Evaluation of Dynamic-Threshold Logic for Low-Power VLSI Design in 0.13μm PD-SOI

Alan J. Drake
Department of EECS
University of Michigan
Ann Arbor, MI 48109
ajdrake@umich.edu

Kevin J. Nowka
Austin Research Lab
IBM
Austin, TX 78758
nowka@us.ibm.com

Richard B. Brown
Department of EECS
University of Michigan
Ann Arbor, MI 48109
brown@umich.edu

Abstract
Dynamic Threshold (DTMOS) circuits have been proposed as a circuit style for low-power VLSI systems that takes advantage of the independent body control in partially-depleted SOI. As SOI technologies have scaled, the increasing body capacitance and body resistance have limited the effectiveness of DTMOS circuits that drive the body at the same speed as the gate. An analysis of DTMOS in 0.13μm PD-SOI shows that a near-static body control independent of the gate is the lowest power DTMOS configuration, provides reduced delay and energy-delay-product over floating-body circuits, and allows for leakage control during low activity. Simulations show that DTMOS is effective at low power supply voltages when used in gates with large transistors, high fan-in, and/or high wire capacitance. Under these conditions, a delay improvement up to 40% while using 20% less energy than floating-body transistors can be achieved.

1. Introduction

The proliferation of battery-powered electronics has made low-power VLSI design an important research topic. Silicon-on-insulator technology is well suited for low-power design because of its reduced junction capacitance due to the buried oxide layer. The buried oxide also isolates the silicon under the transistor’s channel. When the silicon is thick enough, as in Partially-Depleted SOI (PD-SOI), the silicon under the channel does not deplete of charge carriers under strong inversion, leaving a fifth, body terminal that affects threshold voltage and device performance. When the body terminal is connected to the gate or driven during switching from some other signal, a Dynamic Threshold transistor (DTMOS) is created [1]. In traditional DTMOS, Figure 1(a), the gate is tied directly to the body. When the gate switches from high to low, the body also switches from high to low which sets the body-source voltage to 0V. The threshold voltage, which is dependent on the body-source voltage, is at its highest value and leakage currents are low. When the gate, and consequently the body, switches from low to high, the threshold voltage decreases with increasing source-body voltage and the drive current of the transistor increases. In summary, DTMOS provides low-leakage and high-current transistors.

Previous research, using 0.35μm and 0.25μm Partially-Depleted SOI (PD-SOI), has demonstrated several DTMOS inverter topologies that used less energy and had a lower delay than floating-body inverters when operated at supply voltages (VDD) below a few threshold voltages [1]-[5]. The pull-down networks of some of the inverters studied are shown in Figure 1. In inverters (a)-(e) the body switches each time the inverter switches. For inverters (f)-(i), the body switching depends on the behavior of Vref, which can vary from a static reference to switching at the same frequency as the inverter. The concept of DTMOS is independent of topology since DTMOS refers to the ability to change threshold-voltage dynamically during circuit operation, regardless of how quickly that switching occurs. In this work DTMOS will refer to circuits that can dynamically change threshold voltage and “traditional DTMOS” will refer to circuits that switch the body every cycle with the gate. Floating-body gates will simply be referred to as FB-CMOS.

In order for DTMOS to be effective, the body capacitance and body resistance need to be small so that the RC time constant to charge the body is smaller than the switching period driving the body; otherwise, the body voltage will not change until after the transistor has

![DTMOS circuit configurations in the literature.](image-url)
switched which is too late to decrease the transistor’s delay. Unfortunately, in modern SOI processes, technology scaling has increased the body resistance and body capacitance significantly, limiting body contact effectiveness and adding power dissipation. In spite of this scaling trend, DTMOS still maintains good power-delay and delay performance at low voltages as will be shown in this work. Some PD-SOI characteristics and their affect on DTMOS performance will first be examined and a circuit design style developed to examine DTMOS in its lowest power, highest performance configuration. Simulations in an IBM, 0.13µm PD-SOI technology [6] are then described that compare performance of DTMOS gates to comparable FB-CMOS gates over variations in width, fan-out, fan-in, and wire loading with some extensions made to domino logic. Finally an area comparison is made between the two design styles. The simulations show that DTMOS provides up to 40% reduction in delay with a 20% reduction in power-delay-product over FB-CMOS at 0.5V supply, making DTMOS a good option for low-power design.

