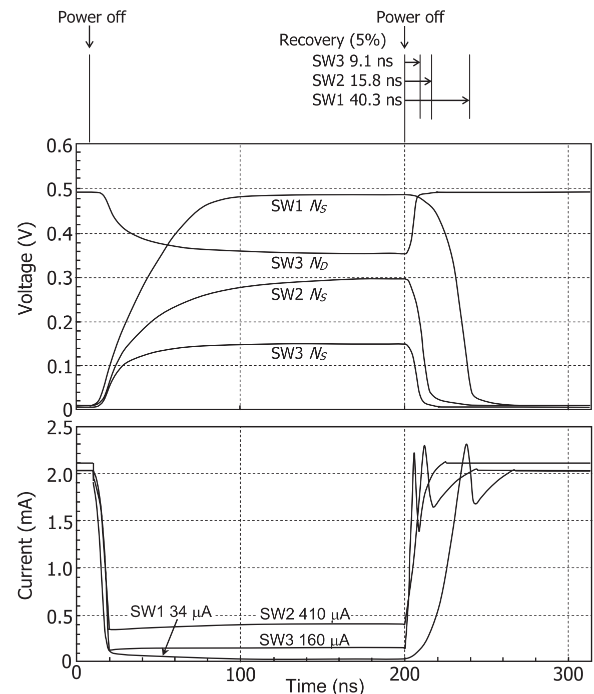


Fig. 16 Simulated internal waveforms of a 65-nm 20-kgate core.

for SW1, SW2, and SW3, respectively, while those of M_D were −0.2 V (depleted) and 3.4%. A $V_{min}(h)$ of 0.2 V was assumed. For example, the recovery time was 40 ns for SW1, 16 ns for SW2, and 9 ns for SW3 for a given transient peak current that was done by adjusting the rise or fall time of the gate pulse applied to the switch MOSFET. The leakage was 34 μA for SW1, 410 μA for SW2, and 160 μA for SW3. Obviously, SW2 is better than SW1 in terms of area, recovery time, and power dissipation despite the larger leakage. SW3 is the fastest, and its power dissipation is the lowest with moderate leakage despite a larger switch area. These switches are thus applicable to various types of power switches corresponding to their respective advantages. All three switches were quite fast even for 60-nm devices. Their performance might be further enhanced if 11-nm devices were used.

	Planar MOSFET	FinFET	
L	$1/\alpha$	$1/\alpha$	$1/\sqrt{\alpha}$
W	$1/\alpha$	α	$\sqrt{\alpha}$
W/L	1	α^2	α
LW	$1/\alpha^2$	1	1
A_{vt}	$1/\sqrt{\alpha}\,(1/\alpha)$	$1/\sqrt{\alpha}\,(1/\alpha)$	$1/\sqrt{\alpha}\,(1/\alpha)$
$\sigma(V_t)$	$\sqrt{\alpha}\,(1)$	$1/\sqrt{\alpha}\,(1/\alpha)$	$1/\sqrt{\alpha}\,(1/\alpha)$
V_{DD}	$\sqrt{\alpha}\,(1)$	$1/\sqrt{\alpha}\,(1/\alpha)$	$1/\sqrt{\alpha}\,(1/\alpha)$
I_{DS}	$\sim\alpha^{1.1}$	$\sim\alpha^{1.9}$	$\sim\alpha^{0.9}$
$\tau(MOS)$	$\sim\alpha^{-2.1}$	$\sim\alpha^{-1.9}$	$\sim\alpha^{-0.9}$
$P\,(=V_{DD} I_{DS})$	$\sim\alpha^{1.6}$	$\sim\alpha^{1.4}$	$\sim\alpha^{0.4}$
$P\tau(MOS)$	$\sim\alpha^{-0.5}$	$\sim\alpha^{-0.5}$	$\sim\alpha^{-0.5}$
W_{min}/L_{min}, F = 45 nm ($\alpha = 1$)	45/45 nm	45/45 nm	45/45 nm
W_{min}/L_{min}, F = 11 nm ($\alpha = 4$)	11/11 nm (aspect ratio = 1)	180/11 nm (16)	90/23 nm (4)

$A_{vt} \propto t_{OX} N_A^{0.25}$, $\sigma(V_t) = A_{vt}/\sqrt{LW}$, $I_{DS} = \beta\,(V_{DD} - V_t)^{1.2}$ for constant N_A, $\tau(MOS) = V_{DD}\,C_G/I_{DS}$

Fig. 17 Comparisons of scaling between planar MOSFET and FinFET [3].

3.2 Low-Voltage MOSFETs

Small-A_{vt} MOSFETs: The most effective way to reduce V_{min} by means of devices is to use small-A_{vt} MOSFETs such as high-*k* metal-gate MOSFETs and/or FD-SOI MOSFETs. Recently, considerable development effort has been directed toward planar FD-SOI devices and fin-type field effect transistors (FinFETs) [38]. Of the many proposals, FD-SOI MOSFETs with an ultra-thin (UT) BOX (buried oxide) layer, called SOTB (silicon on thin box) MOSFETs [18], [39]–[41], [47], are particularly promising because they can be applied with minimal changes to current bulk CMOS devices. In addition to the small A_{vt} and excellent short channel effects, they enable multiple V_{t0} values to be obtained by adjusting the doping of the substrate under the UT-BOX layer and enable the inter-die V_{t0} variation to be compensated for by substrate bias (V_{BB}) application through the UT-BOX layer. Here, the V_{BB} is usually generated by an on-chip V_{BB} generator using a charge pump. To generate a stable V_{BB}, however, the substrate current, I_{BB}, must be small sufficiently, since the charge pump has poor current drivability. Unfortunately, the I_{BB} of conventional bulk CMOS devices is inherently large, making V_{BB} unstable and thus the on-chip V_{BB} approach extremely difficult to achieve. The pn-junction structures of the drain/source and the higher V_{min} and thus a higher V_{DD} of the bulk CMOS, as explained earlier, are responsible for the large I_{BB}. Thus, for the bulk CMOS devices, off-chip compensation may be unavoidable although it is unsuitable for general purpose use. In contrast, for SOTB MOSFETs, the UT-BOX layer stops the I_{BB}, so the pump generates a stable V_{BB} of over 1 V, making the on-chip V_{BB} approach possible.

