Design and CAD Challenges in sub-90nm CMOS Technologies

Kerry Bernstein, Ching-Te Chuang, Rajiv Joshi, Ruchir Puri
IBM Thomas J Watson Research Center, Yorktown Hts, NY
{kbernste,ctchuang,rvjoshi,ruchir}@us.ibm.com

ABSTRACT

This paper discusses design challenges of scaled CMOS circuits in sub-90nm technologies for high-performance digital applications. To continue scaling of the CMOS devices deep into sub-90nm technologies, fully depleted SOI, strained-Si on SiGe, FinFETs with double gate, and even further, three-dimensional circuits will be utilized to design high-performance circuits. We will discuss unique design aspects and issues resulting from this scaling such as gate-to-body tunneling, self-heating, reliability issues, and process variations. As the scaling approaches various physical limits, new SOI design issues such as Vt modulation due to leakage, low-voltage impact ionization, and higher $V_{t,lin}$ to maintain adequate $V_{t,sat}$, continue to surface. With an eye towards the future, design and CAD issues related to sub-65nm device structures such as double gate FinFET will be discussed.

1 INTRODUCTION

Scaling the conventional MOSFET beyond the 90nm technology node requires innovations to circumvent barriers due to the fundamental physics that constrains the conventional MOSFET. The limits most often cited are: quantum-mechanical tunneling of carriers through thin gate oxide, from drain to source, and from drain to body; control of the density and location of dopants to provide high I_{on}/I_{off} ratio; and finite subthreshold slope. These fundamental limits have lead to the pessimistic predictions of the imminent end of technological progress in semiconductor industry. However, the push to scale conventional MOSFET has continued to show remarkable progress. Continued scaling and demand for performance are pushing for lower supply voltage and V_t, shorter channel length, thinner gate oxide, higher body doping concentration, and thinner Si film thickness. In addition, new materials such as strained-Si channel on relaxed SiGe layer are on the semiconductor roadmap to enhance the mobility and current drive. New device structures such as double-gate FinFETs and 3D circuits are being aggressively pursued for 65nm technology and beyond. Such aggressive scaling and new device structures give rise to several unique design issues which must be dealt with before any of these technologies will gain mainstream acceptance.

SOI technology has been widely accepted for use in mainstream high-performance logic applications in sub-90nm technologies. The design issues resulting from the floating-body in partially depleted SOI device structure, such as parasitic bipolar effect and hysteretic V_t variation, are now well-understood and circuit/design techniques to mitigate these effects have been developed. However, as the scaling approaches various physical limits, unique/new SOI design issues continue to evolve/surface.

In this paper, we discuss the design challenges of sub-90nm CMOS circuits with particular emphasis on the implications and impact of each individual device scaling element on circuit design.

2 CMOS DEVICE SCALING AND NEW DEVICES

In CMOS technologies below 90nm, the Field Effect Transistor will remain the fundamental design element of choice for the high speed logic designer. Changes to this device will be evolutionary rather than revolutionary. Improvements will be aimed at increasing mobility and decreasing extrinsic capacitance and resistance. Because a number of factors influence these parameters, multiple mechanisms and structures are proposed for extending CMOS below 90nm. A subset of them are described below.

2.1 Partially Depleted Silicon-on-Insulator

The Partially-depleted floating-body MOSFET was the first SOI transistor adopted for general high speed use, due in part to its processing similarities to common Bulk CMOS. The PD-SOI device is largely identical to the bulk device, except for the addition of a buried oxide ("BOX") layer, isolating the body of the given device from the bodies of other devices. Isolated below by BOX, to the left and right by diffusion junction diodes, and above by gate insulator, the body's potential "floats"; this voltage bias is influenced by both static and dynamic effects. SOI offers three main contributions to improved performance: (1) reduced junction capacitance, (2) lower average threshold, and (3) reduced body effect/source follower action. This performance is at the cost of some design complexity asserted by the floating body of the device, namely parasitic bipolar action and variable drive strength. The reader can reference [1, 2, 3] for a more thorough treatment of these effects. In addition, a number of CAD solutions have been developed to mitigate these problems [4, 6].

2.2 Fully Depleted Silicon-on-Insulator

In order to maintain balance of electric fields and capacitance within the SOI device, the active silicon thickness must be reduced with scaling. In the PD-SOI MOSFET, a "quasi-neutral region" existed deep within the body which remained un-inverted even as the given device went into saturation. A number of device structure alterations have been asserted to preserve and extend this partial depletion. As active silicon scales thinner, however, the depletion region within the body must expand laterally to satisfy the call for minority carriers at the gate oxide interface, to the point where the depletion region completely encompasses or consumes the body. In this *fully depleted* FET, the threshold of the device is now defined by the amount of charge obtainable within the isolated body. Other novel devices are by their nature intrinsically fully depleted, as we will see shortly.

2.3 Strained Silicon Channel

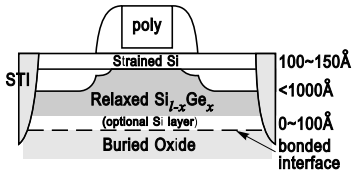

Figure 1: Schematic cross-section of a strained-Si nMOSFET.

