Please click on paper title to view Visual Supplement.

1.2 Adaptive Circuits for the 0.5-V Nanoscale CMOS Era

Kiyoo Itoh

Fellow, Hitachi, Tokyo, Japan

1. Introduction

Low-voltage scaling limitations of memory-rich CMOS LSIs are one of the major problems in the nanoscale era [1-3] because they cause the ever-more-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 [3-5], V_{t0} is still not low enough to reduce the operating voltage, V_{DD}, to the sub-1V region. The second is the variation in V_t (that is, ΔV_t), that becomes more prominent in the nanoscale era [1-3]. 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 enhances various detrimental effects such as variations in delay (and speed) and/or the voltage margins of circuits, and it significantly increases 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 the 1V wall in the 45nm generation, and is expected to rapidly increase with further scaling of poly-Si bulk MOSFETs [1-3], as shown in Figure 1.2.1. To reduce V_{DD}, the minimum usable power supply V_{DD} (that is, V_{min}) determined by the above-described 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}, the power-supply droop and noise in power supply lines/substrate, (that is, ΔV_{ps}), and ΔV, where ΔV_{ps} is usually much higher than ΔV in the nanoscale era. Here, ΔV is the sum of the necessary voltages for compensating for the extrinsic ΔV_t due to short-channel effects and line-edge roughness, and for meeting the speed target. Thus, ΔV depends on the qualities and maturity of the fabrication process and the design target, which cannot be specified here. An associated problem in the nanoscale era is the ever-higher resistance of interconnects [6-8]. This is also closely related to the voltage-limitation problem at the chip and subsystem levels, since it degrades not only the speed of ever-larger chips, but also effects power-supply integrity by increasing ΔV_{ps}. As well, integrity depends on chip packaging such as 3D integration [9].

Analog-rich mixed signal LSIs also presents a similar situation. But, as well, special attention must be paid to the analog block on the chip, because it consists of unique circuit configurations and elements, which differ from the LSIs described above (Figure 1.2.2). Differential and other circuits (such as, cascode amplifiers, comparators, VCOs, low-pass filters, and pipeline ADCs) need an inherently higher V_{DD} to achieve a high gain and/or small offset. This is in addition to the fact that the ever-increasing ΔV_t and thus the offset can be reduced to some extent by enlarging each of the relatively small number of MOSFETs in the analog part of the chip. More particularly, the operation of a VCO is extremely sensitive to ΔV_{ps}, thus calling for high-density on-chip decoupling capacitors to be used throughout a chip, as in high-speed MPUs. As well, some circuits necessitate larger capacitors and high-Q inductors. In any event, for the LSI industry in order to flourish and proliferate, the 1V wall must be breached in the 32nm generation and beyond. 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} and ΔV_{ps} are addressed in this paper. The focus will be on memory-rich LSIs, since such LSIs have usually driven the frontend of scaled device development. Mixed signal LSIs and others will sooner or later encounter similar problems. The V_{min} issue is described in the first part of the paper. After comparing the V_{min}s of logic gates, SRAMs, and DRAMs, state-of-the-art SRAM circuits to tackle the issue are reviewed. Then, devices and circuits to reduce V_{min} to the sub-1V region are described. In the latter part of the paper, the power-supply integrity (ΔV_{ps}) issue is discussed, based on various proposals for high-density cores/chips and compact subsystems. Finally, a scenario for sub-0.5V nanoscale CMOS LSIs is presented, including a brief discussion of mixed-signal LSIs.

2. Low-Voltage Scaling Limitations

2.1 Definition of V_{min}: The V_{min} of an LSI (Figure 1.2.2) including a logic block and RAM (SRAM and/or DRAM) blocks is equal to the highest of the V_{min} values for all blocks [1-3]. Each V_{min} is usually determined by a CMOS inverter for the logic block, a six-transistor (6T) SRAM cell for the SRAM block, and a sense amplifier (SA) for the DRAM block (Figure 1.2.3). Assuming read operation for the SRAM cell and sensing for the twin (2T) DRAM cell along with full-V_{DD} data (bit)-line precharging [3], the delay, τ, of M_1 in these circuits obeys the same formula as $\tau(V_t) \propto V_{DD}/(V_{DD}-V_t)^{\eta}$. Thus, the delay ratio ($\Delta\tau$) of the maximum delay at maximum V_t ($V_{tmax} = V_{t0} + \Delta V_{tmax}$, ΔV_{tmax}: maximum ΔV_t) to the average delay at average V_t ($=V_{t0}$) is given as $\Delta\tau = \tau(V_{t0} + \Delta V_{tmax})/\tau(V_{t0}) = \{(V_{DD} - V_{t0})/(V_{DD} - V_{t0} - \Delta V_{tmax})\}^{\eta}$. If V_{min} is defined as the V_{DD} necessary for a fixed value of $\Delta\tau$ (that is, tolerable $\Delta\tau$), the following relations for M_1 are given by solving the above equation for V_{DD}:

$$V_{min} = V_{t0} + (1+\gamma)\,\Delta V_{tmax},\ \gamma = 1/(\Delta\tau^{1/\eta} - 1),\ \Delta V_{tmax} = m\sigma(V_t) \quad (1)$$

$$\sigma(V_t) = A_{vt}(LW)^{-0.5} = B_{vt}[t_{ox}(V_{t0} - V_{FB} - \phi_S)/LW]^{0.5} \propto t_{ox}N_A^{0.25}(LW)^{-0.5} \quad (2)$$

where $\sigma(V_t)$ is the standard deviation of V_t distribution, A_{vt} and B_{vt} are the Pelgrom constants [10, 11], 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 substrate or well, and LW is the MOSFET size. The $\Delta\tau$ can take on two values, $\Delta\tau$ (+) and $\Delta\tau$ (-), corresponding to plus and minus values of ΔV_t. However, $\Delta\tau$ (+) will be used after this, simply expressed as $\Delta\tau$. The η is as small as 1.2 for deep-sub-100nm MOSFETs due to velocity saturation.