The use of FinFETs [3], [41] enables the use of an ultra low-dose channel and a wide-channel built-in structure. Thus, it achieves not only a higher density and higher drive current but also minimizes $\sigma(V_t)$ with minimized A_{vt} and maximized W. It even enables $\sigma(V_t)$-scalable MOSFETs to be achieved, as explained in the next section. It also enables achievement of high-density MOS capacitors, logic-process-compatible DRAM cells [41], and tiny two-dimensional selection DRAM cells [3]. Thus, it may one day breach the low-voltage, high-density limitations of conventional bulk CMOS devices if relevant devices and processes are developed. The use of FinFETs may increase intra-die V_{t0} variation, which would impose a need for stringent control of shape uniformity on the FinFET. Here, difficulty in compensating for the inter-die V_{t0} variations can be resolved by means of V_{BB} control if a UT-BOX structure [41] is used.

$\sigma(V_t)$-Scalable MOSFETs: If all feature sizes of a planar MOSFET are scaled down by a factor of $1/\alpha$, as illustrated in Fig. 17, V_{min} scaling at $\alpha^{-0.5}$, which is the aim of this work, imposes an intolerable scaling factor of $\alpha^{-1.5}$ on A_{vt} because of the rapid scaling of LW at α^{-2}. Even if the scaling factor of A_{vt} is reduced to the more practical value of $\alpha^{-0.5}$, V_{min} increases by a factor of $\alpha^{0.5}$. Furthermore, V_{min} remains constant even with A_{vt} scaling as large as α^{-1}. However, the vertical structure provided by FinFETs [3], [38], [42] yields a new scaling law for $\sigma(V_t)$, mitigating the requirement to A_{vt}. This is because this structure enables LW to be kept constant or even increased when the fin height (that is, channel width W) is scaled up despite channel length L being scaled down. This can be done without sacrificing MOSFET density. This up-scaling is done in accordance with the degree of A_{vt} scaling, so $\sigma(V_t)$ and thus V_{min} are scaled down. For example, if A_{vt} is scaled down at $\alpha^{-0.5}$, $\sigma(V_t)$ can also be scaled down by the same factor because LW is preserved as a result of the factor of α^{-1} or $\alpha^{-0.5}$ for L, and α or $\alpha^{0.5}$ for W. Such FinFETs enable high-speed operation not only due to the large drive current but also the shorter interconnects deriving from the vertical structures. However, the aspect

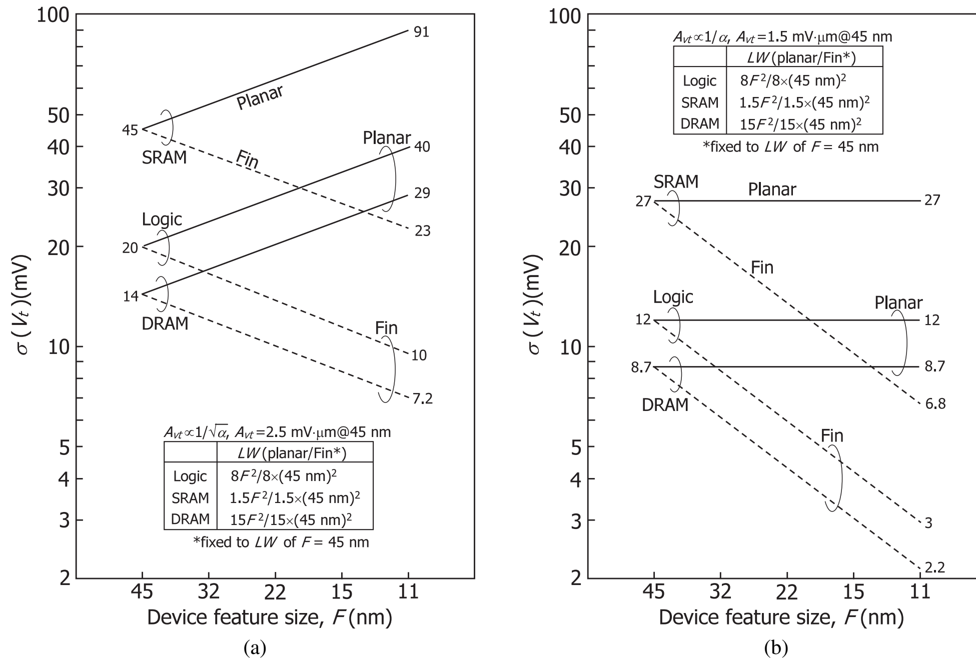


Fig. 18 Expected trends in $\sigma(V_t)$ for (a) low-power designs and (b) high-performance designs [3].