Inducing tension on silicon substrates has been known to improve carrier mobility [7]. Specifically, tensile strain lifts the six-fold degeneracy in Silicon's conduction band. Of these six conduction sub-bands, the D2 and D4 band splits are exaggerated when under tension, which suppresses carrier scattering and enhances mobility and drive current. Tensility has been achieved through various means; mechanical strain and epitaxial strain are two popular approaches. In one approach of mechanical strain, processing of the passivation nitride covering the device is altered to cause strain. In the epitaxial approach, a thin pseudomorphic silicon is

ICCAD'03, November 11-13, 2003, San Jose, California, USA.

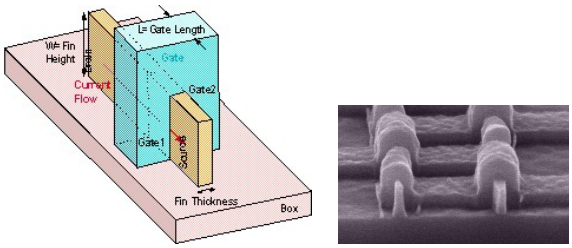

Figure 2: Schematic illustration of a FinFET; inset is a SEM micrograph of a FinFET.

grown on a SiGe substrate, causing lattice strain. Both techniques are capable of improving current drive by perhaps 25% at saturation, and by perhaps 50% in linear mode. Figure 1 shows the schematic cross-section of a strained-Si channel MOSFET.

2.4 Double-Gate MOSFETs

The double-gate (DG) MOSFET offers an appealing but temporary reprieve to scaling. The most popular realization of the DG device is the "FinFET", initially advanced by work published by University of California - Berkeley [8]. The FinFET is a non-planar, fully depleted, double-gate device built upon an insulating oxide layer; its structure is shown in Figure 2. These devices achieve superior drive current through (a) effective suppression of short-channel effect, (b) near-ideal subthreshold swing, and (c) an improved body factor. Further, this device is still eligible for the same material changes advocated for planar devices, namely metal gates and high-permittivity gate dielectrics. The FinFET carries with it, however, a new complication: device-width quantization. In planar devices, the only device width quanta was asserted by the grid step size in the design database used. In FinFET technology, device widths are dispensed in units of whole fins only.

In the following section, we discuss some of the major design issues which will significant influence on circuits in sub-90nm CMOS technologies.

3 MAJOR DESIGN ISSUES

Technologies of the deep-submicron era have responded to the non-scalability of threshold voltages by accommodating higher and higher static IDD currents. For a while, this worked, but now a number of problems arising from continued scaling confront the designer. These issues include excessive power dissipation density, gate oxide tunneling current, self heating of the device and interconnect, and a host of subsequent reliability issues. These issues are reviewed below.

3.1 Gate-to-Body Tunneling/Leakage Current

As the gate oxide thickness is scaled to maintain gate control, V_t, and performance, the oxide tunneling leakage increases (Figure 3) [9, 10]. 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 nitrided 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" [9].

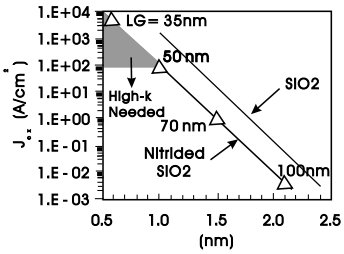

Figure 3: Gate leakage dependence on physically effective oxide thickness for pure and nitrided oxide [9].

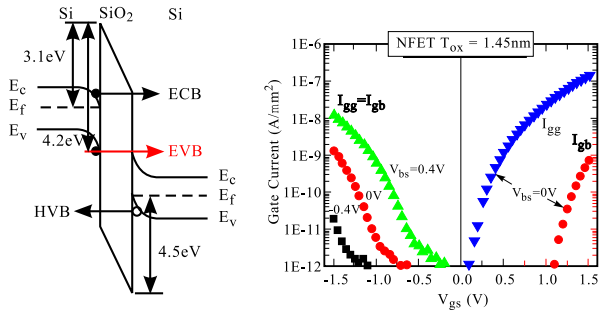

Figure 4: (a) Different tunneling components in a Si/SiO_2/Si structure: electron tunneling from conduction band (ECB), electron tunneling from valance band (EVB), and hole tunneling from valance band (HVB) [10], and (b) Measured nMOS gate current characteristics. I_{gg} is the total gate current, I_{gb} is the gate-to-body tunneling current [11].

The oxide tunneling current consists of several components as shown in Figure 4(a) [10]. The electron tunneling from the valance 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 (Figure 4(b)), and its effect can usually be neglected in bulk CMOS. In partially depleted SOI devices, however, this substrate (body) current charges /discharges the body, thus changing V_t and affecting circuit operation [11]. 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 [12]. Also notice that as the channel length is scaled, the gate-to-source and gate-to-drain tunneling currents become increasingly significant due to the facts (1) the gate-source/drain overlap regions constitute a larger portion of total gate length, and (2) the source/drain and extensions are heavily doped, hence all the applied voltage drops across the gate oxide.

Figure 5(a) shows the change of individual device strength in a PD-SOI 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 μm, t_{OX} = 2.3 nm, and t_{Si} = 150 nm. In the same technology, it causes from 4% slow-down to 6% speed-up at 85 ^{o}C in the critical path delays of a 1.1 GHz, 115 W, 170 million transistor, 64b Power4 PowerPC microprocessor (Figure 5(b)) [12].

