Design Considerations of Scaled Sub-0.1 μm PD/ SOI CMOS Circuits

C. T. Chuang, R. V. Joshi, R. Puri, and K. Kim
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, U. S. A.
E-mail: ctchuang@us.ibm.com

Abstract

This paper reviews the circuit design considerations of scaled sub-0.1 μm partially-depleted SOI (PD/ SOI) CMOS circuits for high-performance digital applications. The impact of technology/device scaling and design challenges are highlighted. Unique design aspects and issues resulting from the scaling of PD/ SOI device structure, such as parasitic bipolar effect and reduced- V_T leakage, hysteric V_T variation, low-voltage impact ionization, higher V_T,lin to maintain adequate V_T, sat , scaling/thinning of Si film, gate-to-body tunneling current, self-heating, soft error rate (SER), and the introduction of strained-Si channel on SOI are addressed.

1. Introduction

Partially-depleted SOI technology has recently emerged for mainstream high-performance logic applications. The unique design aspects resulting from the floating-body in PD/ SOI device structure, such as parasitic bipolar effect and hysteric V_T variation, are now quite well-understood and circuit/design techniques to mitigate these effects have been developed. However, continual scaling and demand for performance call for lower supply voltage and V_T, shorter channel length, thinner gate oxide, higher body doping concentration, thinner Si film thickness, and the introduction of new material such as strained-Si channel on relaxed SiGe layer to enhance the mobility and current drive. As the scaling approaches various physical limits, unique/new SOI design issues continue to evolve/surface. In this paper, we review the design challenges of sub-0.1 μm PD/ SOI CMOS circuits with particular emphasis on the implications and impacts of each individual device scaling element on the circuit/design.

2. Parasitic Bipolar and Reduced- V_T Leakage

Certain circuit topologies, such as stacked devices, pass-gate, and SRAM bitline structure, are susceptible to the parasitic bipolar effect [1, 2, 3, 4, 5]. The topology typically involves a “off” transistor with the source and drain voltage set up in the “High” state (hence body voltage at “High”). When the source is subsequently pulled down, large overdrive is developed across the body-source junction, causing bipolar current to flow through the lateral parasitic bipolar transistor. The parasitic bipolar current and the FET leakage (caused by the lowered V_T due to high body voltage) result in a loss of charge on the precharge (or dynamic) node and can potentially cause circuit failure. The effect is typically more significant at first cycle after long time of dormancy. In SRAM bitline structure, the aggregate parasitic bipolar effect of the unselected cells on the selected bitline causes disturb in the Read/Write operations and limits the number of cells that can be attached to a bitline pair. Various circuit/design techniques to mitigate the parasitic bipolar effect have been developed [4, 5, 6, 7, 8, 9].

While the “base width” of the parasitic bipolar transistor decreases as the channel length is scaled, the reduction in V_DD reduces the overdrive available across the body-source junction. The high doping concentration and steep profile in scaled devices increase the base Gummel number, thus reducing the current gain of the parasitic bipolar transistor. The thinning of the Si film reduces the base-emitter (body-source) junction area. Hence the parasitic bipolar effect becomes less significant with respect to the increased FET current drive. The reduced-V_T FET leakage is also contained, relative to the increased FET current drive, due to the lower V_DD and low body factor in high-performance low-V_T transistor.

3. Hysteric V_T and Low-Voltage Impact Ionization

The hysteric V_T variation due to long time constants of various body charging/discharging mechanisms (impact ionization current, GIDL, and junction leakage/current) and gain/lose of body charges through switching cycles has long been the most challenging task in the design of floating-body PD/ SOI CMOS circuits [4, 6, 10, 11, 12, 13, 14, 15, 16]. Various body voltage estimation/bounding schemes have been developed for circuit simulation and static timing [17, 18].