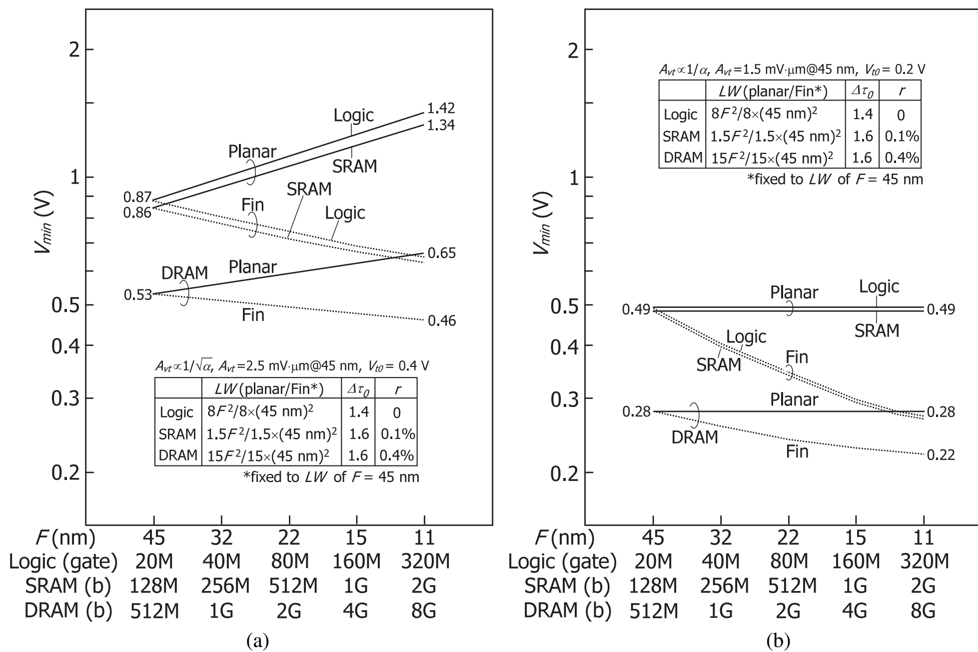


Fig. 19 Expected trends in V_{min} for (a) lower-power designs and (b) high-performance designs [3].

ratio (W/L) of FinFETs increases with device scaling. For example, it is as large as 4 to 16 in the 11-nm generation, as shown in Fig. 17. Note that structures with such large aspect ratios might be possible to achieve, taking the history of DRAM development into account. In DRAMs, the aspect ratio of trench capacitors has increased from about 3 in the early 1980s to as much as 70 for modern 70-nm DRAMs [43], [44]. In addition, the ratio is almost halved for a given W if a sidewall process [41] is used. Even if the resultant W is still unnecessarily large and thus increases power dissipation for a load dominated by the gate capacitance, the sidewall process further halves the W by splitting a MOSFET into two independent MOSFETs [41]. For a load dominated by wiring capacitance, the large W due to FinFETs enables a high speed.

On the basis of the MOSFET scaling, we can predict the V_{min} for future blocks, assuming that the A_{vt} in the 45-nm generation, the A_{vt} scaling factor for further device

scaling, and V_{t0} are 2.5 mV·μm, $\alpha^{-0.5}$, and 0.4 V for low-power designs, and 1.5 mV·μm, α^{-1}, and 0.2 V for high-performance designs (see Fig. 5(b)). The constant LW in Fig. 17 is also assumed for FinFETs. Obviously, the use of FinFETs enables $\sigma(V_t)$ to be scaled down for both designs, as seen in Figs. 18(a) and (b), while planar MOSFETs remain at a fixed $\sigma(V_t)$ even for high-performance designs with α^{-1} scaling, as expected. Therefore, for low-power designs (Fig. 19(a)), FinFETs reduce V_{min} to about 0.65 V for the logic block and SRAMs and to about 0.46 V for DRAMs in the 11-nm generation. Such high V_{min}s result from using a high V_{t0} of 0.4 V. If V_{t0}-scalable, low-leakage circuits, and power switches tolerant to a lower V_{t0}, which were described above, are used, the V_{min}s are effectively reduced to less than 0.5 V. Replacing SRAMs with DRAMs may also be effective to solve the high V_{min} problem of SRAMs. For high-performance designs (Fig. 19(b)), FinFETs reduce V_{min} to as low as about 0.27 V for the logic and SRAMs and 0.22 V for DRAMs.

4. Low-Voltage Analog Circuits

Mixed-signal LSIs are drawing as much attention as memory-rich LSIs. Figure 20 illustrates a mixed-signal LSI comprising analog and a digital circuitry. The analog circuitry includes receiving and transmitting chains for analog signals, a low-jitter clock generator consisting of a PLL and VCO, and a voltage reference generator (VREF). The receiving chain comprises an ADC as well as filters and amplifiers, while transmitting chain has a DAC. The filter bandwidth and amplifier gain are controlled using controller 1 (CTRL1) of the digital circuitry. The receiving chain can process not only external analog signals for communication, medical imaging, and off-chip sensing but also on-chip analog signals coming from various internal nodes of the digital circuitry (Mon_1-Mon_N), thereby serving as an analog sensing circuit. For example, it can perform wireless communications by connecting an off-chip RF-IC at the frontend of this mixed-signal IC. It can also control the operating conditions of MPUs, such as V_{DD}, operating frequency etc., via controller 2 (CTRL2) with on-chip analog signals. In fact, such an on-chip monitoring [48] is becoming vital for tracking and compensating for the process, voltage, and temperature (“PVT”) variations of MPUs. The digital circuitry includes baseband block, an MPU, and memory blocks. The baseband block supports digital filtering and other functions dedicated to the above-mentioned applications.

The minimum operating voltage V_{min} of each analog circuit is reduced by reducing the lowest necessary V_t, V_{t0}, and the maximum variation in V_t, ΔV_{tmax}, which is determined by the circuit count on the chip and the MOS size, as discussed in previous sections. Reducing V_{t0} is achieved by using reduction circuits described previously. Reducing ΔV_{tmax} is relatively easy if the circuit count is small because the MOS sizes can be enlarged. Note that even using a high V_{DD} only dedicated to the interface circuit of the chip may be possible without worrying about the power dissipation.