For pass-transistor based circuits, studies based on a 1.2 V, 0.13 μm PD-SOI technology with L_{poly}= 0.075 μm, physical t_{OX} = 1.5 nm, t_{Si} = 120 nm, and t_{BOX} = 145 nm indicated that the gate-to-body tunneling current can cause delay changes up to 11% - over 13% at room temperature for single-ended and dual-rail CPL circuits [13]. In the same technology, standard CMOS latches exhibit delay changes up to 6% - 7%, while the first cycle latch set-up time

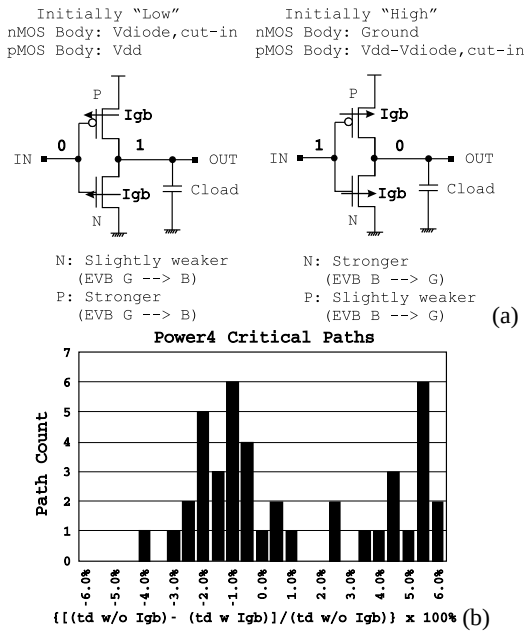

Figure 5: (a) A partially depleted SOI CMOS inverter in quiescent state with input initially at "Low" and "High", respectively, and (b) path counts vs percent changes of path delays due to I_{gb} at $85\,^{o}$C for critical paths in Power4 microprocessor [12].

for a master-slace latch pair can change up to 12% - 20% [14]. The results clearly indicated that the gate-to-body tunneling current has to be carefully accounted for in the timing of scaled SOI CMOS circuits.

Figure 6 depicts the changes in the strength of cell transistors in the quiescent state of a 6T PD-SOI 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 [15].

The gate-to-body tunneling current increases the disparity between the 1st switch and 2nd switch. As shown in Figure 5(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 [11, 12]. The gate-to-body tunneling current can also cause (or "aid") the full-depletion of the body when the device is in accumulation mode (such as in pass-gate configuration with the source and drain at "High" and gate at "Low") [16]. 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 in devices with thin Si film, causing situations of "quasi-depletion".

Excessive gate tunneling also has changed a standing paradigm in power usage. In performance-sorting hardware at the tester, it is expected that the slowest hardware has the longest channels. With low long-channel subthreshold leakage, the slow-sort chips exhibited the lowest static power, but relatively high dynamic power at a fixed frequency (due to larger gate area). With gate tunneling now contributing approximately half of total static power consumption, the slow performance sort hardware now also has high static power. Unfortunately, this high static power from gate tunneling is relatively unresponsive to temperature, eliminating a means formerly available to control leakage. Elevating VDD has also been a powerful means of enriching fast bin performance sorted hardware, but here again the thinner gate oxide becomes a problem; gate oxide intolerance of overvoltage reduces the amount of performance available by supply boost. The implication for products in sub-90nm technologies is that the optimum device design point is much more strongly dependent on application requirements. In some cases, a design is better served by using a previous device generation at higher voltages and / or shorter channels.

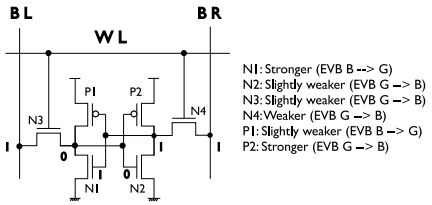

Figure 6: A SOI CMOS SRAM cell in quiescent state. The presence of gate-to-body tunneling current changes the strength of cell transistors [15].

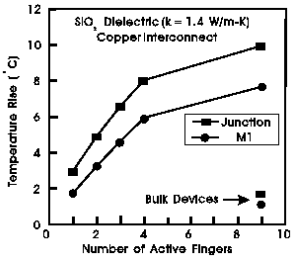

Figure 7: Temperature rise in device junction and at M1 as a function of number of active fingers in a multi-finger devices [17].

3.2 Self Heating of Device and Interconnect

The reduced thermal conductivity of low-K dielectric materials in the interconnect in addition to over two orders of magnitude lower thermal conductivity of the buried oxide layer in SOI devices, results in local self-heating of 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 slew rate (such as clock distribution, I/O driver) [1]. 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 ($(\alpha\tau)^{1/2}$, which is a measure of the length over which the transient temperature fluctuations are significant, where α is the thermal diffusivity of silicon. τ is the clock period. It is typically in the sub-50 nm 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.

Figure 7 shows the temperature rise in the device junction and at M1 based on a detailed 3D thermal analysis for a 0.18 μm, L_{eff} = 0.10 μm SOI technology with tungsten local interconnect and 7 layers of Cu interconnect [17]. Notice that as the number of active finger 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 sin-

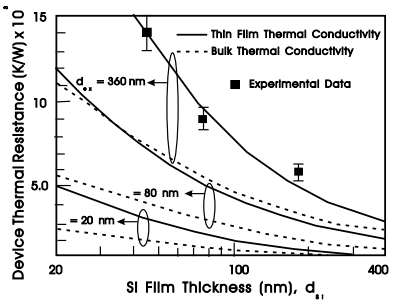

Figure 8: Thermal resistance as a function of Si film thickness d_{Si} with the buried oxide thickness d_{ox} as a parameter [18].

gle 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 occurs 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. Figure 8 shows the thermal resistance as a function of Si film thickness with the buried oxide thickness as a parameter [18]. Notice that the increase is particularly significant for thinner Si film with thick buried oxide.

