SOI for Digital CMOS VLSI: Design Considerations and Advances

CHING-TE CHUANG, FELLOW, IEEE, PONG-FEI LU, MEMBER, IEEE, AND CARL J. ANDERSON

This paper reviews the recent advances of silicon-on-insulator (SOI) technology for complementary metal–oxide–semiconductor (CMOS) very-large-scale-integration memory and logic applications. Static random access memories (SRAM's), dynamic random access memories (DRAM's), and digital CMOS logic circuits are considered. Particular emphases are placed on the design issues and advantages resulting from the unique SOI device structure. The impact of floating-body in partially depleted devices on the circuit operation, stability, and functionality are addressed. The use of smart-body contact to improve the power and delay performance is discussed, as are global design issues.

Keywords**—***CMOS digital integrated circuits, CMOS integrated circuits, CMOS memory circuits, silicon-on-insulator technology.*

I. Introduction

Silicon-on-insulator (SOI) technology has long been used in many special applications, such as radiation-hardened or high-voltage integrated circuits. It is only in recent years, however, that SOI has emerged as a serious contender for low-power, high-performance applications [1]–[4]. The primary reason is the power consumption of scaled bulk complementary metal–oxide–semiconductor (CMOS) technology. While the feasibility and performance of 0.15-μm bulk CMOS technologies with sub-0.1-μm effective channel length have been demonstrated, it is not clear that these bulk CMOS technologies will work satisfactorily within the power constraints of the intended low-voltage applications [2], [4], [5]. By dielectrically isolating the circuit elements, SOI technology significantly reduces the junction capacitances, allowing the circuits to operate at higher speed or substantially lower power at the same speed. The device structure also eliminates latchup in bulk CMOS and improves the short-channel effect and soft error immunity. Although these advantages of the SOI technology are well known, the successful introduction of SOI technology for large-scale mainstream applications faces some key challenges across the entire spectra of material, process, manufacturing, devices, and designs. At the material and process level, neither bonded nor separation by implanted oxygen (SIMOX) SOI are mature enough for mass production of low-cost, low-defect-density substrates [2]. The crucial control of silicon film thickness to accurately control the threshold voltage of fully depleted devices remains a major concern. At the device and circuit level, the floating-body effect in partially depleted devices and the resulting hysteresis and instability during dynamic operations pose major challenges for large-scale designs. While numerous literature has addressed the material/process/device aspects of the SOI technology, and various SOI-based designs have been reported, a comprehensive account of the SOI-specific design issues for the memory and logic circuits has been lacking.

In this paper, we review the recent advances of SOI technology for digital CMOS very-large-scale-integration (VLSI) applications. Particular emphases are placed on the design issues and advantages resulting from the unique SOI device structure. Section II discusses the SOI device structures and the floating-body effect in partially depleted devices. Static random access memories (SRAM's) are then discussed in Section III, and dynamic random access memories (DRAM's) in Section IV. The digital CMOS logic circuit family and some commonly used circuit blocks for fast arithmetic operation in processor data flow are dealt with in Section V, followed by the pass-transistor-based designs in Section VI. The use of "smart" body contact to improve the power and delay performance is discussed in Section VII. Global design issues such as timing considerations, decoupling capacitors, electrostatic discharge (ESD) protection, and heat dissipation are addressed in Section VIII, followed by a general discussion in Section IX. The conclusion of this paper is given in Section X.

II. Device Structure and Floating-Body Effect

The schematic cross section of a basic nMOS field-effect transistor (FET) on SOI is shown in Fig. 1(a). Various device structures and designs have been systematically studied in detail in [4] and [6]. The most fundamental device-

Manuscript received September 10, 1997; revised February 3, 1998.

C.-T. Chuang and P.-F. Lu are with the IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

C. J. Anderson was with the IBM Research Division, T. J. Watson Research Center, Yorktown Heights, NY 10598 USA. He is now with IBM, Austin, TX 78701 USA.

Publisher Item Identifier S 0018-9219(98)02849-7.

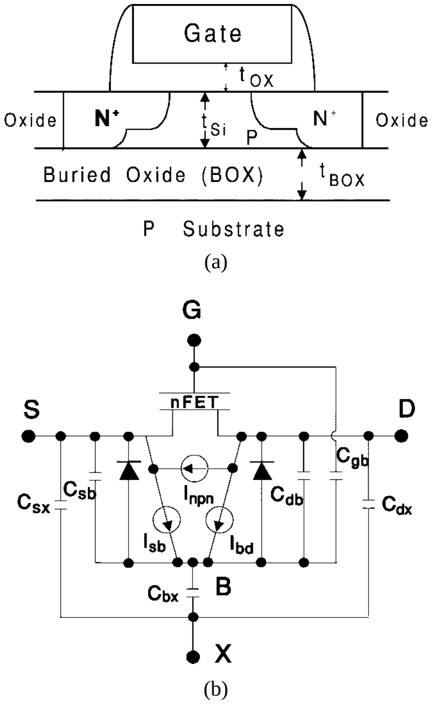


Fig. 1. (a) Schematic cross section of a partially depleted SOI nMOSFET. (b) The equivalent circuit model. (I_{npn} is the parasitic lateral NPN transistor collector current. The two diodes are the internal base-emitter and base-collector junction diodes; I_{sb} and I_{bd} are the impact ionization currents [9], [12], [15].)

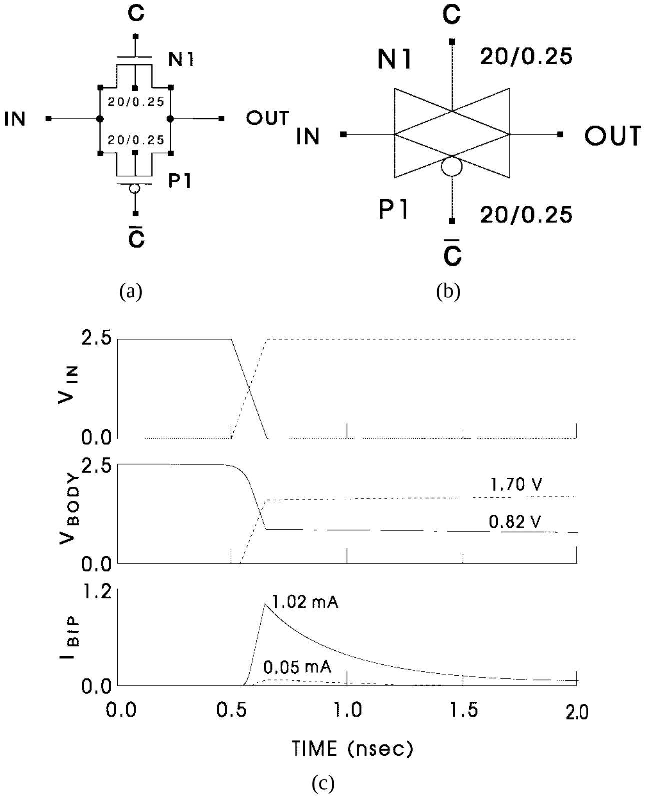


Fig. 2. (a) Basic pass-gate configuration, (b) circuit symbol, and (c) pertinent switching wave forms for parasitic bipolar current through nMOS (solid lines) and pMOS (dashed lines) [10], [12].

design issue is the choice between a fully depleted device versus a partially depleted device. In a fully depleted device, ultrathin (<50 nm or so) silicon film is used so the depletion layer extends through the entire film. The advocated advantages include the elimination of the floating-body effect and better short-channel behavior. However, the claimed better short-channel behavior stems from the reduced source/drain junction depth and is traded against the source/drain series resistance. Furthermore, the requirement that the silicon film thickness always remain well below the depletion width dictates a low device threshold voltage V_T with high sensitivity to process and thickness variations and stringent control of the film thickness to within 5–10 nm [4], [6], [7], which impose severe limitations on the manufacturability of the device. A partially depleted device (with film thickness around 150 nm) alleviates the constraint on V_T and its sensitivity, allowing the channel doping profile to be tailored for any desired V_T and thus easing the manufacturing problem. The major issue of the partially depleted device is the "floating-body effect" and the resulting parasitic bipolar effect. The floating region under the MOS device channel acts as the base of the parasitic lateral bipolar device, with the base current supplied by impact ionization [Fig. 1(b)]. The floating body has been known to introduce a kink in the DC I–V characteristics, lower the V_T at high drain bias, degrade breakdown voltage, and cause hysteresis and instability during dynamic operations [8]–[11]. The global use of body contact in every device to eliminate the floating-body effect may severely degrade the density, and hence the performance, of a large-scale design. While the current gain of the parasitic bipolar device can be suppressed by using source/drain extensions to reduce the emitter/collector area, as well as retrograde channel doping to increase the back interface doping and the effective bipolar Gummel number [4], the existence of the floating body and the possibility of parasitic bipolar leakage inevitably demand that the circuit designers meticulously examine the circuit functionality and margin under various process, supply, and temperature corners.

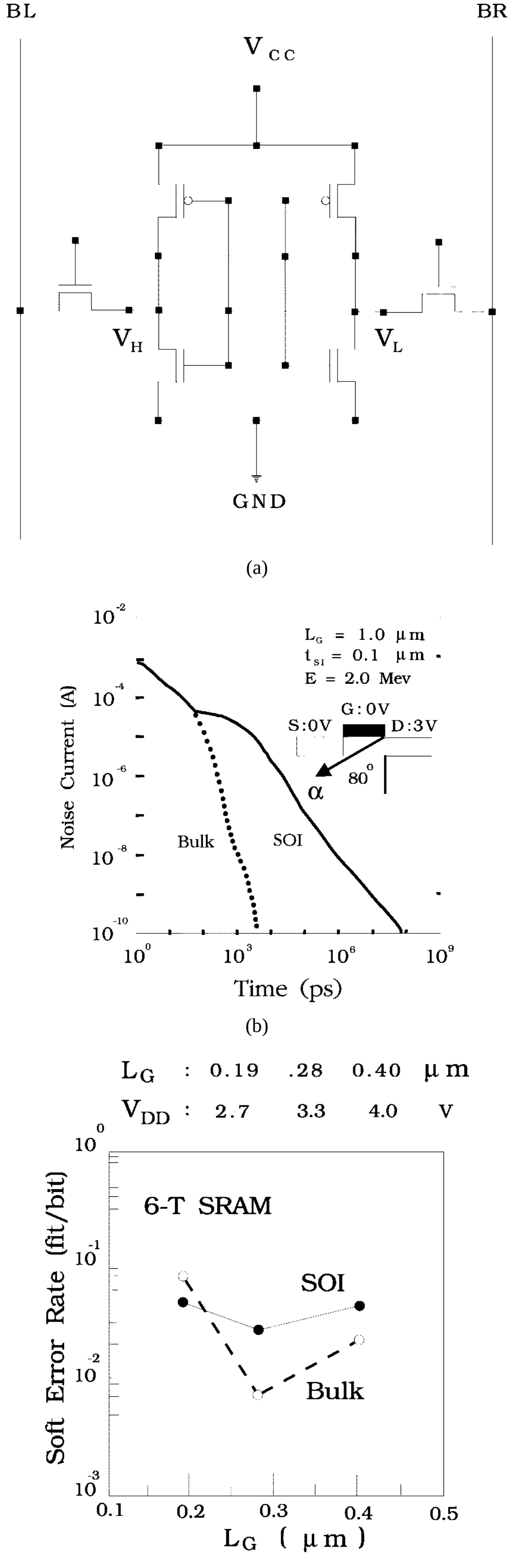
For the parasitic bipolar effect to manifest during the circuit operation, the circuit topologies and switching patterns must be such that a large voltage is created/developed across the base-emitter junction (i.e., the body-source junction) of the parasitic bipolar transistor [12]. In the floating-body configuration, this can only be realized by pulling down the emitter (i.e., source) node. One example is the basic pass-gate configuration depicted in Fig. 2(a) [13]–[15]. The circuit symbol is shown in Fig. 2(b). Consider the situation that after passing the "high" (V_{DD}) state, C switches to "low" (ground), and the input is subsequently pulled down. For the nMOS, with its gate off and both the source and drain nodes at "high," its body will be at "high" as well. Pulling down the input (source) node creates a large forward bias across its body-source junction, resulting in large current through the parasitic bipolar transistor. Since the body (base) is discharged by the current flow, the parasitic bipolar current presents only as a transient phenomenon. The pertinent switching wave forms are shown in Fig. 2(c). For the pMOS, the complementary situation holds. But the parasitic bipolar effect is less pronounced [as shown by the dashed lines in Fig. 2(c)] due to the lower impact ionization rate and smaller current gain of the parasitic pnp transistor.

Other circuit topologies susceptible to the parasitic bipolar effect will be discussed in Section V.

III. Static Random Access Memories

One of the most quoted advantages of the SOI technology is the improvement in soft error rate (SER) [1], mainly because of its long history in the radiation-hardened applications. The α-particle from radioactive elements in packaging has been known to induce soft error and impose severe design constraints in six-transistor (6-T) planar SRAM cells [Fig. 3(a)]. The net charge imbalance Q_{CRIT} necessary to upset/flip the cell state equal to $C\Delta V$, where C is the capacitance seen at the storage node of the cell and ΔV is the cell differential voltage (which equals the supply voltage for a 6-T SRAM cell). The state-of-the-art planar 6-T SRAM cell in 0.5-μm design rules [16], [17] typically has a cell size around 30–35 μm^2 and a Q_{CRIT} around 25–30 fC. With 0.25-μm design rules, the cell size is expected to shrink to about 10–12 μm^2, with a Q_{CRIT} of 10–15 fC. While the reduced cell size (and hence reduced device parasitics) and storage node capacitance do improve cell performance, the sublinear improvement in cell access time is traded against the almost exponential deterioration in SER. At design rules around 0.25 μm, α-induced SER is expected to surpass the cosmic ray (which is relatively insensitive to Q_{CRIT} and whose SER remains relatively constant with technology scaling) and become the major failure mechanism.

In bulk CMOS, the α-generated charges are collected mainly by the funneling effect [18] when an α-particle hits the drain diffusion layer. Due to the buried oxide in an SOI MOSFET, this effect is not significant, and SOI MOSFET's are believed to have excellent soft error immunity [1]. In SOI MOSFET's, appreciable charge collection can only occur when an α-particle hits the channel region [19], [20]. Although the amount of α-generated charges in an SOI MOSFET is substantially less than that in a bulk MOSFET, the total charges collected at the cell storage (drain) node are significantly higher than the α-generated charges due to the parasitic bipolar effect. Detailed three-dimensional simulations have been carried out in [21] to access the SER of an SOI 6-T cell in scaled design rules. The α-induced bipolar current was found to flow over a long period [Fig. 3(b)] [20], [21], and the SER's for the SOI cell are in the same order as the bulk cell [Fig. 3(c)] due to the parasitic bipolar effect. While these results can only be taken qualitatively due to the strong dependence of the parasitic bipolar effect on process/device details and individual cell design, they do point out that the improvement of SER in SOI SRAM cannot be taken for granted. Although the use of body contact is an obvious solution, the density requirement may not warrant such an option. The curve for the SER of SOI SRAM also behaves



(a)

(b)

(c)

Fig. 3. (a) 6-T planar SRAM cell. (b) α-induced noise currents for SOI and bulk nMOSFET's (the difference corresponds to the bipolar current). (c) SER's in SOI and bulk SRAM's as functions of effective gate length [21].

differently from the bulk SRAM as the gate length (cell) is scaled down. Implicit with the gate-length's scaling from 0.4 to 0.19 μm in Fig. 3(c) is the supply voltage's scaling from 4.0 to 2.7 V [21]. For bulk SRAM, scaling of the cell and supply voltage reduce Q_{CRIT}, and the SER rises almost exponentially. For SOI SRAM, scaling the supply voltage reduces (or suppresses) the parasitic bipolar effect since the body, which acts as the base of the parasitic bipolar, will be at lower potential. The reduced (or suppressed) parasitic bipolar effect compensates for the reduction in the Q_{CRIT} of SOI SRAM, and the SER remains relatively flat with technology scaling. This phenomenon, together with the crossover of the curves for bulk and SOI SRAM SER at around 0.20 μm, can be clearly seen in Fig. 3(c).

In SRAM applications, SOI technology offers a significant performance advantage due to device junction capacitance reduction. The benefit is most exemplified in the differential-pair bit-line topology [Fig. 3(a)], as it contains hundreds of source/drain junctions. Collective capacitance reduction of the pass-gate READ/WRITE transistors, in conjunction with the minor contribution of higher current drive of the selected pass-gate transistor induced by raised floating-body potential, results in a substantial reduction of cell access time. This is because the device junction capacitance constitutes a sizable portion of the total bit-line nodal capacitance. Detailed study using an array column of 512 cells per bit line indicated that a 34% reduction of bit-line nodal capacitance could be obtained using 0.25-μm SOI technology as compared with its bulk counterpart [22]. Since the floating-body potential of the READ/WRITE pass-gate transistors depends on the cell content ("0" or "1") and the dynamic coupling of all the internal capacitive elements both during switching and at equilibrium, the device internal C_{sb}'s of the pass-gate transistors become dependent on the array content/pattern, thus causing an imbalance in the nodal capacitance between the two bit lines. This bit-line capacitance disparity becomes more significant at higher supply voltage. Furthermore, during the WRITE operation, a disturbance of half-selected cells may occur due to excess parasitic bipolar current when the bit line is pulled down, especially at the "first cycle" [13], [14], [22]. These effects have to be duly considered in the design phase.

In most high-performance SRAM's, clocked dual-slope sense amplifiers (Fig. 4) are used [17], [23]. During sensing, the narrow device M_{NARROW} is turned on first, so the current increases slowly, allowing differential voltage across the cross-coupled pair to develop and grow. The wide device M_{WIDE} is then turned on when a cell differential voltage of about 150–200 mV is developed for fast pulling down of the bit line. Initial development of the differential voltage is critical. If the current is jammed on (transistor turned on with a large current), the differential voltage will not have a chance to grow, and the output voltage will collapse before the difference is amplified [23]. In floating-body SOI configuration, the uncertainty in the floating-body potential translates into the uncertainty in the transistor threshold voltage. A lower threshold voltage in M_{NARROW} may cause a jam-on of the sense amplifier. The imbalance in the sense transistor threshold voltages further degrades the sense margin and sensing speed. The differential voltage may collapse in its early development stage, resulting in a wrong state in the sense amplifier output. Unlike the case for the cell, it is more manageable to drop body contacts in the sense amplifier transistors. Alternatively, one can tie the bodies of the cross-coupled sense transistors together, forcing equal body potential (and thus equal threshold voltage) on the sense transistors. While this alternative approach saves the area associated with body contacts, the overall sense margin does not improve since sense transistors on different differential bit-line pairs still have different body potential because of the dependency of the bit-line capacitance on the cell contents.

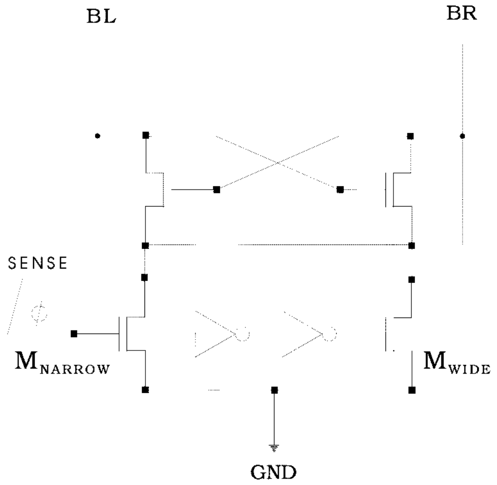
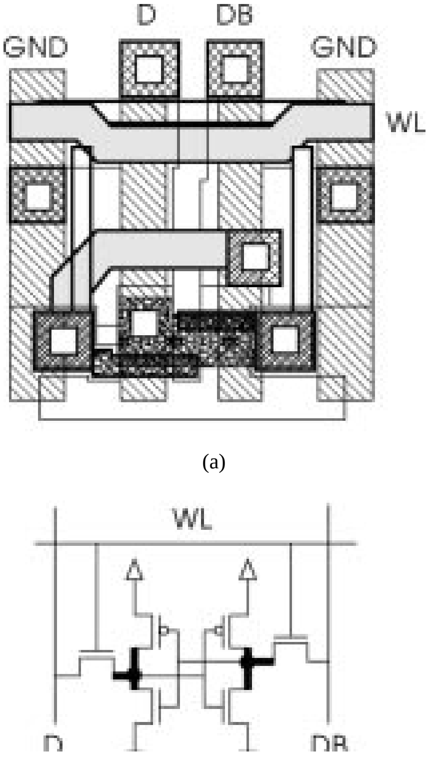


Fig. 4. Dual-slope sense amplifier used in high-performance SRAM's [17], [23].

The cell size and circuit performance of an SOI SRAM can be further improved by using a cell layout with abutted n$^+$ and p$^+$ drain regions [24] (Fig. 5). The n$^+$ and p$^+$ drain regions of the inverter output node and the source/drain region of the nMOS access transistor [shown as bold lines in Fig. 5(b)] are connected by abutting Ti-silicided n$^+$ and p$^+$ regions. This removes the layout constraint of well spacing in the bulk CMOS technology and allows a single contact layout for the cross-coupling connection of the inverters in the memory cell. In 0.35-μm design rules, a cell size reduction of 16% and bit-line capacitance reduction of 39% have been achieved compared with the bulk counterpart. A 128-Kb SRAM macro has been demonstrated with 10–20% improvement in access time over the bulk SRAM macro with comparable yield.

IV. Dynamic Random Access Memories

DRAM density is limited by the minimum cell storage capacitance achievable under the constraints of SER, static and dynamic data-retention time, and sense amplifier sensitivity. The primary advantages of SOI DRAM's are the superior SER and static data-retention time, which promise for higher integration density than bulk-Si DRAM's. This is illustrated in Fig. 6 [25], where the bit-line sense signals



(a)

(b)

Fig. 5. SOI SRAM cell layout with abutted n$^+$ and p$^+$ drain region [24].

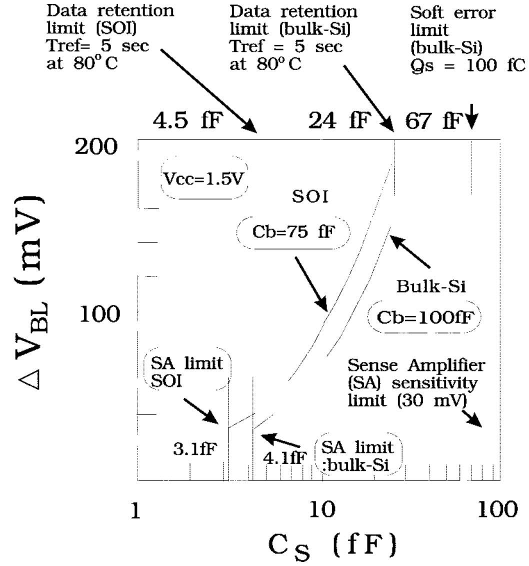
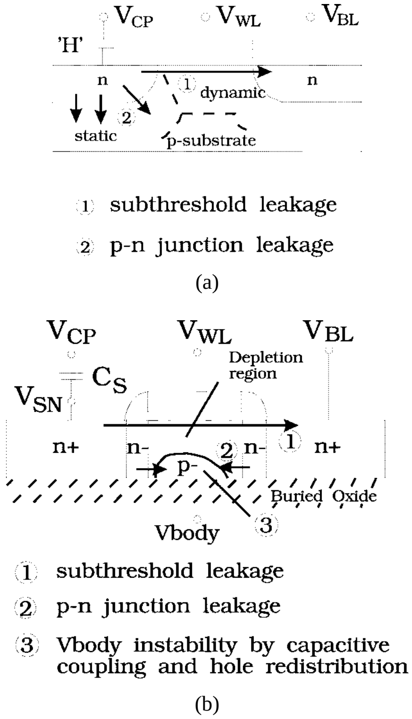


Fig. 6. Limitation of SOI-DRAM and bulk-Si DRAM for 1.5 V operation. The data-retention limit here refers to the “static” data-retention limit [25].