The impact ionization current plays an important role in determining the SOI floating-body behavior. As V_DD is scaled, conventional wisdom based on electric field induced impact ionization mechanism expects significant reduction in the impact ionization current. However, recent study on state-of-the-art SOI devices showed that the onset of the kink in the I-V characteristics is well below the silicon bandgap (E_g ~ 1.2 eV), and the underlying low voltage ionization mechanism could not be explained by the conventional wisdom [19, 20]. Experimental data indicated that while the driving force of impact ionization at high V_DD was the electric field induced by the drain, it switches to lattice temperature as drain voltage is reduced to below 1.2 V [19, 20]. This thermally assisted impact ionization mechanism at low voltage is particularly important in scaled PD/ SOI devices/circuits since self-heating in the thin Si film would significantly enhance this mechanism. Scaling/thinning of Si film has other implication on hysteric V_T variation, which will be discussed later. Furthermore, high doping concentration and steep doping gradient in scaled devices increases the reverse-biased band-to-band tunneling current between the drain and the body, resulting in higher body charging current.

One of the commonly used gauge for hysteric V_T variation (or “history effect” as known in SOI community) is the disparity in the body voltages and delays between the so-called “1st switch” and “2nd” switch [6, 13, 21]. The “1st switch” refers to the case where a circuit (e.g. inverter) starts in an initial quiescent state with input at “Low” and then undergoes an input-rising transition. In this case, the initial DC equilibrium body potential of the switching nMOS is de-
termined primarily by the balance of the back-to-back drain-to-body and body-to-source diodes. The "2nd" switch refers to the case where the circuit is initially in a quiescent state with input at "High". The input first falls, and then rises (hence the name "2nd switch"). For this case, the pre-switch body voltage is determined by capacitive coupling between the drain and the body. In early generations of PD/SOI technology (e.g. 0.25 μm and 0.18 μm), the pre-switch body voltage is typically higher (thus circuit delay lower) for the 1st switch due to high diode balance voltage at high $V_{DD}$. For scaled devices, the lower $V_{DD}$ results in lower diode balance voltage while the capacitive coupling between the drain and body increases due to higher doping concentration and steep doping gradient. Thus, the pre-switch body voltage for 1st switch decreases, while that for the 2nd switch increases, and the 2nd switch tends to become faster than the 1st switch.

4. $V_{T,lin}$

PD/SOI devices are typically designed with a larger $V_{T,lin}$ (threshold voltage at low drain bias) compared with bulk CMOS [22]. This is because as the drain voltage is raised, the floating-body effect causes the threshold voltage to decrease, resulting in significantly lower $V_{T,act}$ (threshold voltage at high drain bias). Thus, higher $V_{T,lin}$ is necessary to maintain adequate $V_{T,act}$ to contain leakage. The higher $V_{T,lin}$ has adverse effects on the performance, especially for circuit configurations where devices spend substantial amount of time in linear region during switching transient, such as pass-gate, stack devices, and SRAM bitline structure, etc. It is also well known that the $V_T$ loss in pasing a "High" state through a nMOS-only pass-gate degrades both the performance and noise margin, especially at low supply voltage. While full transmission-gate can and should be used to alleviate this problem for logic circuits, it is not practical and offers no benefit for SRAM read/write pass-transistors due to impact on density. Fortunately, it has been shown that as $V_{DD}$ is scaled, the decrease in $V_{T,act}$ due to floating body effect becomes much less due to reduction in the electric field induced impact ionization. Thus, for low $V_{DD}$, the requirements for higher $V_{T,lin}$ in PD/SOI devices is relaxed. Furthermore, the optimum SOI device design matches the $I_{off}$ to those of bulk CMOS at the shortest channel length of the given technology at the chip operating temperature. This allows higher $I_{off}$ at the nominal channel length (since PD/SOI device has better short-channel roll-off) and room temperature for the SOI devices, thus alleviating the requirement for higher $V_{T,lin}$ as well [22].

5. Scaling/Thinning of Si Film

The major benefits of scaling/thinning of Si film are: (1) reduction of junction capacitance for performance improvement, (2) better short channel roll-off, (3) better soft error rate (SER) due to less charge generation/collection volume. In addition, the history effect (disparity between 1st switch and 2nd switch) is also reduced. The reduced junction capacitance improves delays of both 1st and 2nd switches. However, for the 2nd switch (which tends to be the faster one in scaled technologies as mentioned previously), the reduced junction capacitance reduces the capacitive coupling between the drain and body, causing a decrease in the pre-switch body voltage for the 2nd switch, thus partially offsets the performance improvement [21].