The low thermal conductivity of enhanced-permittivity dielectrics also causes problems with the interconnect's performance and reliability. The interconnect's Thermal Coefficient of Resistance ("TCR") can range anywhere from 0.1 to 1% per degree C. This resistance change can be an issue in high speed or long distance busses. As power dissipated in the wire as well as in the FEOL increases, average temperatures in the wire should be anticipated when calculating RC delay. The reliability implications are reviewed in the next section.

3.3 Reliability Issues

Sub-90nm technologies do not introduce new reliability exposures, but instead show mechanisms that have evolved with reduced film thicknesses and structure dimensions. Wear-out mechanisms of concern include Hot Carrier Injection (HCI), Negative Bias Threshold Instability (NBTI), Time-Dependent Dielectric Breakdown (TDDB) in the front-end-of line and Electromigration (EM) in the back-end-of-line; advanced models and CAD checking tools have been developed to anticipate the magnitude of these effects specific to the application conditions of a given product. HCI, or commonly "Hot electrons" occur when the lateral electric field in a MOSFET elevates the energy of minority carriers above the ionization potential of a silicon lattice atom, observed mainly in N-MOSFETS [19]. Impact Ionization generates electron-hole pairs; of the pairs that do not spontaneously recombine, these free charges can alter the threshold of the device over time, either by remaining resident, creating negative trap sites in the gate oxide, or by accumulating positive charge in the SOI floating body. The so-called "prompt shift" observed in devices at stress is actually HCI occurring out at the gate-overlap region of the rain. Hot carrier effects are diminishing in sub-90nm technologies due to the reduction in operating voltage. NBTI has become a first-order concern in CMOS [22]. Affecting predominantly the P-MOSFET, NBTI is thought to be caused by moisture originating in photoresist materials becoming trapped below the device's nitride layer, where operating conditions such as high voltage and/or temperature can ionize it. This mobile charge has the ability over time to increase the absolute value of P-MOSFET threshold. As overdrive has become more precious with voltage reduction, the impact of NBTI has become profound. While incompletely understood, NBTI has been thoroughly characterized and modeled, allowing designers to anticipate long term performance impacts. As gate insulator thicknesses have been reduced to maintain influence over channel inversion, its ability to maintain high dielectric integrity has been diminished. Failure of thicker gate insulators (2.5nm and above) tended to be sudden in onset, and catastrophic. The advent of thin gate insulators, however, has presented a new failure mode, *soft breakdown*. [20]. High electric fields sustained over extended periods of time ages the dielectric, causing steadily increasing time-dependent breakdown currents. At a cross-over point, the breakdown reverts to the more traditional *hard breakdown*. It is believed that sustained fields slowly establish defect centers in the gate insulator. Given that these insulators are now less than 6 atoms thick, defect tolerance is understandably low. The addition of other species in the SiO2 to mitigate *direct tunneling* currents is known to modulate these effects. TDDB currents are at least 2 decades higher than, and quite different from, typical tunneling currents. TDDB is also accurately captured in reliability lifetime modeling software. Electromigration effects in sub-90nm CMOS are exacerbated by the rising average temperatures in CMOS, and by the inferior thermal generations, except that scaling has induced dramatic increases in instantaneous current densities and joule heating. EM is well modeled and can be checked as a chip release criteria. Other reliability mechanisms are more closely associated with the products application environment. Gate oxide damage from Electrostatic Discharge (ESD) during component handling, assembly, or system malfunction can cause failure modes post-testing that are not discovered until the part is in the field. Soft Errors induced by materials (alpha particles) or extraterrestrially (cosmic rays) used to be confined to upsets of the 6-device SRAM memory cell. In sub-90nm technologies, logic circuitry and latches are becoming increasingly vulnerable. CAD solutions for addressing SER in logic are only now becoming available, and only can address a portion of this complex problem.

Precision of on-chip parametrics is clearly compromised with scaling, and requires designer vigilance to assure functionality. *Gate oxide thickness, device channel length, threshold voltage*, and *overlap capacitance* are among the important parameters influencing delay variability which are growing harder to control as CMOS approaches quantum-mechanical boundaries. A number of excellent references address contributors to delay variation, and circuit design means to mitigate the effects [21].

4 DESIGN ISSUES SPECIFIC TO NEW DEVICES

In this section, we will discuss design issues that are specific to partially depleted SOI, fully depleted SOI, strained-Si devices, and double-gate MOSFETs device structures.

4.1 Partially Depleted SOI

With continued scaling of partially depleted SOI devices, new design issues such as Vt modulation due to leakage, low-voltage impact ionization, and higher $V_{t,lin}$ to maintain adequate $V_{t,sat}$, continue to surface. These issues and their impact on circuits designs is discussed as follows.

4.1.1 Parasitic Bipolar Effect, Reduced-V_t Leakage

Certain circuit topologies, such as stacked devices, pass-gate, and SRAM bitline structure, are susceptible to the parasitic bipolar effect [23, 24, 1, 25, 26]. 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 [25, 26, 2, 27, 28, 29].