(a)

(b)

Fig. 7. Leakage mechanisms for (a) bulk-Si DRAM cell and (b) SOI-DRAM cell [26].

are plotted against the cell storage capacitance for both bulk-Si and SOI DRAM’s, together with limitations due to SER, static data retention, and sense amplifier sensitivity. These limitations are for densities of 256 Mb and beyond and 1.5-V operation. The bit-line capacitance C_b for the an SOI DRAM is assumed to be 25% smaller than that for a bulk-Si DRAM (75 versus 100 fF). Consider first the bulk DRAM. A sense amplifier sensitivity limit of 30 mV requires that the cell storage capacitance be larger than 4.1 fF. The static data-retention requirement that the mean retention time be larger than 5 s at 80°C places a lower bound of 24 fF on the cell storage capacitance. The SER requirement ($<$100 FIT[1]) dictates a Q_{CRIT} larger than 100 fC, and therefore a cell storage capacitance larger than 67 fF at 1.5 V. Hence, in a bulk-Si DRAM, the density is limited by SER. For an SOI DRAM, the limit due to the sense amplifier sensitivity is 3.1 fF due to the smaller bit-line capacitance. The static data-retention limit on the cell storage capacitance is 4.5 fF, substantially smaller than the 24 fF for a bulk-Si DRAM due to the order-of-magnitude smaller cell p-n junction leakage area. Last, SOI DRAM’s experimentally have been found to be essentially soft-error free. Therefore, the cell storage capacitance limit on an SOI SRAM is set by the static data-retention time requirement to 4.5 fF, which is an order of magnitude smaller than 67 fF in bulk-Si’s DRAM set by the SER requirement.

The data-retention limit just described is the static data-retention limit. It refers to the case where, during the refresh period, the bit line of the unselected cell is held steady (e.g., at 1/2 V_{CC}), and the leakage is primarily due to the p-n junction leakage (Fig. 7) [26]. Due to the significantly reduced p-n junction area, this leakage in an SOI DRAM is typically an order of magnitude smaller than that in a bulk-Si DRAM. The static data-retention time also has been experimentally verified to improve as the silicon film thickness, and hence the p-n junction area, is reduced [27].

A much more severe limitation is imposed by the dynamic data retention. In this case, the bit line of the

[1] 1 FIT = 1×10^{-9} failures per hour.

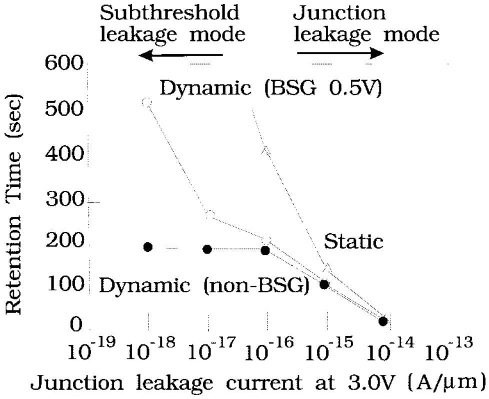


Fig. 8. Simulated data-retention time at 27°C as a function of junction leakage current [26].

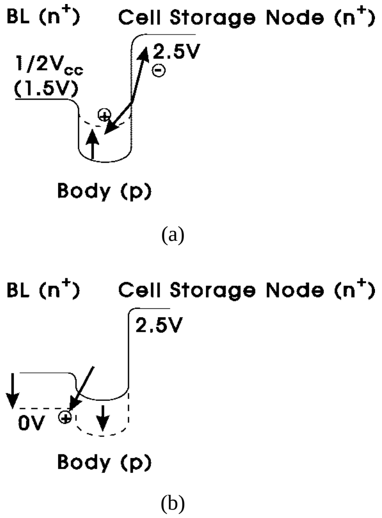


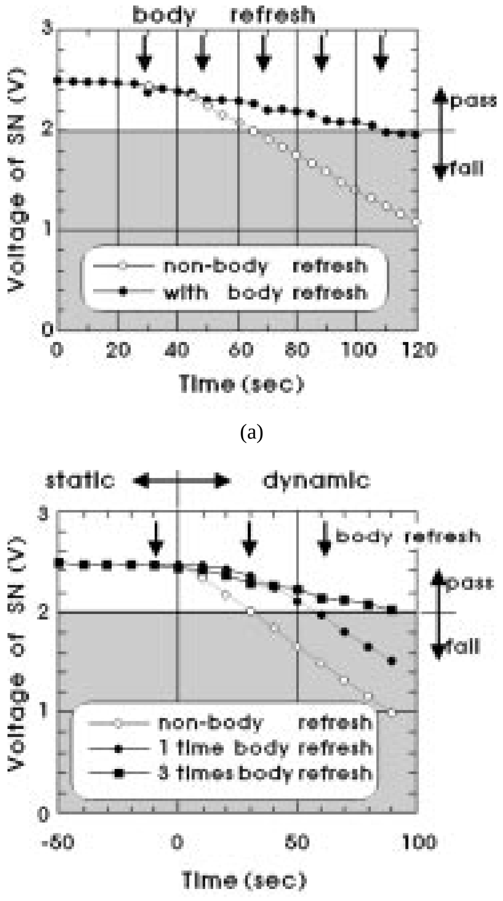
Fig. 9. Potential of SOI DRAM cell. (a) Static retention condition. (b) Body refresh condition [28].

half-selected cell swings from "high" to "low" (e.g., 1/2 $V_{\rm CC}$ to ground), thus increasing the gate-to-source voltage of the cell transistor and causing subthreshold leakage [26]. In an SOI DRAM, although the static p-n junction leakage is small, holes are injected into the floating body. The body potential inevitably rises due to the hole injection/redistribution and capacitive coupling, thus reducing the threshold voltage and degrading the subthreshold leakage. The worst case scenario is the dynamic data retention after a long static data retention, since the dynamic retention begins when the body is being charged up by the leakage current for a sustained long period [28]. Poor dynamic-retention characteristics have been experimentally observed and pose a major design challenge for SOI DRAM's [29], [30]. Since the use of body contact in the DRAM cell is not an option due to the density requirement, one has to resort to other process/device modifications and/or circuit techniques. Process/device modifications such as a lightly doped source/drain region (to reduce the current gain of the parasitic bipolar transistor and the impact ionization near the drain) [29] and the use of a pMOS cell transistor (lower impact ionization rate and current gain for the parasitic pnp transistor) [30] have been explored. One very effective circuit technique is the boosted sense-ground (BSG) scheme [31]–[33], where the "low" bit-line level is raised (to, say, 0.5 V) above the unselected word-line level to suppress the subthreshold leakage by the negative $V_{\rm GS}$ of the cell transistor. The BSG scheme has been found very effective in improving the dynamic retention time of an SOI DRAM (Fig. 8). Using the BSG scheme with $V_{\rm BSG} = 0.5$ V, a dynamic retention time of 520 s at 27°C has been shown to be achievable for an SOI DRAM, compared with 200 s for a 16-Mb bulk-Si DRAM [26].

The dynamic-retention time can be further improved by employing a "body refresh" scheme [28], where the bit-line level is lowered momentarily (from 1/2 $V_{\rm CC}$ to ground in BSG scheme) to remove the accumulated holes from the body region through the forward-biased body-source p-n junction. As shown in Fig. 9(a), in the static retention condition, the body potential increases due to the junction leakage. When the bit-line level is lowered to ground [Fig. 9(b)], the accumulated holes are removed from the body region to the bit line, and the body potential decreases. It is crucial in this scheme to make sure that the body refresh/discharge current does not destroy the stored data in the cell. Detailed two-dimensional device simulation has been performed on an SOI DRAM cell with 0.5 μm channel length, 2.5 V supply, and a body refresh period of 20 ns [26]. The decrease of the stored "high" data has been found to be only 0.04 V, thus alleviating the concern of degrading and even destroying the data. The body refresh function also has to be embedded in the normal DRAM operations and timings [26]. Fig. 10(a) shows the dynamic data-retention characteristics after "high" data writing. The body refresh after a few operations keeps the subthreshold leakage low and improves the dynamic data retention time by about two times. The dynamic retention characteristics for the worst case scenario (dynamic data retention after long static data retention) are shown in Fig. 10(b). The body refresh before the dynamic data retention suppresses any increase in the subthreshold leakage and improves the data-retention time by about two times. If the two body-refresh operations (before dynamic data retention and during dynamic data retention) are combined, an improvement factor of about three times can be obtained.

The floating-body-induced degradation in dynamic retention time can be alleviated by raising the threshold voltage to reduce the subthreshold leakage (at the expense of cell performance) [34]. One can also try to create a leaky body-source junction (to lower the current gain of the parasitic bipolar transistor) and reduce the drain-body coupling to lower the floating-body voltage [34]. This requires detailed device design and process window tradeoff, especially for the case of leaky body-source junction, where the junction leakage must remain substantially lower than the subthreshold leakage of the MOSFET.

The SOI structure is also expected to improve the cosmic-ray-induced soft error in high-density DRAM's. Cosmic



(a)

(b)

Fig. 10. (a) Estimated dynamic data retention for SOI DRAM. (b) Worst case scenario (dynamic data retention after long static data retention) [28].

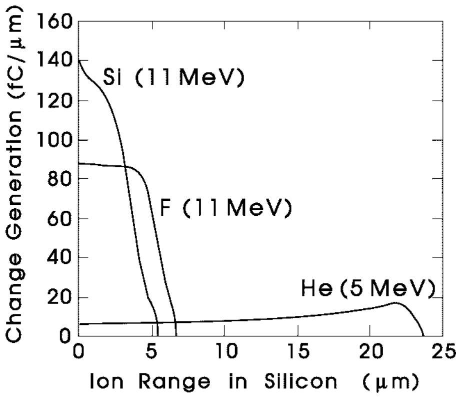
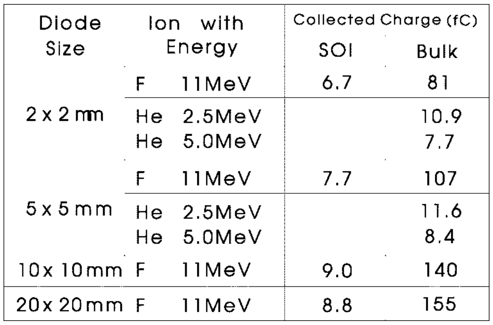


Fig. 11. Charge generation curves for ions traveling through silicon [37].



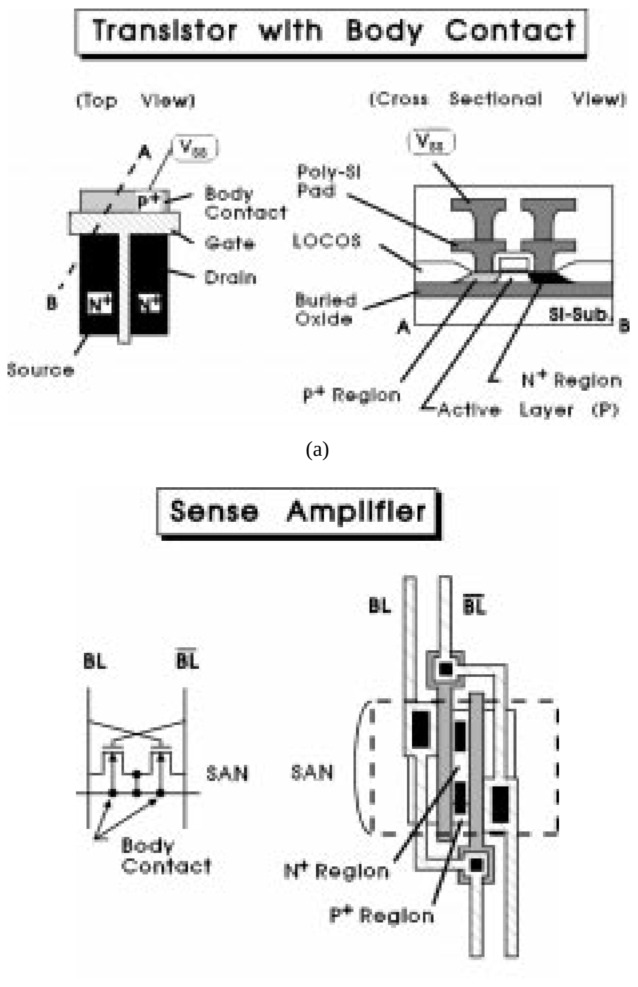
Diode Size	Ion with Energy		Collected Charge (fC)	
			SOI	Bulk
2 x 2 mm	F	11MeV	6.7	81
	He	2.5MeV		10.9
	He	5.0MeV		7.7
5 x 5 mm	F	11MeV	7.7	107
	He	2.5MeV		11.6
	He	5.0MeV		8.4
10x 10mm	F	11MeV	9.0	140
20x 20mm	F	11MeV	8.8	155

Fig. 12. Summary of charge collection results for α-particle and fluorine ion strike [37].

rays hitting the atmosphere generate neutrons. These neutrons, with a small probability, interact with silicon nuclei. The resulting events, while few in number, each have a large probability of causing an error [35]–[37]. The recoiling heavy ions generate a large number of electron-hole pairs over a short path. Due to the small neutron-nucleus cross section for silicon, energy required for the neutron source, and difficulty in controlling the neutron beam, it is hard to experimentally observe the neutron event. One approach to studying the neutron event is to directly strike devices with energetic heavy ions that produce a charge track similar to that of the nuclei (i.e., silicon ions) recoiling from neutrons. An experiment has been carried out using fluorine ion, which produces a track charge density more than half that of silicon ions and about ten times that of α-particles (He ions) near the silicon surface, as shown in Fig. 11 [37]. Collected charges are measured using diodes formed from the inner and outer diffusion of a ring transistor, with the collection nodes monitored by source-follower transistors. The results (Fig. 12) indicate that for a fluorine ion strike, the charge collected in the SOI structure is more than one order of magnitude smaller than for the bulk (7–9 fC for SOI versus over 100 fC for bulk). The difference is primarily due to the silicon film thickness, noting that the generated charge density remains relatively constant for ion range in silicon up to about 4.0 μm (see Fig. 11). The results clearly show the advantage of the SOI structure in avoiding the cosmic-ray neutron-induced soft errors.

In DRAM sensing, one typically waits for the initial (critical) bit-line signal to slowly develop to about 50–100 mV before speeding up the splitting of the bit-line voltages. Due to the stringent sensing requirements and the smaller initial bit-line signal compared with SRAM's, it is almost inevitable that body contacts be used in the sense transistors. Fitting the sense transistors with body contacts to the cell pitch is more challenging than in SRAM's due to the tighter cell pitch. Fig. 13(a) shows an SOI nMOSFET with body contact. An example of the sense amplifier layout with body contacts is shown in Fig. 13(b) [25].

The word line in DRAM is typically boosted to about 1.5 V_{DD} for storing full V_{DD} data in the cell. Circuitry dealing with boosted level generation and the word-line driver requires higher drain-to-source breakdown voltage. Consequently, body contacts are necessary for these circuitry to suppress the early breakdown due to the floating body. Body contacts are also needed in the output drivers to counter the large supply/ground bounce and avoid single transistor latchup. Fig. 14 shows a typical data path for a READ operation. Circuits where body contacts are needed/used are enclosed in dashed lines [25].



(a)

(b)

Fig. 13. (a) An SOI nMOSFET with body contact. (b) A DRAM sense amplifier with body contact [25].

Improvement in t_{RAC} by using SOI technology typically ranges 20–35%.

V. Digital CMOS Logic Circuits

We now discuss the specific design issues for digital CMOS logic circuits using SOI technology. Certain circuit topologies and switching patterns are susceptible to the parasitic bipolar effect resulting from the floating-body configuration with partially depleted SOI devices. Furthermore, because the time constants for body charging by the impact ionization current and charging/discharging by various leakage mechanisms range from several nanoseconds to several tens of nanoseconds [11], the body potential during the switching transient is primarily determined by the external biasing and capacitive coupling. The circuit behavior thus depends upon the prior state (hysteresis) and switching patterns. A thorough understanding of the complex interactions among the device behavior, circuit topologies, and switching patterns is necessary to allow proper design/sizing of various circuits and selective use of body contact to achieve the full potential of the SOI technology.

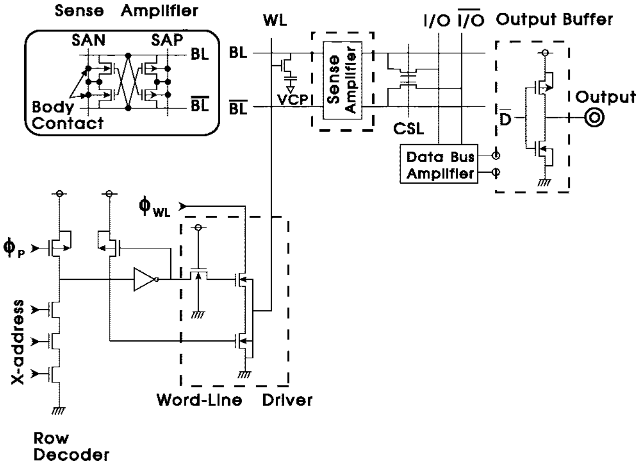


Fig. 14. DRAM data path for a READ operation. Circuits where body contacts are needed are enclosed in dashed lines [25].

One circuit topology susceptible to the parasitic bipolar effect is the basic pass-gate configuration discussed in Section II. Another case is illustrated by the stacked OR-AND static CMOS circuit in Fig. 15(a) [12], [15]. Consider the situation in which the input to N1 is at "high" (V_{DD}) and the inputs to N2, N3, and N4 are all at "low" (ground) at $t = 0$ [Fig. 15(b)]. The output node at $t = 0$ is at V_{DD} because the input to P4 is "low." Node 1, the common source node of N1/N2/N3, sits at a voltage one V_T below the input to N1. The body voltages of N1/N2/N3 sit between their drain voltages and their source voltages and hence are at "high" as well. When the input to N1 switches from "high" to "low" (at $t = 0.6$ ns), the common source node (Node 1) is capacitively coupled down slightly by the gate-to-source capacitance. The body voltage of N1 ($V_{\mathrm{B.N1}}$) is capacitively coupled down significantly by the large gate-to-body capacitance of N1. The body voltages of N2 and N3, on the other hand, are only down slightly because their respective gate voltages remain unchanged and the voltage at the common-source node (Node 1) is down only slightly. Hence, when the input to N4 subsequently switches (at $t = 1.1$ ns) to pull the common-source node (Node 1) to ground, large base-emitter voltages (i.e., body-source voltages) are developed for N2 and N3 (not N1, since the body voltage of N1 has been capacitively coupled down significantly), and significant parasitic bipolar currents flow through the supposedly off devices N2 and N3 [Fig. 15(b)].

For the static CMOS circuit, the pMOS path restores and holds the output by construct. If the circuit has been properly sized, the net effect is only a very small dip in the output voltage wave form and the extra power consumption due to the parasitic bipolar current.

For dynamic circuits [38], the consequence can be much more severe. Fig. 16(a) shows a dynamic four-way OR circuit [12]. Notice that the stack formed by the logic transistors and the evaluation transistor resembles the OR-AND stack for the static circuit in Fig. 15(a). Assume that in the precharge phase, the input to N1 is at "high" and the inputs to N2/N3/N4 are at "low." The dynamic Node 2 is at V_{DD}, and the common-source Node 1 is at $V_{\mathrm{DD}} - V_T$. The

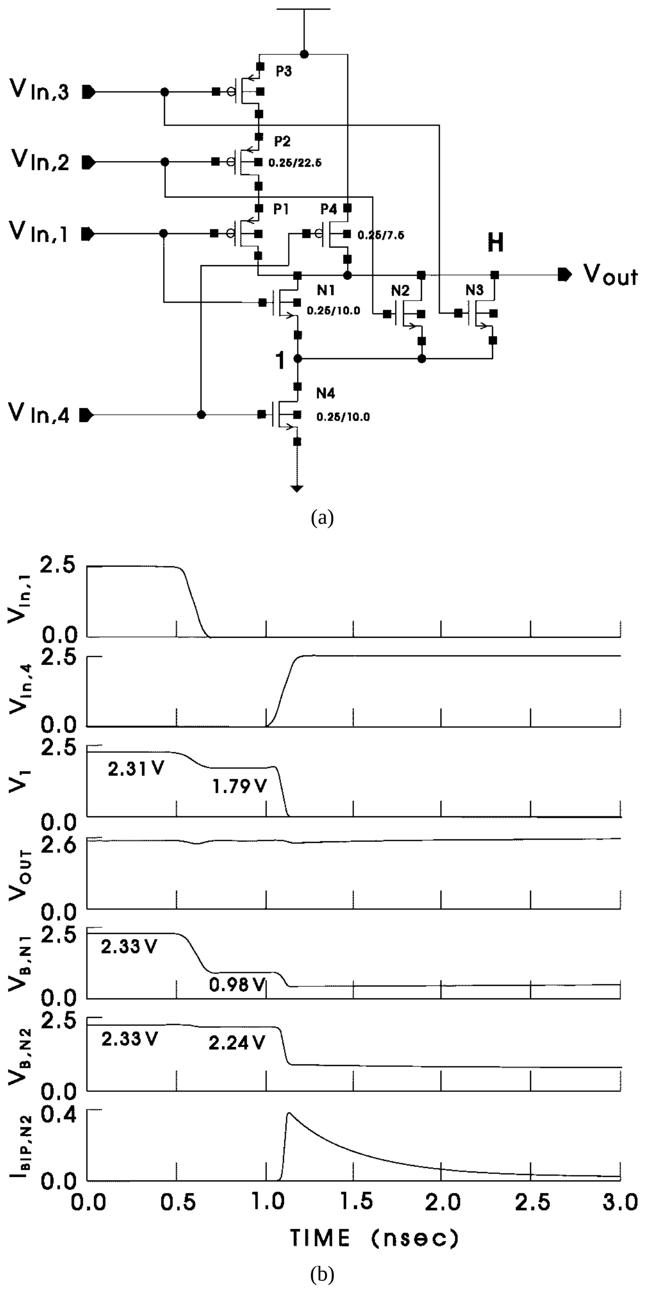


Fig. 15. (a) Static three-way OR-AND circuit. (b) Pertinent switching wave forms (voltage unit: volts; current unit: milliamperes) [12], [15].

input to N1 switches at $t = 0.6$ ns from "high" to "low," and the circuit subsequently evaluates at $t = 1.1$ ns [Fig. 16(b)]. These switching patterns set up N2/N3/N4 in a condition similar to that just described for the static circuit, and large parasitic bipolar currents flow through these off devices to pull down the dynamic Node 2 when the circuit evaluates. Depending on the strength of the feedback half-latch P1, the parasitic bipolar currents may produce a disturbance at Node 2 and the output node or completely upset and invert the logic state [Fig. 16(b)]. Hence, in the dynamic circuits, the feedback half-latch has to be sized up (at the expense of circuit speed) to overcome this effect in the worst case situation.