V_{t0} depends on subthreshold-leakage specifications. Figure 1.2.4(a) plots the leakage versus V_{t0} for a device feature size, F, of 130nm. Note that V_{t0} is an extrapolated V_t, that is, the sum of constant-current V_t (nA/µm) and 0.3V. This figure was prepared using previously reported SRAM data [21]. If the V_{t0} is between 0.2V (for high-speed designs) and 0.4V (for low-power designs), the leakage ranges from about 100mA to 1mA for a 1Mgate logic block, 70mA to 0.5mA for a 1Mb SRAM, and 20mA to 0.15mA for a 64k DRAM SA. Parameter γ strongly depends on tolerable $\Delta\tau$, exemplified by γ = 6.09, 3.09, and 2.09 for $\Delta\tau$ = 1.2, 1.4, and 1.6, respectively. For the logic block, a $\Delta\tau$ of 1.4 will be used hereafter despite need for smaller $\Delta\tau$ (that is, larger γ) to quickly and stringently control the timing at every gate. Using such a large $\Delta\tau$ is justified by the following features of the logic block: The ΔV_{tmax} in Eq. (1) is valid only if each of the logic gates are of the same size and operate randomly. In particular, the logic gates which use transistors whose sizes range from from $4F^2$ to $12F^2$ are assumed to consist entirely of transistors of $8F^2$, all with threshold V_{t0}. However, this is not the case for the actual logic block. For example, each gate does not work independently. Some gates form logical configurations with deep logical depth and small fan out (see Figure 1.2.2(b)), allowing the $\sigma(V_t)$ to be reduced due to the averaging effect of random variations. The well-known dual-V_t logic block reduces ΔV_{tmax}. The critical path reduces the $\sigma(V_t)$ due to the low-V_{t0} and large MOSFETs necessary to attain high speed. The small total MOS width of the path (typically, about 10% of the total for the whole block) reduces the m. Thus, the non-critical path, which inherently tolerates large $\Delta\tau$, eventually determines the V_{min} of the whole block. However, MOSFETs in the specific gate in the path can be enlarged without a substantial area increase, allowing the $\sigma(V_t)$ to be reduced. Therefore, the ΔV_{tmax} calculated by using $m\sigma(V_t)$ in Eq. (1) is much higher than the actual ΔV_{tmax}. To more accurately predict the actual V_{min} by using Eq. (1), the $\Delta\tau$ in Eq. (1) must be larger than the actually required $\Delta\tau$, as explained earlier. Recently-reported delay-error

 978-1-4244-3457-2/09/$25.00 ©2009 IEEE

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

detection and/or correction circuits for logic blocks [14 to 16] will reduce *m*, as in RAMs. As will be explained later, CMOS dynamic circuits will also reduce V_{t0}. For the RAM blocks, $\Delta\tau$ is assumed to be 1.6 because even such a large $\Delta\tau$ makes operations reliable. If the output signals from SRAM cells or SAs to the I/O line are aligned with a column clock (clk' in Figure 1.2.2(a)), that is, waiting for the signal from the slowest cell or SA, the signal transferred to I/O can be successfully discriminated. If the resulting slow speed is intolerable, the data line can be shortened despite the area overhead involved, so the short discharge time despite the variations makes contribution to the total speed of the block negligible.

The number, *m*, depends on the circuit count on the block. This ranges from 4.9 to 6.0 for the 0.6Mgate to 320Mgate logic blocks, from 5.2 to 6.3 for the 4Mb to 2Gb SRAMs, and from 4.8 to 5.9 for the 16Mb to 8Gb DRAMs connecting 64 cells to an SA [3]. 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 ECC and redundancy, *m* is reduced to about 3.29 for SRAMs and to about 2.88 for DRAMs [1, 2]. It should be noted that $\sigma(V_t)$ depends on V_{t0}. For a poly-Si gate having V_{FB} = -0.9V and ϕ_S = 0.8V, $\sigma(V_t)$ is reduced to 0.89, 0.77, 0.63, and 0.45 times that for V_{t0} = 0.4V when V_{t0} is reduced from 0.4V to 0.3V, 0.2V, 0.1V, and 0V, respectively. Furthermore, $\sigma(V_t)$ depends on A_{vt} and MOSFET size, as expected from Eq. (2). The expected trends in t_{OX} and A_{vt} are plotted in Figure 1.2.4(b). For 130nm poly-Si gate bulk nMOSFETs [12, 13], A_{vt} is about 4.2 mVμm when V_{t0} and t_{OX} are 0.30 to 0.45V and 2.1 to 2.4nm. Most-advanced planar MOSFETs in the 45nm generation attain low A_{vt}s of 1.0 to 2.5mVμm [29, 30, 43] with high-κ metal-gate materials for a thinner t_{OX} and/or FD-SOIs for a lighter N_{sub}. Moreover, the expected lower limit of A_{vt} is about 0.4mVμm with a non-doped-channel FD-SOI and an EOT of 0.5nm. Figure 1.2.5 plots trends in the $\sigma(V_t)$ of various A_{vt}s. Note that nMOSFET sizes for the inverter, the transfer MOSFET that is the smallest in the SRAM cell, and the DRAM SA are $8F^2$, $1.5F^2$, and $15F^2$, respectively. Obviously, $\sigma(V_t)$ of each block rapidly decreases with decreasing A_{vt}.

2.2 Comparisons of V_{min}s for Logic Block, SRAMs, and DRAMs: The V_{t0} of transfer MOSFETs in the SRAM cell is almost the same as that of cross-coupled MOSFETs since their leakage current must be comparable. Thus, the V_{min}s of all blocks can be calculated with Eq. (1). Figure 1.2.6(a) compares the V_{min}s for the logic block and RAMs using repair techniques for various A_{vt}s, exhibiting their strong dependencies on A_{vt}. For A_{vt} = 4.2mVμm, the V_{min}s of the logic and SRAM blocks are almost the same but still high, reaching an intolerable level of about 1.5V in the 32nm generation. For A_{vt} = 1.5mVμm, however, they are reduced to below 1V even in the 22nm generation. Note that the V_{min} of DRAMs is lowest due to the smallest $\sigma(V_t)$ and fewer SAs.

2.3 State-of-the-Art SRAM Circuits: Recent developments in high-speed 6T SRAMs have focused on managing to remain at around the 1V wall rather than reducing V_{min} and thus V_{DD}. Managing the power of the cell is an effective way of tackling the rapidly degraded voltage margin caused by an ever-larger $\sigma(V_t)$, despite a lithographically symmetric cell layout being used [3]. Figure 1.2.7 summarizes practical 6T cells using power management and an 8T cell. The combination of a low-V_t (V_{tL}) transfer MOSFET and a negative word-line scheme [17] (Figure 1.2.7(a)) increases the read margin more than another combination of a high-V_t (V_{tH}) transfer MOSFET and a boosted word-line scheme [18]. This is because the low V_t reduces the $\sigma(V_t)$. In this scheme, as the data (bit)-line voltage can be scaled down according to the scaled MOSFET in peripheral circuits, high density and low power are achieved for data-line-relevant circuits. Dynamic power controls of driver nMOSFETs (Figure 1.2.7(b)) [19 to 22] or load pMOSFETs [23, 24] reduce the V_t of the MOSFETs in the active mode (ACT) while reducing leakage in the standby mode (STB) with increased V_t (= δV_t) due to the body effect. Power control of pMOSFET loads (Figure 1.27(c)) [25, 26] to increase load impedance during write periods improves the write margin. In addition to such power management, shortening the data (bit) line [27] effectively prevents data from being destroyed in half-selected cells along the word line during writing. It has been reported that 8T SRAM cells (Figure 1.2.7(d)) [28] widen the read and write margins due to separated read and write functions in a cell, despite the need for an additional read signal detector and/or rewrite circuit on each data (bit) line. The read static-noise margin becomes wider than that of the 6T cell due to eliminating the ratioed operation. Moreover, larger MOSFETs can be used for M_4 and M_5, enabling a higher read speed and lower V_{min} defined by speed variations. Furthermore, the resulting small driver MOSFETs (M_2) increase the write margin and offset increase in the area due to additional MOSFETs (M_4, M_5).