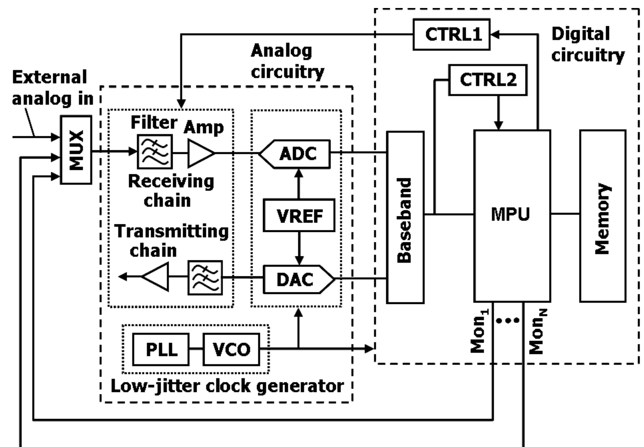


Fig. 20 A Mixed-signal LSI.

Such solutions, however, are invalid for analog circuits necessitating a large circuit count, thus calling for other solutions. In addition to V_{t0} and ΔV_{tmax}, special attention must be paid to unique specifications and circuit configurations of analog circuits, which may further increase the V_{min}. In this sense, the ADC of the receiving chain is most crucial for reducing the V_{min} of the mixed-signal LSIs, although the preceding analog signal pre-processing by the amplifiers and filters is nevertheless important for relaxing the speed and resolution required on ADC. The topology of the ADC is determined in accordance with the requirements. In particular, the pipeline ADC should have both a high resolution (up to 12 bits) and high speed (up to 100 MS/s). A successive approximation register (SAR) ADC and a sigma-delta ADC have higher resolution (more than 12 bits) and lower speed ($<$ 1 MS/s), while flash ADC has much higher speed ($>$ 1 GS/s) and lower resolution (up to 6 bits). SAR and flash ADCs use comparators which need a smaller offset voltage. Pipeline and sigma-delta ADCs and the amplifiers and filters use operational amplifiers (op-amps) which usually need a high gain and wide dynamic range. Therefore, comparator and op-amp are the two basic types of analog circuit cores for mixed-signal LSIs. Table 1 compares the circuit count and MOSFET size. About 100 comparators (CPs) using an about $500F^2$ MOSFET and 10 op-amps using an about $5{,}000F^2$ MOSFET have been used to implement either 1-GS/s 6-bit flash ADC or 100-MS/s 10-bit pipeline ADC. These implementations are much smaller in circuit count and larger in MOS size compared with those of memory-rich LSIs, resulting in a much smaller ΔV_{tmax} and thus a smaller offset voltage, as exemplified by the 11-nm FinFET in Table 1. However, the offset still affects V_{min}, as discussed in Sect. 4.1. In addition to the cores described above, the leak current from the analog switches used throughout the analog circuitry, caused by reduced V_t, needs to be suppressed [49] by using gate-source reverse biasing. Also, the power supply noise of all analog circuits must be minimized for low-voltage operation, which calls for high-density on-chip decoupling capacitors. Furthermore, some analog circuits

Table 1 Comparison between typical digital and analog cores. $\sigma(V_t)$s for the 11-nm FinFET in Fig. 18(a), and use of MOSFETs with $V_{t0} = 0$ V are assumed. $V_{OFS} = 2^{0.5}\Delta V_{tmax}$ for a pair of MOSFETs.

	Inv.	SRAM cell	DRAM SA	CP	op-amp
Count	640M	2G	128M	100	10
$LW\,(F^2)$	8	1.5	15	500	5,000
ΔV_{tmax} (mV)	27	34	9	4.3	0.9

require a high-Q inductor and highly linear capacitors. The capacitors and inductors can be made using SOI structures. In any event, few papers [50] have consistently and systematically described low-voltage analog circuits. Although 0.5-V, 0.18-μm active filters [51] using substrate bias control of bulk MOSFETs to reduce V_t have been reported, the technique is not sufficient for 0.5-V high-speed high-resolution ADC. FinFETs have also been reported to affect analog circuit design [52] through achieving higher gain of the inverter op-amp described below due to having a smaller output conductance [52] as well as to have superior matching performance. However, the details remain unknown.

In the following, the V_{min} of the comparators and op-amps is investigated to reduce the V_{min} of analog circuits to less than 0.5 V using the expected values in Table 1. Note that the V_{min} of analog circuits is not defined in terms of the speed variation but defined as V_{DD} at which the cores start to operate. However, it is almost the same as the V_{min} previously defined for memory-rich LSIs, since the ΔV_{tmax} of analog circuits becomes negligible if FinFETs are used.

4.1 Comparators

The V_{min} of a comparator depends on the type of ADC used. Figure 21(a) shows an equivalent circuit and waveforms of the first-stage preamp of a comparator for a flash ADC [53], [54] in which v_i and v_r are the input signal and reference voltages, respectively. In this comparator, a large input common-mode swing, which is equal to the full-scale range V_{FS} (i.e., half the ADC full-scale range), is required. The upper and lower limits of the input common-mode voltage must be $V_{DD} - V_{OV} + V_t$ and $V_t + 2V_{OV}$, respectively, to ensure that the MOSFETs operate in the saturation region, where V_{OV} is the minimum gate-overdrive voltage required for strong-inversion operation of M_1, M_2, and M_3. Thus, the V_{min} of a flash ADC is given as $V_{min} = V_{FS} + 3V_{OV} = V_{OFS}\,2^N + 3V_{OV}$, where V_{OFS} is the offset voltage, and N is resolution, usually less than 6. To realize $V_{min} <$ 0.5 V, V_{OFS} must be as small as 0.8 mV for $V_{OV} = 0.15$ V and $N = 6$, since it must be smaller than the half LSB voltage step, i.e., $V_{FS}/2^N$. Unfortunately, however, V_{min} results in as high as 1.0 V since V_{OFS} is expected to be 8.6 mV for an 11-nm FinFET (see Table 1), considering that four input MOSFETs [53], [54] are actually used. Thus, even for FinFETs, digital calibration techniques for the offset [55] will be indispensable to reduce V_{min} to below 0.5 V. In principle, the techniques are expected to reduce V_{OFS} to a negligible value, so such a low V_{min} may be realized. Alternatively, the MOS sizes could be larger than $500F^2$.