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.

4.1.2 Hysteretic V_t Variation and Low-Voltage Impact Ionization

The hysteretic 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 the switching cycles has long been the most challenging task in the design of floating-body SOI CMOS circuits [25, 2, 30, 31, 32, 33, 34, 35, 36] Various body voltage estimation and bounding schemes have been developed for circuit simulation and static timing [37, 4].

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 ($E_g \sim 1.2$ eV), and the underlying low voltage ionization mechanism could not be explained by the conventional wisdom [38, 39]. 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 [38, 39]. This thermally assisted impact ionization mechanism at low voltage is particularly important in scaled 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 hysteretic V_t variation, which will be discussed later. Furthermore, high doping concentration and steep doping gradient in scaled devices increases the reversed-biased band-to-band junction tunneling current between the drain and the body, resulting in higher body charging current.

One of the commonly used gauge for hysteretic 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 [2, 33, 11]. 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 determined 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 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.

4.1.3 Higher $V_{t,lin}$

SOI devices are typically designed with a larger $V_{t,lin}$ (threshold voltage at low drain bias) compared with bulk CMOS [40]. This is because as the drain voltage is raised, the floating-body effect causes the threshold voltage to decrease, resulting in significantly lower $V_{t,sat}$ (threshold voltage at high drain bias). Thus, higher $V_{t,lin}$ is necessary to maintain adequate $V_{t,sat}$ 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 passing 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 $V_{t,sat}$ 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 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 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 [40].

4.2 Scaling of Si-Film: from PD-SOI to FD-SOI

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 volumn. 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 [11].

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-to-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.

4.3 Strained Silicon Channel

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} [41, 7]. 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 (Figure 1) complements the improved I_{on} of strained-Si channel device with the benefit of SOI [42, 43] However, there are quite a few design implications. The narrower bandgap of SiGe layer causes heterostructural band offset, which reduces V_t and increases I_{off}. The mobility enhancement for nMOS and pMOS may be quite different due to device design and process integration constraints [41, 7], which may upset the β (p/n 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^{o} 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 narrower bandgap, resulting in higher junction capacitance [41, 43]. 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 [41].

The presence of the SiGe layer also significantly affects the floating-body effect [43]. For 20% Ge content, the bandgap is about 90% of that for Si. This narrower bandgap results in a higher ($\sim$ 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.

4.4 Double-Gate MOSFETs

Although the FinFET is still a CMOS transistor, its physical realization requires design accommodation. Section 2.4 alluded to the quantized nature of non-planar double gated MOSFETs. The device width quantum for the FinFET is the height H of the fin. Each fin provides 2H of device width. Designers in planar technologies have been relatively unconstrained in selected device widths, such that appropriate ratios of drive strength in N-MOSFET and P-MOSFET devices will achieve desired trade-offs in performance, power consumption, and noise immunity. To achieve comparable flexibility using FinFETs, more fins of potentially longer channel lengths may be needed to achieve a given beta ratio. If not monitored, active and standby power can be sub-optimal. Conversely, selected functions such as digital delay lines or programmable current sources may exploit the fixed increments that fins enable. Restrictions in device orientation to reduce across-chip linewidth variation are becoming common in VLSI, and the fin enjoys those same benefits. By orienting fin channels in parallel, image clock distribution and power conventions comparable to those in planar devices may be realized.

5 CAD Challenges and Opportunities

Design and CAD issues related to both partially and fully depleted SOI circuits have been discussed at length in the literature [2][6]. In addition, recent research has also focused on the issue of gate-leakage reduction [5]. However the CAD challenges arising with the new device structures such as double-gate MOSFETs in sub-65nm circuit technologies has not been discussed. In this section, we focus on the impact of double-gate MOSFETs on Design automation tools.

Design automation has an important role in the introduction of FinFETs, and also must address new complexities. Dual-gated devices differ significantly from their traditional CMOS device counterparts by providing two gate controls for the same channel. Thus, they provide approximately twice the drive current but also present twice the capacitance to previous stage. If the two gates are always tied together or the back gate is used as a bias terminal for modulating device threshold, the impact on DA tools would be no more severe than the introduction of a typical scaled technology with multiple threshold voltages. The major innovation that will cause DA tools to become inadequate is the use of two gates of double-gate MOSFET to be controlled by independent signals. In that case, a single device can provide an OR function and many of the DA tools would be impacted. In the following, we discuss the impact of double-gate MOSFET technologies on various tools starting from fron-end synthesis tool.

Synthesis: Standard-Cell synthesis will not be impacted extensively since it does not require selection of transistor topologies which are preselected in the standard-cell library. However, transistor level synthesis must be aware of the flexibility of dual gated devices in order to exploit functionality of both the gates for implementing single device OR function. Since the threshold voltage in double-gate devices can be dynamically varied by changing the back gate signal, this presents a new opportunity for logic synthesis to exploit the performance vs power consumption tradeoff. For new designs, synthesis of logical functions from higher level descriptions may be performed in FinFET technologies, but with appropriate accommodation for quantization.

Placement and Routing: Because of the ability to perform OR-ing of two signals with a single device, CMOS cells may not have an equal number of P and N devices. Thus the cells may become non-rectangular. Effective block placement should have the ability to handle the new shapes. The impact on routing will probably not be severe. The ability to provide two signals to the same device may place more constraints on the location of cell pins.