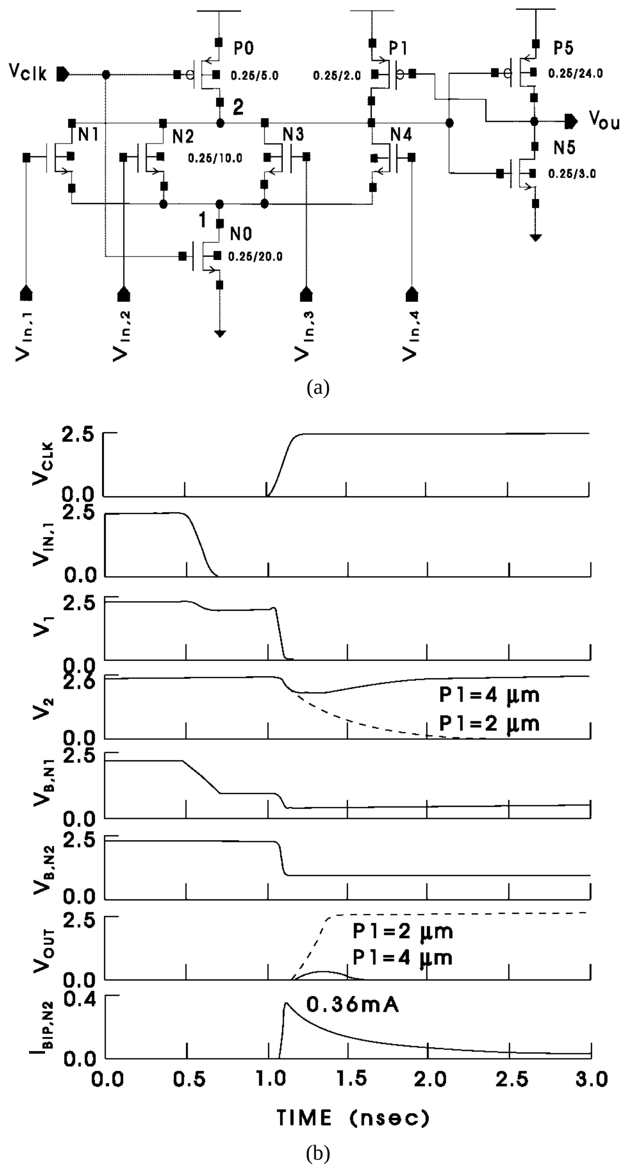
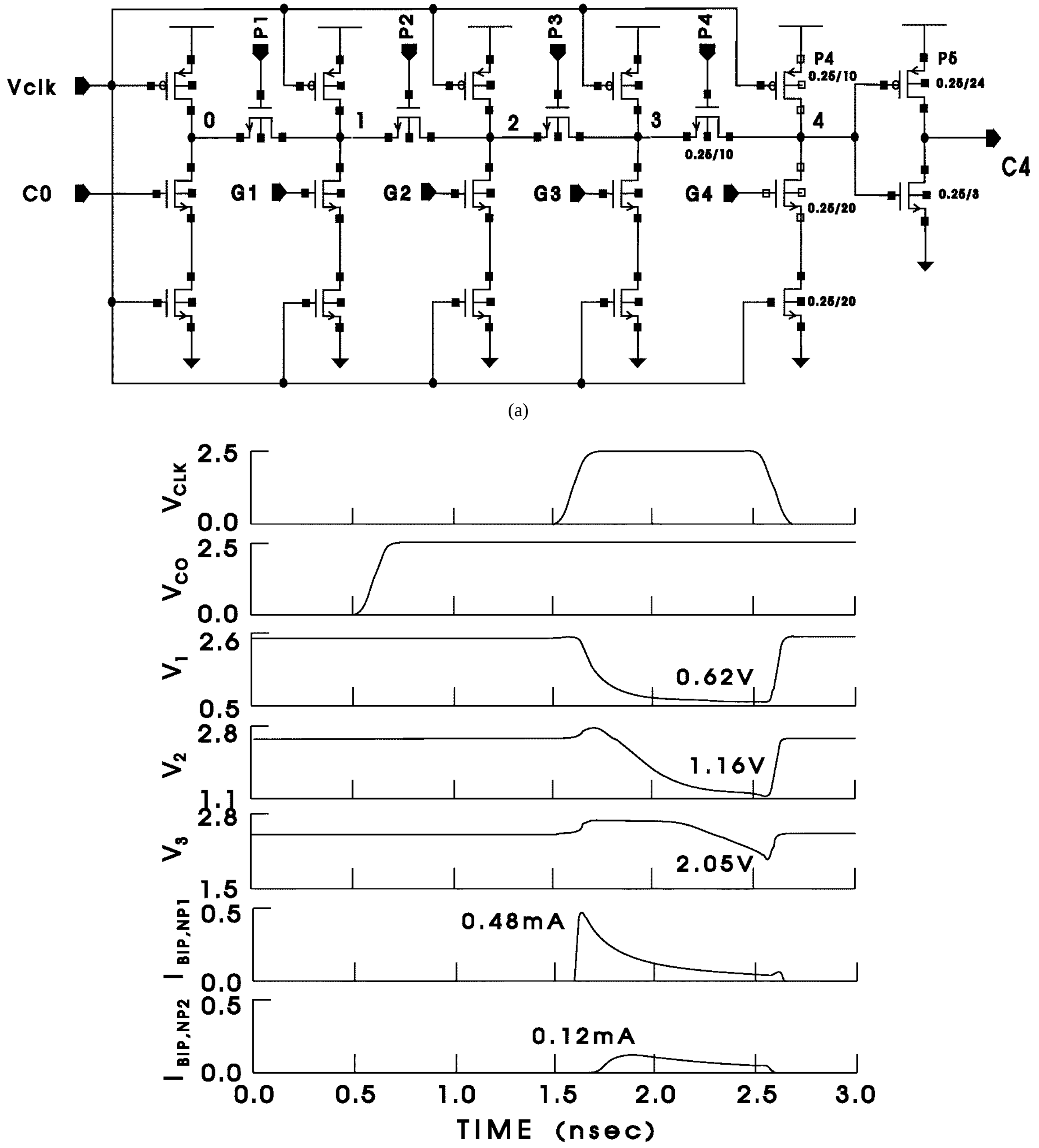


Fig. 16. (a) Dynamic four-way OR circuit. (b) Pertinent switching wave forms (voltage unit: volts; current unit: milliamperes) [12], [15].

Similar effects are present in other circuit families such as the static and dynamic cascade voltage switch logic (CVSL) circuits [12].

Some commonly used circuit building blocks for fast arithmetic operations in processor data flow can also experience problem with the parasitic bipolar effect. Fig. 17(a) shows a Manchester carry chain circuit [38] for fast carrier propagation/generation, where the propagate signal (P_i) is used to gate the previous $\overline{carry}$ ($\overline{C}_{i-1}$) in a pass-gate configuration. Assume that in the precharge phase (G_4, G_3, G_2, G_1) = (0, 0, 0, 0), (P_4, P_3, P_2, P_1) = (0, 0, 0, 0), and $C_0 = 1$. Node 0 ($\overline{C}_0$), Node 1 ($\overline{C}_1$), Node 2 ($\overline{C}_2$), Node 3 ($\overline{C}_3$), and Node 4 ($\overline{C}_4$) are all precharged to "high" ($V_{\rm DD}$). Hence, all the pass-gate transistors N_{P1}, N_{P2}, N_{P3}, and N_{P4} are set up with their gate inputs at "low" and their drain nodes and source nodes at "high" ($V_{\rm DD}$). When the circuit evaluates (at



(a)

(b)

Fig. 17. (a) Manchester carry chain circuit and (b) switching wave forms for input patterns $(G_4, G_3, G_2, G_1) = (0, 0, 0, 0)$, $(P_4, P_3, P_2, P_1) = (0, 0, 0, 0)$, and $C_0 = 1$ (voltage unit: volts; current: milliamperes) [12], [15].

$t = 1.6$ ns), Node 0 is pulled down ($C_0 = 1$), and parasitic bipolar current flows through the off pass-gate transistor N_{P1} to pull down Node 1 (Fig. 17). As a result, parasitic bipolar current flows through the off pass-gate N_{P2} to pull down Node 2. This chain parasitic bipolar effect, stemming from the series-connected pass-gate configuration, fades as it propagates down the pass-gate chain. Consequently, no significant parasitic bipolar current can be observed in N_{P3} and N_{P4}. Node 1 can be seen to be pulled down to 0.62 V and Node 2 pulled down to 1.16 V, both low enough to cause errors in their logic states if they are buffered by inverters for use in the subsequent logic. Node 3, because of the fading chain parasitic bipolar effect, is pulled down only to 2.05 V. Many other input patterns will also cause parasitic bipolar effect and result in logic state errors for this circuit [12].

Multilevel voltage-switch current-steering-type circuits may encounter much more complicated situations [39]. Fig. 18 shows the schematics of a three-input dynamic CVSL XOR circuit. In the precharge phase, all branches are "nonactive" because the clocked evaluation transistor N0 is off. Node 1 and Node 2 are precharged to V_{DD}. Due to the differential input configuration, all common-source nodes in all cascade levels are at "high." Consequently, all logic transistors with "low" inputs (i.e., half of the total logic transistors) are set up in a condition with both their drains and sources at "high" (thus, bodies at "high" as well if given enough time to settle to their steady-state values). When the circuit evaluates, the common-source nodes in the active branches are pulled down to "low" and parasitic bipolar currents flow through these off transistors. Because of the criss-cross drain connection, each common-source node is also a common-drain node of the lower cascade level. Pulling down these common-drain nodes can cause "inverse-mode" parasitic bipolar current to flow from the source to the drain in the off transistors in the nonactive branches at the lower cascade level [39]. This is illustrated in Fig. 19, which shows the pertinent switching wave forms for input pattern $(A, B, C) = (0, 1, 1)$. In this case, transistors with inputs connected to A (N1), $\overline{B}$ (N5/N6), and $\overline{C}$ (N9/N10) are susceptible to the parasitic bipolar effect in the evaluation phase. As can be seen, normal-mode parasitic bipolar currents flow through the off transistors N1, N6, and N10 when the circuit evaluates at $t = 1.5$ ns. Inverse-mode parasitic bipolar currents flow through the off transistors N5 and N9 because their drain nodes are pulled down by cascade active transistors and therefore come down much faster and earlier than their source nodes, which are pulled down by the parasitic bipolar currents at the lower cascade levels. The normal-mode parasitic bipolar current can also been seen to pull down the supposedly high Node 2, resulting in an erroneous state where both Node 1 and Node 2 are at "low." The parasitic bipolar currents in the second evaluation cycle (at $t = 4.0$ ns) are substantially smaller than those in the first evaluation cycle. This hysteresis behavior stems from the fact that the time interval for the second precharge cycle (from $t = 2.5$ to $t = 4.0$ ns) is not long enough for the body voltages to charge up to their steady-state values.

Up to now, we have focused our discussion on the parasitic bipolar effect. An equally important effect is the floating-body-induced transient threshold voltage variation in partially depleted SOI-CMOS devices [40]. This effect is present even when the parasitic bipolar current is not significant enough to affect circuit operation. The threshold voltage variation has been shown to cause a frequency-dependent pulse-stretching effect in partially depleted SOI-CMOS inverter chains [40]. Such behavior is attributed to the charge imbalance between logic states during switching. Consider the situation in Fig. 20, where the equilibrium distribution of holes is shown schematically for "low" [Fig. 20(a)] and "high" [Fig. 20(b)] input gate voltage. With gate input at "low," there are more majority charges (holes) in the floating body compared with the case when the gate input is at "high." This is because a larger volume of majority charge is depleted by the gate at "high" voltage than by the drain at "high" voltage. Hence, if the input to the inverter chain is at "low" for a long time prior to any switching, the first-stage nMOS (gate input at "low") has more majority charges in the floating body compared with the second stage nMOS (gate input at "high"). When the chain switches, the first-stage nMOS switches with a higher current drive (lower V_T) than the second-stage nMOS. The same scenario holds for all the subsequent odd- and even-stage nMOS devices. From a circuit point of view, all the odd-stage nMOS devices have their gate inputs at "low" and their drain nodes at "high." The bodies of these devices are charged by the off-state impact ionization current and the leakage through the reverse-biased drain-body PN junctions, and discharged by the leakage through the forward-biased body-source PN junctions. Equilibrium body voltage is reached when the charging and discharging currents are equal. Thus, the body voltages of these devices sit at one diode (body-source PN junction diode) cut-in voltage above the source nodes, resulting in lower threshold voltage. For the even-stage nMOS devices, with gate inputs at "high" and drain nodes at "low," the body voltages (sitting between the drain and source voltages) are at "low," resulting in higher threshold voltage. The pMOS devices operate in a complementary fashion. The odd-stage pMOS devices have less majority body charges (electrons) and thus higher V_T, while the even-stage pMOS devices have more majority body charges and lower V_T. Consequently, when the input pulse rises from a sustained period of "low" state to "high," the rising edge of the input pulse propagates down the chain faster than the falling edge of the pulse. This is because all the devices involved with the propagation of the rising edge (odd-stage nMOS and even-stage pMOS) have lower V_T, and all involved with the falling edges have higher V_T. The pulse, therefore, stretches as it propagates through the chain, as demonstrated by the measurement results shown in Fig. 20(c). This pulse-stretching effect depends on the input frequency and the supply voltage $V_{\rm DD}$ (Fig. 21). As the input pulse frequency increases, the pulse stretching decreases. This is because as devices switch more often, there is less time for devices to recover to

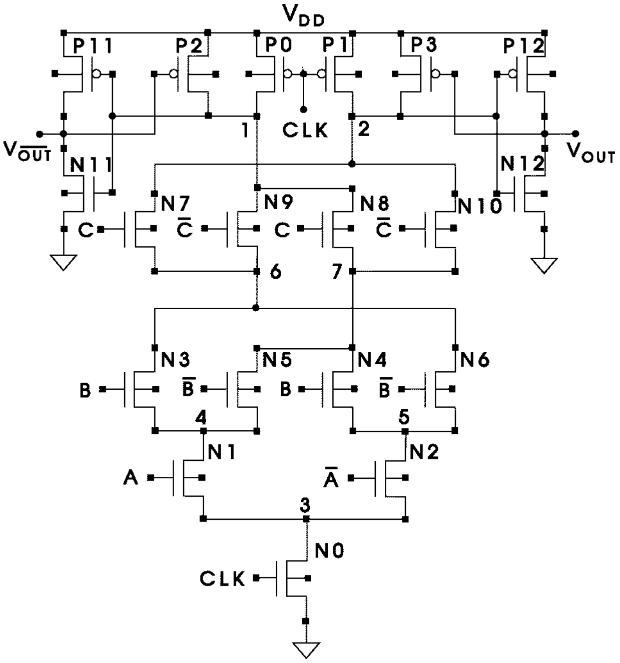
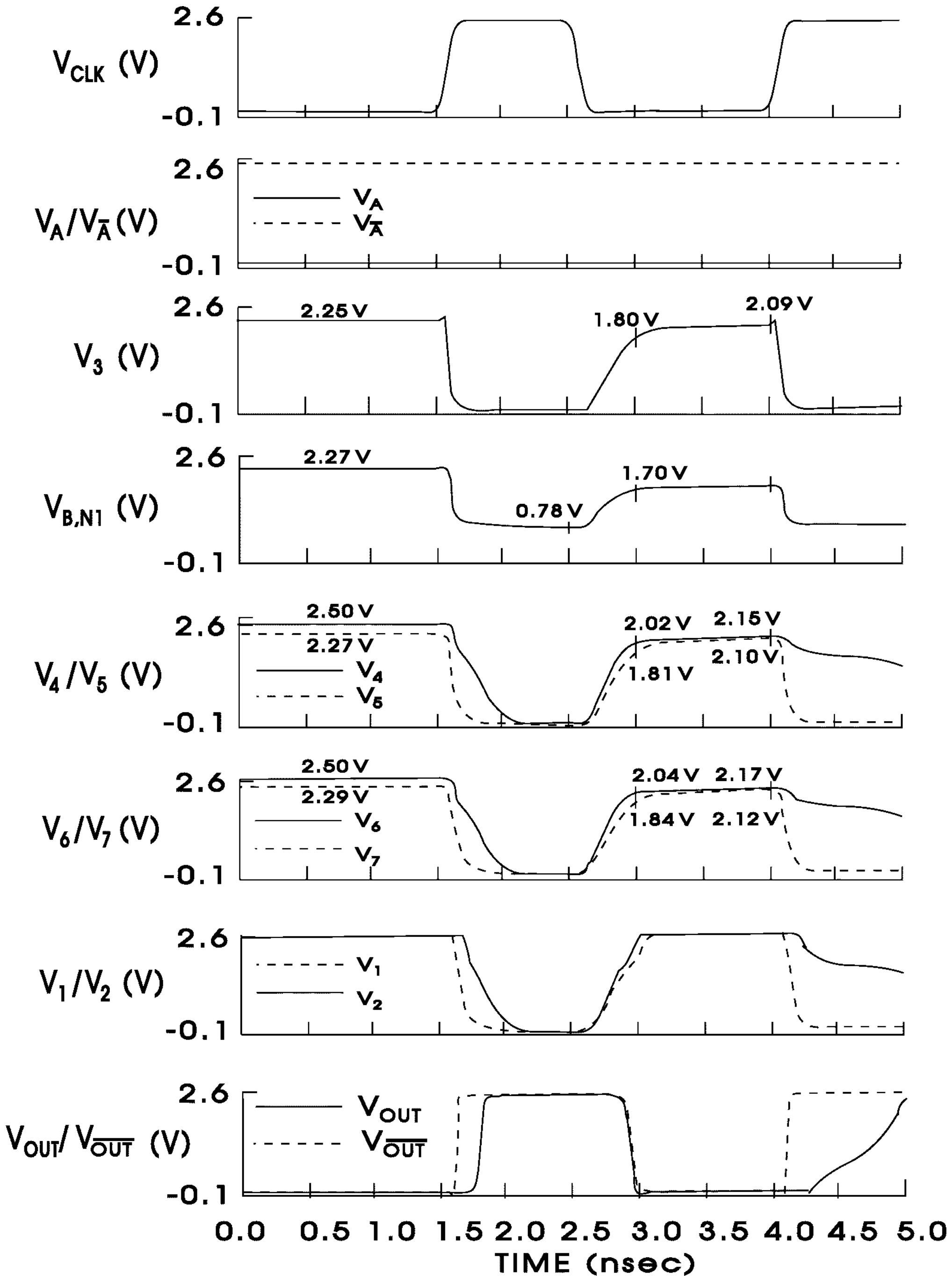


Fig. 18. Schematics of three-input dynamic CVSL XOR circuit [39].

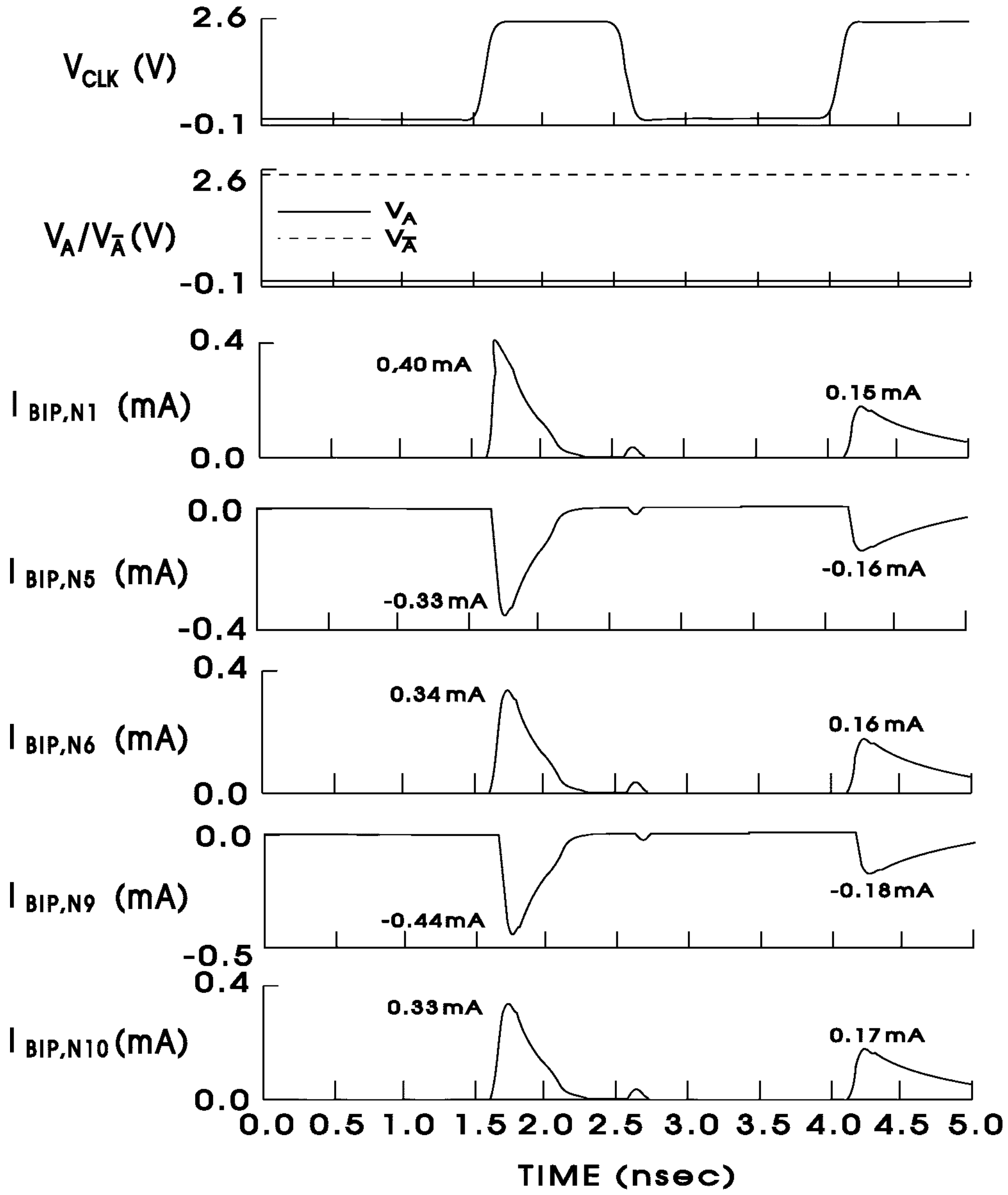


(a)

Fig. 19. Pertinent (a) voltage wave forms for dynamic CVSL XOR circuit with inputs (A, B, C) $= (0, 1, 1)$ [39].

the starting (equilibrium) state that they began with prior to any switching. The pulse stretching also decreases with increasing $V_{\rm DD}$. At higher $V_{\rm DD}$, the drain-induced depletion of the body at output "high" state [gate input at "low"; see Fig. 20(a)], which reduces the equilibrium number of holes in the body, becomes more significant. Thus, the difference of equilibrium number of holes between the two states shown in Fig. 20(a) and (b) diminishes (and pulse stretching decreases) with increasing $V_{\rm DD}$.

The pulse-stretching effect has several implications on the circuit operation. First, it affects the duty cycle and degrades the clock skew and jitter in a clock distribution network. Second, circuit timing rules would have to take this effect into consideration, thus complicating the timing methodology and degrading the circuit performance as well. Last, it complicates the design and degrades the margin of the self-timed type of circuits [41]. This is illustrated in Fig. 22, where we show the schematics of a so-called "self-resetting CMOS" (SRCMOS) circuit [17], [41]. This type of self-timed circuit utilizes a delayed feedback signal derived from the circuit output to reset the circuit to precharge state, thus eliminating the skew associated with the distribution



(b)

Fig. 19. *(Continued.)* Pertinent (b) current wave forms for dynamic CVSL XOR circuit with inputs (A, B, C) = $(0, 1, 1)$ [39].

of the global clock and improving the cycle time. The input and output signals are pulses. When the input pulses arrive and the circuit evaluates, the dynamic Node A is pulled down and the output rises. The rising edge of the output signal is delayed via a "reset timing chain" and then applied to the precharge (reset) transistor P0, thus initializing the precharge process [Fig. 22(b)]. When Node A is precharged "high," the output falls. The falling edge of the output signal then goes through the delay chain to turn off the precharge transistor P0. The pulse-stretching effect broadens the precharge pulse $V_{\rm PCH}$ since the rising edge of the output pulse propagates through the reset timing chain faster than the falling edge, thus squeezing the evaluation cycle and degrading the cycle time. Furthermore, in this type of circuit, input pulses must align properly to ensure enough overlap among pulses. One also must ensure proper separations among signal pulses that are not supposed to overlap so that they will not "collide" in the worst case. The "pulse alignment" and "avoidance of pulse collision" become much more complicated in the presence of the pulse-stretching effect.