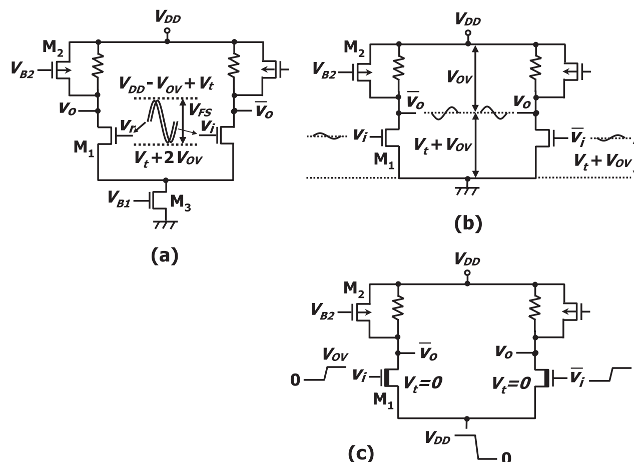


Fig. 21 (a) Preamp with common-mode rejection [54], (b) preamp without common-mode rejection, and (c) preamp with gate-source reverse biasing.

Figure 21(b) shows a circuit for the first-stage preamp of a comparator for an SAR ADC. The preamp does not need a common-mode rejection or tail current source because the input common-mode voltage can be set to the V_{GS} ($=V_t + V_{OV}$) of the input MOSFET's M_1. Therefore, an SAR ADC is more suitable for low-voltage operation than a flash ADC. In this case, the V_{min} of the comparator is the higher of the two V_{min} values (V_{min1}, V_{min2}) determined by V_{OFS} and the circuit configuration, respectively. V_{min1} is given as $V_{min1} = V_{OFS}\ 2^N$ if $V_{FS} = V_{DD}$ is assumed. To realize $V_{min1} < 0.5$ V, V_{OFS} must be less than 0.1 mV for $N = 12$. If combined with digital calibration of the offset described above, even such a small value may be attained by maximizing the MOSFET sizes. The maximization is justified by the fact that only one comparator is used in a chip. On the other hand, $V_{min2} = V_t(M_1) + 2V_{OV}$ if the second-stage preamp is assumed to have the same topology and if the output common-mode level of the first stage is assumed to be equal to the V_{GS} $(= V_t(M_1) + V_{OV})$ of the second-stage input MOSFET. This means that V_{min2} should be reduced to 0.3 V if $V_t(M_1) = 0$. The resultant increase in the leakage current of M_1 can be cut by using gate-source reverse biasing, as shown in Fig. 21(c). Therefore, V_{min} is equal to V_{min1} and thus expected to be lower than 0.5 V.

Although reduction in V_{OV} is also crucial for further reducing V_{min} for both comparators, the details remain unknown.

4.2 Op-Amps

An op-amp usually needs a wide dynamic range (that is, a large V_{FS}) and a high enough gain, causing the V_{min} usually much higher than that of the comparator. Figure 22(a) shows a circuit for a conventional regulated cascode op-amp for a pipeline ADC. Four stacked MOSFETs (M_1–M_4) help attain a gain of more than 60-dB. To ensure that they op-

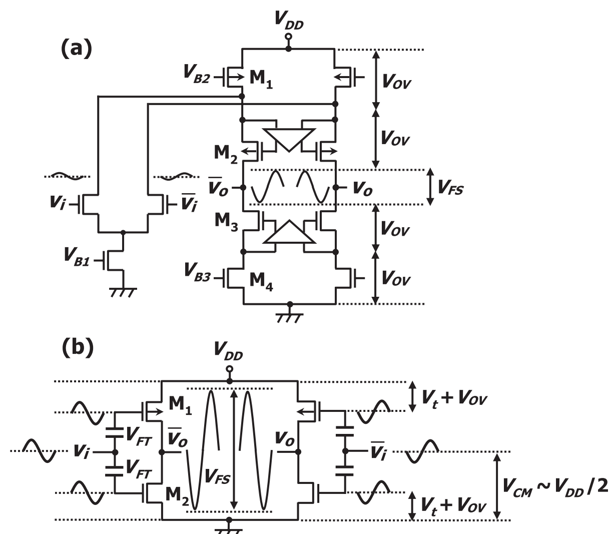


Fig. 22 (a) Regulated-cascode op-amp, and (b) inverter-based op-amp [56].

erate in the saturation region, V_{FS} is set to $V_{DD} - 4V_{OV}$, so $V_{min} = V_{FS} + 4V_{OV}$. Therefore, V_{min} cannot be lower than 0.6 V for $V_{OV} = 0.15$ V. To solve the high V_{min} problem, a simple CMOS inverter-based op-amp [56], as shown in Fig. 22(b), has been proposed. Because the output of the op-amp can continuously cross over the saturation and the linear regions to give almost rail-to-rail V_{FS}, V_{FS} can be close to V_{DD}, so $V_{min} = V_{FS}$. For pipeline ADC, V_{FS} is expressed as $K(f_S N/I)^{1/2}\, 2^N$, where f_S, N, and I are the sampling rate, resolution and current consumption, respectively, and K is a constant around 10^{-9} [57]. Therefore, V_{min} is limited by only the speed, accuracy, and power consumption. For example, $V_{min} = 0.1$ V for a 100-MS/s 10-bit ADC with a 100-mA current. Instead, a lower gain and nonlinearity inevitably arise. However, another digital calibration technique [58] can solve the problems. The discussion above assumes that the gate overdrives of M_1 and M_2 are appropriately set to V_{OV} to maintain the gain for any V_{DD}. In fact, the floating DC biasing with V_{FT}, which can be implemented using capacitors for example [59], makes the V_{GS} of M_1 and M_2 equal to $V_t + V_{OV}$ if V_{FT} is set to $V_{DD}/2 - V_t - V_{OV}$. Note that V_{FT} can be set to a negative value to achieve a V_{DD} smaller than $2(V_t + V_{OV})$. Also note that a common-mode regulation circuit [59] must be implemented though not shown in the figure. It finely tunes the input common-mode voltage V_{CM} to the appropriate value close to $V_{DD}/2$ so that the output common-mode level should become $V_{DD}/2$.