On the down side, the thinning of Si film degrades the body resistance, rendering body contact less effective and eventually useless. Self-heating becomes more severe. Furthermore, as the film thickness is scaled to below 50 nm, the device may become dynamically fully-depleted (or quasi-depleted), where the body would become fully-depleted under certain bias conditions or during certain circuit switching transient. This necessitates an unified partially-depleted/full-depleted device model with smooth and seamless transition among different modes of operation. Typically, this is modeled by varying the built-in potential between the body and source junction, thus changing the amount of body charges the body-source diode can sink for a given change in the body potential. The presence of dynamic full depletion also complicates the static timing methodology, where the various body voltage bounds established based on the assumption of partial-depletion need to be extended to cover this new phenomena. Notice that dynamic depletion tends to occur first in long channel, low-$V_T$ devices. For short channel devices, the proximity of the heavily doped HALO increases the effective body doping, and the device is less likely to be dynamically fully-depleted.

6. Gate-to-Body Tunneling Current

As the gate oxide thickness is scaled to maintain gate control, $V_T$, and performance, the oxide tunneling leakage increases (Fig. 1) [23, 24]. Nitrided oxide, which reduces the leakage by order of magnitude, has been widely used in the industry to contain this leakage. Nevertheless, the oxide tunneling leakage increases by 2.5X for every 0.1 nm decrease in oxide thickness. This amounts to over 30X increase per technology generation. On the contrary, the channel leakage increases by about 3X - 5X per technology generation. As such, the oxide tunneling leakage has quickly approached $I_{off}$, and will surpass $I_{off}$ at room temperature for oxide thickness around 1.0 nm or below, thus becoming a serious concern for overall chip leakage. Furthermore, at 1.0 nm, the tunneling leakage for nitried oxide reaches 100 A/cm$^2$, while the traditional reliability criterion for oxide leakage is 1.0 A/cm$^2$. Recent study showed that at 100 A/cm$^2$, static CMOS and domino circuits in bulk CMOS still exhibit "acceptable functionality and noise margin" [23].

The oxide tunneling current consists of several components as shown in Fig. 2(a) [24]. The electron tunneling from the valence band (EVB) generates the substrate current in both nMOS and pMOS. This substrate current component is significantly less than the tunneling current between the gate and the channel (Fig. 2(b)), and its effect can usually be neglected in bulk CMOS. In PD/SOI devices, however, this substrate (body) current charges/discharges the body, thus changing $V_T$ and affecting circuit operation [21]. As this gate-to-body tunneling current has a weaker temperature dependence than the channel current, and other leakage and body charging/discharging current components, its effect is more pronounced at lower temperature [25].

Fig. 3(a) shows the change of individual device strength in a static CMOS inverter in different initial quiescent states due to the presence of the gate-to-body tunneling current. Depending on the initial condition and input transition, the inverter delay can slow-down or speed-up up to 10% - 15% in
a 1.5 V, 0.18 μm PD/SOI technology with $L_{eff} = 0.075 \mu m$, $t_{OX} = 2.3 \text{ nm}$, and $t_{Si} = 150 \text{ nm}$. In the same technology, it causes from 4% slow-down to 6% speed-up at 85°C in critical path delays of a 1.1 GHz, 115 W, 170 million transistor, 64b Power4 PowerPC microprocessor (Fig. 3(b)) [25].

For pass-transistor based circuits, studies based on a 1.2 V, 0.13 μm PD/SOI technology with $L_{poly} = 0.075 \mu m$, physical $t_{OX} = 1.5 \text{ nm}$, $t_{Si} = 120 \text{ nm}$, and $t_{BOX} = 145 \text{ nm}$ indicated that the gate-to-body tunneling current can cause delay changes up to 11% [26]. In the same technology, standard CMOS latches with a “trickl” feedback inverter exhibits delay changes of up to 6% - 7%, while the latch set-up time can change up to 15% - 17%. The results clearly indicated that the gate-to-body tunneling current has to be carefully accounted for in the timing of scaled PD/SOI CMOS circuits.