2. DTMOS Performance Considerations

Adoption of DTMOS has been limited by the following design challenges: the body capacitance is large and adds power dissipation and gate loading; the body-source and body-drain diodes limit the power supply voltage in order to prevent exponentially growing static currents; the body resistance limits the effectiveness of the body contact due to the body RC charging delay; the body-drain capacitor has a feed-forward effect that increases gate capacitance; and the body contacts cost a lot of area [7]. The feed-forward effect, extra gate loading, and limited supply voltage are designed around by disconnecting the body from the gate and driving the body from a voltage source which does not exceed a diode drop, about 0.7V. The area penalty can only be reduced by careful design and will be discussed later. Decreasing the impact of increased gate capacitance, large power dissipation from switching the body capacitance, and body biasing will now be examined in some detail.

2.1. Body Capacitance

Figure 2 shows the body-contacted gate capacitance normalized to the floating-body gate capacitance. It is normalized to the floating-body transistor of the same width, to scale. The body contact is enclosed in the dashed square directly influenced by gate capacitance, they will demonstrate the same exponential decay in their performance as seen in Figure 2.

2.2. Frequency of switching body voltage

Most DTMOS circuits switch the body at the same frequency as the gate. The reason for switching the body with the gate is to limit static power when the transistor is off and increase current when the transistor is switching. The problem with this approach is the added dynamic power that results from switching the body capacitance. The dynamic power dissipation in DTMOS can be approximated by:

\[
P_{dyn} = C_L V_{dd}^2 f_L + C_{body} V_{body}^2 f_{body}
\]

where \(C_L\) is the gate and wire capacitance and \(C_{body}\) is the capacitance associated with switching the body. \(V_{dd}\) and \(V_{body}\) are the load and body supply voltages and \(f_L\) and \(f_{body}\) are the load and body switching frequencies. In traditional DTMOS, \(V_{body}=V_{DD}\) and \(f_{body}=f_L\) which means that the power nearly doubles by switching the body since \(C_{body}\) is the same order of magnitude as the gate capacitance.

From (1) it is obvious that to reduce power, \(V_{body}\) and \(f_{body}\) should be reduced since \(C_{body}\) is fixed by the circuit design. Some power reduction is obtained from reducing \(V_{body}\), but at least 0.5V is needed for DTMOS to perform better than FB-CMOS as will be described later. With operating voltages in modern technologies around 1V, reducing the supply voltage to 0.5V only reduces the body switching power by 1/4. More power reduction is needed which is obtained by reducing \(f_{body}\) to near-static levels compared to \(f_L\).

A body-source bias that is only high when the circuit is on, and then turns off when the circuit is in standby essentially reduces the body-switching power to zero. A near-static body control signal also provides better body control than a fast body control. The body’s RC charging
time-constant low-pass filters signals that attempt to switch the body too quickly, resulting in a lower body-source voltage than expected. The drawback to using a near-static body control is an increase in static power since the body voltage is always high during circuit operation. Simulations on a DTMOS inverter show that the active power to switch the gate capacitance is $10^3$ times the static power for a 0.5V body-source voltage and $10^2$ times the static power for 0.6V body-source voltage. Since body capacitance can be as high as gate capacitance, the small increase in static power is at least two orders of magnitude lower than the dynamic power dissipated switching the body at the same frequency as the gate. It should be noted that since the gate capacitance of DTMOS is always higher than FB-CMOS, DTMOS will never dissipate less power than FB-CMOS when both styles are operated at the same frequency and supply voltage.

2.3. Effective body-bias voltage

Since a near-static bias is the lowest-power configuration for DTMOS, it is important to determine the most effective bias value. Figure 3 charts the delay and static-power dissipation of an FO4 inverter. The static power grows exponentially with body-bias, as expected, but is still two orders of magnitude less than dynamic power at 0.6V body-source voltage. The upper bound of the body-source voltage is limited by static current because at about 0.7V the poor, parasitic bipolar transistor turns on. While the parasitic BJT aids switching, the static power dissipation is intolerable. The lower bound for the body-bias voltage is about 0.5V because biases below that don’t provide enough delay improvement over floating-body transistors to compensate for the trouble of adding gate capacitance and layout area. Biases from 0.5 to 0.6 volts provide the best delay improvement for added power dissipation. The following circuit simulations are based on 0.5V body-source bias as the low operating bound.

Figure 4: DTMOS Circuit configurations for simulations. The FB-CMOS gates have the same topology minus the body contacts.