5. Conclusion

The minimum operating voltage, V_{min}, of nanoscale CMOS LSIs was investigated in an effort to reduce to below 0.5 V. Use of a new method for evaluating V_{min} on the basis of speed variation revealed that V_{min} is very sensitive to the lowest necessary threshold voltage, V_{t0}, of MOSFETs and to threshold-voltage variations, ΔV_t, which become more significant with device scaling. There is thus a need for low-V_{t0} circuits and ΔV_t-immune MOSFETs. The SRAM block is particularly problematic for memory-rich LSIs because it has the highest V_{min}. As a result of investigating various techniques for reducing the V_{min} of the SRAM block, it turned out that using RAM repair techniques, shortening the data line, up-sizing, and using more relaxed MOSFET scaling are effective. Also investigated were new low-V_{t0} circuits — dynamic logic circuits enabling the power-delay product to be reduced to 0.09 at a 0.2-V supply owing to gate-source reverse biasing — and a DRAM dynamic sense amplifier and power switches operable at below 0.5 V. The low-V_{t0} circuits use a dual-V_{t0}, dual-V_{DD} scheme. In addition, the use of a fully-depleted structure (FD-SOI) and fin-type structure (FinFET) for ΔV_t-immune MOSFETs was evaluated in terms of their low-voltage potential and challenges. As a result, the height up-scalable FinFETs turned out to be quite effective to reduce V_{min} to less than 0.5 V, if combined with the low-V_{t0} circuits. For mixed-signal LSIs, investigation of low-voltage potential of analog circuits, especially for comparators and operational amplifiers, revealed that simple inverter op-amps, in which the low gain and nonlinearity are compensated for by digitally assisted analog designs, are crucial to 0.5-V operations. In addition to such adaptive circuits, the development of relevant devices and fabrication processes should lead to the achievement of 0.5-V nanoscale LSIs.

Acknowledgements

We are grateful for the invaluable contributions of many colleagues at Hitachi Central Research Laboratory, especially S. Kimura, D. Hisamoto, N. Sugii, R. Tsuchiya, T. Sekiguchi, and M. Saen for their stimulating discussions and helpful suggestions. Special thanks go to M. Kokubo for his critical reading of the analog section.

References

[1] K. Itoh, et al., "Low-voltage limitations of memory-rich nano-scale CMOS LSIs," ESSCIRC Dig., pp.68–75, Sept. 2007.
[2] K. Itoh and M. Horiguchi, "Low-voltage scaling limitations for nano-scale CMOS LSIs," Solid-State Electron., vol.53, no.4, pp.402–410, April 2009.
[3] K. Itoh, "Adaptive circuits for the 0.5-V nanoscale CMOS era," ISSCC Dig., pp.14–20, Feb. 2009.
[4] K. Itoh, M. Horiguchi, and H. Tanaka, Ultra-Low Voltage Nano-Scale Memories, Springer, 2007.
[5] K. Itoh, VLSI Memory Chip Design, Springer-Verlag, 2001.
[6] Y. Nakagome, et al., "Review and prospects of low-voltage RAM circuits," IBM J. R & D, vol.47, no.5/6, pp.525–552, Sept./Nov. 2003.
[7] J.A. Davis, et al., "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," Proc. IEEE, vol.89, no.3, pp.305–324, March 2001.
[8] W. Haensch, et al., "Silicon CMOS devices beyond scaling," IBM J. Res. Dev., vol.50, no.4/5, pp.339–361, July/Sept. 2006.
[9] T.C. Chen, "Where CMOS is going: Trendy hype vs. real technology," ISSCC Dig. Tech. Papers, pp.22–28, Feb. 2006.
[10] A.W. Topol, et al., "Three-dimensional integrated circuits," IBM J.

Res. Dev., vol.50, no.4/5, pp.491–506, July/Sept. 2006.

[11] M.J.M. Pelgrom, et al., "Matching properties of MOS transistors," J. Solid-State Circuits, vol.24, no.5, pp.1433–1439, Oct. 1989.

[12] K. Takeuchi, et al., "Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies," IEDM Dig., pp.467–470, Dec. 2007.

[13] K. Itoh, et al., "Reviews and future prospects of low-voltage embedded RAMs," CICC2004 Dig., pp.339–344, Oct. 2004.

[14] H. Masuda, et al., "Approach for physical design in sub-100 nm era," ISCAS, pp.5934–5937(6), May 2005.

[15] S. Mukhopadhyay, et al., "Statistical characterization and on-chip measurement methods for local random variability of a process using sense-amplifier-based test structure," ISSCC Dig., pp.400–401, Feb. 2007.

[16] K.J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," IEDM Dig., pp.471–474, Dec. 2007.

[17] Y. Morita, et al., "Smallest V_{th} variability achieved by intrinsic thin channel on thin BOX (SOTB) CMOS with single metal gate," Symp. VLSI Tech. Dig., pp.166–167, June 2008.

[18] O. Weber, et al., "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding," IEDM Dig., vol.10.4, pp.245–248, Dec. 2008.