Fig. 4 depicts the changes in the strength of cell transistors in the quiescent state of a 6T CMOS SRAM cell. Detailed study on a 34Kb L1 directory SRAM showed that the presence of gate-to-body tunneling current resulted in much more significant degradation in the ”Write” operation compared with the ”Read” operation. On the other hand, the initial cycle parasitic bipolar disturb resulting from the aggregate effect of unselected cells in the same bitline was reduced [27].

The gate-to-body tunneling current increases the disparity between the 1st switch and 2nd switch. As shown in Fig. 3(a), for 1st switch with input initially at "Low", the body of nMOS sits at a diode cut-in voltage. So, there is a "small" negative bias across the gate and the body, resulting in a "small" body-to-gate tunneling current to discharge the body and therefore lower pre-switch body voltage. For the 2nd switch with the input initially at "High", the body sits at "Ground", and there is full $V_{DD}$ across the gate and the body, resulting in a "large" gate-to-body tunneling current to charge up the body and therefore higher pre-switch body voltage [21, 25].

The gate-to-body tunneling current can also cause (or "aid") full-depletion of the body when the device is in accumulation mode (such as in pass-gate configuration with source and drain at "High" and gate at "Low") [28]. In accumulation mode, the gate-to-body tunneling current flows from the body to the gate, thus discharging the body. This extra body discharging current can potentially result in full-depletion of the body with thin Si film, causing situations similar to the "quasi-depletion" discussed previously.

7. Self-Heating

The thermal conductivity of the buried oxide is about 2 orders of magnitude lower than that of Si, resulting in local self-heating of SOI devices. This is particularly a concern for devices that are on most or all the time (e.g. biasing elements, current source, current mirror, bleeder, etc.), and for circuits with high duty cycle and slow skew rate (such as clock distribution, I/O driver) [3]. If the device channel is considered as a heat source, the bell-shaped spatial temperature distribution due to local self-heating has a characteristic width determined by the thermal diffusion length in silicon $(\alpha r)^{1/2}$, which is a measure of the length over which the transient temperature fluctuations are significant, where $\alpha$ is the thermal diffusivity of silicon. $\tau$ is the clock period. It is typically in the sub-0.5 μm range. For a multi-finger device within the same body, the spatial temperature distributions due to individual active finger overlap each other (local self-heating affecting nearby neighbors). This close thermal coupling among nearby fingers increases the effective thermal resistance, resulting in much more severe self-heating than that predicted for a single isolated device. Fig. 5 shows the temperature rise in the device junction and at 31 based on a detailed 3D thermal analysis for a 0.18 μm, $L_{eff} = 0.10 \mu m$ PD/SOI technology with tungsten local interconnect and 7 layers of Cu interconnect [29]. Notice that as the number of active fingers increases, the temperature rise (or equivalently, the thermal resistance) increases. The increase saturates at about 9 active fingers, where the temperature rise is about 3 times that for the single finger. Notice that this saturation occurs when the distance between the center finger and the far-away finger is on the order of the characteristic width of the bell-shaped spatial temperature distribution. For scaled technology with smaller and tighter groundrules, the saturation will occur at higher number of active fingers and the increase in thermal resistance will be larger since more fingers will be thermally coupled. This large increase in thermal resistance is particularly important for clock driver, line drive, and I/O drivers, where large multi-finger devices are typically used.

Scaling/thinning of the Si film degrades the thermal conductivity and increases the thermal resistance due to phonon-boundary scattering. Fig. 6 shows the thermal resistance as a function of Si film thickness with the buried oxide thickness as a parameter [30]. Notice that the increase is particularly significant for thinner Si film with thick buried oxide.

The self-heating also enhances the thermally-assisted impact ionization at low voltage as discussed early [19, 20].

8. Soft Error Rate (SER)

The α-generated charges in SOI device is substantially less than that in bulk CMOS device due to the presence of the buried oxide, and appreciable charge generation can only occur when an α-particle hits the channel region [3, 31]. However, the total charges collected at the cell storage node can be significantly higher than the α-generated charges due to the parasitic bipolar effect (Fig. 7(a)) [31, 32]. While scaling of the device reduces the charge generation volume, the $Q_{crit}$ is also reduced due to reduced capacitance at cell storage node and reduced $V_{DD}$. In properly scaled PD/SOI device, the parasitic bipolar gain is reduced, thus improving SER. Furthermore, scaling/thinning of Si film reduces the charge generation volume and the base-emitter (body-source) junction area of the parasitic bipolar transistor, thus improving SER as well (Fig. 7(b)) [32].

