Design Challenges for High-Performance SOI Digital CMOS VLSI

C. T. Chuang
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, U. S. A.

Abstract—This paper reviews the recent advances of SOI for high-performance digital CMOS VLSI applications. The technology/device requirements and design issues/challenges for high-performance, general-purpose microprocessor applications are differentiated with respect to low-power portable applications. Particular emphases are placed on the impact of floating-body in partially-depleted devices on the circuit operation, stability, and functionality. Unique SOI design aspects such as parasitic bipolar effect and hysteretic $V_T$ variation are addressed. Circuit techniques to improve the noise immunity and global design issues are discussed.

I. INTRODUCTION

Silicon-on-insulator (SOI) technology has long been used in radiation-hardened and high-voltage applications. In the past few years, due to the impressive progress in materials, process/device, and product research/development, SOI technology has emerged as a strong contender for low-power portable applications. It is only until very recent years, however, that SOI has been seriously considered for mainstream high-performance for circuit design applications [1, 2]. The ever-demanding performance of general-purpose microprocessors dictates far more stringent technology/device requirements and poses more complicated design issues/challenges compared with low-power portable applications. The underlying reasons and the resulting divergence in the technology/device choice/design are discussed, and the design issues differentiated in Section II. We then focus on the impact of floating-body in partially-depleted devices on the circuit operation, stability, and functionality. Unique SOI design aspects such as parasitic bipolar effect and hysteretic $V_T$ variation, and circuit techniques to improve the noise immunity are addressed in Section III and IV. The conclusion of the paper is given in Section V.

II. LOW-POWER VS HIGH-PERFORMANCE

The slow appearance of SOI in mainstream digital CMOS VLSI applications is primarily due to the rapid progress and scaling of bulk CMOS technology. As the scaling of the bulk CMOS approaches the end of the roadmap (0.12 µm - 0.07 µm), the technology evolution de-elerates to a slower pace, and SOI stands a better chance of becoming the mainstream technology than before [1]. Beyond the limit of bulk CMOS scaling (50 nm - 25 nm), thin-body SOI and dual-gate SOI with raised source/drain and gate workfunction engineering hold the promise to be the "future" technology/device choices [1]. At present and in the near future, nevertheless, the primary SOI technology/device choice is between a fully-depleted (FD) device and partially-depleted (PD) devices. Use of fully-depleted device significantly reduces the floating-body effect. However, its large source/drain series resistance limits the performance. The process and device design are less compatible with the bulk CMOS. Its very low device $V_T$ with thin gate-oxide and large process sensitivity pose manufacturability concern and scalability limit. Partially-depleted device alleviates the constraint on the source/drain series resistance and $V_T$, offering higher performance and easing the manufacturing problem by allowing the doping profiles to be tailored for any desired $V_T$. The process and device design are also much more compatible with the bulk CMOS [3, 4, 5]. However, the floating-body in partially-depleted device, and the resulting parasitic bipolar effect and hysteretic $V_T$ variation represent serious design issues/challenges for circuit designers [2].

For low-power portable applications, power and cost are the primary considerations. There is less demand on performance and scalability, thus no need to use the most advanced technology. The transistor count is low, thus manufacturability and design resource are of less concern. The low-voltage operation alleviates the reliability issues and suppresses the parasitic bipolar effect. Highly regular circuit structures, such as the MAC (Multiply-Accumulate-Cell) in DSP applications, result in less demand on design methodology and resource. The choice of technology/device (fully-depleted vs partially-depleted) is less crucial. For these applications, decision to use the fully-depleted device tends to be based on its cost of design. When partially-depleted device is chosen, body contacts (for rail-to-rail suppress the floating body effect or for smart body contacts to dynamically control the threshold voltage) can be easily incorporated since there is much less area constraint. The operating frequency is low, so body contacts are easily effective. These application-specific features result in the prolific use of smart body contacts in CPL and SIMOX-MTCMOS type of circuits for ultra-low-power deep sub-1.0 V applications [6, 7].