Using the largest MOSFET possible [21, 22] in the 6T or 8T cell also effectively and simply widens the margin with reduced $\sigma(V_t)$ despite the increased area. For example, for a 6T cell for which all transistors scale down channel lengths with fixing channel widths to the same sizes such as in the 90nm generation (where $LW \propto F$ with W fixed at 90nm in Figure 1.2.8(a)), increase in the V_{min} can be suppressed. In contrast, the V_{min} of the conventional scaling (that is, $LW \propto F^2$) rapidly increases with decreasing F. The cell size (Figure 1.2.8(b)) of the W-fixed approach is gradually reduced since all Ws in the cell are fixed at each generation. Thus, it becomes equal to that of the 8T cell having the cell size of $156F^2$ to $185F^2$ before the 32nm generation, while the conventional scaling reduces cell size more rapidly while maintaining the cell size of $120F^2$. In practice, MOSFET sizes in the 6T cell can be adjusted between the two approaches, and the V_{min} is thus between about 1V and 0.6V in the 32nm generation for A_{vt} = 2.5mVμm, as seen in Figure 1.2.8(a). This suggests that multiple cell sizes/types combined with multi-V_{DD} on a chip are feasible, depending on the required memory chip capacity. For a small-capacity SRAM, in which an overhead due to ECC is intolerable, enlargement of MOSFETs in the cell enables low-V_{DD} operations. However, for a large capacity SRAM which necessitates a small cell size, repair techniques and/or a dedicated high-voltage supply for the SRAM will be a solution. In any event, although V_{DD} has been managed so that it will remain at about 1V even in 45 to 32nm generations, it will continue to increase after this, especially for conventional scaling aiming at higher density, as long as such poly-Si-gate bulk MOSFETs are used.

3. Challenges to Low-Voltage Devices and Circuits

If V_{min} needs to be lowered by a factor of at least $\alpha^{-0.5}$ (α: scaling factor > 1) with device scaling, taking the past trends (Figure 1.2.1) into account, both $\sigma(V_t)$ (= $A_{vt}LW^{-0.5}$) and V_{t0} must be scaled down at the same factor, as predicted by Eq. (1). Thus, the challenge is to develop new $\sigma(V_t)$-scalable MOSFETs. Indeed, conventional poly-Si gate MOSFETs having an A_{vt} as large as 4.2 to 2.5mVmm are pessimistic for reducing V_{min}, as has been discussed thus far. Another challenge is to create new $\sigma(V_t)$-scalable circuits.

3.1 $\sigma(V_t)$-Scalable MOSFETs: If all feature sizes of a planar MOSFET are scaled down by a factor of $1/\alpha$, as outlined in Figure 1.2.9, V_{min}-scaling at $\alpha^{-0.5}$ imposes an intolerable scaling factor of $\alpha^{-1.5}$ on A_{vt}, because of rapidly-scaled LW at α^{-2}. Even if the scaling factor of A_{vt} is reduced to the practical value of $\alpha^{-0.5}$, V_{min} increases by a factor of $\alpha^{0.5}$. Furthermore, V_{min} manages to remain constant even with an A_{vt} scaling as large as α^{-1}. However, the vertical structure provided by FinFETs [31, 32], regardless of fully-depleted (FD) or partially-depleted (PD) MOSFETs, yields a new scaling law for $\sigma(V_t)$, mitigating the requirement to A_{vt}. This is because this structure allows LW to be constant or even increase when the fin height (that is, the channel width W) is scaled up despite the channel length L being scaled down. This can be done without sacrificing MOSFET density. This up-scaling is done according to 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, respectively. Such FinFETs enable high speed not only due to the large drive current but also the shorter interconnects deriving from the vertical structures. However, the aspect ratio (W/L) of FinFETs increases with device scaling. For example, this is as large as 4 to 16 in the 11nm generation, as shown in Figure 1.2.9. However, such large aspect-ratio structures might be possible when the history of DRAM development is taken into consideration. In DRAMs, the aspect ratio of

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

trench capacitors has increased from about three in the early 1980s to as large as 70 for modern 70nm DRAMs [33, 34].

Based on the MOSFET scaling, let us try to predict the V_{min}s for the blocks of the future, assuming that the A_{vt} in the 45nm generation, the A_{vt} scaling factor for further device scaling, and V_{t0} are 2.5mVμm, $\alpha^{-0.5}$, and 0.4V for low-power designs, and 1.5mVμm, α^{-1} and 0.2V for high-performance designs (see Figure 1.2.4(b)). The constant LW in Figure 1.2.9 is also assumed for FinFETs. Obviously, FinFETs allow $\sigma(V_t)$ to be scaled down for both designs, as seen in Figures 1.2.10(a) and 1.2.10(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 (Figure 1.2.11(a)), FinFETs reduce V_{min} to about 0.65V for the logic block and SRAMs, and 0.46V for DRAMs in the 11nm generation, while for high-performance designs (Figure 1.2.11(b)) they reduce to as low as about 0.27V and 0.22V, respectively. It should be noted that FinFETs reduce the V_{min} of SRAMs to such low levels even with $1.5F^2$ transfer MOSFETs. However, the V_{min}s for low-power designs are still higher than 0.5V due to a high V_{t0} of 0.4V, calling for further reductions in V_{t0} and A_{vt} or further increases in the fin height. Here, V_{t0} is reduced by V_{t0}-scalable low-leakage circuits that will be explained later and power switches tolerating a lower V_{t0}. Note that FD-FinFETs are compulsory for RAM cells to ensure high density and robust design. Even for the logic block, they can be used throughout the chip. However, unless inter-die V_t-variations are confined to a tolerable level, PD-FinFETs must be used for major logic gates on a chip to compensate for the variations by controlling the substrate bias. Fortunately, substrate noise is lowered by the reduced pn-junction area due to using SOI structures. Multiple (deep and shallow) fins and/or their combinations may overcome the inconvenience that the width of MOSFETs can only be controlled in multiples of fins.