VI. Pass-Transistor-Based Circuits

Pass-transistor logic has been known for its efficiency in device use. The lower transistor count required to implement a given function improves the density, power, and delay. It has long been a popular circuit choice for fast arithmetic operations such as arithmetic logic unit (ALU), multiplier, and processor data-flow elements such as multiplexer, barrier shifter, etc. [38], [42]–[44]. Implementation of this low-power circuit style in the low-power SOI technology would potentially result in substantial power

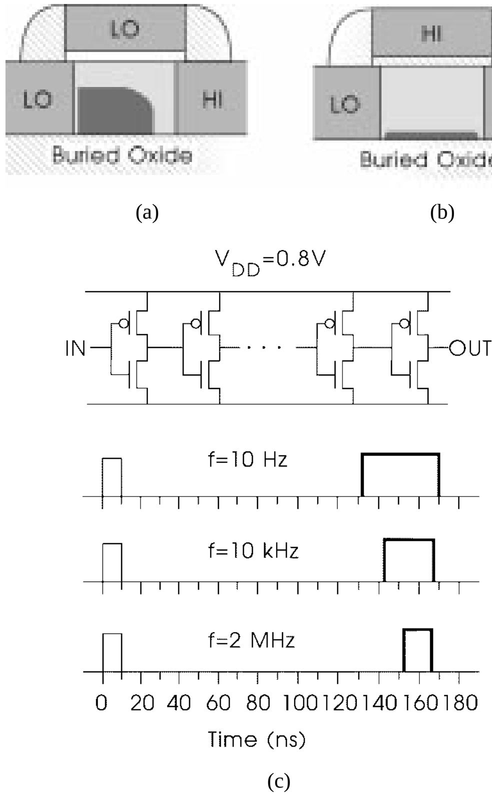


Fig. 20. Cross section of a partially depleted SOI nMOSFET schematically showing the equilibrium distribution of holes (dark-shaded region) (a) in the output-HI state and (b) in the output-LO state. (c) Measured variation in output pulse width of a 480-stage PD-SOI-CMOS inverter chain at $V_{\rm DD} = 0.8$ V versus input frequency shown schematically. The input pulse width is 10 ns. The devices have $L_{\rm eff} = 0.45$ μm, $t_{\rm OX} = 9$ nm, and $t_{\rm Si} = 110$ nm [40].

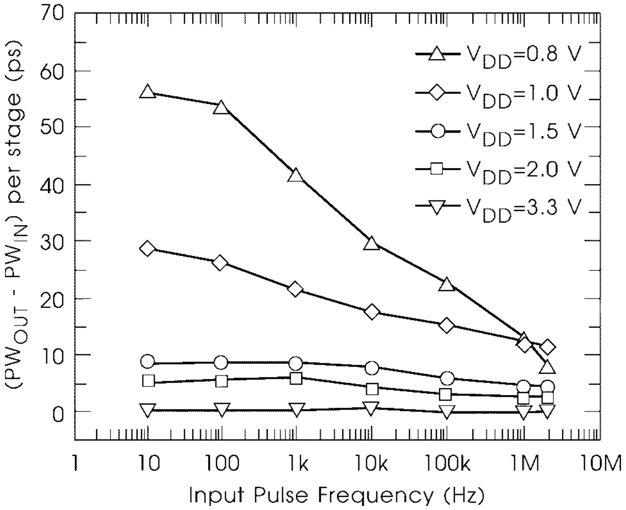


Fig. 21. Measured pulse stretch per stage in a 480-stage PD-SOI-CMOS inverter chain versus input pulse frequency and $V_{\rm DD}$ [40].

reduction for low-power applications. This circuit style, however, is particularly vulnerable to the parasitic bipolar effect resulting from the floating body, as discussed in Section II.

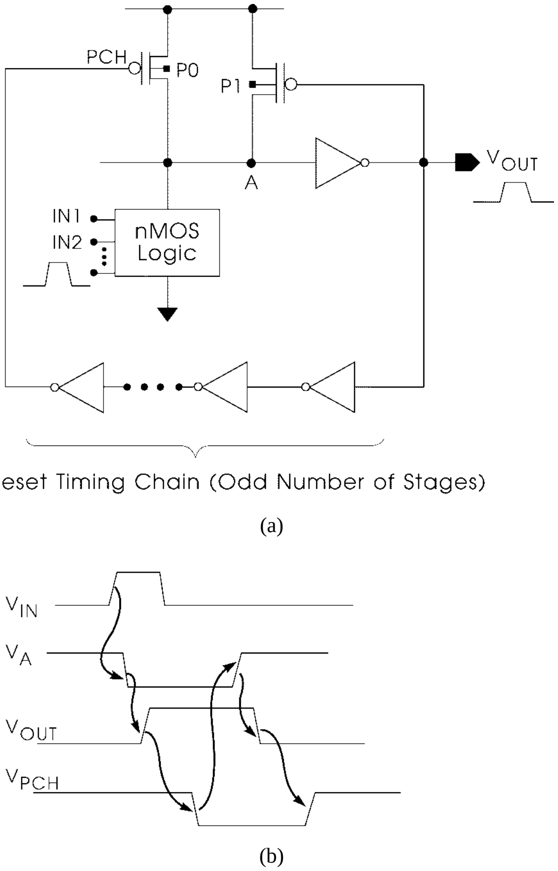
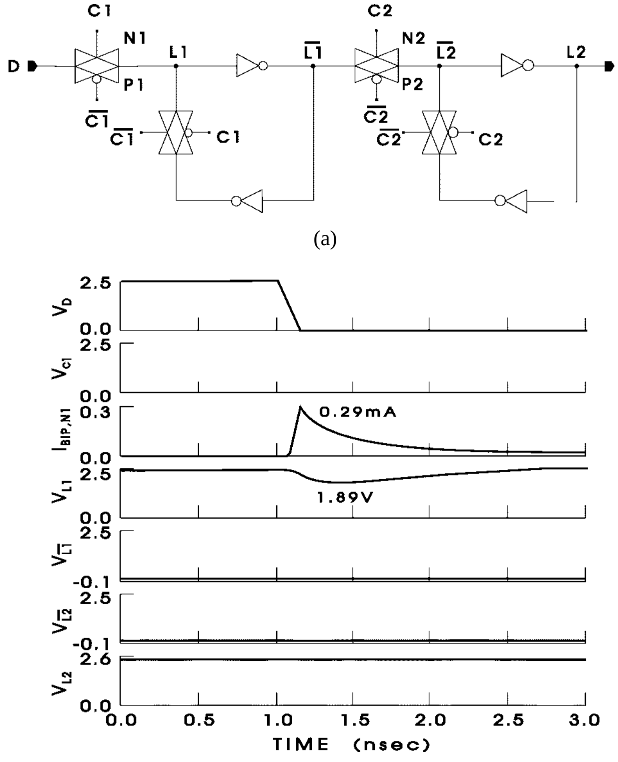


Fig. 22. (a) Schematics of a generic SRCMOS circuit. P0 is the precharge (reset) transistor, P1 is the feedback half-latch to improve the noise margin of dynamic Node A. (b) Relationship between timing edges of various signals in SRCMOS circuit [41].

One of the most common and important applications of the pass gate is for the clock/timing control in various latch designs [38]. Fig. 23 depicts an L1/L2 type of latch with two nonoverlapping clocks, C1 and C2. Consider the situation that after passing the “high” state, C1 switches to “low” at $t = 0$ ns, and the input D is subsequently pulled down at $t = 1.0$ ns, as shown in Fig. 23(b). With C1 at “low,” the pass gate at the input to the L1 latch is supposed to be off, while the pass gate in the feedback loop of the L1 latch is on to hold the state of the latch. Significant parasitic bipolar current, however, flows through the off nMOS of the input pass gate [Fig. 23(b)] to pull down Node L1. Since the pass gate in the feedback loop is on, the pMOS in the feedback inverter fights the parasitic bipolar current to restore Node L1. The result is a transient voltage dip (large enough to be a design concern) of Node L1 voltage, as shown in Fig. 23(b) [15]. The complementary situation is less of a concern because of the low impact ionization of pMOS and the lower current gain of the parasitic pnp transistor.

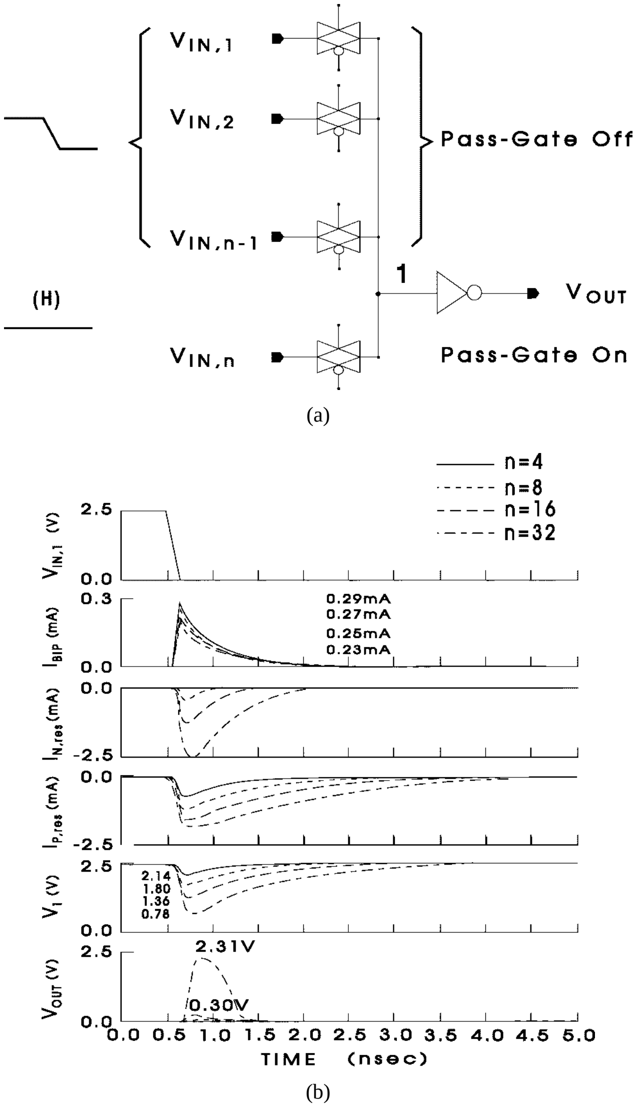
Notice that in some high-density designs, the pass gate in the feedback loop is removed and a “trickle” inverter with small β devices is used [38]. It is important in this case to make sure that the trickle inverter has enough strength to overcome the parasitic bipolar current and restore Node L1; otherwise, the latch may flip and latch into the wrong state.



(a)

(b)

Fig. 23. (a) L1/L2 latch with two nonoverlapping clocks, C1 and C2. (b) Pertinent switching wave forms for parasitic bipolar current through the nMOS of the L1-latch input pass gate. [12], [15].



(a)

(b)

Fig. 24. An n-to-1 multiplexer. (a) Circuit schematic and (b) pertinent switching wave forms for $n = 4, 8, 16$, and 32 [12], [15].

Pass-transistor-based wide multiplexers are important for critical data-flow elements such as rotators and shifters and for the control portion of a processor. The schematics of a pass-gate-based n-to-1 multiplexer are shown in Fig. 24(a). In most applications, the control signals are "orthogonal," selecting one and only one input at a time. Consider the worst case scenario for the parasitic bipolar effect as follows. Assume all inputs are at "high" to start with, the selected input passes the "high" state to Node 1 (and continues to hold the state of Node 1 afterwards), and all the unselected inputs ($n-1$ of them) are then pulled down. As a result, parasitic bipolar currents flow through the $n-1$ nMOS in the $n-1$ unselected pass gates to pull down Node 1, which is being held/restored only by the single selected pass gate. Fig. 24(b) shows the pertinent switching wave forms for $n = 4$, 8, 16, and 32 [15]. For $n = 16$, Node 1 is pulled down to 1.36 V, close to the threshold of the output buffer (inverter), and a "bump" starts to surface in the output voltage wave form $V_{\rm OUT}$. For $n = 32$, Node 1 is pulled down to 0.78 V, decisively crossing the threshold of the output buffer, and the output voltage rises to 2.31 V. Since the parasitic bipolar current is a transient phenomenon, the selected pass gate eventually restores the Node 1 (and hence output) voltage. If the output is sampled and latched into the subsequent logic stages when it is "high," however, a wrong logic state would result.

The parasitic bipolar effect can also potentially lead to logic-state errors in pseudo-two-phase dynamic logic [38], as shown in Fig. 25(a). Consider the time period when C2 is "low" (the other case when C1 is "low" is symmetrical). The pass gate to the second stage is off and the second stage is evaluating. If the data stored in the gate capacitance is "high" and the first stage is evaluated to be "low" following the falling edge of C1, parasitic bipolar current flows through the nMOS in the off pass gate, discharging the input (gate) node (Node INT3) from "high" to "low." In Fig. 25(b) [15], this nMOS parasitic bipolar current (peak current = 0.27 mA) can be seen to discharge the gate node completely from 2.5 to 0 V.

VII. Smart Body Contact

The fact that each SOI device has a fully isolated, individually accessible body actually offers an additional degree of freedom for design and can be exploited to enhance the power and delay performance. The body contact can be used to control dynamically the threshold voltage of the device to achieve high-speed, low-voltage operation

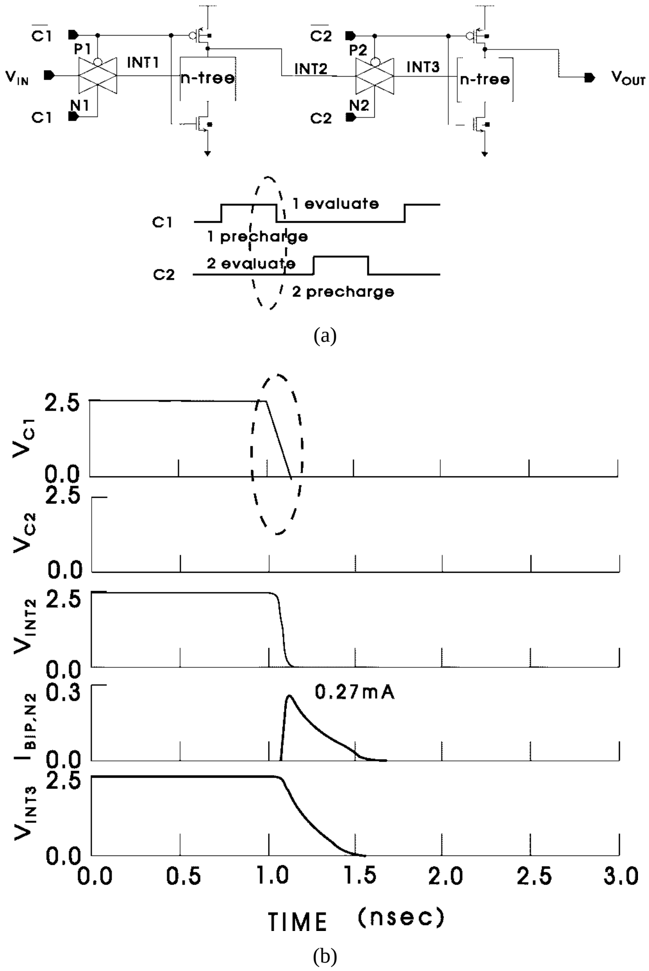


Fig. 25. A pseudo-two-phase dynamic logic circuit. (a) Circuit schematic and (b) pertinent switching wave forms for parasitic bipolar current through nMOS N2 (voltage unit: volts; current unit: milliamperes) [12], [15].

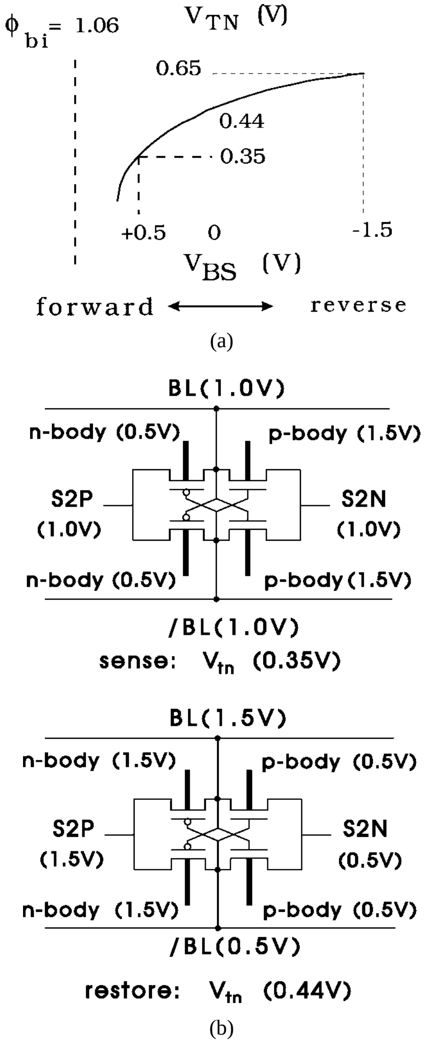


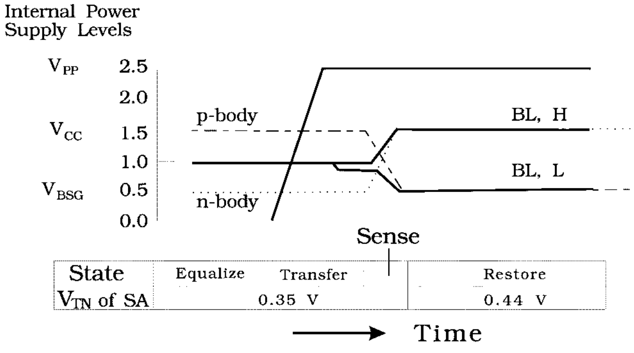
Fig. 26. Super body-synchronous sensing scheme for DRAM. (a) $V_{\rm BS}$ controls $V_{\rm TN}$ and (b) pertinent bit-line and body-voltage levels for sense transistors in sense and restore states [46].

in active mode and low-leakage, low-power operation in standby or sleep mode.

One example is the dynamic threshold voltage MOSFET (DTMOS) [45], where the body is directly tied to the gate. The threshold voltage is thus reduced, and the current drive improves when the circuit is active. In standby mode, the body potential is low and the threshold voltage remains high to reduce the leakage. This scheme has been shown to provide significant leverage for low-voltage operation. The power-supply voltage, however, is limited to less than one diode voltage to avoid turning on the parasitic bipolar transistor. A circuit limiter to limit the body-source junction voltage to less than one diode drop is necessary for higher supply voltages.

Dynamic threshold voltage control has also been proposed for high-density, low-voltage DRAM applications. Figs. 26 and 27 illustrate the so-called super body-synchronous sensing scheme for a 4-Gb 1.5-V SOI DRAM application [46]. In this circuit, the word line is boosted to $V_{\rm PP} = 2.5$ V, and a BSG level [31], [32] $V_{\rm BSG} = 0.5$ V is applied. In the equalizing state, the p- and n-bodies of the sense amplifier are biased at 1.5 and 0.5 V, respectively [Fig. 26(b)], resulting in a body-to-source bias of 0.5 V and hence a low threshold voltage of $V_{\rm TN} = 0.35$ V [Fig. 26(a)] for the nMOS sense transistor for fast sensing. After sensing, the p-body is pulled down to 0.5 V. Since in the restore state the low level of the bit line is at $V_{\rm BSG} = 0.5$ V, the body-to-source voltage for the nMOS sense transistor is 0 V; thus, $V_{\rm TN}$ is increased to 0.44 V to reduce the standby current induced by the subthreshold leakage. This sensing scheme has been found to improve the sensing speed by 2.7 ns. Overall, a $t_{\rm RAC}$ improvement of about 35% over bulk-Si DRAM can be achieved [Fig. 27(a)]. Notice that dynamic threshold voltage control through body bias is practical only for SOI DRAM. Due to large well-to-substrate and other parasitic capacitances in bulk-Si DRAM, the current required for body-bias control is about 20 times larger than that in SOI DRAM [Fig. 27(b)] [46].

Notice that the dynamic threshold voltage-control techniques typically utilize partially depleted (PD) SOI devices. Nevertheless, fully depleted (FD) devices have significantly lower leakage current in the off state. It is possible to design/optimize the SOI film thickness and channel doping so that the FD/PD transition occurs in a small range of



(a)

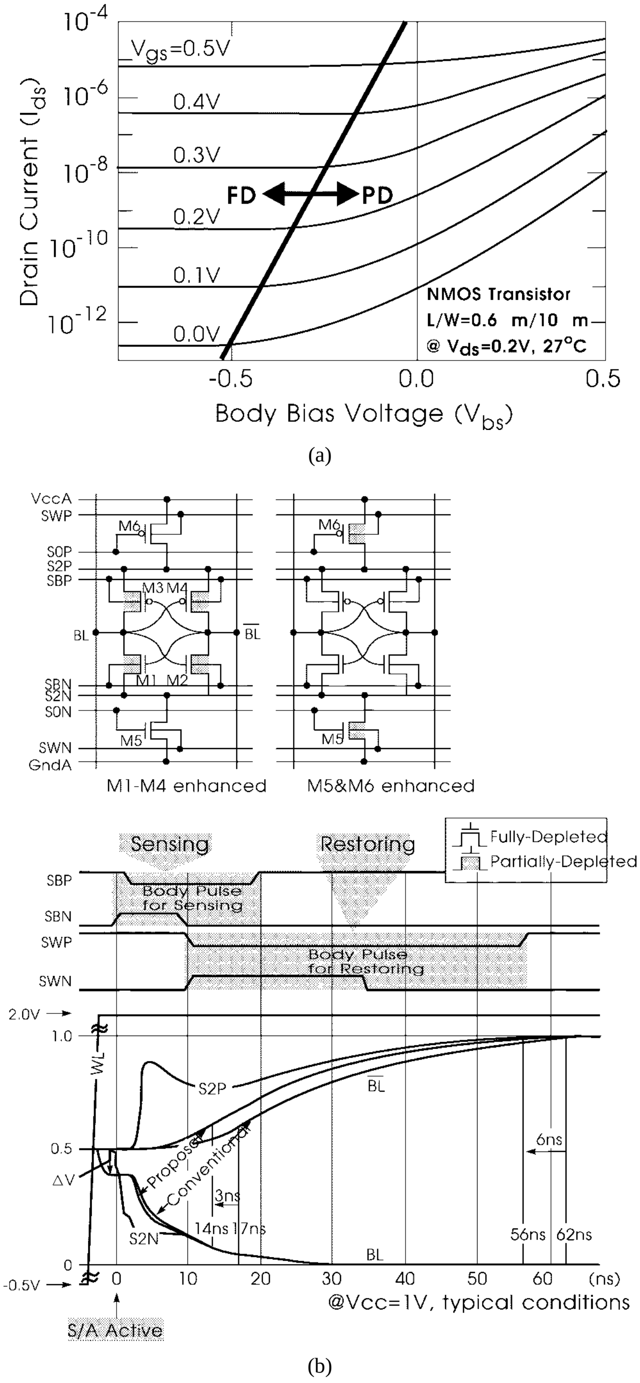
	Current for body-bias change (peripheral circuit) (ratio)	tRAC (ratio)
Bulk	31mA (1)	35ns (1)
SOI	1.7mA (0.05)	23ns (0.65)

(b)

Fig. 27. Super body-synchronous sensing scheme for DRAM. (a) Timing sequence and (b) comparison of current required for body-bias control and t_{RAC} for bulk-Si and SOI DRAM [46].

body-to-source bias V_{bs}. This FD/PD transition by the body-bias control is particularly useful for a low-voltage circuit because the on-state current is enhanced by the forward body bias in the PD mode, while the off-state current is suppressed by the better subthreshold slope and lower leakage in the FD mode. This technique was first demonstrated in a 1.0-V, 46-ns, 16-Mb DRAM [47]. Fig. 28(a) illustrates the FD/PD mode transition controlled by the body bias. To enhance the speed at low voltage, the design employs a body-pulsed sense amplifier (BPS), as shown in Fig. 28(b). Four body voltages (SBP, SBN, SWP, SWN) are independently controlled to accelerate both sensing and restoring. In the initial sensing, the driving capability of transistors M1–M4 is enhanced by asserting SBP and SBN pulses. As a result, the sensing time is reduced from 17 to 14 ns. During restoring, the driving capability of M5 and M6 is enhanced by SWP and SWN pulses; thus, the time for restoring to full swing is shortened from 62 to 56 ns. A similar technique is also applied to improve the bit-line equalization. This body-driven equalizer (BDEQ) scheme is depicted in Fig. 29. This scheme overcomes the drawback of the conventional bit-line equalizing circuit, where the driving capability of equalizing transistors decreases as equalizing proceeds owing to the back bias effect. In the BDEQ scheme, bias of 0.5 V is applied to the body, and V_{bs} remains positive through the equalizing process [the "EQBODY" maintains 0.5 V as long as "BLEQ" (bit-line equalization) is asserted]. The circuit is thus free from the back bias effect, and equalizing time is reduced from 15 to 9 ns.