Simulation and Tuning: Transistor level simulation must consider an additional gate terminal of double-gate devices. Tuning of existing layout topologies will probably not be impacted severely, but tuning of schematics will have to consider the new option of alternate ways of fingering an OR device.

Layout and Migration: Creating new cell layouts with dual gated devices will present new challenges. However, since the trend is toward fewer cell topologies (e.g., primitive gates such as NAND2, NAND3, NOR2, simple AOIs etc), manual design should be able to address the new challenges of how to create rectangular cells and how to utilize the additional interconnect capability provided by the back gate poly. Migration from a single gate technology to a dual gate technology will present considerable challenges since it violates one of the premises of migration, namely the preservation of topology. The conversion of planar designs into discrete fins may be achieved "behind the covers", for the most part transparent to the designer. Transistor layout conversion and round off to the closest number of Fins may be automated. This capability will be important for designs migrated into FinFET technologies.

Checking: Extraction must recognize new types of devices defined with two gates. A new level of interconnect must be handled and back and front gate signal connectivity must be determined.

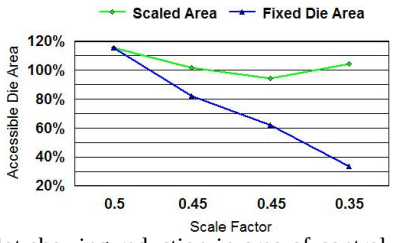

Figure 9: Plot showing reduction in area of control with scaling. Cross-chip latency with absolute die size held constant is observed to increase each generation.

LP checking must recognize OR-ing that occurs within a device as logically equivalent to OR-ing that occurs between parallel devices. Ground rule checking requires rules for the new layer, but should not present difficulties unless the new rules require a new type not supported in current technologies.

6 EMERGING TECHNOLOGIES

Both transistor and interconnect are confronting fundamental limitations which challenge conventional scaling. Determination of next-generation dielectric integrity, power dissipation density, delay variability, and manufacturing defects is becoming dominated by quantum-mechanical considerations, obviating the need to explore alternative paradigms. Two potential future transistor and interconnect technologies are described.

6.1 3-Dimensional Integration and Interconnects

Generation over generation, leveraging the improvement of the MOSFET transistor is increasingly compromised by interconnect impedance. It has only been through the heroic contributions of back-end-of-line process engineers that wire delays have even closely kept up with device improvements. The primary means have been through (a) the continued introduction in every generation of advanced materials (copper wires, low K dielectrics, and better liners), (b) the scaling of wire dimensions and (c) addition of more wiring levels. Despite these efforts, impedance in the back end of line has reduced the span of control [44]; chip crossing latencies, once under a cycle long are at approximately 4 cycles now, and projected to grow to 20 cycles within 3 years. Figure 9 below shows that if die size was scaled (nothing added from generation to generation), latency would have remained roughly constant. The introduction of architectural throughput enhancements such as out-of-order execution and speculative execution has caused die to grow however, as illustrated in the plot in which absolute die size remained constant. In this case, the area of the die accessible in one cycle decreases. This degrade is caused by many factors; non-scaling of the wire liner thickness, roll-off in effectiveness of added wiring levels, and inductive effects at high speeds are three examples. The case then for 3-dimensional integration of VLSI logic and memory is compelling. Potential benefits include:

- Increased device density
- Reduced wire delay
- Access to a greater number of devices in a fixed cycle time
- More effective use of capacitive gain
- Less lateral coupling noise issues
- Less power spent in interconnect
- Ability for integration of incompatible wafer processes

By adding a third degree of freedom in the placement and partitioning of logical elements, the electrical wire length distribution for the implementation of a given function may be reduced, allowing more of the MOSFET's performance advantage to be exploited. A goal of 3D integration is impedance reduction. A demonstrated technique to implement 3D integration calls for the independent fabrication of each layer followed by handle-wafer removal and bonding [45]. Figure 10 illustrates the technique. Multiple technologies must converge to realize this structure in manufacturing:

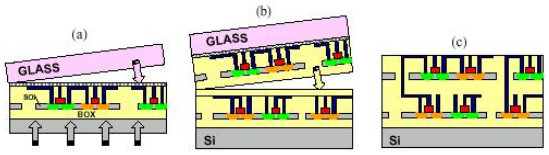

Figure 10: Layer transfer process for 3D integration : (a) Glass handle wafer secures first functional device layer and substrate is removed. (b) Independent layers are aligned, bonded. (c) Original handle wafer is removed, vertical interconnects formed [45].

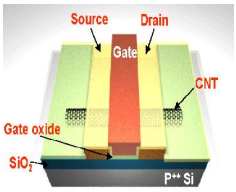

Figure 11: Diagram representation of the Carbon Nanotube NFET.

- Stacked substrate alignment capability
- Temperature management via electrical or mechanical means
- 3D Modeling of device and interconnect, CAD productivity
- Defect management, rework capability
- Manufacturing cost

3D integration may be achieved with existing processes and tooling, with only minor additional innovation necessary. New advances needed before this concept is embraced by the industry reside in the management of manufacturing cost and defect density.