3.2 $\sigma(V_t)$-Scalable Circuits: $\sigma(V_t)$ is reduced by reducing V_{t0}, as described previously. For example, the $\sigma(V_t)$ of poly-Si gate bulk MOSFETs is reduced to 45% when V_{t0} is changed from 0.4V to 0V. Combined with a high-V_{t0} circuit, such low-V_{t0} circuits effectively reduce the V_{min} of the whole logic block if the leakage involved is sufficiently reduced. Although the well-known dual-V_t circuit is an example of this idea, it is not very effective, since there is little difference (such as, 0.1V) between the two V_{t0}s. Figure 1.2.12 shows other dual-V_{t0} and dual-V_{DD} (V_{DD} and $V_{DL} < V_{DD}$) circuits using the concept of gate-source back-biasing [35]. They work with a large difference in V_{t0}, exemplified by a high V_t (V_{tH}) of 0.4V and a low V_t (V_{tL}) of 0V. In the basic concept shown in Figure 1.2.12(a), back-biasing is applied to a V_{tL}-pMOSFET during inactive periods with the help of the higher power supply, V_{DD}. As a result, an effective high V_{t0} (V_{teff}), despite a low-actual V_{tL}, is developed to reduce leakage during inactive periods. Even so, the gate-over-drive (V_{geff}) is maintained at a high level during the active periods. Thus, V_{t0} becomes scalable by adjusting the back-bias, also making $\sigma(V_t)$ almost scalable. Figures 1.2.12(b) and 1.2.12(c) show applications to a dynamic inverter with V_{DD}-precharge clock P [1, 36] and a self-resetting inverter [37 to 40]. M_2 as well as M_1 are back-biased during inactive periods. Even for such circuits, Eq. (1) can be applied to high- and low-V_t MOSFETs if the A_{vt}s of the pMOSFET and the nMOSFET are equal. Figure 1.2.13 shows trends in $V_{min}(V_{DD})$ and $V_{min}(V_{DL})$, which are for the V_{DD} sub-block and the V_{DL} sub-block in the low-power-designed FinFET logic block, assuming that $\sigma(V_{tL} = 0V)/\sigma(V_{tH} = 0.4V) = 0.45$. Obviously, $V_{min}(V_{DD})$, which is the same as in Figure 1.2.11(a), is gradually reduced, while $V_{min}(V_{DL})$ remains at an extremely low value. The V_{min} of the whole block is between $V_{min}(V_{DD})$ and $V_{min}(V_{DL})$. It becomes equal to $V_{min}(V_{DL})$ where no V_{DD} circuits are used, while it becomes equal to $V_{min}(V_{DD})$ where no V_{DL} circuits are used. Therefore, using V_{DL} circuits as much as possible effectively reduces the V_{min} of the logic block to less than 0.5V. A similar circuit can be seen in a recently presented DRAM [44], in which a low-V_{t0} (V_{tL}) temporarily activated dynamic preamplifier is backed up by a high-V_{t0} (V_{tH}) sense amplifier. If $V_{tH} = 2V_{tL}$, even half-V_{DD} data-line precharging [3, 4] achieves the same V_{min} as in Figure 1.2.11(a). Eventually, for low-power designs, the V_{min} of the SRAM block might be highest after applying such $\sigma(V_t)$-scalable circuits to the logic block and DRAMs. It should be noted that high-performance designs do not accept such dynamic circuits but only static circuits for robust designs because of the low V_{t0} of 0.2V. Thus, the V_{min} is kept the same as in Figure 1.2.11(b), which is higher than the $V_{min}(V_{DL})$.

4. Challenges to High-Density Cores, Chips and Compact Subsystems

Small cores and chips, new architectures such as multi-core MPUs, and 3-D thermally conscious chip integration [9] for compact subsystems are keys to alleviating the interconnect-delay problem with reduced wire-length distributions. As they will also ensure power-supply integrity throughout the subsystem, low-V_{DD} operation is made possible with a reduced difference between V_{DD} and V_{min}. For these, a drastic reduction in the memory array area is particularly vital since the array dominates the core or chip.

4.1 Logic-Process-Compatible FinFET DRAM Cells: Unique two-dimensional (2D) selection minimizes the array area of DRAMs. For example, each data line (DL) connecting 512 cells (p = 512) in the conventional selection (Figure 1.2.14(a)) is formed into a sub-array of p' = 16 and q' = 32 with only one SA, so only one cell at the cross point of a selected row line (WL) and a selected column line (YS) is selected while activating only one data-line (DL), unlike conventional selection. Consequently, this enables a simple cell capacitor, a simple layout for SA, and negligible capacitive-coupling noise from adjacent data lines. For half-V_{DD} data-line precharging, the read signal is given as $v_s' = C_s'V_{DD}'/2(C_s' + p'C_d + q'C_{i/o})$ for 2D selection, and $v_s = C_sV_{DD}/2(C_s + pC_d)$ for conventional selection, where C_d and $C_{i/o}$ correspond to data-line and i/o-line capacitances per cell. If $v_s = v_s'$, $C_s' \ll p'C_d + q'C_{i/o}$, $C_s \ll pC_d$, and $C_d = 4C_{i/o}$, the selection reduces necessary capacitance C_s' to $0.05C_s$ for $V_{DD} = V_{DD}'$. This selection minimizes the cell area and simplifies the cell structure, if two FD-FinFETs and a FinFET capacitor are combined in a cell, and a buried YS line forms two MOS gates at the inner side walls of adjacent fins, as outlined in Figure 1.2.15. The resulting cell is as small as $5F^2$, which is smaller than existing stand-alone DRAM cells with sophisticated capacitor structures and about 1/32 of that of a 6T SRAM cell (that is, > $160F^2$ as seen in Figure 1.2.8). Note that the V_{min} is the same as that in Figure 1.2.11(a) as long as a low-V_{t0} preamplifier [44] is used, as was previously explained.

4.2 Low-Power High-Density Sub-Systems: If small FinFET cores, each embedding a large-capacity DRAM, are connected with low-resistive global interconnects and meshed power-supply lines, as seen in the multi-divided array of modern DRAMs [4], high-speed multi-core LSIs [41,45] will be achieved. For example, a hypothetical 0.5V 16k-core LSI accommodating as many as 320Mgate and 8Gb DRAMs in a 10×10mm^2 chip would be feasible in the 11nm generation. Each homogeneous core including 20kgate and 512Kb DRAM with a $5F^2$ cell, as previously mentioned, would be less than 60×60μm^2. The V_{t0} of FinFETs in each core may be reduced (for example, by 0.1V or more from V_{t0} = 0.4V) to keep speed high even at a low V_{DD} (for example, < 0.5V); the resulting increased leakage can be reduced if high-speed low-noise power-switches are available. A real challenge is to find appropriate applications fully utilizing such a powerful multi-core chip. The compact 3D integration of small chips with high-density through silicon vias (TSVs) [9] will ensure excellent power-supply integrity and low noise in signal lines throughout the subsystem.