Dynamic threshold voltage control has also been used in multithreshold CMOS (MTCMOS) circuit [48] combining low V_T-CMOS logic gates and variable-V_T sleep-mode-control MOSFET's for sub-1.0-V battery-operated portable applications. Fig. 30(a) shows a MTCMOS circuit on SIMOX, which operates for supply voltage as low as 0.5 V [49]. The circuit utilizes fully depleted low-V_T MOSFET's for the logic gates to achieve high speed at low supply voltage and a partially depleted variable-V_T pMOS for sleep-mode control. In the active mode, the body voltage of the sleep-mode-control pMOS is lowered, thus reducing its V_T to allow a voltage closer to the supply voltage to be applied to the logic circuits. In the sleep mode, the body voltage of the pMOS is raised, thus increasing its



(a)

(b)

Fig. 28. (a) Body-controlled FD/PD mode transition and (b) BPS. (0.5-μm SIMOX with modified MESA isolation.) Film thickness for t_{ox}/SOI/BOX = 10/100/400 nm [47].

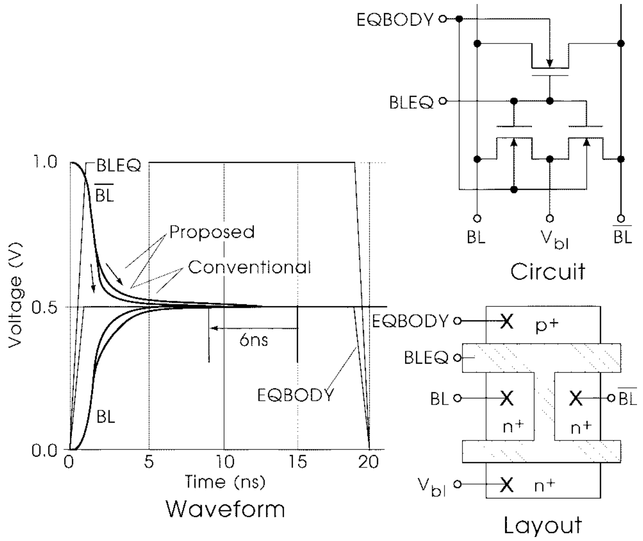
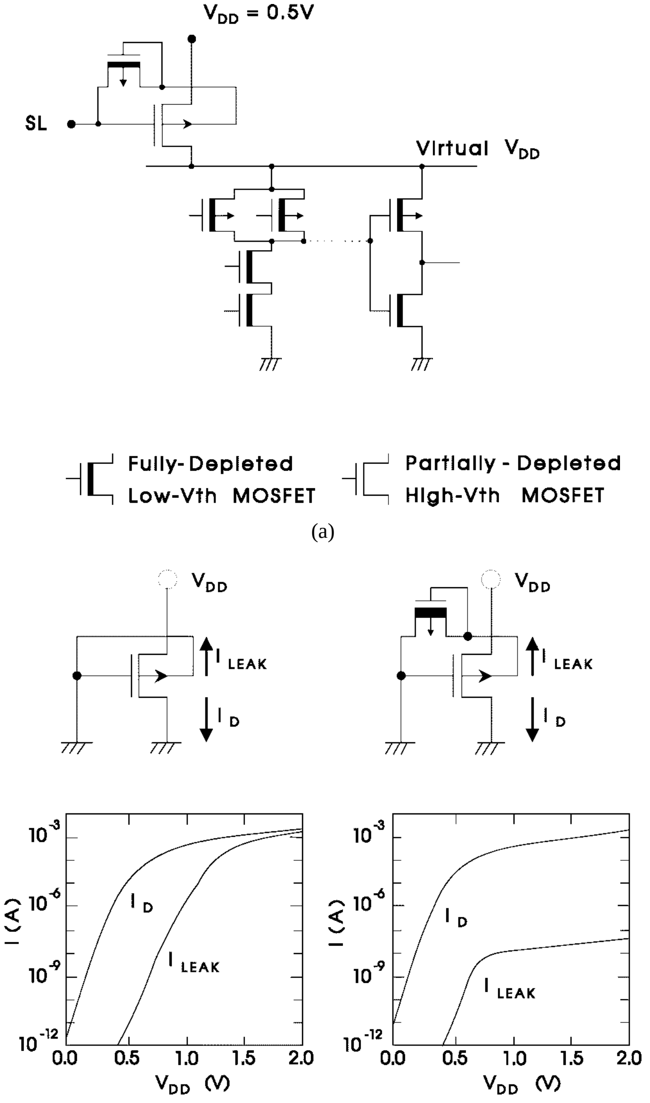


Fig. 29. BDEQ [47].

V_T to reduce the leakage. Tying the pMOS body directly to its gate would result in large leakage from the drain-body pn junction to the gate when the sleep-mode control signal SL goes "low" (active mode) if the supply voltage is larger than one diode drop. By connecting the pMOS body to its gate through a reverse-biased diode composed of a low-V_T pMOSFET, this leakage is significantly suppressed [Fig. 30(b)]. Notice that when SL is "high" (sleep mode), the body of the sleep-mode-control pMOS will sit at one (low) diode drop below SL. Fig. 31 compares the circuit delays of bulk-MTCMOS and SIMOX-MTCMOS as functions of the supply voltage. The SIMOX-MTCMOS offers more than two times delay improvement for supply voltages less than 1.0 V. The SIMOX-MTCMOS can also been seen to operate for supply voltages down to 0.4 V, while bulk-MTCMOS loses its functionality around 0.5 V.

Similar techniques have also been applied to pass-transistor logic [50]. Fig. 32(a) depicts a conventional complementary pass-transistor logic (CPL). Low-voltage applications of this circuit are constrained by the V_T loss in passing the "high" signal and degradation in driving capability. By using gate-body connected SOI pass gates [Fig 32(b)], low-threshold voltage for the "on" pass-gate and high-threshold voltage for the "off" pass-gate are realized, thus minimizing the V_T loss, improving the driving capability, and suppressing the leakage. Body-bias control can also be applied to the buffer section composed of a pMOS latch and two CMOS inverters, as shown in Fig. 32(b) [50]. The bodies of the inverter MOSFET's are connected to their respective gates. Due to the low-threshold voltage of the "on" MOSFET's, this scheme reduces the delay of a full adder to one-third that of the conventional SOI CPL at a supply voltage of 0.5 V (Fig. 33). At a fixed delay of 2.0 ns, the lowest operation voltage is improved by 0.17 V. The buffer section can be further improved as shown in Fig. 34, where the pullup pMOSFET's are cross coupled. The bodies of the pullup pMOSFET's can be connected either to their respective gates or to the inputs, as shown in Fig. 34. With the bodies connected to the gates and the gates cross coupled to the output nodes of the opposite phases, the threshold voltage of the "on" pullup pMOS remains high until the output nodes respond. It is therefore advantageous to connect the body to the input so that the threshold voltage of the "on" pMOS decreases immediately once the input changes, thus improving the circuit speed and reducing the short-circuit current. Experimentally, it has been verified to provide about 36% improvement in circuit speed [50].



(a)

(b)

Fig. 30. (a) SIMOX-MTCMOS circuit and (b) static current characteristics of DTCMOS and variable high V_T-MOSFET with reverse-biased MOS diode [49].

Notice that in practice, it is desirable to minimize the number of supply voltages and other circuits in the system may require supply voltage higher than one diode voltage. To extend the operating voltage of the body-bias controlled SOI pass-gate circuits to larger than one diode voltage, one can exploit the boosted ground scheme as shown in Fig. 35(a) [51]. In this scheme, a reference voltage generator, composed of a drain-body-connected nMOS SOI

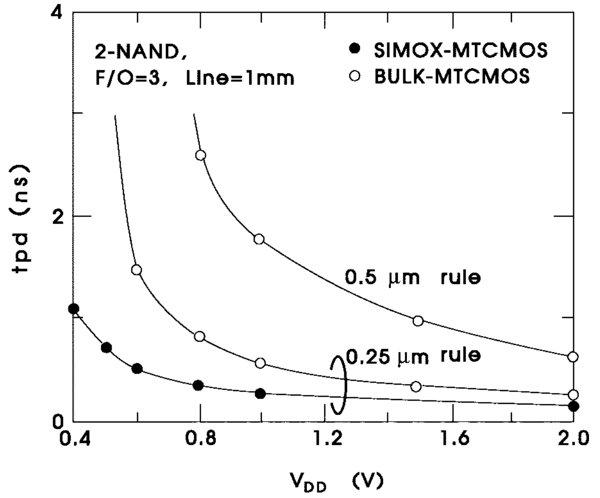
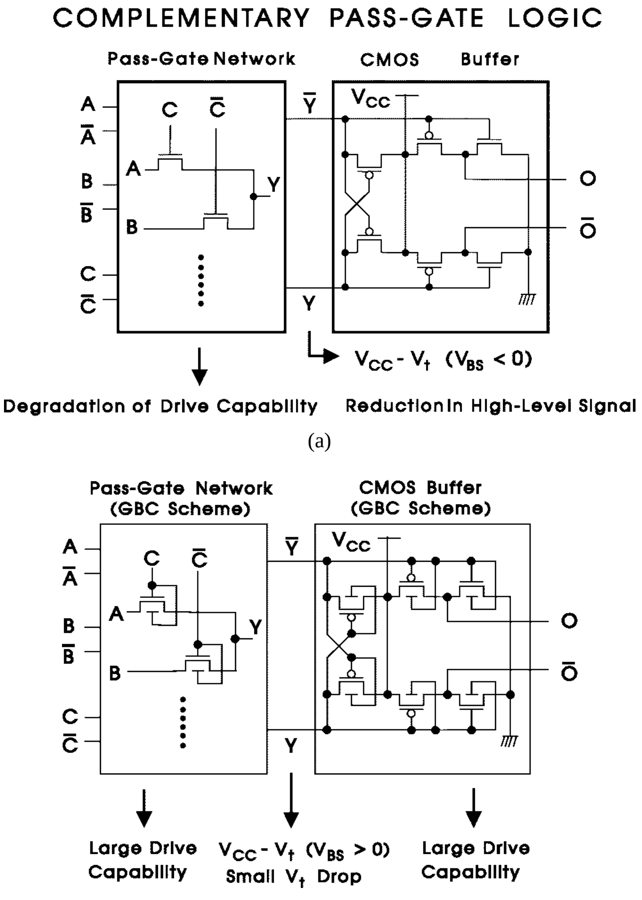


Fig. 31. Comparison of basic logic gate delays of bulk-MTCMOS and SIMOX-MTCMOS. The threshold voltages are 0.38 V for high V_T nMOS, 0.13 V for low V_T nMOS, 0.44 V for high V_T pMOS, and 0.18 V for low V_T pMOS [49].


COMPLEMENTARY PASS-GATE LOGIC
Pass-Gate Network
CMOS Buffer
A
Ā
B
B̄
C
C̄
C
C̄
A
B
Y
Ȳ
Y
Vcc
O
Ō
Vcc - Vt (VBS < 0)
Degradation of Drive Capability
Reduction in High-Level Signal
Pass-Gate Network (GBC Scheme)
CMOS Buffer (GBC Scheme)
Large Drive Capability
Vcc - Vt (VBS > 0) Small Vt Drop
Large Drive Capability

(a)

(b)

Fig. 32. (a) Conventional complementary pass-gate logic. (b) Gate-body connected SOI pass-gate logic [50].

transistor and a series resistor, generates V_{ref}. The bias applied to the main circuitry, $V_{\mathrm{dd}} - V_{\mathrm{ref}}$, is therefore lower than the soft breakdown voltage of the body-bias controlled SOI devices, thus reducing the leakage. The boosted ground is driven by a body-source tied SOI nMOS transistor to avoid the floating-body effect and to assure high enough V_t to suppress the standby leakage current of the main circuitry. A charge pump boosts the supply voltage for stable operation of the analog portion of the circuits. Fig. 35(b) shows the maximum frequency and active power as functions of the supply voltage for a one-stage, 32-b ALU using the boosted ground scheme [51]. The operating voltage can be seen to extend to 1.5 V while containing the power. Notice that the maximum frequency remains relatively constant for supply voltage above 0.5 V since the highest voltage applied across the main circuitry is limited to 0.5 V in this implementation.

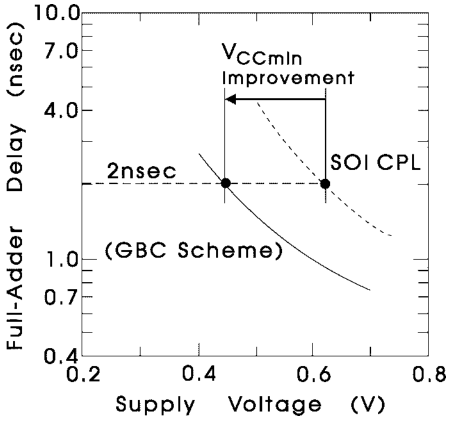


Fig. 33. Simulated full-adder delay of conventional complementary pass-gate logic and gate-body connected pass-gate logic on SOI [50].

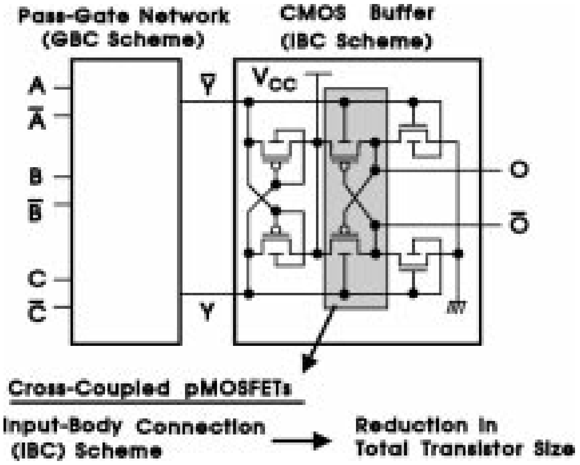

Pass-Gate Network (GBC Scheme)
CMOS Buffer (IBC Scheme)
A
Ā
B
B̄
C
C̄
Ȳ
Y
Vcc
O
Ō
Cross-Coupled pMOSFETs
Input-Body Connection (IBC) Scheme
Reduction in Total Transistor Size

Fig. 34. SOI-CMOS buffer with cross-coupled pullup pMOSFET's and body-input connections [50].

A capacitive coupling technique has also been used to maintain the advantage of body-bias control and extend the operating voltage range. Fig. 36(a) shows the schematic of a so-called double-gate-driven CMOS (DGMOS) circuit where the body of the switching device is dynamically connected to the gate by a capacitor [52]. Taking the pMOS as an example [Fig. 36(a)], a coupling capacitor C_b and a pn junction diode D_1 are added. The operation of the circuit is illustrated in Fig. 36(b). When the input switches from V_{dd} to ground (GND), the pMOS is turned on and its body

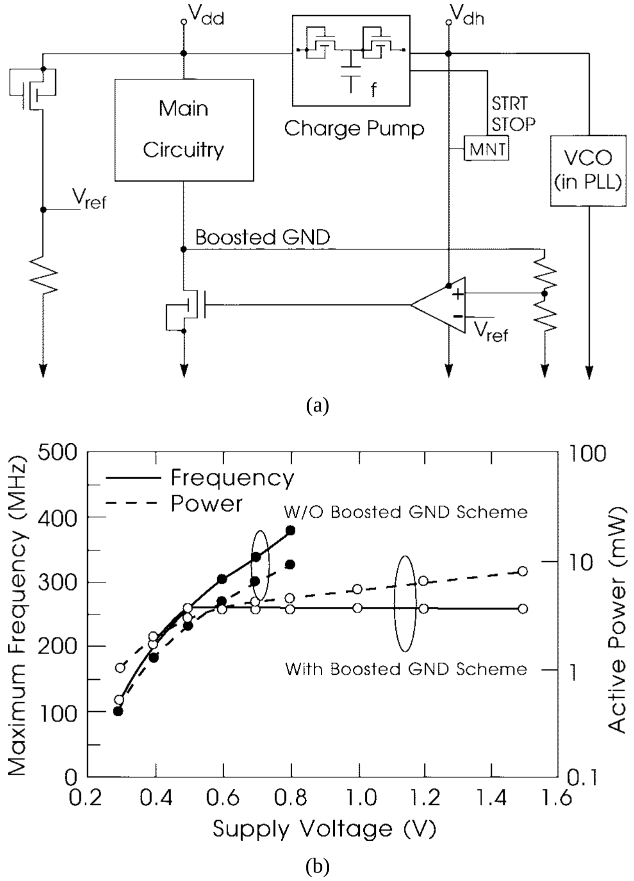


Fig. 35. (a) Body-bias controlled SOI pass-gate logic with boosted ground scheme and (b) maximum operating frequency and active power versus supply voltage for a one-stage, 32-b ALU using body-bias controlled SOI pass-gate logic [51].

voltage $V_{\rm bP}$ is capacitively coupled down, thus reducing the threshold voltage and resulting in a large drain current. When $V_{\rm bP}$ drops down to a value larger than a diode drop from $V_{\rm dd}$, the pMOS source-body pn junction will turn on to charge the body and C_b. Hence, $V_{\rm bP}$ eventually settles at $V_{\rm dd} - V_{\rm cut-in, P}$, where $V_{\rm cut-in, P}$ is the cut-in voltage of the source-body diode. Similarly, when the input switches from GND to $V_{\rm dd}$, the pMOS is turned off. The pMOS body voltage $V_{\rm bP}$ rises from $V_{\rm dd} - V_{\rm cut-in, P}$ to $V_{\rm dd} - V_{\rm cut-in, P} + V_{\rm dd}$ and eventually drops down to $V_{\rm dd} + V_{\rm cut-in, D1}$ (where $V_{\rm cut-in, D_1}$ is the cut-in voltage of the diode D_1) through the discharge diode D_1. This circuit can operate at any supply voltage compatible with the device design without suffering the leakage in other body-bias controlled schemes. Fig. 36(c) compares the inverter delay and leakage of a conventional CMOS circuit and DGCMOS circuit as functions of the coupling capacitor C_b. As a large C_b increases $V_{\rm bP}$ modulation, the full effect is achieved and the leakage is reduced significantly. A large C_b increases the input capacitance, however, resulting in a larger load on the previous stage. The value of C_b therefore represents a design tradeoff between the circuit speed and leakage current.

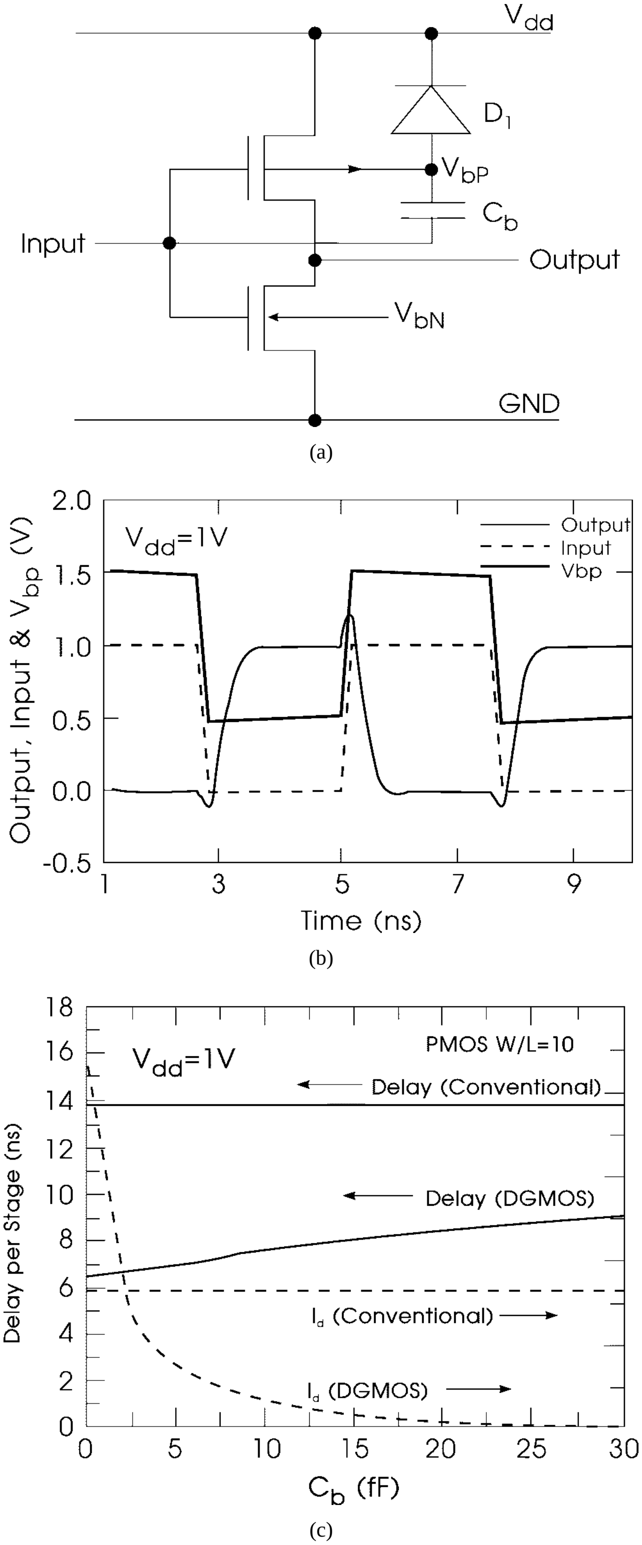
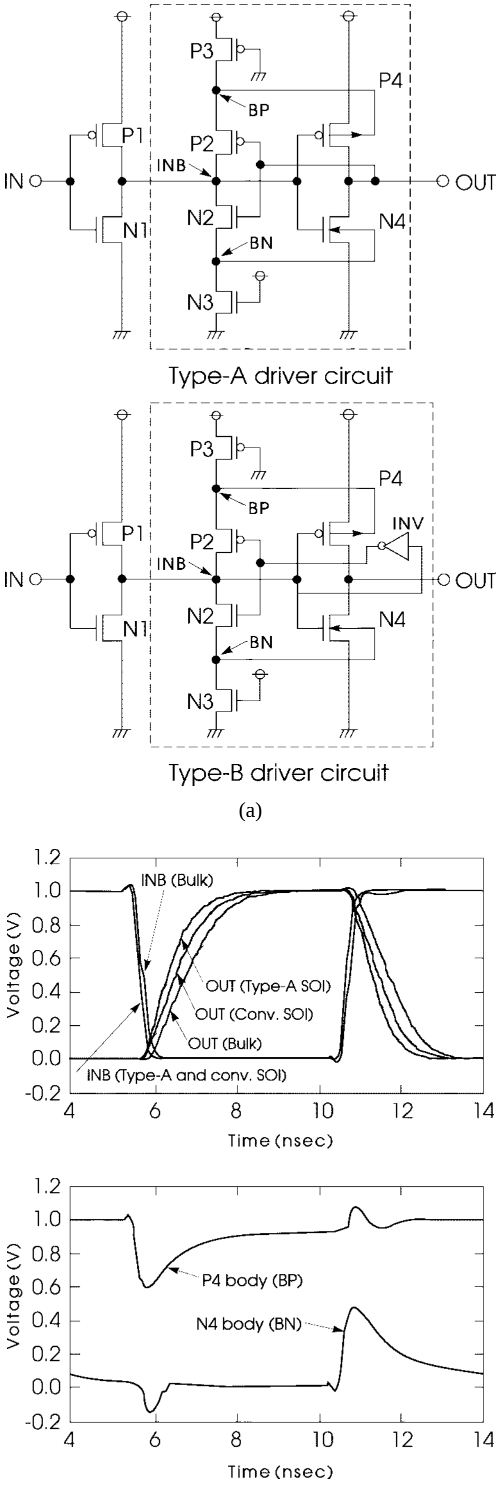


Fig. 36. (a) Double-gate-driven pMOS in an inverter. (b) DGMOS inverter input, output, and node $V_{\rm bP}$ wave forms. (c) DGMOS ring oscillator delay and leakage current [52].