6.2 Molecular Computing

The development of small switches exhibiting gain represents one of the most exciting frontiers in our industry. Building transistors at molecular scales has long been a goal in the drive to achieve biological computation efficiency. The value of such a system is in its ability to respond to each of the challenges presented by scaling - power, performance, area, reliability, and cost. A number of organic molecules have been examined [46], and much of the work in nano-electronics and micro-electronic-machining (MEMS) is applicable. The Carbon nanotube Field-effect-Transistor (CNFET) is the farthest along of any molecular device. Carbon, when properly formed into self-closing tubes, presents field-effect behavior, and exhibits gain, comparable or superior to state-of-the-art Silicon-based MOSFETs. Figure 11 shows a schematic representation of a CNFET As expected there are substantial hurdles which first need to be overcome.

- Diameter and Chirality of the nanotube must be tightly controlled. Bandgap of the nanube is a function of its diameter.
- Nanotube semiconductors need to be differentiated from nanotube metals, which form regularly during processing.
- Nanotubes like to grow in bundles. Separation and distribution of nanotubes is a key requirement
- CNT placement and device connectivity must be determined by the designer. A number of techniques have been proposed. Proposed devices using other organic molecules have advocated "self-organizing systems" to overcome the question of how to assert structure and discipline into these devices.
- Interconnects still remain to be defined between CNTs. The contact to a CNT is a Schottky barrier diode, so innovative techniques will be needed to achieve good level transfer.

Along with common scaling, designer productivity has had to also increase in order to harness the larger transistor count made available. At the CNFET count which may some day be integrated on a common substrate, CAD productivity tools will be an essential enablement.

7 Conclusion

We have discussed the implications and impact of device scaling on the circuit design of sub-90nm CMOS circuits. Major design issues such as gate-to-body tunneling, self-heating, reliability issues, and process variations were discussed. The gate oxide tunneling leakage has emerged to become a serious concern and has to be carefully considered for proper circuit operation and timing. Proper modeling and 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. New SOI design issues such as Vt modulation due to leakage, low-voltage impact ionization, and higher $V_{t,lin}$ to maintain adequate $V_{t,sat}$, that will surface due to device scaling were also discussed. Strained Si-channel on SOI improves the device mobility and current drive with new material properties, device design considerations, and circuit implications. Looking beyond 65nm technologies, design and CAD issues related to device structures such as double gate FinFETs were also discussed and some emerging trends were outlined.

References

[1] C. T. Chuang et al. "SOI for digital CMOS VLSI: design considerations and advances," Proc. IEEE, vol. 86, no. 4, April 1998.

[2] C. T. Chuang and R. Puri, "SOI digital CMOS VLSI - a design perspective," Design Automation Conf., 1999, pp. 709-714.

[3] K. Bernstein, et al, "SOI Circuit Design Concepts", Kluwer Academic Publishers, February, 2000.

[4] K. L. Shepard et al., "Body-voltage estimation in digital PD-SOI circuits and its application to static timing analysis," ICCAD, 1999.

[5] D. Lee, et al., "Analysis and Minimization Techniques for total Leakage Considering Gate Oxide Leakage," Design Automation Conf., 2003.

[6] S. R. Nassif et al., "SOI Technology and Tools," ICCAD, 1999.

[7] K Rim, et al., "Characteristics and device design of sub-100 nm strained Si N- and PMOSFETs," Symp. VLSI Technology, 2002, pp. 98-99.

[8] J. Kedzierski, et al., "High-performance symmetric-gate and CMOS-compatible Vt asymmetric-gate FinFET devices", IEDM, 2001, p. 19.5.1.

[9] T. Ghani, et al., "Scaling challenges and device design requirements for high performance sub-50 nm gate length planar CMOS transistors," Symp. VLSI Technology, 2000, pp. 174-175.

[10] W. C. Lee and C. Hu, "Modeling gate and substrate currents due to conduction- and valence-band electron and hole tunneling," Symp. VLSI Technology, 2000, pp. 198-199.

[11] S. K. H. Fung, et al., "Controlling floating-body effects for 0.13 μm and 0.10 μm SOI CMOS," IEDM, 2000, pp. 231-234.

[12] C. T. Chuang, R. Puri, and K. Bernstein, "Effect of gate-to-body tunneling current on PD/SOI CMOS circuits," Int'l Conf. on Solid State Devices and Materials, Tokyo, Japan, 2001, pp. 262-263.

[13] C. T. Chuang and R. Puri, "Effects of gate-to-body tunneling current on pass-transistor based PD/SOI CMOS circuits," Proc. IEEE Int'l SOI Conf., 2002, pp. 121-122.

[14] C. T. Chuang and R. Puri, "Effects of gate-to-body tunneling current on PD/SOI CMOS Latches," Submitted to Conference on Simulation of Semiconductor Processes and Devices (SISPAD'03), 2003.

[15] R. V. Joshi, et al., "Effects of gate-to-body tunneling current on PD/SOI CMOS SRAM," Symp. VLSI Technology, 2001, pp. 75-76.

[16] H. Wan, et al., "Tendency of full depletion due to gate tunneling current," Proc. IEEE Int'l SOI Conf., 2002, pp. 140-141.

[17] R. V. Joshi, et al., "3D thermal analysis for SOI and its impact on circuit performance," Proc. International Conf. on Simulation of Semiconductor Processes and Devices (SISPAD'01), 2001, pp. 242-245.

[18] M. Ashegi, et al., "Thermal conductivity model for thin silicon-on-insulator layers at high temperatures," Proc. IEEE Int'l SOI Conf., 2002, pp. 51-52.