[19] K. Itoh, et al., "A deep sub-V, single power-supply SRAM cell with multi-V_T, boosted storage node and dynamic load," Symp. VLSI Circuits Dig., pp.132–133, June 1996.

[20] J. Pille, et al., "Implementation of the CELL broadband engine in a 65 nm SOI technology featuring dual-supply SRAM arrays supporting 6 GHz at 1.3 V," ISSCC Dig., pp.322–323, 606, Feb. 2007.

[21] H. Akamatsu, et al., "A low power data holding circuit with an intermittent power supply scheme for sub-1 V MT-CMOS LSIs," pp.14–15, June 1996.

[22] M. Yamaoka, et al., "A 300 MHz 25 μA/Mb leakage on-chip SRAM module featuring process-variation immunity and low-leakage-active mode for mobile-phone application processor," ISSCC Dig. Tech. Papers, pp.494–495, Feb. 2004.

[23] M. Khellah, et al., "A 4.2 GHz 0.3 mm^2 256 kb dual-Vcc SRAM building block in 65 nm CMOS," ISSCC Dig., pp.624–625, Feb. 2006.

[24] F. Hamzaoglu, et al., "A 153 Mb-SRAM design with dynamic stability enhancement and leakage reduction in 45 nm high-k metal-gate CMOS technology," ISSCC Dig., pp.376–377, Feb. 2008.

[25] H. Pilo, et al., "A 450 ps access-time SRAM macro in 45 nm SOI featuring a two-stage sensing scheme and dynamic power management," ISSCC Dig., pp.378–379, Feb. 2008.

[26] M. Yamaoka, et al., "Low-power embedded SRAM modules with expanded margins for writing," ISSCC Dig., pp.480–481, Feb. 2005.

[27] K. Zhang, et al., "A 3-GHz 70 MB SRAM in 65 nm CMOS technology with integrated column-based dynamic power supply," ISSCC Dig., vol.611, pp.474–475, Feb. 2005.

[28] L. Chang, et al., "A 5.3 GHz 8T-SRAM with operation down to 0.41 V in 65 nm CMOS," Symp. VLSI Circuits Dig., pp.252–253, 2007.

[29] M. Yamaoka, et al., "A Cell-activation-time controlled SRAM for low-voltage operation in DVFS SoCs using dynamic stability analysis," ESSCIRC Dig., pp.286–289, Sept. 2008.

[30] Y. Nakagome, et al., "Sub-1-V swing bus architecture for future low-power ULSIs," Symp. VLSI Circuits Dig., pp.82–83, June 1992.

[31] T. Chappel, et al., "A 2-ns cycle, 3.8-ns access 512 Kb CMOS ECL SRAM with a fully pipelined architecture," IEEE J. Solid-State Circuits, pp.1577–1584, Nov. 1991.

[32] T. Mori, et al., "A 1 V 0.9 mW at 100 MHz 2x16b SRAM utilizing a half-swing pulsed-decoder and write-bus architecture in 0.25 μm Dual-Vt CMOS," ISSCC Dig., pp.354–355, Feb. 1998.

[33] G. Braceras, et al., "A 940 MHz data-rate 8 Mb CMOS SRAM," ISSCC Dig., pp.198–199, Feb. 1999.

[34] T. Kirihata, et al., "A 390 mm^2 16 bank 1 Gb DDR SDRAM with hybrid bitline architecture," ISSCC Dig., pp.422–423, Feb. 1999.

[35] K. Itoh, et al., "Low-voltage limitations of deep-sub-100-nm CMOS LSIs-view of memory designers," Proc., GLSVLSI2007, pp.529–533, March 2007.

[36] S. Akiyama, et al., "Low-V_T small-offset gated preamplifier for sub-1-V gigabit DRAM arrays," ISSCC Dig., pp.142–143, Feb. 2009.

[37] A. Kotabe, et al., "CMOS Low-V_T preamplifier for 0.5-V gigabit-DRAM array," A-SSCC2009 Dig., pp.213–216, Nov. 2009.

[38] D. Hisamoto, et al., "A fully depleted lean-channel transistor (DELTA) — A novel vertical ultra thin SOI MOSFET," IEDM Tech. Dig., pp.833–836, Dec. 1989.

[39] R. Tsuchiya, et al., "Silicon on thin BOX: A new paradigm of the CMOSFET for low-power and high-performance application featuring wide-range back-bias control," IEDM Tech. Dig., pp.631–634, Dec. 2004.

[40] R. Tsuchiya, et al., "Low Voltage (V_{dd}~0.6 V) SRAM operation achieved by reduced threshold voltage variability in SOTB (silicon on thin BOX)," VLSI 2009 Tech., pp.150–151, June 2009.

[41] K. Itoh, et al., "FD-SOI MOSFETs for the low-voltage nanoscale CMOS era," International SOI Conference, Oct. 2009.

[42] J. Kavalieros, et al., "Tri-gate trnsistor architecture with high-k gate dielectrics, metal gates and strain engineering," Symp. VLSI Tech. Dig., pp.62–63, June 2006.

[43] H. Sunami, "The role of the trench capacitor in DRAM innovation," IEEE SSCS News, Winter 2008, pp.42–44, Jan. 2008.

[44] J. Amon, et al., "A highly manufacturable deep trench based DRAM cell layout with a planar array device in a 70 nm technology," IEDM Dig. pp.73–76, Dec. 2004.

[45] D. Truong, et al., "A 167-processor 65 nm computational platform with per-processor dynamic supply voltage and dynamic clock frequency scaling," Symp. VLSI Circuits Dig., pp.22–23, June 2008.

[46] S. Vanbal, et al., "An 80-tile 1.28TFLOPS network-on-chip in 65 nm CMOS," ISSCC Dig., pp.98–99, Feb. 2007.