3. Circuit Simulations

The analysis on body-contact show that the lowest-power DTMOS uses a near-static body-bias decoupled from the gate in the range of 0.5V to 0.6V. To characterize the performance of DTMOS with respect to FB-CMOS, simulations were performed in IBM’s 0.13um PD-SOI technology [6] on the gates in Figure 4. In the figure $w_{min}$ is the net minimum width, $C_{min}$ is the input capacitance of a minimum sized inverter, $m$ is the width multiplier, and $\alpha$ determines the wire capacitance. Body conditions are set by static, independent nfet and pfet biases that limit the body-source voltage ($V_{BS}$) to less than 0.7V. The body-bias inputs are not shown on the inverter chain in Figure 4. Delay and power were measured across the third gate as shown where the gates in the chain were either inverters or NAND gates. Variations in fan-out, device width, $V_{DD}$, $V_{BS}$, fan-in, and wire load were simulated to canvass the logic design space. An analysis of the area cost of using DTMOS was also made.

3.1. Fan-Out

Figure 5 shows the DTMOS delay and power-delay-product (PDP), normalized to FB-CMOS, for the inverter in Figure 4 over variations in width, $V_{DD}$, and fan-out with $V_{BS} = 0.5V$ and no wiring capacitance ($\alpha = 0$). Fan-out was simulated by adding gates to the output of each stage of the chain of gates in Figure 4. Each of the added gates also has the appropriate fan-out to properly shape switching waveforms. Because delay and power are directly influenced by gate capacitance, both normalized delay and power-delay curves have the same shape as the normalized gate capacitance curve in Figure 2. Increasing the fan-out increases the loading of the DTMOS inverters and their delay due to the body-contact.
overhead added at each fan-out gate. The added capacitance swamps the increase in drain current for all but low-voltage operation. DTMOS is not faster than FB-CMOS at nominal operating voltages because the FB-CMOS steady-state threshold voltage is higher at high voltages than at low voltages due to impact ionization currents and other body charging mechanisms. Thus, at nominal VDD, the DTMOS current drive is not that much higher than the FB-CMOS and only at very large transistor widths is the added gate capacitance small enough to see a delay improvement in DTMOS.

At low-voltages, DTMOS has a higher threshold voltage than FB-CMOS and better performance. At 0.5V VDD and 0.5V VBS, delay improves from 10% to 20% over FB-CMOS delay for 50% more power at 20X widths and a 25% delay improvement for widths larger than 40X. The power-delay-product is also lower than FB-CMOS at these widths for a more energy efficient transistor. 20X and larger transistors are reasonable sizes for bus or clock drivers in digital logic but large for inverters in logic functions.

3.2. Wire loading

Since DTMOS is more suited for driving buses and clock lines an FO1 inverter was simulated over increasing capacitive wire loads. The capacitance of the wire load, see Figure 4, is controlled by the parameter α, which is swept from 0 to 4, representing a wire load from 0 to 4 times that of the gate capacitance, mCmin. This value is scaled with width so that the wire capacitance is a constant percentage of gate capacitance. Figure 6 shows the simulated normalized delay and PDP of an FO1, DTMOS inverter at VBS = 0.5V. Only supply voltages of 0.5V and 1.2V are shown in the PDP graph to make it legible.

Normalized delay is reduced by as much as 15% as the wire load, αmCmin, of the transistor is increased from 0 to 4 times the inverter input capacitance. Increasing the wire load decreases the overhead of the body contact since more of the gate loading is in the wires. This makes the DTMOS gate more efficient which is why the normalized delay and PDP drop as wire loading increases. At 0.8V VDD the normalized DTMOS delay drops to less than one for large gates and high wire loads but the effect is most pronounced at low voltages. At 0.5V VDD, DTMOS delay is 25%-30% lower than FB-CMOS delay and PDP is 14%-21% lower than FB-CMOS for inverters larger than 20X. Gate capacitance is estimated to be between 30% to 50% of the total inverter load with the remaining capacitance coming from wiring and junction capacitance. This corresponds to setting α between 1 and 2. At these wire loads, the DTMOS is faster and more energy efficient than FB-CMOS at low voltages.
3.3. Fan-in