High-performance general-purpose microprocessor applications place ultimate demand on performance and scalability. The state-of-the-art processors contain hundreds of millions of transistors. The performance requirement and considerations for manufacturability, scalability, and compatibility with the bulk CMOS device design and process strongly favor the partially-depleted device [8, 9]. On the other hand, circuit designers are forced to face the design complexity resulting from the floating-body. Wide variety of circuits (from static, dynamic pass-transistor based, to self-timed circuits) and design style (from full-custom, semi-custom, to synthesis) are employed in these microprocessors. Because of the area limitation, the use of body contacts have to be highly selective. Furthermore, for the body contact to be effective at the high operating frequency, low RC time constant is essential, yet difficult to achieve under the process/device constraints. Smart body contacts, if ever intended, are containable only in drivers, arrays, and some highly regular structures [2]. Global use of smart body contacts require sophisticated tool capability/development. Designs with floating body and/or smart body contact require detailed circuit characterisation different from the bulk-CMOS. Layouts with body contacts are incompatible with bulk-CMOS. Circuit simulation with floating body requires substantially more memory and simulation time. The problem is aggravated by the long time for the DC convergence of the initial solutions and the prohibitively long time to reach steady-state for large scale circuits. The timing methodology is complicated by the topology and history/pattern dependency. Finally, floating-body induced noise and globally-coupled noise have to be properly accounted for in the noise and checking methodology.

III. PARASITIC BIPOLAR EFFECT

Certain circuit topologies and switching patterns are susceptible to the parasitic bipolar effect, and logic state error can occur if the effect is not properly accounted for. Examples are: (1) stack OR-AND-like structure (e.g., dynamic OR, Fig. 1), (2) pass-transistor based design (e.g., wide multiplexer and generic 2-phase dynamic logic circuit), and (3) multi-level voltage-switch current-steering circuit (e.g., dynamic CVSL XOR circuit) [10, 11]. The topology typically involves a "off" transistor situated high in a stack or in pass-gate configuration, with the source and drain voltage set up in the "High" state (hence the body voltage at "High"). When the source is subsequently pulled down either by the clocked evaluation transistor (in dynamic circuit) or by the input signal (in pass-gate configuration), 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 current (caused by noise and aggravated by the lower $V_T$) result in a loss of charge on the precharge (or dynamic) node. This has been shown to cause failure in a dynamic adder circuit (Fig. 2) in a 580 MHz 32B RISC microprocessor (PowerPC 750) in a 0.12 µm $L_{MIN}$ SOI technology with Cu interconnect [8]. One can size up the keeper device (at the expense of few % in performance) or selectively dropping body contacts (at the expense of few % in area) to overcome the effect. Alternative implementation of a given function and circuit techniques have been used to improve the noise immu-
nity of dynamic circuits on SOI with minimal impact on their delays [9]. These techniques exploit the reduced charge sharing effect and reduced delay dependency on stack ordering in SOI technology. Examples are: (1) "pre-discharging" technique (Fig. 3) that discharges intermediate nodes in a stack so that the transistors high in the stack are prevented from charging to a high potential, (2) "re-arrange the pull-down tree" to position the widest parallel group of transistors at the bottom of the stack, (3) "re-ordering (cross-connecting) inputs" in multiple-fingered stacked transistors such that the signal connecting to the gate of the higher transistor in one stack connects to the lower transistor in the other stack, thus preventing one-half of the transistors from conducting parasitic bipolar current, (4) "early discharge" to force parasitic bipolar current to occur during the precharge phase, (5) "re-mapping" boolean logic to reduce the parallel stacks, (6) "complex domino" structures with the output inverter replaced by a static NAND or NOR gate to break up large parallel logic trees. During stress testing, "conditional" feedback, where a additional feedback half-latch is conditionally gated by the test signal, can be used to enhance the noise immunity. These techniques have been used to achieve robust operation in a 550 MHz 64b PowerPC microprocessor in a 0.12 μm L2∪/SOI technology [9].

For on-chip cache SRAM, the primary concerns due to the floating body effect are to control capacitance disparity, bit-line leakage during read operation, half-selected cell disturb during write operation, and the sense transistor Vth mismatch [2, 6]. In general, body controls should be used in the sense transistors to improve the sense margin. In well-designed state-of-the-art devices, the parasitic bipolar leakage can be suppressed to well below 10 μA/μm2 [3, 4]. Experimentally, the parasitic bipolar effect does not appear to increase as the channel length (hence "base width" of the parasitic bipolar transistor) is reduced since the transistor structure- and doping-wise is far from an ideal bipolar transistor. The effect decreases significantly with scaled supply.