5. Scenario for Reaching the Sub-0.5V Nanoscale Era

For memory-rich CMOS LSIs, planar metal-gate MOSFETs, regardless of bulk or FD-SOI, will be used for low-cost medium-voltage LSIs because of their simple structure, despite the more severe requirement on scaling in $\sigma(V_t)$, as discussed earlier. In fact, an intensive study focusing on the sources and controls of diverse V_t-variation components is in progress, especially for planar FD-SOI MOSFETs [42, 43]. However, in the long run, FinFETs are promising candidates as nanoscale LSI devices because of their great potential for voltage scaling and interesting applications previously described, despite the need for more sophisticated control of their dimensions and impurities, and so on. The dual-V_{t0} and dual-V_{DD} approach using

978-1-4244-3457-2/09/$25.00 ©2009 IEEE

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

gate-source back-biasing, as exemplified previously, will be a new circuit style in the nanoscale era. Repair techniques against variations will continue to be crucial even for logic blocks. 6T SRAM cells as well as 8T SRAM cells will continue to be used for small-memory applications. However, for large-memory applications, simple-capacitor FinFET DRAM cells might replace SRAM cells because of their higher density and lowest V_{min}. New multi-core architectures and 3D integration are expected to solve ever-larger local interconnect delays in the nanoscale era, and reduce the operating voltage with excellent power-supply integrity.

For mixed signal LSIs, the above-described low-voltage technologies are applicable to the logic/memory block occupying 20 to 70% of the total chip area. For the analog block, FinFETs also lower the V_{min} of differential circuits with a reduced ΔV_t/offset and no body effect. In particular, combined with the above-described gate-source back-biasing and non-doped MOSFETs (i.e., V_{tL} = 0V), the V_{min} of the cascode amplifier that usually has the highest V_{min} in the analog block due to multi-stacked MOSFETs is greatly reduced while maintaining a low leakage power. FinFETs also reduce the power supply/substrate noise of the analog block due to the reduced pn-junction area of all the MOSFETs in the chip. In addition, they can be used to make not only high-Q inductors due to the reduced substrate loss, but also high-density capacitors for low pass filters and pipeline ADCs, and can be used to ensure power supply integrity. The 1V wall can thus be broken and the 0.5V nanoscale era will open the door to lower power dissipation, if the necessary devices and fabrication-process technologies (which are beyond the scope in this paper) are developed. Disruptive inventions and technologies expected in the future will make such an era a reality.

6. Conclusion

The V_{min}s of logic, SRAM, and DRAM blocks were compared with a newly proposed methodology for evaluating V_{min} based on speed variations, taking repair techniques into account. State-of-the-art 6T SRAM cells were then discussed in terms of V_{min} and cell size. After that, many adaptive circuits and relevant technologies needed to break the 1V wall were proposed and evaluated, while taking the interconnect problem into account. Finally, 0.5V nanoscale LSIs including mixed signal LSIs were predicted to be feasible, if relevant devices and fabrication processes are developed.

Acknowledgement:
The author acknowledges contributions from many research colleagues at Hitachi Central Research Laboratory, especially Drs. M. Yamaoka, S. Kimura, D. Hisamoto, N. Sugii, R. Tsuchiya, T. Kawahara and T. Oshima, and Mrs. T. Sekiguchi, M. Saen and T. Yamawaki for their valuable discussions and suggestions.