One problem of connecting the body to the gate through a reverse-biased diode [Fig. 30(b)] is the high-frequency operation. While the scheme allows the circuit to operate at supply voltage higher than a diode drop, the device operation becomes unstable at high frequency because the excessive charges in the body can not be discharged quickly as the diodes are reverse biased. An active body-bias scheme has been proposed to overcome this high-frequency limitation [53]. Fig. 37(a) shows two active body-biased driver circuits. The Type-A driver consists of six transistors (enclosed in the broken-line box), where the branch containing P3/P2/N2/N3 is used to actively and dynamically control the body bias of switching transistors P4/N4. Consider the case when the input "in" is "low." Node "INB" is at "high" and the output "out" is "low." The feedback signal from "out" turns on P2 and turns off N2. In this state, P2 functions like the feedback half-latch in a dynamic circuit [38]. When "in" switches from "low" to "high," P1 turns off and N1 turns on. However, P3 and P2 are still on at the beginning of the switching period; thus, there is a current path through P3, P2, and N1. Transistor P3 is sized such that its current is substantially smaller than that through N1/P2. Thus, the excess current through N1/P2 pulls down the body potential of P4, reducing the threshold voltage and speeding up the output pullup transition via P4. When the output becomes "high," P2 turns off and the current path is cut off. The body potential of P4 is pulled up to the supply voltage by P3, thus increasing the threshold voltage and reducing the standby leakage. Since the body potential is charged/discharged through active transistors, the circuit can operate at higher frequency than the circuit using a reverse-biased diode [Fig. 30(b)]. In the Type-B circuit [Fig. 37(a)], the feedback signal is provided through a small-size inverter "INV" to obtain a longer period of low-threshold voltage for P4 or N4. Fig. 37(b) shows the operating wave forms of the bulk, conventional SOI (body tied to gate through a reverse-biased diode), and Type-A SOI circuit at 1.0 V, 60 pF load, and 100 MHz operating frequency. The body potential of the output transistor (P4 and N4) can be seen to change synchronously with the change of the "INB" signal. The delay time and power as functions of the supply voltage are shown in Fig. 38. At a supply voltage of 1.0 V, the Type-A SOI circuit operates 23% faster than the conventional SOI circuit and 37% faster than the bulk. The extra power consumption in the Type-A circuit due to the active body bias is only 2.4% compared with the conventional SOI circuit.



(a)

(b)

Fig. 37. (a) Active body-bias SOI-CMOS driver circuits and (b) simulated wave forms for the bulk, conventional SOI, and Type-A SOI circuit at 1.0 V, 60 pF, and 100 MHz [53].

A relatively simple dynamic body-bias control scheme for an SOI-CMOS inverter using subsidiary MOSFET transistors is shown in Fig. 39(a) [54]. In this scheme, the bodies of the primary driving transistors M_{ND} and M_{PD} are driven/controlled by the subsidiary transistors M_{NS} and M_{PS}. The drains of the subsidiary transistors are connected to the output, while the sources are connected to the bodies of the primary driving transistors. When the input switches from "low" to "high," the body voltage of M_{ND} is raised rapidly through M_{NS} in a source-follower configuration, thus reducing the threshold voltage of M_{ND} to improve the pulldown delay. As the output voltage goes down, the body voltage of M_{ND} also goes down (to ground) since the gate of the subsidiary transistors M_{NS} is at "high," and the threshold voltage is increased. Notice that in the initial pulldown state, the current through the subsidiary transistors M_{NS} not only charges up the body of M_{ND} but also helps to pull down the output node directly. The threshold voltage of the subsidiary transistors can be designed to be substantially lower than the primary driving transistor since they do not provide a direct leakage current path in the standby condition. This scheme works for any supply voltage compatible with the device design. Fig. 39(b) compared the measured gate delays of

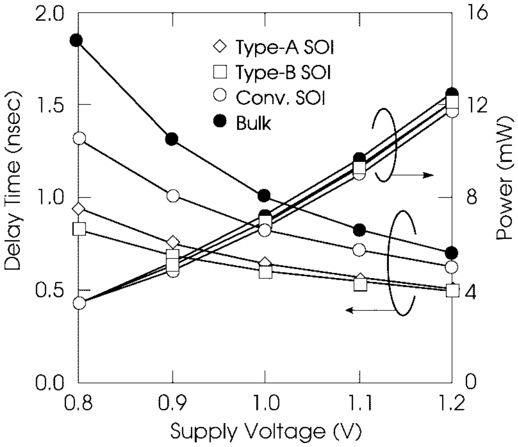
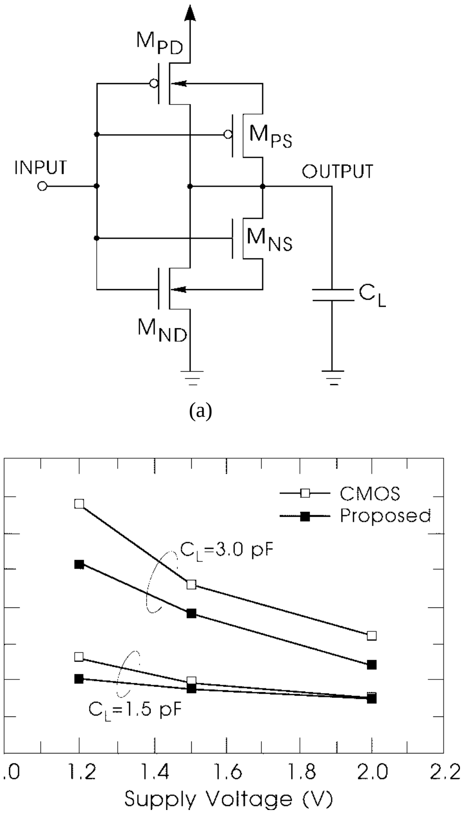


Fig. 38. Supply-voltage dependence of delay time and power dissipation at 60 pF for the bulk, conventional SOI, and Type-A SOI circuit [53].

a conventional CMOS inverter with the present scheme. When the load is small, the delay improvement of the present scheme is not significant due to the increased input capacitance caused by the subsidiary devices. As the load capacitance increases, the improvement becomes significant (about 25–30% with 3.0-pF load).

Notice that for the circuit scheme in Fig. 39(a), an nMOS subsidiary transistor $M_{\rm NS}$ is used for the nMOS pulldown transistor $M_{\rm ND}$, and a pMOS subsidiary transistor $M_{\rm PS}$ is used for the pMOS pullup transistor $M_{\rm PD}$. The body voltages of the primary driving transistor—$M_{\rm ND}$ for example—will not be charged until the input voltage rises above the threshold voltage of the subsidiary transistor $M_{\rm NS}$ and is one threshold voltage below the gate input voltage once the body charging starts. Furthermore, the delay in modulating the body voltages through the subsidiary transistors may become comparable to the primary gate delay. Thus, the effectiveness of modulating the body voltage to reduce the delay is degraded. Last, the imbalance between the pullup and pulldown delays becomes worse because of the nMOS-nMOS (primary-subsidiary) and pMOS-pMOS (primary-subsidiary) configuration.

The aforementioned drawbacks can be overcome by the improved configuration shown in Fig. 40(a), where a pMOS subsidiary transistor (MP2) is used to drive the body of the primary nMOS pulldown transistor (MN1) and vice versa [55], [56]. This scheme works only for "noninverting" buffers since the gate input to the primary transistors and the gate input to the subsidiary transistors have to be derived from opposite voltage phases, and an extra inverter at the input becomes necessary. Since the subsidiary transistors charge the bodies of the primary transistors in the common-source configuration, the drawback associated with the threshold voltage drop for Fig. 39(a) is eliminated. Also, the modulation of the body voltages occurs before the switching of the primary driving transistors. The nMOS-pMOS (primary-subsidiary) and pMOS-nMOS (primary-subsidiary) configuration balances the pullup and pulldown delays. This scheme also offers the extra benefit of turning on the parasitic bipolar transistor early and harder to help reduce the gate delay. Full supply voltage is applied to the bodies [bases of the parasitic bipolar transistors (QN1 and QP1)] during the initial phase of the switching because of the common-source configuration of the subsidiary transistors, resulting in larger transient parasitic bipolar current to help the transition of the output node. Fig. 40(b) compares the pulldown delays for noninverting buffers driving 2.0-pF load using the two schemes in Figs. 39(a) and 40(a) as functions of the supply voltage. The scheme in Fig. 40(a) can be seen to offer significant delay improvement (about 30% at 1.5 V).



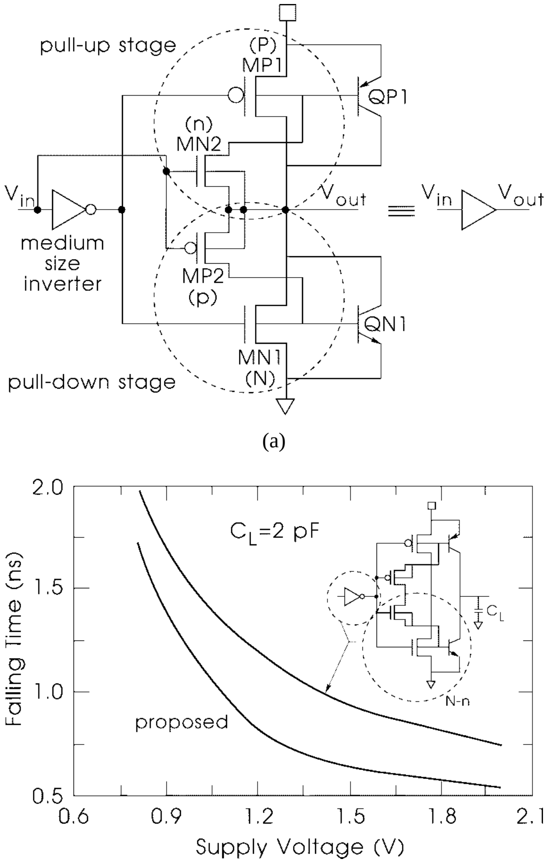
(a)

(b)

Fig. 39. (a) An SOI-CMOS inverter with dynamic body-bias control subsidiary MOSFET's. $M_{\rm ND}$ and $M_{\rm PD}$ are the driving nMOS and pMOS transistors. $M_{\rm NS}$ and $M_{\rm PS}$ are the subsidiary nMOS and pMOS transistors. (b) Measured gate delays versus supply voltage for 1.5- and 3.0-pF load [54].

VIII. Global Design Issues

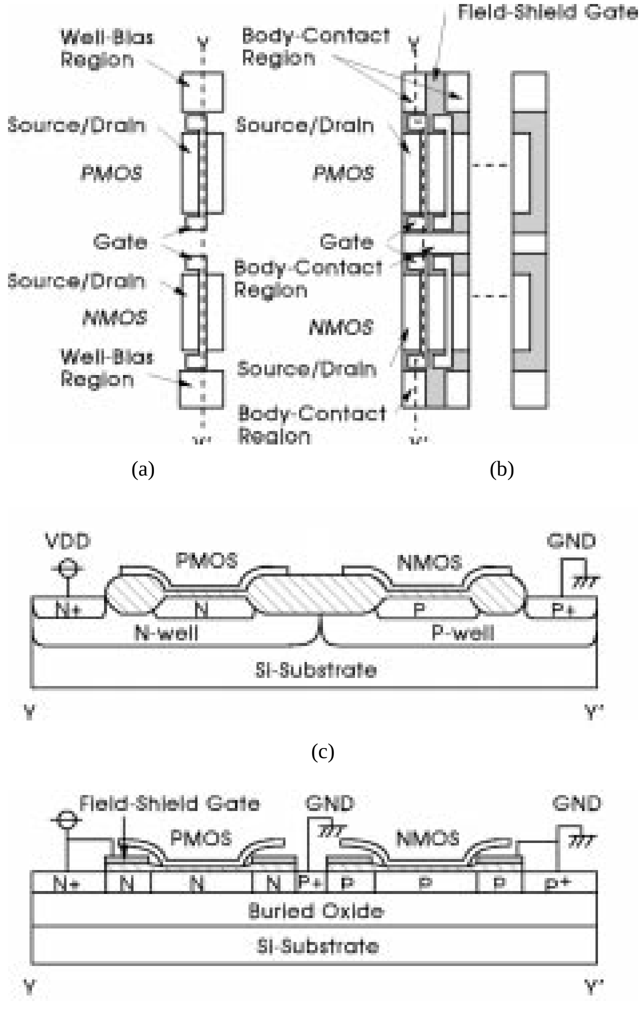
Compatibility of SOI cell libraries and design methodology with the bulk CMOS tends to be an issue. Designs with floating-body and/or smart-body contact require detailed circuit characterization and layout incompatible with the bulk CMOS. This is basically an issue of design resource and design time rather than a fundamental roadblock. It is possible to have cell libraries and design methodology compatible with the bulk CMOS if SOI device bodies are tied to the supply rails. Fig. 41 show a basic cell structure for a 0.35-μm, 220-kG SOI-CMOS gate array [57]. This gate array uses partially depleted devices with the bodies tied to the supply rails. The cell layout and power-line



(a)

(b)

Fig. 40. (a) A noninverting SOI-CMOS buffer with dynamic body-bias control subsidiary MOSFET's. MN1 and MP1 are the driving nMOS and pMOS transistors. MP2 and MN2 are the subsidiary pMOS and nMOS transistors. QN1 and QP1 are the parasitic bipolar transistors. (b) Simulated pulldown delay characteristics of the buffer circuit in (a). Also shown for comparison is the delay of a buffer circuit using the inverter circuit in Fig. 39(a) [55], [56].



(a)

(b)

(c)

(d)

Fig. 41. Basic cell structures for 0.35-μm gate arrays on bulk CMOS and SOI CMOS [57]. (a) Bulk CMOS. (b) SOI CMOS. (c) Cross section of bulk CMOS. (d) Cross section of SOI CMOS.

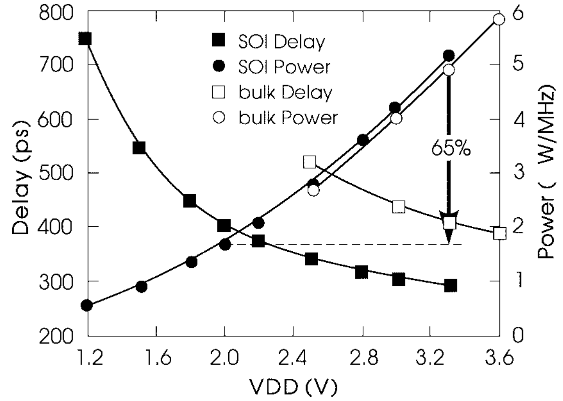


Fig. 42. Delay time and power dissipation of two-input NAND gate in bulk CMOS and SOI-CMOS gate arrays (FO = 2-, 3-mm metal wires) [57].

wiring are optimized to allow use of cell libraries and design methodology compatible with the bulk CMOS gate arrays. The basic cell utilizes field-shield gates to isolate the pMOS and nMOS transistors [58]. The field-shield gates also provide the current paths between the bodies and the body-contact regions in the SOI layer when forward biased. To improve the breakdown voltage for SOI nMOS devices, additional body-contact regions for nMOS transistors are formed between the pMOS and nMOS transistors. This structure has no area penalty, and the cell pitch is the same as that for the bulk CMOS, since body contacts are formed in regions the well bias would occupy in the bulk CMOS case. Fig. 42 compares the delay and power as functions of the supply voltage for a two-input NAND gate with FO = 2- and 3-mm metal wires in SOI-CMOS and bulk CMOS gate arrays. The SOI-CMOS gate array consumes 65% less power than the bulk CMOS gate array operating at the same speed.

In VLSI designs, timing rules are typically generated based on worst case considerations of process corners, supply, temperature, slew rate, and switching patterns. For bulk CMOS, the pattern dependency in stacked configuration has long been known to degrade the timing rules. To illustrate, consider a simple dynamic two-way NAND circuit in Fig. 43. Switching the bottom transistor N1 (i.e., V_{in1} switches from "low" to "high" with V_{in2} staying "high")

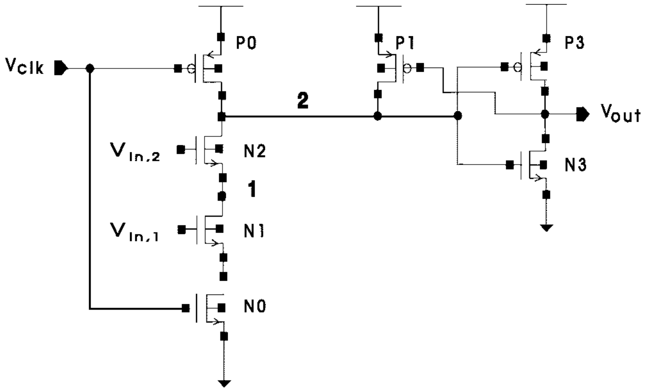


Fig. 43. A dynamic two-way NAND circuit.

to pull dynamic Node 2 down is substantially slower than switching the top transistor N2. This is because:

a) the voltage V_1 at Node 1 is at $V_{\mathrm{in2}} - V_T$ (less than V_{DD}) to start with (hence, smaller V_{DS} and I_{DS} when N1 switches "on");

b) the top transistor N2 suffers from the reverse body effect and has a high V_T to start with (with its source, Node 1, sitting at "high" and its body grounded, there is a large reverse bias between the body and the source of N2, resulting in high threshold voltage).

For the case of switching the top transistor N2, Node 1 is at ground to start with. Hence, full V_{DD} is present as V_{DS} for N2. Also, current to pull down Node 2 is immediately available once N2 switches "on." While the reduced parasitic in the SOI structure improves both switching cases, the improvement for the bottom transistor switching case is particularly significant due to three reasons. First, the body voltage of N2 sits between Node 2 and Node 1 voltages and so is very close to V_{DD} before N1 switches. The V_T of N2 is thus significantly reduced. Second, Node 1 voltage is closer to V_{DD} than the bulk case, resulting in larger V_{DS} across N1 to start with. Last, the body voltage of N1 sits between Node 1 voltage and ground, thus reducing the V_T of N1. The timing rules are therefore improved for the NAND-like (nMOS stack) and NOR-like (pFET stack) configurations. Similarly, timing rules for pass-transistor-based circuits such as CPL [42]–[44] are improved due to the lack of reverse body effect in floating-body configuration in SOI. The timing rules for some other circuit topologies, however, are degraded by the parasitic bipolar effect and the transient threshold voltage variation discussed earlier. The necessity to size up holding devices, drop body contacts, or widen the design margin to overcome the effects under worst case patterns degrades the circuit speed and area and hence timing rules. The parasitic bipolar effect also worsens while CMOS devices are degraded at an elevated temperature.

The pattern and history dependency and the resulting nonuniform speedup (over the bulk) complicate the circuit timing and design methodology. They may also degrade the overall chip timing. One example is a single clock, single latch design as shown in Fig. 44. In this kind of design, the latches are typically "open" only for a prescribed time to allow the evaluated data to be written into the latches. The latches are then closed to avoid feed-forward of the data from latches at the preceding pipeline stage. The lower bound of the duration for the latches to remain open is dictated by the longest path so valid data can be written into the latches, while the upper bound is determined by the shortest path to ensure that feed-forward will not occur. Short paths are "padded" to ensure that feed-forward (or racing) will not occur under worst case timing conditions. The padding tends to hurt the long paths and degrade the cycle time in most cases. Due to the nonuniform, pattern-dependent speedup in the floating-body configuration, some short paths may get "shorter," thus requiring more padding and hurting the long paths and cycle time more.

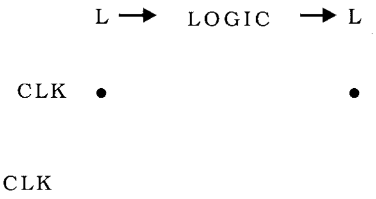
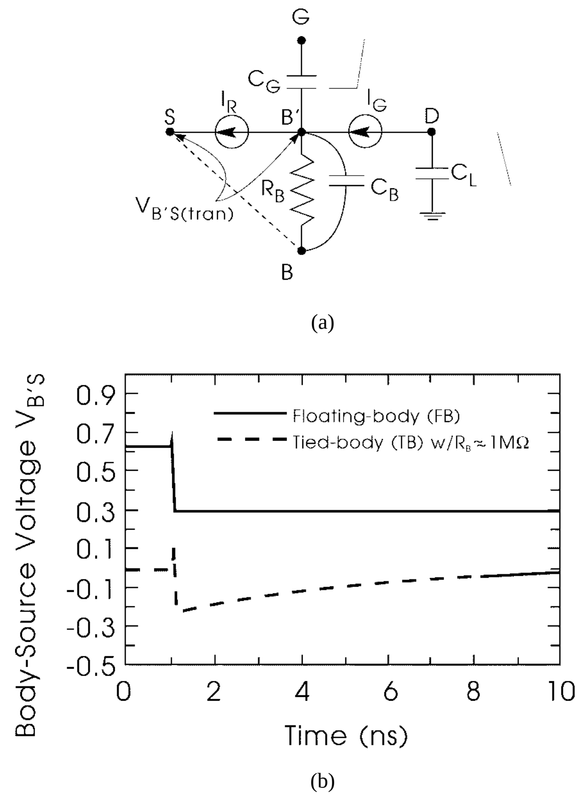


Fig. 44. A single-clock, single-latch design.

Clock distribution and skew/jitter minimization have been a major challenge in high-performance designs. In floating-body SOI configuration, the uncertainty about the floating-body potential and the hysteresis effect translates into clock skew and jitter, thus degrading the performance. This is exemplified by the pulse-stretching effect discussed in Section V. Since it is impractical, if not impossible, to design the clock distribution tree to balance out these effects, body contacts should be used in the clock distribution network to eliminate these effects and simplify the design process.

On-chip decoupling to reduce supply bounce/noise represents another challenge in SOI technology, primarily for high-performance applications. State-of-the-art, high-performance microprocessors operate at clock frequencies ranging from 200 to above 400 MHz with power consumption in the range of 20–30 W [59]–[61]. The total effective switching capacitance can be easily estimated from $P = C\ \mathrm{V}^2 f$. For a 200-MHz/30-W/3.3-V processor, the total effective switching capacitance is about 13 nF [60]. A similar amount of total effective switching capacitance is present for a 433-MHz/25-W/2.0-V microprocessor [61]. To achieve effective supply decoupling and limit the supply voltage reduction to 10%, the ratio of the on-chip decoupling capacitance to the total effective switching capacitance should be greater than 10 : 1 [60], thus requiring a decoupling capacitor in the range of 130 nF and above. In bulk CMOS technology, about 40–50% of the required decoupling capacitance is supplied by the "built-in" nonswitching capacitances such as well-to-substrate capacitance, diffusion-to-well capacitance, etc. In SOI technology, with the faster circuit speed and absence of these "built-in" decoupling capacitances, the supply and



(a)

(b)

Fig. 45. (a) Equivalent circuit of the body of a PD-SOI nMOSFET under transient. C_B represents the effective capacitance of the body node. The time constant of the body R_BC_B must be smaller than the switching time of the gate for the body tie to be efficient in transient and (b) simulated transient body voltage in response to a rising gate input in an inverter. The gate input rises at $t = 1.0$ ns with a rise time of about 50 ps. Note the long body recovery time with a high-resistance body tie [62].

ground bounce is more severe. The decoupling capacitors have to be distributed around the chip, especially under the data bus. Due to the fast on-chip slew rate (rise/fall time in the order of 200–300 ps), a very low resistance–capacitance (*RC*) time constant ($\leq$100 ps) for the decoupling capacitor is essential to achieve effective decoupling. In bulk CMOS, the n-well sheet resistance is typically about 500 $\Omega/\square$, so the *RC* requirement can be met with thin oxide capacitors built on the n-well with reasonable device aspect ratios. In an SOI structure, the body region is thin and has substantially higher sheet resistance. A viable decoupling capacitor structure with low *RC* time constant therefore becomes a process and device design challenge.