IV. HYSTERETIC Vth VARIATION

Due to the long time constants for various body charging/discharging mechanisms (impact ionization current and junction leakage current), the body potential during the switching transient is determined primarily by the external biasing and capacitive coupling. Depending on the circuit topology, process/device design details, and the relative magnitude of the various charging/discharging mechanisms, the body may gain or lose charge through the switching cycle, causing the body voltage to drift until steady-state is reached. The hysteresis has been known to result in frequency-dependent delay variation and shown to cause pulses to stretch or shrink [12, 13] as they propagate down an inverter chain. Since the majority of CMOS circuits are inverting logic, the effect is pervasive.

The hysteretic delay variations of a basic inverter in a L2∪/SOI is 0.145 μm PD/SOI technology are illustrated in Fig. 4(a), where the output-rise-delay (TDrise), output-fall-delay (TDfall), and average delay (TDave, average of TDrise and TDfall) are shown as functions of time when the inverter is continuously cycled at 50% duty cycle with VDD = 1.8 V and T = 27°C [14]. Two cases are shown, corresponding to two different initial states (input at "Low" or "High") for a sustained period of time) for the circuit. The body-to-source voltages of the nMOS and pMOS device through the switching cycles are shown in Fig. 4(b). Notice that different initial states result in delay disparity at beginning of the switching activity, yet converge as the circuit approached the steady-state (since the steady-state is determined only by the net charges gained/lost through the switching cycle and is reached when the net charges gained/lost through the switching cycle equal to zero). A detailed study of the dependence of the hysteretic gate delay on the supply voltage, Vref/Vth, ratio, duty cycle, slew rate, output load, and initial state of the circuit is presented in Ref. [14].

Pass-transistor based circuits and circuits using stacked transistors are known to be of high performance in SOI due to the lack of reverse-body effect [2]. These "hysteresis" circuits, however, tend to exhibit highly-hysteretic behavior. The performance leverage comes about because the body is rarely reverse-biased (and actually tends to be forward-biased) with respect to the source. Since there are always biasing conditions or switching patterns that reduce the forward-bias of the body/source junction, the performance leverage comes hand-in-hand with large hysteretic delay variations. This is exemplified by the single-ended CML circuit (known as LEAP) shown in Fig. 5(a) [15], where the pass-transistor QA is initially set up with both its input (IN1) and output at "High" (thus its body at VDD), and its gate select signal (A) at "Low". The gate select signal A then switches to "High" and the data input IN1 is subsequently pulled down. As the gate select signal A switches "High", the body voltage of nMOS QA is capacitively coupled up. With both the source and drain of transistor QA at VDD, the inversion layer in the channel never forms, and the current coupling between the body and the channel is much reduced, completing the entire gate signal ramping. The body voltage is then boosted to 2.44 V, significantly above VDD (1.8 V), resulting in fast first switch. Fig. 5(b) shows the variation of body voltage and delays for "rising input" transition as functions of time, and with the frequency as a parameter. The period for which IN1 signal is "Low" remains constant at 1.0 ns in every switching cycle. Hence higher frequency (higher duty cycle) implies shorter duration for which input IN1 remains "High", and thus less time for the body to get charged (by the reverse-biased drain-to-body and source-to-body PN junction current, noting that both the drain and source will be at VDD after input IN1 rises). This effect can be clearly seen in Fig. 5(b), where the body voltage falls only slightly at 50 MHz over 200 MHz comparison, while the body voltage drops significantly when IN1 is switching at 50 MHz or 5 MHz. The hysteretic variation is particularly significant for the "rising input" delay due to the body effect in passing the "High" state. The "rising input" delay changes from 0.90 ns at t = 0 ns to 0.10 ns at t = 4,000 ns (a change of 44%). Use of a dual-rail CML with cross-coupled pMOS load devices (Fig. 6(a)) significantly reduces the delay variation (Fig. 6(b)) and its frequency dependence [15]. This is because the "failing" transition through the nMOS tree helps the rising transition of the complimentary nMOS tree through the cross-coupled pMOS. Thus, the "rising input" delay and its variation are compensated by the "falling input" of the other branch.