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 Electronics, to be published in 2008.
[3] K. Itoh, M. Horiguchi, and H. Tanaka, *Ultra-Low Voltage Nano-Scale Memories*, Springer, 2007.
[4] K. Itoh, *VLSI Memory Chip Design*, Springer-Verlag, 2001.
[5] Y. Nakagome, et al., "Review and prospects of low-voltage RAM circuits," IBM J. R & D, vol. 47, no. 5/6, pp. 525-552, Sep./Nov. 2003.
[6] J. A. Davis, et al., "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," Proc. of the IEEE, vol. 89, no.3, pp.305-324, March 2001.
[7] W. Haensch, et al., " Silicon CMOS devices beyond scaling," IBM J. Res. Dev., vol. 50, no. 4/5, pp. 339-361, July/Sept. 2006.
[8] T.C. Chen, "Where CMOS is going: Trendy Hype vs. Real Technology," ISSS Dig. Tech. Papers, pp.22-28, Feb. 2006.
[9] A.W. Topol, et al., "Three-dimensional integrated circuits," IBM J. Res. Dev., vol. 50, no. 4/5, pp. 491-506, July/Sept. 2006.
[10] M.J.M Pelgrom, et al., "Matching properties of MOS transistors," J. SSC Oct. 1989; 24(5):1433-1439.
[11] K. Takeuchi, et al., "Understanding Random Threshold Voltage Fluctuation by Comparing Multiple Fabs and Technologies," IEDM Dig. pp. 467-470, Dec. 2007.
[12] H. Masuda, et al., "Approach for physical design in sub-100 nm era," ISCAS pp. 5934-5937(6), May 2005.
[13] 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.
[14] S. Das, et al., "A self-tuning DVS processor using delay-error detection and correction," J. SSC April 2006; 41(4): 792-804.
[15] T. Nakura, et al., "Fine-Grain Redundant Logic Using Defect-Prediction Flip-Flops," ISSCC Dig., pp. 402-403, Feb. 2007.
[16] D. Blaauw, et al., "In Situ Error Detection and Correction for PVT and SER Tolerance," ISSCC Dig., pp. 400-401, Feb. 2008.
[17] 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.
[18] J. Pille, et al., "Implementation of the CELL Broadband Engine in a 65nm SOI Technology Featuring Dual-Supply SRAM Arrays Supporting 6GHz at 1.3V," ISSCC Dig., pp. 322-323, 606, Feb. 2007.
[19] H. Akamatsu, et al., A low power data holding circuit with an intermittent power supply scheme for sub-1V MT-CMOS LSIs", pp. 14-15, June. 1996.
[20] M. Yamaoka, et al., "A 300MHz 25mA/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.
[21] K. Itoh, et al., "Reviews and future prospects of low-voltage embedded RAMs," CICC2004 Dig., pp. 339-344, Oct. 2004.
[22] M. Khellah, et al., "A 4.2GHz 0.3mm^2 256kb Dual-Vcc SRAM Building Block in 65nm CMOS," ISSCC Dig., pp.624-625, Feb. 2006.
[23] F. Hamzaoglu, et al., "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High-κ Metal-Gate CMOS Technology," ISSCC Dig., pp. 376-377, Feb. 2008.
[24] H. Pilo, et al., "A 450ps Access-Time SRAM Macro in 45nm SOI Featuring a Two-Stage Sensing Scheme and Dynamic Power Management," ISSCC Dig., pp. 378-379, Feb. 2008.
[25] M. Yamaoka, et al., "Low-power embedded SRAM modules with expanded margins for writing," ISSCC Dig., pp. 480-481, Feb. 2005.
[26] K. Zhang, et al., "A 3-GHz 70MB SRAM in 65nm CMOS technology with integrated column-based dynamic power supply," ISSCC Dig., pp. 474-475, 611, Feb. 2005.
[27] 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.
[28] L. Chang, et al., "A 5.3GHz 8T-SRAM with Operation Down to 0.41V in 65nm CMOS," Symp. VLSI Circuits Dig., pp. 252-253, 2007.
[29] 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.
[30] 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.
[31] D. Hisamoto, et al., "A Fully Depleted Lean-Channel Transistor (DELTA)-A Novel Vertical Ultra Thin SOI MOSFET," IEDM Dig., pp. 833-836, 1989.
[32] J. Kavalieros et al., "Tri-Gate Trnsistor Architecture with High-κ Gate Dielectrics, Metal Gates and Strain Engineering," Symp. VLSI Tech. Dig., pp. 62-63, June 2006.
[33] H. Sunami, "The role of the Trench Capacitor in DRAM Innovation," IEEE SSCS News, Winter 2008, pp. 42-44, Jan. 22, 2008.
[34] J. Amon, et al., "A highly manufacturable deep trench based DRAM cell layout with a planar array device in a 70nm technology," IEDM Dig. pp.73-76, Dec. 2004.
[35] Y. Nakagome et al., "Sub-1-V swing bus architecture for future low-power ULSIs", Symp. VLSI Circuits Dig., pp. 82-83, June 1992.
[36] K. Itoh et al., "Low-Voltage Limitations of Deep-Sub-100-nm CMOS LSIs-View of Memory Designers-," GLSVLSI2007, Proc., pp. 529-533, March 2007.
[37] T. Chappel, et al., "A 2-ns Cycle, 3.8-ns Access 512Kb CMOS ECL SRAM with a Fully Pipelined Architecture," IEEE JSSC, pp. 1577-1584, Nov. 1991.
[38] T. Mori, et al., "A 1V 0.9mW at 100MHz 2x16b SRAM utilizing a Half-Swing Pulsed-Decoder and Write-Bus Architecture in 0.25µm Dual-Vt CMOS," ISSCC Dig., pp. 354-355, Feb. 1998.
[39] G. Braceras, et al., "A 940MHz Data-Rate 8Mb CMOS SRAM," ISSCC Dig., pp. 198-199, Feb. 1999.
[40] T. Kirihata, et al., "A 390mm^2 16 Bank 1Gb DDR SDRAM with Hybrid Bitline Architecture," ISSCC Dig. pp. 422-423, Feb. 1999.
[41] 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.
[42] 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.
[43] O. Weber, et al., "High Immunity to Threshold Voltage Variability in Undoped Ultra-Thin FDSOI MOSFETs and its Physical Understanding," IEDM Dig., 10.4, Dec. 2008.
[44] S. Akiyama et al., "Low-V_T Small-offset Gated Preamplifier for Sub-1-V Gigabit DRAM Arrays," ISSCC Dig., Feb. 2009.
[45] S. Vanbal et al., "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS," ISSCC Dig., pp.98-99, Feb. 2007.

Please click on paper title to view Visual Supplement.
Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

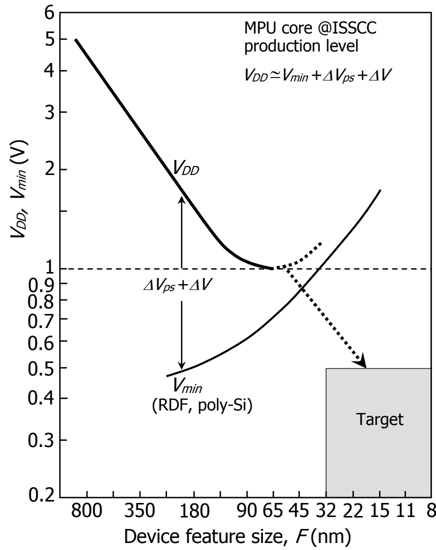
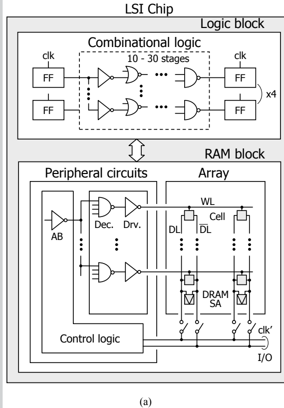


Figure 1.2.1: Trends in V_{DD} and V_{min} of high-performance MPUs.



(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)

Figure 1.2.2: (a) LSI composed of logic block and RAM block and (b) features of blocks [1]. RAM block denotes SRAM block or DRAM block.

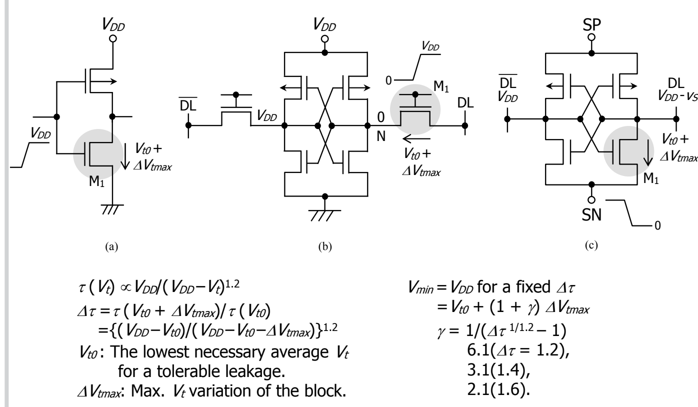


Figure 1.2.3: (a) Inverter, (b) 6-T SRAM cell, and (c) DRAM sense amplifier, and definition of their V_{min}s.

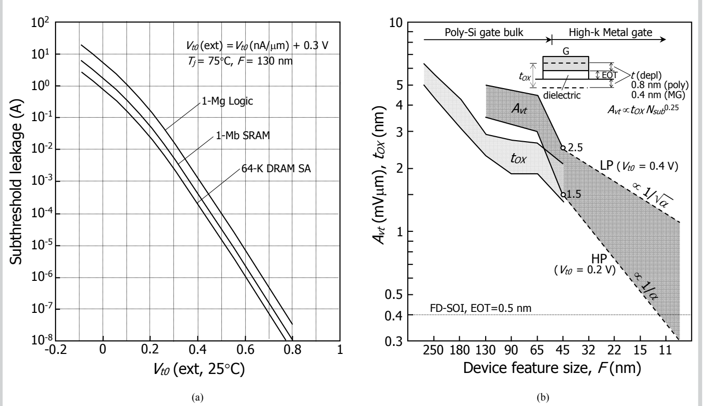


Figure 1.2.4: (a) Leakage vs V_{t0} for various blocks and (b) trends in t_{ox} and A_{vt} [29,30].