The series resistance of the body contact is crucial in determining its effectiveness in suppressing the floating-body effect (when used as a body tie) or in modulating the threshold voltage (when used as a smart contact) [62]. The time constant of the body [Fig. 45(a)] must be substantially lower than the switching time of the gate for the body contact to be efficient under dynamic conditions. This is illustrated in Fig. 45(b), where the simulated body voltages in response to a rising gate input are shown for the floating-body configuration and a high-resistance body-tied configuration. The device response for the body-tied configuration can be seen to be similar to the floating-body configuration even though the initial DC levels are different. The body contact typically requires significant area and consumes the device width. Also, with an ohmic body tie, the source and drain cannot be swapped, thus disallowing bidirectional operation of the device. A compact Schottky body-contact scheme has been developed to overcome these drawbacks in a 0.35-μm, partially depleted SOI technology with cobalt silicided source/drain and polysilicon [63], [64]. In this scheme, a Schottky diode forms between the source/drain and body wherever the source/drain implant into the silicided source/drain region is selectively not performed. Fig. 46(a) and (b) shows an example of how the source implant is masked to provide a Schottky contact to the body, while Fig. 46(c) shows a schematic of the device. The forward bias turn-on voltage of the Schottky diode (about 0.3 V) is substantially lower than that of the body-source PN junction, thus limiting the body potential and suppressing the floating-body effects. The structure allows bidirectional operation of the device if Schottky diodes are placed at both the source and drain terminals. It is also self-aligned along the device width direction since misalignment in the width direction does not affect the device performance. It thus requires less area than typical ohmic body contacts and consumes less device width. Fig. 47 compares the device width consumed for a minimum-sized body contact for the Schottky contact and the ohmic contact. Since for the ohmic contact case one has to absorb the P^+ to N^+ alignment overlay requirement (not required for the Schottky case) and the lateral diffusion of the P^+ region, the ohmic contact consumes two times (overlay + lateral diffusion) more device width than the Schottky contact, which can be substantial in deep submicron technology. This scheme has been used in a state-of-the-art microprocessor (DEC StrongArm-110) [63], [64] using a 2.0-V, 0.35-μm, partially depleted SOI technology and proven effective in suppressing the floating-body effect.

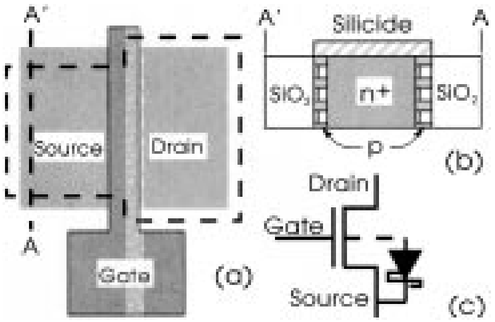


Fig. 46. A PD-SOI nMOS with Schottky body contact at the source. (a) Top view. The dashed line indicates the N^+ source/drain implant mask. (b) Source cross-sectional view. (c) Device schematic [63], [64].

Process modification for more efficient body tie has also been pursued [65]. One example is shown in Fig. 48(a), where the field oxide does not consume the silicon film completely, thus leaving a thin silicon film between the field oxide and the buried oxide. The body can then be contacted through the remaining thin silicon film beneath the field

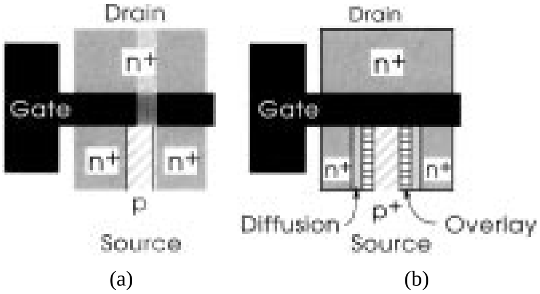


Fig. 47. Comparison of the device width consumed for a minimum-sized body contact for (a) the Schottky case and (b) the ohmic case [63], [64].

oxide. Due to the capacitance between the source/drain diffusion and the remaining silicon film, this device structure has larger junction capacitance than the normal SOI structure. Compared with bulk CMOS, the structure eliminates the well-to-substrate capacitance in bulk CMOS and has lower source/drain junction capacitance since the silicon film is typically more lightly doped than the well in bulk CMOS. The performance of this structure therefore lies between the normal SOI and the bulk CMOS devices, as shown in Fig. 48(b). Notice that once the silicon film is completely depleted, the source/drain junction capacitance reaches its minimum value and remains constant. The structure has been applied to 1.0-Gb DRAM in a 0.17-μm bulk-compatible SOI technology [65].

ESD susceptibility is a major reliability issue for SOI technology. Most ESD protection schemes developed for bulk CMOS—such as the use of a large-area, low-series-resistance vertical PN junction or thick-field-oxide devices—are either unavailable or impractical in SOI technology. The thermal resistance of the buried oxide further degrades the ESD reliability. A grounded-gate MOSFET operating in the bipolar breakdown/snapback mode (Fig. 49) has been studied/proposed for ESD protection in SOI structures [66]–[68]. The second snapback for SOI nMOSFET occurs at a smaller drain current (indicating more serious Joule heating) with lower holding voltage compared with the bulk case. Also, the high-current holding voltage is relatively independent of the body bias, so similar ESD performance is expected for floating- and grounded-body SOI for positive-polarity stress. For positive-polarity Human Body Model (HBM) discharge, the ESD sustained voltage for SOI has been found to be about half that for the bulk case. For negative-polarity HBM discharge, the discharge mechanism is significantly different in the bulk and SOI structures. For the bulk case, the negative-polarity discharge pulse is absorbed by the large forward-biased drain/substrate PN junction; thus, the ESD sustained voltage is about 25–30% higher than for positive-polarity discharge. For SOI, the negative-polarity discharge path is restricted to the thin active silicon film, and the discharge current is clamped by the nMOSFET in the transistor-diode mode (the nMOSFET is driven into deep saturation in the inverse direction, with its gate and "drain" tied together and its "source" pulled down by the negative ESD stress pulse). The high series resistance in the transistor-diode mode, together with the high current density in the silicon film, results in serious local heating and an ESD sustained voltage about 10% lower than for the positive polarity discharge. The use of a body-tie configuration provides a discharge path for ESD current with lower series resistance through the forward-biased body-drain PN junction, resulting in better negative-polarity ESD performance over the floating-body configuration. Because the breakdown/snapback voltage in SOI pMOSFET is significantly higher than that for the nMOSFET, the pMOSFET (in CMOS buffer configuration) does not help to improve the ESD performance.

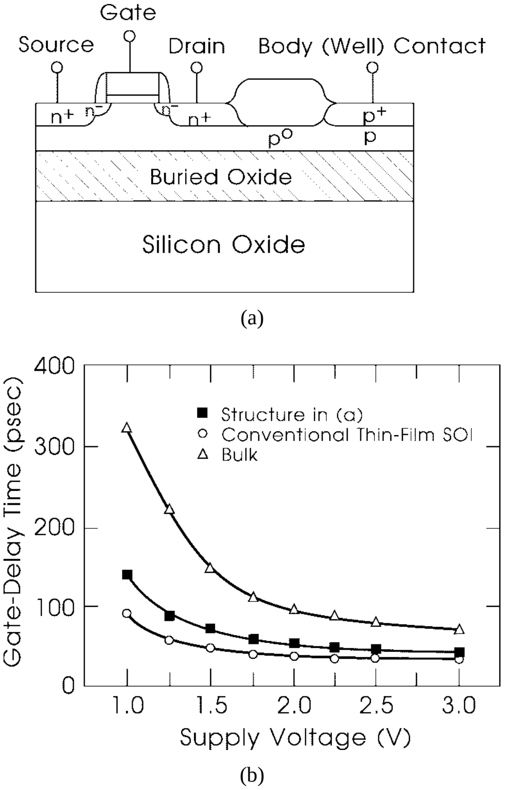


Fig. 48. (a) Cross section of a PD-SOI nMOSFET where the field oxide does not consume the silicon film completely. The device has a silicon film thickness of 200 nm and semirecessed field oxide thickness of 250 nm. A thin silicon film of 75-nm thickness remains beneath the field oxide. (b) Ring oscillator gate delays versus supply voltage for bulk CMOS, conventional SOI, and the structure in (a) [65].

Various process/device and circuit approaches have been explored to improve the ESD performance for SOI. A through-oxide ESD protection scheme [68], where the ESD protection circuitries are fabricated on the underlying bulk substrate, has been proposed to enhance the ESD performance. By doing so, most of the ESD protection schemes developed for bulk MOSFET's can be directly applied to SOI circuits. Fig. 50 shows an example of a CMOS output buffer utilizing such a scheme. Thin buried oxide is required in this scheme for good step coverage in processing and depth of focus in optical lithography systems. The lateral PN junction (e.g., body-source junction) in SOI, operated in parallel with the MOSFET, can be pursued for ESD applications. The high series resistance associated

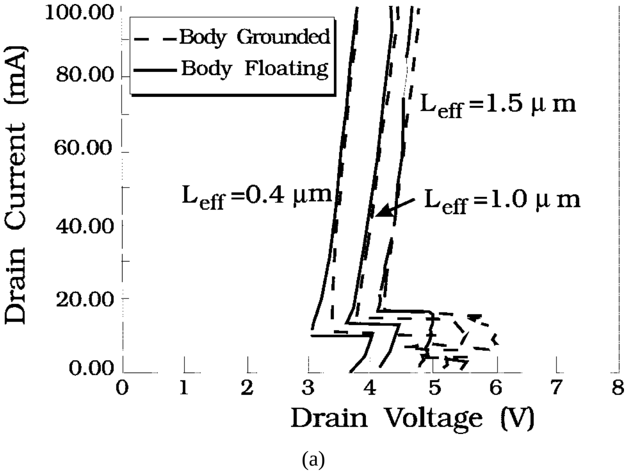


Fig. 49. Breakdown/snapback characteristics of (a) bulk nMOSFET's and (b) SOI nMOSFET's with body floating and grounded [68].

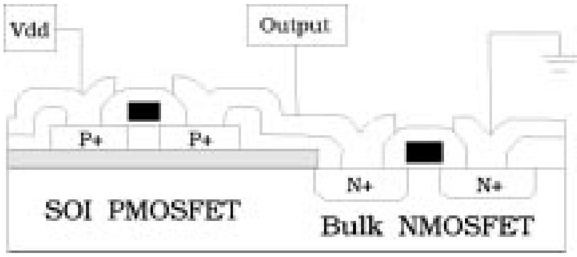


Fig. 50. A CMOS output buffer utilizing the through-oxide ESD protection scheme with the pMOSFET on the thin silicon film and nMOSFET on the substrate [68].

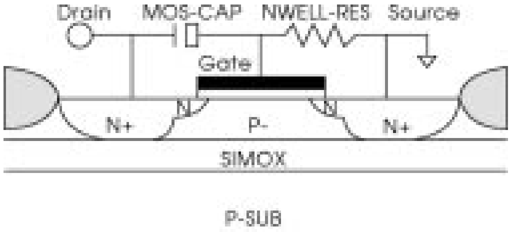


Fig. 51. Cross section of the gate-biased SOI nMOS ESD protection design [69].

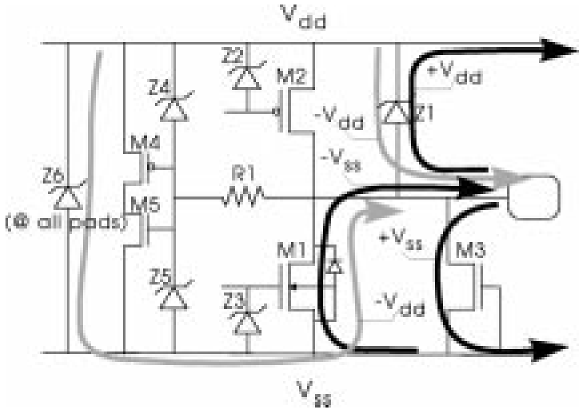


Fig. 52. I/O pad schematic with ESD current discharge paths [70].

with the body remains a limitation on ESD performance. Circuit means and ESD protection network design without resort to technology modification are more practical and cost effective. Fig. 51 shows the cross section of a gate-biased SOI nMOS ESD protection design [69], where the gate is biased above V_T during ESD stress to achieve more uniform turn-on of the wide multifinger device. A MOS capacitor controls the gate coupling during ESD stress, while a resistor allows the gate to discharge for normal operation. This scheme has been shown to offer high ESD voltage (>2 kV), good symmetric ESD performance for positive- and negative-polarity stresses, and tighter ESD failure distribution in a 0.35-μm SOI technology. Protection network designs can also enhance the ESD performance. One example is shown in Fig. 52 [70]. The circuit elements that form the main ESD protection network are the grounded-gate protection MOSFET M3, Zener diode Z1, bus Zener Z6, and the drain diode of the nMOS output driver M1. Zener diodes Z5 and Z4 and input resistor R1 protect the gate oxides of M4 and M5 (input buffer). Zener diodes Z3 and Z2 protect the gate oxides of M1 and M2. This protection network has been demonstrated to provide >3.75 kV HBM ESD performance for all stress modes.

Self-heating in SOI represents another design constraint. The thermal conductivity of silicon dioxide is about two orders of magnitude lower than that of silicon. If the power/heat is uniformly distributed across the chip, the high thermal resistance of the buried oxide does not appear to be a problem. This is because the silicon substrate thickness for the state-of-the-art technology is about 650 μm for 6-in wafers and about 740 μm for 8-in wafers. A buried oxide of 400 nm, with two orders of magnitude high thermal resistance, would have the equivalent thermal resistance of 40 μm of silicon and thus add only about 6% thermal resistance for heat removal through the silicon substrate. So, the majority of the heat dissipation is still through the substrate, while the source, drain, gate, and interconnects of the device serve as cooling fins, spreading out the area over which heat flows downward [71]–[73]. Problems arise when there are local hot spots resulting from circuits that are constantly on (such as current source, current mirror, bleeder devices, etc.) or have high switching activity (such as clock distribution network). Detailed thermal modeling and measurements have shown that under static operating conditions with the device always on, the rise in the channel temperature for SOI devices can be an order of magnitude higher than that for the bulk Si, sometimes in excess of 100°C [73]. The higher channel operating temperature causes negative differential conductance at high gate biases [71], degrades the device mobility, and reduces the maximum drain saturation current. This imposes a severe constraint for analog applications, where most devices can be on all the time. In digital applications, devices are typically on only for a small fraction of the clock cycle. Because the thermal time constants are much longer than typical electrical periods (e.g., clock cycle), the temperature rise will not follow the instantaneous power dissipation and will average out over a period on the scale of the thermal time constants. Thus, the temperature rise is significantly lower than the static case, in a range (typically 5–10°C, assuming devices are on 5–10% of the time) that can be reasonably handled. Notice that this also implies that nonself-heated device characteristics and model parameters (from high-speed pulse measurement, for example) are needed for typical digital circuit applications [73], [74]. The channel-substrate thermal resistance is roughly proportional to the square root of $t_{\mathrm{BOX}}/t_{\mathrm{Si}}$, where t_{BOX} is the thickness of the buried oxide and t_{Si} is the thickness of the silicon film. Thus, increasing the silicon film thickness or decreasing the buried oxide thickness appears equally effective in reducing the thermal resistance. Since the thermal diffusion length $(\alpha\tau)^{1/2}$ (which is a measure of the length over which the transient temperature fluctuations are significant)[2] is typically in the sub-0.5-μm range, the heat dissipation and cooling occurs primarily in the active area [73]. Heat removal means that are introduced with a physical distance substantially larger than the thermal diffusion length will not be effective. In state-of-the-art sub-0.25-μm SOI-CMOS technologies [75], [76], self-heating has been shown to degrade the current drive of nMOSFET by 10–12% and pMOSFET by about 5%.

IX. Discussion

The floating-body and hysteresis effects have long impeded the use of partially depleted SOI for mainstream logic and memory applications. Better understanding and technology advances have allowed the parasitic bipolar leakage to be accurately measured/modeled and suppressed [10], [77]. The parasitic bipolar effect can be suppressed by using a nonideal (or "leakage") body-source junction to reduce the current gain of the parasitic bipolar transistor and the floating-body voltage. It is important to confine the nonideal body-source junction leakage to well below the subthreshold leakage of the MOSFET so as not to degrade the leakage of the device. Scaling the supply voltage reduces the parasitic bipolar effect significantly since there is less voltage across the body-source PN junction. One might expect the parasitic bipolar effect to be more significant as the effective channel length (hence, "base width" of the parasitic bipolar transistor) is reduced. Experimentally, however, it has been found that the parasitic bipolar effect actually decreases with decreasing channel length [76] as the transistor structure- and doping-wise is far from an ideal bipolar transistor. In well-designed state-of-the-art devices [75], [76], the pass-gate leakage resulting from the parasitic bipolar effect can be well below 10 μA/μm. Nevertheless, detailed circuit design/simulation to ensure enough margin is still crucial.

In contrast with the parasitic bipolar effect, the floating-body-induced hysteretic (switching-history-dependent) V_T variation becomes more serious as the supply voltage and V_T are scaled if the device is not properly designed. The hysteretic V_T variation effect can be minimized through proper device design [40], [79]. Consider the case in Fig. 53, where the body voltage of an nMOS device in a CMOS inverter is shown as a function of time through one switching period. Before the start of the switching, the gate input is at "high" and the drain output is at "low," so the body voltage is at $V_{B0} = 0$ V. As the gate voltage switches down, the drain voltage rises to V_{DD}, and the body voltage rises by an amount ΔV_{BD} determined by the drain-body capacitive coupling. The body voltage is then charged by the diode current through the reverse-biased drain-body PN junction and off-state impact ionization current and discharged by the diode current through the forward-biased body-source PN junction. When the gate switches up again, the body voltage follows the gate voltage rises initially due to the gate-body capacitive coupling, resulting in a "spike" ΔV_{BG}, as shown in Fig. 53. The spike increases the leakage through the forward-biased body-source PN junction, resulting in a loss of body charges. During this period, the body is also charged by the on-state impact ionization current. The body voltage then decreases as the drain voltage goes down. To minimize the hysteretic V_T variation, it is essential to balance the gain and loss of body charges in different logic states. This can be achieved through a process/device design with decreased gate-body capacitive coupling and nonideal (leaky) body-source/drain diodes [79]. Nonideal diode characteristics result in higher diode currents, which reduces the significance of impact ionization through the switching period. It also increases the symmetry between forward- and reverse-biased diode currents, thus providing more reverse diode current to replenish the loss of body charges due to the forward-biased diode. Diode nonideality also reduces the loss of charge through capacitive coupling by decreasing the change in forward current per change in the voltage and more symmetric forward- and reverse-biased diode currents. Operating in the regime where the "leaky" body-source/drain diodes dominate device operation to minimize hysteretic V_T variation implies a higher value of body voltage (when the gate input is "low"), which increases the off-state current and degrades the transient subthreshold leakage, thus representing a design tradeoff. Scaled device design for lower supply voltage tends to alleviate the hysteretic V_T variation since higher peak dopings in devices

[2] α is the thermal diffusivity of silicon; τ is the clock period.

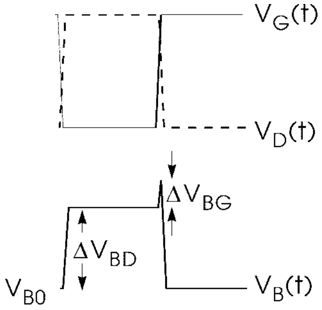


Fig. 53. Body voltage of a charge-balanced nMOS device in a CMOS inverter versus time through one switching period. For simplicity, the switching period is assumed to start with gate input at "high" (drain output at "low"), so that the reference $V_{B0} = 0$ V. $\Delta V_{\rm BG}$ is due to the gate-body capacitive coupling and $\Delta V_{\rm BD}$ is due to the drain-body capacitive coupling [79].

with highly nonuniform doping profiles result in more nonideal diode characteristics and less gate-body capacitive coupling due to increased body-source/drain capacitance. Experimentally, hysteretic V_T variation has been shown to cause 5–6% frequency-dependent delay variation in simple circuits (inverter, NAND, NOR, etc.) in state-of-the-art, sub-0.25-μm SOI-CMOS technologies [75], [76].

The performance advantage of SOI over bulk CMOS technology depends on several factors. To maintain the same transistor off-state leakage current, SOI MOSFET's are typically designed to have higher threshold voltage than their bulk counterparts. The higher threshold voltage reduces the available current drive and the performance leverage of SOI devices, especially at low supply voltages. Furthermore, in typical logic applications, interconnect accounts for $>$30% of the total chip load capacitance and can easily contribute to more than 60% of the delays in heavily loaded circuits. If the threshold voltage of SOI devices is set too high because of process technology and device design window concerns, the advantage of SOI diminishes and can even disappear completely [78]. With proper process/device design to keep the threshold voltage from being undesirably high while containing the leakage, 0.25-μm SOI-CMOS technology with minimum effective channel length down to the 0.1-μm range [75], and sub-0.15-μm SOI-CMOS technology with effective channel length down to the 0.06-μm range [76] have been demonstrated with excellent device characteristics and significant performance improvement over bulk CMOS. Raw inverter delays of 7.9 and 5.5 ps have been achieved at room temperature and liquid nitrogen temperature, approaching that of Josephson junctions.

Design strategy can also affect the leverage of SOI. In a "dual-design" approach requiring the design to work on both bulk and SOI technologies with only minor mask-level adjustments, the performance in both the bulk and SOI versions is likely to be sacrificed. The necessity to reserve area in the bulk version to accommodate body contacts in the SOI version increases the area in the bulk design. On the other hand, the SOI version may suffer from more severe supply and ground bounce (thus hurting the performance) due to inadequate amount of decoupling capacitor, as discussed earlier. The long paths in the SOI design may differ significantly from the bulk due to all the effects discussed earlier. Thus, timing optimization for the bulk version may not improve (and may even hurt) the performance of the SOI version. The wire delay also degrades the advantage of the SOI version substantially, since the use of smart driver configurations described in the previous section may not be an option in a dual-design approach. In contrast, in a SOI-only design, a global tradeoff can be pursued to enhance the overall performance. For example, one can put in more decoupling capacitors and recoup the lost area by using a smaller SRAM cell (as shown in Fig. 5) for the on-chip cache memory. One can also explore the smart-body contacts and driver configurations to reduce the wire delay significantly, thus achieving much better performance.

Circuit simulations with floating bodies require substantially more memory and simulation time, and simulation convergence for the initial DC solutions tends to be a problem since the initial DC potential of floating nodes in the circuits is set by the (small) leakage currents with long time constants. The floating-body-related effects further complicate the identification and definition of various timing modes (early mode, late mode, etc.) and corners (nominal case, best case, worst case, etc.). For memory applications, the circuits are more confined. The circuit design, timing, optimization, and use of selective and/or smart-body contacts to improve the performance appear more straightforward than for mainstream logic (i.e., microprocessor) applications. Timing large-scale logic circuits with all the floating-body-related effects represents the single most challenging task in bringing SOI into mainstream microprocessor applications. Simple-minded approaches by "blindly" adding design margins to account for various effects (e.g., adding, say, a 5% margin to the early or late mode to account for hysteretic V_T effect) will quickly "eat away" the performance leverage for SOI. A viable timing strategy, based on the thorough understanding of the device behavior and circuit topology, to incorporate and "bound" various floating-body effects in the timing rules is crucial in exploiting and maintaining the advantage of SOI over bulk CMOS technology.