[19] C. Hu, et al., "Hot-Electron-Induced MOSFET Degradation-Model, Monitor, and Improvement", IEEE Transactions on Electron Devices, Vol. ED32, No. 2, pp-375-382, February, 1985

[20] A.M. Yassine, et al., "Time dependent breakdown of ultrathin gate oxide", Electron Devices, IEEE Transactions on , Volume: 47 Issue: 7 , Jul 2000, Page(s): 1416 -1420

[21] K. Bernstein, et al, "High Speed Circuit Design Styles", Chapter 1, Kluwer Academic Publishers, January, 1998

[22] V. Reddy, et al., "Impact of negative bias temperature instability on digital circuit reliability", Reliability Physics Symposium Proceedings, 2002. 40th Annual , 2002, Page(s): 248 -254

[23] P. F. Lu, et al., "Floating body effects in partially-depleted SOI CMOS circuits," IEEE J. Solid-State Circuits, vol. 32, no. 8, August 1997, pp. 1241-1253.

[24] C. T. Chuang, et al., "Dual-mode parasitic bipolar effect in dynamic CVSL XOR circuit with floating-body partially-depleted SOI devices," Proc. Tech. Papers, Int'l Symp. on VLSI Tech., Syst., and Applications, Taipei, Taiwan, 1997, pp. 288-292.

[25] M. Canada, et al., "A 580MHz RISC microprocessor in SOI," ISSCC, 1999, pp. 430-431.

[26] D. H. Allen, et. al., "A 0.20 μm 1.8 V SOI 550MHz 64b PowerPC microprocessor with Cu interconnects," ISSCC, 1999, pp. 438-439.

[27] J. B. Kuang, et al., "A dynamic body discharge technique for SOI circuit applications," Proc. IEEE Int'l SOI Conf., 1999, pp. 77-78.

[28] J. B. Kuang, D. H. Allen, and C. T. Chuang, "Dynamic body charge modulation for sense amplifiers in partially depleted SOI technology," IEEE J. Solid-State Circuits, vol. 36, no. 4, April 2001, pp. 597-604.

[29] J. B. Kuang and C. T. Chuang, "A tri-state body charge modulated SOI sense amplifier," Proc. IEEE Int'l SOI Conf., 2001, pp. 135-136.

[30] A. Wei, et al., "Minimizing floating-body-induced threshold voltage variation in partially depleted SOI CMOS," IEEE Elec. Dev. letters, vol. 17, no. 8, Aug. 1996, pp. 391-394.

[31] A. Wei and D. Antoniadis, "Design methodology for minimizing hysteretic V_t-variation in partially-depleted SOI CMOS," IEDM, 1997, pp. 411-414.

[32] T. W. Houston and S. Unnikrishnan, "A guide to simulation of hysteretic gate delays based on physical understanding," Proc. IEEE Int'l SOI Conf., 1998, pp. 121-122.

[33] M. M. Pelella, et al., "Hysteresis in floating-body PD/SOI CMOS circuits," Symp. on VLSI Tech., Syst., and Applications, Taipei, Taiwan, 1999, pp. 278-281.

[34] R. Puri and C. T. Chuang, "Hysteresis effect in pass-transistor based partially-depleted SOI CMOS circuits," Proc. IEEE Int'l SOI Conf., 1998, pp. 103-104.

[35] R. Puri and C. T. Chuang, "Hysteresis effect in floating-body partially-depleted SOI CMOS domino circuits," Proc. Int'l Symp. on Low Power Electronics and Design, 1999, pp. 223-228.

[36] K. A. Jenkins, R. Puri, C. T. Chuang, and F. L. Pesavento, "Measurement of history effect in PD/SOI single-ended CPL circuit," Proc. IEEE Int'l SOI Conf., 2001, pp. 57-58.

[37] I. Aller and K. E. Kroell, "Detailed analysis of the gate delay variability in partially depleted SOI CMOS circuits," Proc. IEEE Int'l SOI Conf., 1999, pp. 40-41.

[38] P. Su, et al., "Self-heating enhanced impact ionization in SOI MOSFETs," Proc. IEEE Int'l SOI Conf., 2001, pp. 31-32.

[39] P. Su, et al., "An impact ionization model for SOI circuit simulation," Proc. IEEE Int'l SOI Conf., 2002, pp. 201-202.

[40] E. Leobandung, et al., "Scalability of SOI technology into 0.13 μm 1.2 V CMOS generation," IEDM, 1998, pp. 403-406.

[41] K. Rim, "Strained Si surface channel MOSFETs for high-performance CMOS technology," ISSCC, 2001, pp. 116-117.

[42] L. J. Huang, et al., "Carrier mobility enhancement in strained Si-on-insulator fabricated by wafer bonding," Symp. VLSI Technology, 2001, pp. 57-58.

[43] K. Kim, et al., "Performance assessment of scaled strained-Si channel-on-insulator (SSOI) CMOS device/circuit," Proc. IEEE Int'l SOI Conf., 2002, pp. 17-19.

[44] D. Matzke, "Will physical scalability sabotage performance gains?", Computer , Volume: 30 Issue: 9 , Sep 1997, Page(s): 37 -39

[45] K. Guarini, et al., "The Impact of Wafer Level Layer Transfer on High Performance Devices and Circuits for 3D IC Fabrication", ECS, Paris France, May 2003

[46] M. Reed, et al., "Computing with Molecules", Scientific American, June, 2000.