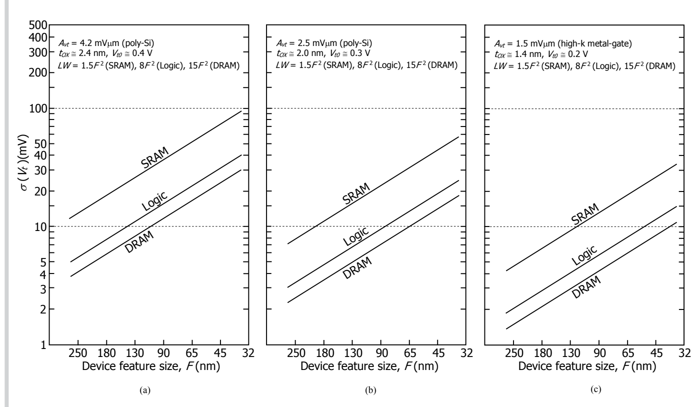


Figure 1.2.5: Trends in $\sigma(V_t)$ for (a) A_{vt} = 4.2mVμm, (b) A_{vt} = 2.5mVμm, and (c) A_{vt} =1.5mVμm.

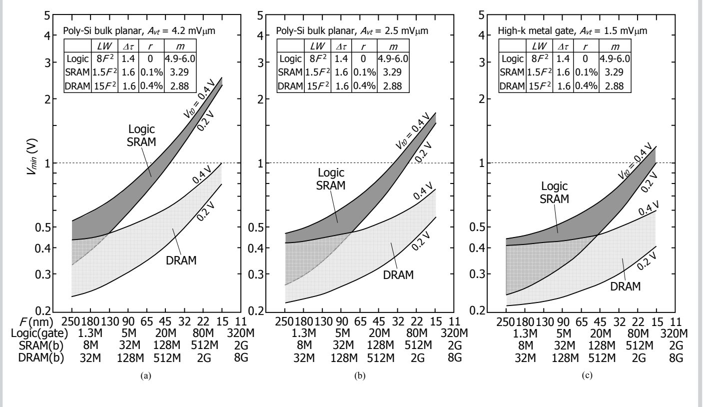


Figure 1.2.6: V_{min}s for the logic block and repaired RAMs for various MOSFETs having (a) A_{vt} = 4.2mVμm, (b) A_{vt} = 2.5mVμm, and (c) A_{vt} = 1.5mVμm.

 978-1-4244-3457-2/09/$25.00 ©2009 IEEE

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

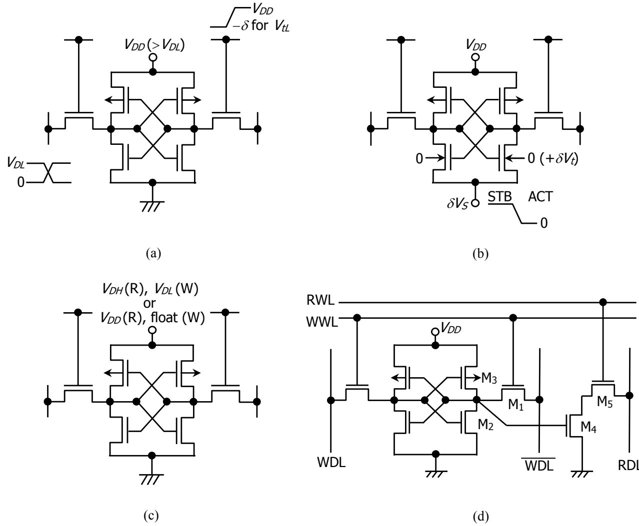


Figure 1.2.7: Practical schemes to maintain voltage margin of SRAM cells. (a)–(c) for 6-T cell, and (d) for 8-T cell.

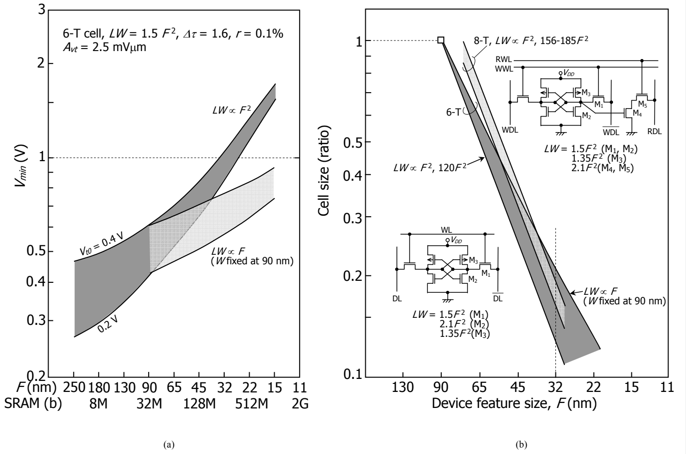


Figure 1.2.8: (a) V_{min} of 6-T cell and (b) cell size of 6-T and 8-T cells.

	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}	$\sim\alpha^{1.1}$	$\sim\alpha^{1.9}$	$\sim\alpha^{0.9}$
τ (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 (α = 1)	45/45 nm	45/45 nm	45/45 nm
W_{min}/L_{min}, F = 11 nm (α = 4)	11/11 nm (aspect ratio = 1)	180/11 nm (16)	90/23 nm (4)

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

Figure 1.2.9: Comparisons of scaling between planar MOSFET and FinFET.

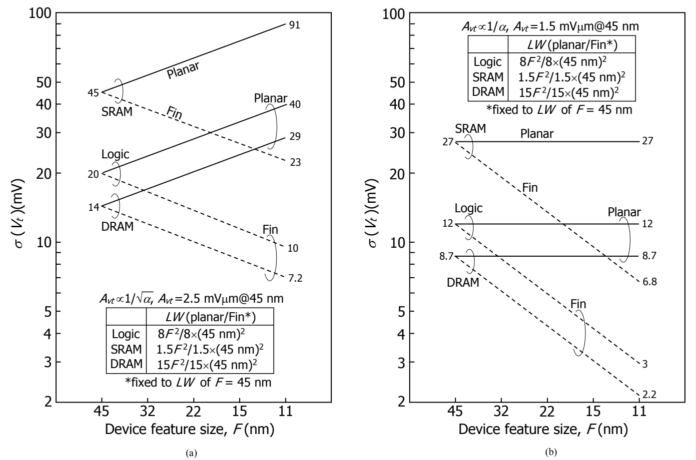


Figure 1.2.10: Expected trends in $\sigma(V_t)$ for (a) low-power designs and (b) high-performance designs.