The device lifetime does not appear to be a problem for SOI. Adequate (and similar) hot carrier lifetimes have been demonstrated for 0.35-μm partially depleted devices at 2.0 V with body floating or body grounded [64]. Device lifetime (to 10% reduction in the saturation current) similar to bulk technology has also been demonstrated in a 0.25-μm SOI-CMOS technology, and ten-year operation at 1.8 V with effective channel length down to 0.1 μm appears feasible [75].

As of December 1997, a 4.0-Mb SOI SRAM with yield comparable to bulk CMOS and over 20% performance improvement has been demonstrated [75]. A similar performance improvement has been demonstrated for DRAM at the 16-Mb level [29]. Also reported was 16-Mb SOI DRAM with body-voltage-control technique aiming at 1.0 V operation [47]. Small DRAM arrays exploring SOI circuit techniques for 1.0- and 4.0-Gb applications have been abundant. A 1.0-Gb SOI DRAM has been "successfully fabricated" (although no performance figures have been quoted) [65]. The "core" of a state-of-the-art microprocessor (with phase-locked loop and ESD protection circuits disabled) for embedded applications, StrongArm-110, has been built on SOI and showed over 20% performance improvement [64]. The application of SOI technology for mainstream high-performance, general-purpose microprocessors is on the horizon.

X. Conclusion

We have reviewed the design considerations and recent advances of SOI for CMOS VLSI memory and logic applications. For SOI SRAM's, the improvement in SER cannot be taken for granted. For SOI DRAM's, process/device modifications or circuit techniques to suppress the subthreshold leakage are necessary to achieve superior dynamic retention time and maintain density leverage over the bulk-Si DRAM's. Body contacts in general are needed in boosted word-line drivers and output drivers to prevent premature device breakdown and in sense amplifiers to maintain enough sense margin and sensing speed. Logic circuit topologies and switching patterns susceptible to the parasitic bipolar effect resulting from the floating body were discussed. Stacked OR-AND configurations, pass-transistor-based designs, and multilevel voltage-switch current-steering type circuits are particularly vulnerable to the parasitic bipolar effect. Proper design/sizing of these circuits and judiciously dropping body contacts in selected devices/circuits are necessary to achieve the full potential of partially depleted SOI devices. Hysteretic V_T variation and the impact on circuit operations were addressed. The smart use of body contacts enhances the performance and lowers the operating voltage and power. Global design issues such as chip timing, on-chip decoupling capacitor, electrostatic discharge protection, and heat dissipation were discussed. By taking advantage of the unique device structure and characteristics, SOI-specific circuit design extends the operating voltage, power, and performance into a regime unattainable with the bulk CMOS technology.

References

[1] J. P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI.* Boston, MA: Kluwer, 1991.

[2] Z. J. Lemnios, "Manufacturing technology challenges for low power electronics," in *Dig. Tech. Papers, Symp. VLSI Technology,* 1995, pp. 5–8.

[3] G. G. Shahidi *et al.*, "SOI for a 1-Volt CMOS technology and applications to a 512-Kb SRAM with 3.5 ns access time," in *IEDM Tech. Dig.,* 1993, pp. 813–816.

[4] ——, "CMOS scaling in the 0.1-μm, 1.X-Volt regime for high-performance applications," *IBM J. Res. Develop.,* vol. 39, no. 1/2, pp. 229–244, Jan./Mar. 1995.

[5] Y. Taur *et al.*, "CMOS scaling into the 21st century: 0.1 μm and beyond," *IBM J. Res. Develop.,* vol. 39, no. 1/2, pp. 245–260, Jan./Mar. 1995.

[6] G. G. Shahidi *et al.*, "A room temperature 0.1 μm CMOS on SOI," *IEEE Trans. Electron Devices,* vol. 41, pp. 2405–2412, Dec. 1994.

[7] ——, "Fabrication of CMOS on ultrathin SOI obtained by epitaxial lateral overgrowth and chemical-mechanical polishing," in *IEDM Tech. Dig.,* 1990, pp. 587–590.

[8] A. Wei, M. J. Sherony, and D. A. Antoniadis, "Transient behavior of the kink effect in partially-depleted SOI MOSFET's," *IEEE Electron Device Lett.,* vol. 16, pp. 494–496, Nov. 1995.

[9] D. Suh and J. G. Fossum, "Dynamic floating-body instabilities in partially depleted SOI CMOS circuits," in *IEDM Tech. Dig.,* 1994, pp. 661–664.

[10] M. M. Pelella, J. G. Fossum, D. Suh, S. Krishnan, and K. A. Jenkins, "Low-voltage transient bipolar effect induced by dynamic floating-body charging in PD/SOI MOSFET's," in *Proc. IEEE Int. SOI Conf.,* Oct. 1995, pp. 8–9.

[11] J. Gautier and J. Y. C. Sun, "On the transient operation of partially depleted SOI NMOSFET's," *IEEE Electron Device Lett.,* vol. 16, pp. 497–499, Nov. 1995.

[12] P. F. Lu, J. Ji, C. T. Chuang, L. F. Wagner, C. M. Hsieh, J. B. Kuang, L. Hsu, M. M. Pelella, S. Chu, and C. J. Anderson, "Floating body effects in partially-depleted SOI CMOS circuits," in *Proc. 1996 Int. Symp. Low Power Electronics and Design,* Monterey, CA, Aug. 12–14, 1996, pp. 139–144.

[13] A. Wei and D. A. Antoniadis, "Measurement of transient effects in SOI DRAM/SRAM access transistors," *IEEE Electron Device Lett.,* vol. 17, pp. 193–195, May 1996.

[14] M. M. Pelella, J. G. Fossum, D. Suh, S. Krishnan, K. A. Jenkins, and M. J. Hargrove, "Low-voltage transient bipolar effect induced by dynamic floating-body charging in scaled PD/SOI MOSFET's," *IEEE Electron Device Lett.,* vol. 17, pp. 196–198, May 1996.

[15] P. F. Lu, C. T. Chuang, J. Ji, L. F. Wagner, C. M. Hsieh, J. B. Kuang, L. Hsu, M. M. Pelella, S. Chu, and C. J. Anderson, "Floating body effects in partially-depleted SOI CMOS circuits," *IEEE J. Solid-State Circuits,* vol. 32, pp. 1241–1253, Aug. 1997.

[16] H. Pilo, S. Lamphier, F. Towler, and R. Hee, "A 300 MHz, 3.3 V 1 Mb SRAM fabricated in a 0.5 μm CMOS process," in *ISSCC Dig. Tech. Papers,* 1996, pp. 148–149.

[17] A. Pelella, P. F. Lu, Y. Chan, W. Huott, U. Bakhru, S. Kowalczyk, P. Patel, J. Rawlins, and P. Wu, "A 2 ns access, 500 MHz 288 Kb SRAM Macro," in *Dig. Tech. Papers, Symp. VLSI Circuits,* 1996, pp. 128–129.

[18] C. M. Hsieh, P. C. Murley, and R. R. O'Brien, "A field-funneling effect on the collection of alpha-particle-generated carriers in silicon devices," *IEEE Electron Devices Lett.,* vol. EDL-2, pp. 104–106, Apr. 1981.

[19] S. E. Kerns, L. W. Massengill, D. V. Kerns, Jr., M. L. Alles, T. W. Houston, H. Lu, and L. R. Hite, "Model for CMOS/SOI single-event vulnerability," *IEEE Trans. Nucl. Sci.,* vol. 36, pp. 2305–2310, Dec. 1989.

[20] H. Iwata and Ohzone, "Numerical analysis of alpha-particle-induced soft errors in SOI MOS devices," *IEEE Trans. Electron Devices,* vol. 39, pp. 1184–1190, May 1992.

[21] Y. Tosaka, K. Suzuki, and T. Sugii, "α-particle-induced soft errors in submicron SOI SRAM," in *Dig. Tech. Papers, Symp. VLSI Technology,* 1995, pp. 39–40.

[22] J. B. Kuang, S. Ratanaphanyarat, M. J. Saccamango, L. Hsu, R. C. Flaker, L. F. Wagner, S. Chu, and G. G. Shahidi, "SRAM bitline circuits on PD SOI: Advantages and concerns," *IEEE J. Solid-State Circuits,* vol. 32, pp. 837–844, June 1997.

[23] L. A. Glasser and D. W. Dobberpuhl, *The Design and Analysis of VLSI Circuits.* Reading, MA: Addison-Wesley, 1988, p. 288.

[24] K. Kumagai, T. Yamada, H. Iwaki, H. Nakamura, and H. Onishi, "A new SRAM cell design using 0.35 μm CMOS/SIMOX technology," in *Proc. IEEE Int. SOI Conf.,* 1997, pp. 174–175.

[25] K. Suma *et al.*, "An SOI-DRAM with wide operating voltage range by CMOS/SIMOX technology," in *ISSCC Dig. Tech. Papers,* 1994, pp. 138–139.

[26] F. Morishita *et al.*, "Leakage mechanism due to floating body and countermeasure on dynamic retention mode of SOI-DRAM," in *Dig. Tech. Papers, Symp. VLSI Technology,* 1995, pp. 141–142.

[27] T. Tanigawa, A. Yoshino, H. Koga, and S. Ohya, "Enhancement of data retention time for giga-bit DRAM's using SIMOX technology," in *Dig. Tech. Papers, Symp. VLSI Technology,* 1994, pp. 37–38.

[28] S. Tomishima *et al.*, "A long data retention SOI-DRAM with the body refresh function," in *Dig. Tech. Papers, Symp. VLSI Circuits,* 1996, pp. 198–199.

[29] H. S. Kim *et al.*, "A high performance 16 M DRAM on a thin film SOI," in *Dig. Tech. Papers, Symp. VLSI Technology,* 1995,

pp. 143–144.

[30] T. Nishihara, H. Moriya, N. Ikeda, H. Aozasa, and Y. Miyazawa, “Data retention in ultra-thin-film-SOI DRAM with buried capacitor cell,” in *Dig. Tech. Papers, Symp. VLSI Technology*, 1994, pp. 39–40.

[31] M. Asakura *et al.*, “A 34 ns 256 Mb DRAM with boosted sense-ground scheme,” *ISSCC Dig. Tech. Papers*, 1994, pp. 140–141.

[32] ——, “An experimental 256-Mb DRAM with boosted sense-ground scheme,” *IEEE J. Solid-State Circuits*, vol. 29, pp. 1303–1309, Nov. 1994.

[33] T. Ooishi *et al.*, “An automatic temperature compensation of internal sense ground for sub-quarter micron DRAM’s,” in *Dig. Tech. Papers, Symp. VLSI Circuits*, 1994, pp. 77–78.

[34] M. Terauchi and M. Yoshimi, “Analysis of floating-body-induced leakage current in 0.15 μm SOI DRAM,” in *Proc. IEEE Int. SOI Conf.*, 1996, pp. 138–139.

[35] G. R. Srinivanson *et al.*, “Accurate, predictive modeling of soft error rate due to cosmic rays and chip α radiation,” in *Proc. IRPS*, 1994, pp. 12–16.

[36] E. Normond *et al.*, “Single event upset and charge collection measurements during high-energy protons and neutrons,” *IEEE Trans. Nucl. Sci.*, vol. 41, pp. 2203–2209, 1994.

[37] T. Aton *et al.*, “Direct measurement for SOI and bulk diodes of single-event-upset charge collection from energetic ions and alpha particles,” in *Dig. Tech. Papers, Symp. VLSI Technology*, 1996, pp. 98–99.

[38] N. Weste and K. Eshraghian, *Principles of CMOS VLSI Design—A System Perspective.* Reading, MA: Addison-Wesley, 1988.

[39] C. T. Chuang, P. F. Lu, J. Ji, L. F. Wagner, S. Chu, and C. J. Anderson, “Dual-mode parasitic bipolar effect in dynamic CVSL XOR circuit with floating-body partially-depleted SOI devices,” in *Proc. Tech. Papers, Int. Symp. VLSI Technology, Systems, and Applications*, Taipei, Taiwan, June 3–5, 1997, pp. 288–292.

[40] A. Wei, D. A. Antoniadis, and L. A. Bair, “Minimizing floating-body-induced threshold voltage variation in partially depleted SOI CMOS,” *IEEE Electron Device Lett.*, vol. 17, pp. 391–394, Aug. 1996.

[41] T. I. Chappell, B. A. Chappell, S. E. Schuster, J. W. Allan, S. P. Klepner, R. V. Joshi, and R. L. Franch, “A 2 ns cycle, 4 ns access 512 kb CMOS ECL SRAM,” *ISSCC Dig. Tech. Papers*, 1991, pp. 50–51.

[42] K. Yano *et al.*, “A 3.8-ns CMOS 16 × 16-b multiplier using complementary pass-transistor logic,” *IEEE J. Solid-State Circuits*, vol. 25, pp. 388–395, Mar. 1990.

[43] M. Suzuki *et al.*, “A 1.5 ns 32 b CMOS ALU in double pass-transistor logic,” in *ISSCC Dig. Tech. Papers*, 1993, pp. 90–91.

[44] Y. Sasaki, K. Yano, S. Yamashita, and H. Chikata, “Multi-level pass-transistor logic for low-power ULSI,” in *Dig. Tech. Papers, Symp. Low Power Electronics*, 1995, pp. 14–17.

[45] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, “A dynamic threshold voltage MOSFET (DTMOS) for ultra-low voltage operation,” in *IEDM Tech. Dig.*, 1994, pp. 809–812.

[46] S. Kuge, T. Tsuruda, S. Tomishima, M. Tsukude, T. Yamagata, and K. Arimoto, “SOI-DRAM circuit technologies for low power high speed multi-giga scale memories,” in *Dig. Tech. Papers, Symp. VLSI Circuits*, 1995, pp. 103–104.

[47] K. Shimomura *et al.*, “A 1 V 46 ns 16 Mb SOI-DRAM with body control technique,” *ISSCC Dig. Tech. Papers*, 1997, pp. 68–69.

[48] S. Mutoh *et al.*, “1-V high-speed digital circuit technology with 0.5 μm multi-threshold CMOS,” in *Proc. IEEE Int. ASIC Conf.*, 1993, pp. 186–189.

[49] T. Douseki, S. Shigematsu, Y. Tanabe, M. Harada, H. Inokawa, and T. Tsuchiya, “A 0.5 V SIMOX-MTCMOS circuit with 200 ps logic gate,” in *ISSCC Dig. Tech. Papers*, 1996, pp. 84–85.

[50] T. Fuse, Y. Oowaki, M. Terauchi, S. Watanabe, M. Yoshimi, K. Ohuchi, and J. Matsunaga, “0.5 V SOI CMOS pass-gate logic,” in *ISSCC Dig. Tech. Papers*, 1996, pp. 88–89.

[51] T. Fuse *et al.*, “A 0.5 V 200 MHz 1-stage 32b ALU using a body bias controlled SOI pass-gate logic,” in *ISSCC Dig. Tech. Papers*, 1997, pp. 286–287.

[52] L. S. Y. Wong and G. A. Rigby, “A 1 V CMOS digital circuit with double-gate-driven MOSFET,” in *ISSCC Dig. Tech. Papers*, 1997, pp. 292–293.

[53] Y. Wada *et al.*, “Active body-bias SOI-CMOS driver circuits,” in *Dig. Tech. Papers, Symp. VLSI Circuits*, 1997, pp. 29–30.

[54] I. Y. Chung, Y. J. Park, and H. S. Min, “A new SOI inverter for low power applications,” in *Proc. IEEE Int. SOI Conf.*, 1996, pp. 20–21.

[55] J. H. Lee and Y. J. Park, “High speed SOI buffer circuit with the efficient connection of subsidiary MOSFET’s for dynamic threshold control,” in *Proc. IEEE Int. SOI Conf.*, 1997, pp. 152–153.

[56] T. W. Houston, “A novel dynamic V_t circuit configuration,” in *Proc. IEEE Int. SOI Conf.*, 1997, pp. 154–155.

[57] K. Ueda *et al.*, “A CAD-compatible SOI/CMOS gate array having body-fixed partially-depleted transistors,” in *ISCC Dig. Tech. Papers*, 1997, pp. 288–289.

[58] T. Iwamatsu *et al.*, “High-speed SOI 1/8 frequency divider using field-shield body-fixed structure,” *Jpn. J. Appl. Phys.*, vol. 35, pp. 965–968, 1996.

[59] D. Dobberpuhl *et al.*, “A 200MHz 64b dual-issue CMOS microprocessor,” in *ISCC Dig. Tech. Papers*, 1992, pp. 106–107.

[60] ——, “A 200-MHz 64-b dual-issue CMOS microprocessor,” *IEEE J. Solid-State Circuits*, vol. 27, pp. 1555–1567, Nov. 1992.

[61] P. E. Gronowski *et al.*, “A 433 MHz 64 b quad-issue RISC microprocessor,” in *ISSCC Dig. Tech. Papers*, 1996, pp. 222–223.

[62] S. Krishnan, “Efficacy of body ties under dynamic switching conditions in partially-depleted SOI CMOS technology,” in *Proc. IEEE Int. SOI Conf.*, 1997, pp. 140–141.

[63] J. Sleight and K. Mistry, “A compact Schottky contact technology for SOI transistors,” in *IEDM Tech. Dig.*, 1997, pp. 419–422.

[64] K. Mistry, G. Grula, J. Sleight, L. Bair, R. Stephany, R. Flatley, and P. Skerry, “A 2.0 V, 0.35 μm partially depleted SOI-CMOS technology,” in *IEDM Tech. Dig.*, 1997, pp. 583–586.

[65] Y. H. Koh *et al.*, “1 giga bit SOI DRAM with fully bulk compatible process and body-contacted SOI MOSFET structure,” in *IEDM Tech. Dig.*, 1997, pp. 579–582.

[66] K. Verhaege, G. Groeseneken, J. P. Colinge, and H. E. Maes, “Analysis of snapback in SOI NMOSFET’s and its use for an SOI ESD protection circuit,” in *Proc. IEEE Int. SOI Conf.*, 1992, pp. 140–141.

[67] ——, “Double snapback in SOI NMOSFET’s and its application for SOI ESD protection,” *IEEE Electron Device Lett.*, vol. 14, pp. 326–328, July 1993.

[68] M. Chan, S. S. Yuen, Z. J. Ma, K. Y. Hui, P. K. Ko, and C. Hu, “ESD reliability and protection schemes in SOI CMOS output buffers,” *IEEE Trans. Electron Devices*, vol. 42, pp. 1816–1821, Oct. 1995.

[69] C. Duvvury, A. Amerasekera, K. Joyner, S. Ramaswamy, and S. Young, “ESD design for deep submicron SOI technology,” in *Dig. Tech. Papers, Symp. VLSI Technology*, 1996, pp. 194–195.

[70] J. C. Smith, M. Lien, and S. Veeraraghavan, “An ESD protection circuit for TFSOI technology,” in *Proc. IEEE Int. SOI Conf.*, 1996, pp. 170–171.

[71] L. J. McDaid, S. Hall, P. H. Mellor, W. Eccleston, and J. C. Alderman, “Physical origin of the negative differential resistance in SOI transistors,” *Electron. Lett.*, vol. 25, no. 13, pp. 827–828, June 22, 1989.

[72] K. E. Goodson and M. I. Flik, “Effect of microscale thermal conduction on the packing limit of silicon-on-insulator electronic devices,” *IEEE Trans. Comp., Hybrids, Manufact. Technol.*, vol. 15, pp. 715–722, Oct. 1992.

[73] L. T. Su, J. E. Chung, D. A. Antoniadis, K. E. Goodson, and M. I. Flik, “Measurement and modeling of self-heating in SOI NMOSFET’s,” *IEEE Trans. Electron Devices*, vol. 41, pp. 69–75, Jan. 1994.

[74] K. A. Jenkins and J. Y. C. Sun, “Measurement of I–V curves of silicon-on-insulator (SOI) MOSFET’s without self-heating,” *IEEE Electron Device Lett.*, vol. 16, pp. 145–147, Apr. 1995.

[75] D. J. Schepis *et al.*, “A 0.25 μm CMOS SOI technology and its application to 4 Mb SRAM,” in *IEDM Tech. Dig.*, 1997, pp. 587–590.

[76] F. Assaderaghi *et al.*, “A 7.9/5.5 psec room/low temperature SOI CMOS,” in *IEDM Tech. Dig.*, 1997, pp. 415–418.

[77] ——, “Accurate measurement of pass-transistor leakage current in SOI MOSFET’s,” in *Proc. IEEE Int. SOI Conf.*, 1996, pp. 66–67.

[78] R. Chau *et al.*, “Scalability of partially depleted SOI technology for sub-0.25 μm logic applications,” in *IEDM Tech. Dig.*, 1997,

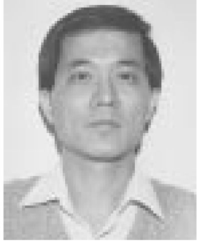
pp. 591–594.

[79] A. Wei and D. Antoniadis, "Design methodology for minimizing hysteretic V_T-variation in partially-depleted SOI CMOS," in *IEDM Tech. Dig.*, 1997, pp. 411–414.

Ching-Te Chuang (Fellow, IEEE) received the B.S.E.E. degree from National Taiwan University, Taipei, Taiwan, R.O.C., in 1975 and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1982.

From 1977 to 1982, he was a Research Assistant in the Electronics Research Laboratory, University of California, Berkeley, working on bulk and surface acoustic wave devices. He joined the Bipolar Devices and Circuits Group at the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1982, working on scaled bipolar devices, technology, and circuits. He studied the scaling properties of epitaxial Schottky barrier diodes, did pioneering work on the perimeter effects of advanced double-poly self-aligned bipolar transistors, and designed the first subnanosecond 5-Kb bipolar emitter-coupled logic SRAM. From 1986 to 1988, he was Manager of the Bipolar VLSI Design Group, working on low-power bipolar circuits, high-speed, high-density bipolar SRAM's, multi-Gb/s fiber-optic data-link circuits, and scaling issues for bipolar/BiCMOS devices and circuits. Since 1988, he has been Manager of the High Performance Circuit Group, investigating high-performance logic and memory circuits. Since 1993, his group has been primarily responsible for the circuit design of a 400-MHz IBM/ES-390 CMOS microprocessor. He has been personally responsible for evaluating SOI for high-performance logic and memory applications since 1996. He has received six U.S. patents and has authored or coauthored more than 115 papers. He was on the Device Technology Program Committee for the International Electron Devices Meeting in 1986 and 1987 and the Program Committee for the Symposium on VLSI Circuits from 1992 to 1998. He was Publication/Publicity Chairman for the Symposium on VLSI Technology and the Symposium on VLSI Circuits in 1993 and 1994.

Dr. Chuang is a member of the New York Academy of Science.



Pong-Fei Lu (Member, IEEE) received the B.S.E.E degree from National Taiwan University, Taipei, Taiwan, R.O.C., in 1978 and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ, in 1986.

From 1981 to 1985, he was a Research Assistant at Princeton University, investigating electron transport properties in heterojunctions of III–V compound semiconductors. He joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1985, working on the device design for high-speed Si bipolar transistors. Since 1989, he has been with the High-Performance Circuit Group. His current interests include cache design for microprocessors and various scaling-related device-design issues.

Carl J. Anderson received the Ph.D. degree in physics from the University of Wisconsin, Madison, in 1979.

His thesis research was on atomic collisions and the neutralization of high-energy negative hydrogen beams. In 1979, he joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member in the Josephson Technology Department. He became Manager of the GaAs Circuit Design Group in 1983 and Manager of the GaAs Process and Circuit Development area in 1987. From 1990 to 1991, he was Manager of the Optical Communication Technology Development area with responsibilities of optical link circuit design, packaging, optical device analysis, and switch design. From 1991 to 1997, he was Senior Manager of the VLSI Design area, responsible for advanced digital design and future technology evaluation. Since 1997, he has been with IBM, Austin, TX, where he is responsible for high-performance PowerPC development.