The hysteretic Vth variation affects the duty cycle and degrades the clock skew and jitter in clock distribution network [2]. The clock distribution network typically consists of multiple levels of trees/buffers, so the effect is cumulative. Extra "guard band" is necessary to absorb the delay variations in clocks and latches, and to protect against early mode (or race condition). If "gated clock" is used, additional margin is required to properly account for the variations in clock duty cycle, skew, and path-delay between the first switch (when the clock first becomes active) and the steady-state (when the clock is continuously running). Similarly, margins have to be widened to account for transitions from "low power mode" (where the entire global clock distribution or portion of it is shut down) and "test mode" (where the system clock is stopped, and test clocks at much lower frequency are used). As in scan-in/scan-out test systems) to normal operation [2].

In the 580 MHz PowerPC 750 SOI microprocessor, the latch hold-times were increased by 25-40 ps to protect against early-mode. Latch race margins were increased by another 5.0 ps to account for additional skew in local clock generation. Margin of 25 ps is added to half-cycle paths between globally gated latches for transition from low power mode to normal mode, and 25 ps race margin budgeted for latches interfacing between the globally gated and un gated clock domains. Many shortpath margin were increased to overcome variations in path delays [8].

The hysteresis effect complicates the timing methodology, degrades the timing rules, and complicates the design and degrades the margin of self-timed type of circuits widely used in the state-of-the-art on-chip cache SRAM [2]. Fig. 7 illustrates the situation where the race condition between a self-timed path (which fires every cycle) and a pull-down discharge path (which starts "fresh" after a long period of "no discharge") causes a failure in PowerPC 750 SOI microprocessor [8]. The self-timed path was slowed down to overcome this problem.

For on-chip cache SRAM, repetitive reads of one data state in a bitcolumn can cause offsets in the body voltage, ren can cause mismatches between the source and the body of the gate transis-

271
creases. Scaled device design for lower supply voltage also tends to alleviate the hysteretic $V_T$ variation [16]. However, the balance point between the body-to-source forward junction current and the drain-to-body reverse junction leakage, which determines the body potential at DC equilibrium, is relatively independent of $V_{DD}$ due to the weak dependence of the reverse junction leakage on $V_{DP}$. Furthermore, variation in $V_T$ becomes a larger portion of the total supply voltage. The hysteretic $V_T$ variation, therefore, remains a serious design concern/challenge even at low supply voltage.

V. CONCLUSION

We have reviewed the design considerations for high-performance SOI digital CMOS VLSI. It was shown that the ever-demanding performance of general-purpose microprocessor dictated far more stringent technology/device requirements and posed much more complicated design issues/challenges compared with low-power portable applications. Unique SOI design aspects resulting from the floating-body in partially-depleted devices, such as parasitic bipolar effect and hysteretic $V_T$ variation, were discussed. Circuit topologies susceptible to parasitic bipolar effect and circuit techniques to improve the noise immunity were addressed. The impact of hysteretic $V_T$ variation on various circuits (static CMOS, CPL, self-timed, clock distribution, latch timing, and global timing methodology) were discussed. 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. It is crucial for circuit designers to understand and quantify/contain the hysteretic delay and noise margin variations to fully exploit the performance leverage of a scaled PD/SOI technology.

REFERENCES

Fig. 4: Hysteric delay variation for a static CMOS inverter: (a) $T_{D_{rise}}$, $T_{D_{fall}}$, and $T_{D_{ave}}$ as functions of time, and (b) body-to-source voltage of nMOS and pMOS as functions of time. ($W_{p}/W_{n} = 2$, 1.0 ns period, 50% duty cycle, initial input at “Low” (L-H), and initial input at “High” (H-L)) [14].

Fig. 6: (a) A dual-ended CPL circuit, and (b) comparison of “rising input” delay variation with time for CPL and LEAP circuit in a 1.8 V, $L_{eff} = 0.12 \mu m$ PD/SOI technology [15].

Fig. 5: (a) A single-ended LEAP circuit, and (b) variation of body voltage and delay as functions of time for “rising input” with the input pulse frequency as a parameter in a 1.8 V, $L_{eff} = 0.12 \mu m$ PD/SOI technology. (The input pulse width is fixed at 1.0 ns) [16].

Fig. 7: A 27b dynamic OR with self-timed mask. Hysteric $V_T$ variation causes race condition between the self-timed path and the pull-down discharge path [8].