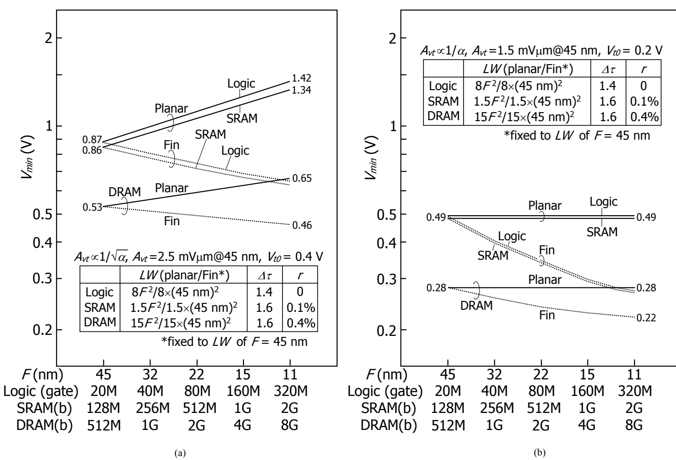


Figure 1.2.11: Expected trends in V_{min} for (a) lower-power designs and (b) high-performance designs.

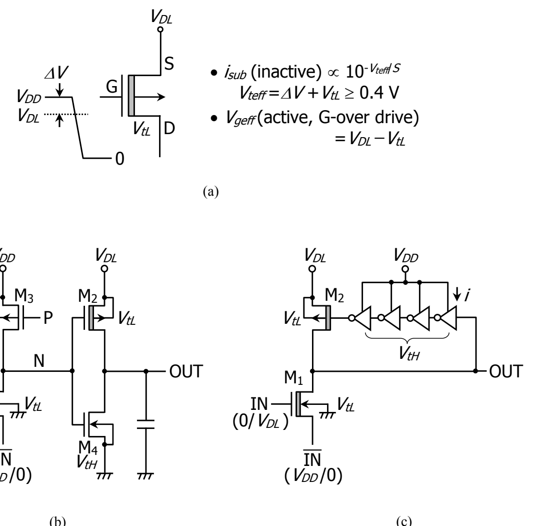


Figure 1.2.12: Dual-V_{DD} dual-V_t circuits using gate-source offset driving. (a) Concept behind gate-source offset driving [1, 35], (b) inverter, and (c) self-resetting inverter.

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.

Please click on paper title to view Visual Supplement.

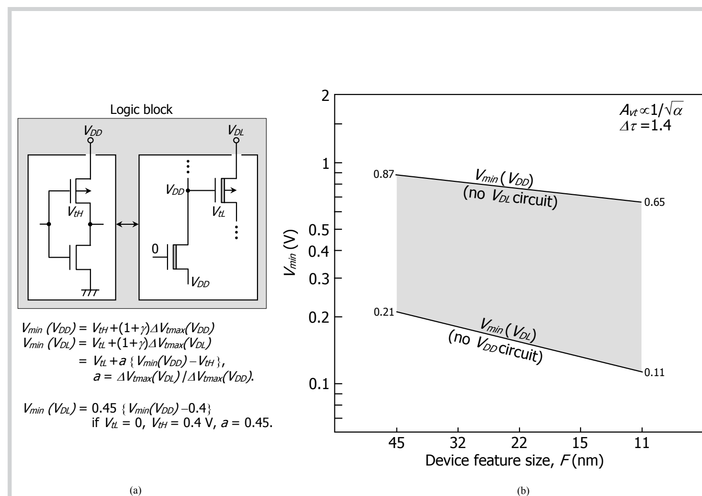


Figure 1.2.13: (a) Low-power FinFET logic block using dual-V_{t0} dual-V_{DD} approach, and (b) trends in V_{min}.

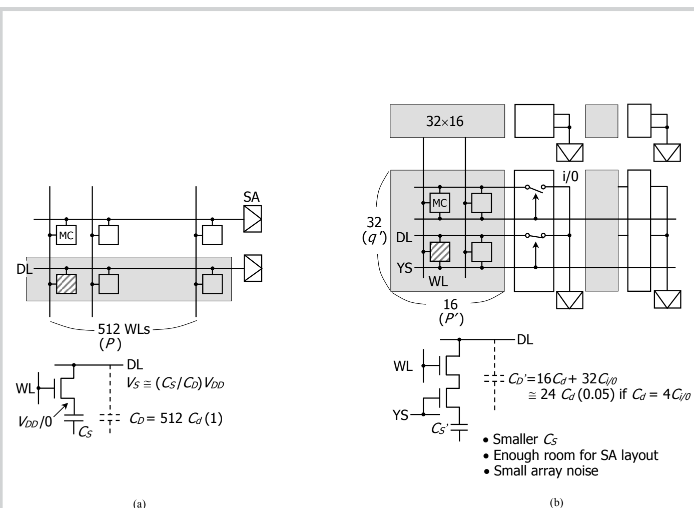


Figure 1.2.14: (a) Conventional and (b) 2-D selections of DRAM cell array.

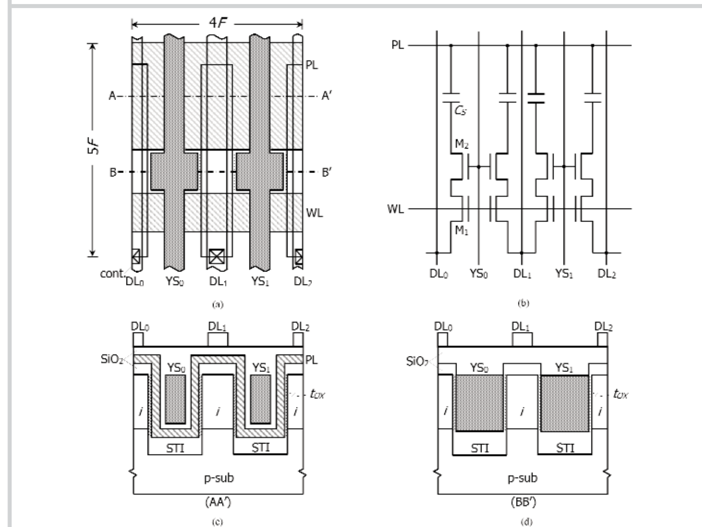


Figure 1.2.15: FinFET DRAM cell structures for 2-D selection. (a) Layout, (b) circuit, and (c) and (d) cross sections.

978-1-4244-3457-2/09/$25.00 ©2009 IEEE

Please click on paper title to view Visual Supplement.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on January 27, 2010 at 20:12 from IEEE Xplore. Restrictions apply.