[47] N. Sugii, et al., "Comprehensive study on Vth variability in silicon on thin BOX (SOTB) CMOS with small random-dopant fluctuation: Finding a way to further reduce variation," IEDM Dig., 10.5, Dec. 2008.

[48] R. McGowen, et al., "Power and temperature control on a 90 nm Itanium family processor," IEEE J. Solid-State Circuits, pp.229–237, Jan. 2006.

[49] D. Daly, et al., "A 6b 0.2-to-0.9 V highly digital flash ADC with comparator redundancy," ISSCC Dig., pp.554–555, Feb. 2008.

[50] B. Nauta, et al., "Analog/RF circuit design techniques for nanometerscale IC technologies," ESSCIRC Dig., pp.45–54, Sept. 2005.

[51] S. Chatterjee, et al., "0.5-V analog circuit techniques and their application in OTA and filter design," IEEE J. Solid-State Circuits, pp.2373–2387, Dec. 2005.

[52] P. Wambacq, et al., "The potential of FinFETs for analog and RF circuit applications," IEEE Trans. Circuits Syst. I, pp.2541–2551, Nov. 2007.

[53] M. Choi, et al., "A 6-b 1.3-Gsample/s A/D converter in 0.35-μm CMOS," IEEE J. Solid-State Circuits, pp.1847–1858, Dec. 2001.

[54] K. Deguchi, et al., "A 6-bit 3.5-GS/s 0.9-V 98-mW flash ADC in 90-nm CMOS," IEEE J. Solid-State Circuits, pp.2303–2310, Oct. 2008.

[55] G. Plas, et al., "A 0.16 pJ/conversion-step 2.5 mW 1.25 GS/s 4b ADC in a 90 nm digital CMOS process," ISSCC Dig., pp.566–567, Feb. 2006.

[56] M. Uno, et al., "Optimum design considerations for a CMOS amplifier and efficiency of a class-AB structure," IEICE Technical Report, ICD2008-64, Oct. 2008.

[57] M. Miyahara, et al., "A study on a pipeline ADC," IEICE Technical Report, ICD2004-51, July 2004.

[58] T. Oshima, et al., "23-mW 50-MS/s 10-bit pipeline A/D converter with nonlinear LMS foreground calibration," ISCAS Dig., pp.960–963, May 2009.

[59] S. Kawahito, et al., "A 15b power-efficient pipeline A/D converter

using non-slewing closed-loop amplifiers," CICC Dig., pp.117–120, Sept. 2008.
[60] M. Yabuuchi, et al., "A 45 nm low-standby-power embedded SRAM with improved immunity against process and temperature variations," ISSCC Dig., pp.326–327, Feb. 2007.

Kiyoo Itoh received the B.S. and Ph.D. Degrees in Electrical Engineering from Tohoku University, Japan, in 1963 and 1976. He is currently a Fellow in Hitachi Ltd. He was a Visiting MacKay Lecturer at U.C. Berkeley in 1994, a Visiting Professor at the University of Waterloo in 1995, and a Consulting Professor at Stanford University in 2000–2001. He served on the IEEE Solid-State Circuits Award Committee from 1998 to 2000. He was a Member of the IEEE Fellow Committee from1999 to 2002, and an elected AdCom Member of IEEE Solid-State Circuits Society from 2001 to 2003. He is an IEEE Solid-State Circuits Society Distinguished Lecturer. Since 1972 he has led DRAM circuit technology and low-power/low-voltage CMOS circuits at Hitachi Ltd. As the lead designer of the first prototype for eight generations of Hitachi DRAMs ranging from 4Kb to 64Mb, he led the development of world's first DRAM chips, and invented and developed many de-facto standard circuits such as the concept of folded data-line architecture and on-chip voltage down-converters. As early as 1988 as a pioneer, he started to invent and develop subthreshold-current reduction circuits for the standby and active modes, such as power switches, gate-source offset driving, gate-source self-reverse biasing, and dual Vth circuits. He holds about 440 patents in Japan and US. He has authored and co-authored four books and two book chapters on memory designs, and has given 168 IEEE-related technical papers and presentations. Dr. Itoh has won many honors, including the IEEE Paul Rappaport Award in 1984, the Best Paper Award of ESSCIRC1990, the 1993 IEEE Solid-State Circuits Award (Now named D.O. Pederson Award), and the 2006 IEEE Jun-ichi Nishizawa Medal. He is an IEEE Fellow. In Japan, his awards include a National Medal of Honor with Purple Ribbon.

Masanao Yamaoka received the B.E., M.E., and Ph.D. degrees in Electronics and Communication Engineering from Kyoto University, Kyoto, Japan, in 1996, 1998, and 2007 respectively. In 1998, he joined the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan, where he has been engaged in the research and development of low-power embedded SRAMs and CMOS circuits for mobile processors, including CMOS-process non-volatile memories, low-voltage SRAMs, and low-leakage SRAMs. His current research interests include V_t-variation-tolerant CMOS circuits and V_t-variation immune devices.

Takashi Oshima received B.S., M.S. and Ph.D. in Physics from University of Tokyo in 1996, 1998 and 2001, respectively. He joined Central Research Laboratory of Hitachi Ltd., Tokyo, Japan in 2001, where he has been engaged in the design of analog circuits including ADC, filter, PLL and RF circuits and also in the development of wireless transceiver architectures for Bluetooth, RF-ID and other wireless applications. From 2005 to 2006 he was a visiting researcher at Berkeley Wireless Research Center (BWRC) of University of California at Berkeley, where he made a research on digitally assisted ADC. Dr. Oshima is currently an associate editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, also serving as a treasurer of IEEE Solid-State Circuits Society Japan Chapter. His current interests are in digitally assisted ADCs and future low-voltage mixed-signal ICs.