9. Strained-Si Channel on SOI

Strained-Si surface channel CMOS has recently emerged as a strong contender for future high-performance applications due to higher mobility and improved $I_{on}$ [33, 34]. The lattice mismatch between the Si channel and the underlying relaxed SiGe layer results in biaxial tensile strain, which reduces the intervalley scattering by increasing subband splitting and enhances carrier transport by reducing conductivity effective mass. Combining strained-Si channel and SOI (Fig. 8) complements the improved $I_{on}$ of strained-Si channel device with the benefit of SOI [35, 36]. However, there are quite a few design implications. The smaller bandgap of SiGe
layer causes heterostructural band offset, which reduces \(V_T\) and increases \(I_{pp}\). The mobility enhancement for nMOS and pMOS may be quite different due to device design and process integration constraints [33, 34], which may upset the \(\beta\) (n/p strength) ratio while migrating existing designs. The tensile strain is "biaxial," so mobility enhancement (therefore \(I_{on}\) improvement) are the same along X- and Y-axis. However, in some high density design (e.g., SRAM cell), "bent gates" at 45\(^\circ\) angle are sometimes used, which would result in disparity in mobility enhancement and \(I_{on}\) improvement. The SiGe layer has 7\% higher dielectric constant and 10\% lower built-in potential due the narrow bandgap, resulting in higher junction capacitance [33, 36]. Furthermore, higher body doping density could be needed to compensate for the \(V_T\) reduction, which further increases the junction capacitance. The thermal conductivity of the SiGe layer is about 15X lower than that for Si, thus aggravating the self-heating effect [33].

The presence of the SiGe layer also significantly affects the floating-body effect [36]. For 20\% Ge content, the bandgap is about 90\% of that for Si. This narrower bandgap results in a higher (~10X) intrinsic carrier density \(n_i\), and thus proportionally higher recombination current at the body-to-source junction. However, the narrower bandgap and higher dielectric constant of the SiGe layer, and the higher body doping to compensate for the lowered \(V_T\) caused by the band offset, give rise to larger band-to-band tunneling current and trapped-assisted tunneling current at the drain-to-body junction. The latter effect may overpower the increase in recombination current at the body-to-source junction, resulting in more significant floating-body effect.

10. Conclusions

We have discussed the implications and impact of device scaling on the circuit design of sub-0.1 \(\mu m\) PD/SOI CMOS circuits. The parasitic bipolar effect tends to be well contained in scaled devices. The thermally-assisted impact ionization at low-voltage need to be carefully considered for SOI devices due to self-heating effect. The requirement for higher \(V_{TR, \text{lin}}\) is alleviated due to reduce electric field induced impact ionization at low voltage and by proper device design. Scaling/thinning of Si film improves the performance, short channel roll-off, and SER, while aggravates the self-heating effect, degrades the quality and usefulness of body contact, and may cause dynamic depletion (quasi-depletion) of the body. The gate oxide tunneling leakage has emerged to become a serious concern, and gate-to-body tunneling current in PD/SOI devices has to be carefully considered for proper circuit operation and timing. Proper modeling and duly consideration of thermal resistance increase due to thermal coupling in multi-finger devices and in thin Si film are crucial for accurately predicting the self-heating effect. SER improves with proper scaled device design. Strained Si-channel on SOI improves the device mobility and current drive with new material properties, device design considerations, and circuit implications. Full understanding of these effects are crucial for robust circuit design to exploit the performance potential of scaled PD/SOI technology.

References


Fig. 6: Thermal resistance as a function of Si film thickness $d_{Si}$ with the buried oxide thickness $d_{ox}$ as a parameter [30].

Fig. 7: (a) Charge collection normalized to the generated charge for 0.18 μm PD/SOI devices. Values larger than 1.0 indicate more charge collected than generated (due to parasitic bipolar gain). $\theta$ is the incident angle with respect to surface normal; and (b) effect of Si film thinning on α-particle induced SER [32].

Fig. 8: Schematic cross-section of a strained-Si channel on SOI nMOSFET.