Next to consider is the affect of fan-in on the delay of DTMOS circuits by simulating the NAND gate in Figure 4 with an increasing number of inputs. Since higher fan-in gates are used in logic and will have multiple fan-out, FO4 outputs were simulated to measure the effects of fan-in. Figure 7 shows the normalized delay and PDP of FO4 DTMOS NAND gates with fan-ins of 1, 2, 3, and 4 at $V_{BS} = 0.5\, V$ and $\alpha = 0$. The low wire-load was chosen as the worst case for delay and power since it has been shown already that increasing wiring capacitance favors DTMOS by reducing the impact of the body contact on gate input capacitance. The x-axis is the effective pull-down strength of the gate: a NAND2 gate with an effective pull-down width of 2X has nfets with a physical width of 4X. The delay measured is the worst case pull-down delay when the bottom transistor switches last. Adding fan-in adds about the same amount of junction capacitance for DTMOS and FB-CMOS since junction capacitance depends little on the body contact. The added junction capacitance offsets the body-contact capacitance overhead much like wire capacitance does. The large change in normalized delay as fan-in is increased is mostly due to the transistors being much wider to achieve acceptable delay. Even at 1.0V, the high fan-in DTMOS gates are faster than FB-CMOS, but the transistors are large. The DTMOS delay is 20%-40% lower than FB-CMOS delay and the DTMOS PDP is 10%-22% lower than FB-CMOS for FO3 and FO4 gates with effective pull-downs larger than 20X. A large part of the improvement in delay for the higher fan-in gates is due to the physical length of the transistors being large to reduce pull-down delay. This is typical for most logic gates so high fan-in gates have transistors whose sizes may approach the widths at which DTMOS is effective.

3.4. Domino logic

Domino logic is a popular dynamic logic style used in high-performance VLSI. Domino circuits have a pull-down network driving the input of an inverter. Feed-back from the inverter controls a weak-keeper that prevents leakage and charge sharing in the pull-down from incorrectly evaluating the gate. The simulations performed on fan-out and fan-in can be extended to domino logic in a fairly straightforward manner by adding the pull-down delay of the fan-in to the pull-up delay of an inverter of a similar size and normalizing the sum to a floating-body gate built in the same manner. There are some differences that limit this comparison: the multiple nfets in the NAND gates simulated reduce pull-down time, although the weak keeper in domino gates has this same effect. Since domino gates only drive the nfet pull-down network a similar power comparison can’t be made. The estimation of domino delay provides a rough measure of how DTMOS will perform in domino logic. Figure 8 shows the estimated domino delay for AND gates composed of the FO4 NANDs and the FO4 inverters previously simulated with $\alpha=0$. This rough estimate shows that domino gates using DTMOS may be faster than FB-CMOS even up to 0.8V $V_{DD}$ for large transistors. Since the transistors in the pre-charge path do not need the extra speed of the body-contact and can be regular floating-body transistors, DTMOS domino can be designed with DTMOS only in the evaluate path, resulting in power savings and area savings. It may also be beneficial to
only place DTMOS in the clocked transistors to make them more efficient.

3.5. Area penalty

Several DTMOS and FB-CMOS gates with the same netlist were laid out to compare the area of the two logic styles. The DTMOS gates have dual body contacts and external nfet and pfet body-biases. For the domino gates, DTMOS transistors are only used in the input pull-down and output inverter pull-up. The results of the area comparison are shown in Table 1. Small gates such as the inverter, NAND, XOR, and AND cover nearly twice the area in DTMOS than in FB-CMOS. The more complex gates range from 1.2 to 2 times larger than FB-CMOS. A 24-bit, floating-point, radix-2 Booth multiplier would contain roughly 12 decoders, 321 muxes, 260 full-adders, 42 PGZ generators, 105 PGZ merge gates, and 42 PGZ sum gates. This results in a DTMOS multiplier 1.58 times larger than a comparable FB-CMOS multiplier. This is a significant amount of area and represents the main drawback to implementing DTMOS.

Table 1: Layout area comparison

<table>
<thead>
<tr>
<th>Gate</th>
<th>FB(µm²)</th>
<th>DT(µm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static 20X inverter</td>
<td>11.8</td>
<td>22.4</td>
</tr>
<tr>
<td>Static 20X NAND2</td>
<td>19.4</td>
<td>36</td>
</tr>
<tr>
<td>Domino XOR2</td>
<td>17.3</td>
<td>35.9</td>
</tr>
<tr>
<td>Domino AND2</td>
<td>12.1</td>
<td>22.4</td>
</tr>
<tr>
<td>Domino Booth decoder</td>
<td>146.2</td>
<td>176.3</td>
</tr>
<tr>
<td>Domino Booth mux</td>
<td>48.2</td>
<td>97.8</td>
</tr>
<tr>
<td>Dual-Rail Domino full-addr</td>
<td>86.1</td>
<td>132.1</td>
</tr>
<tr>
<td>Domino PGZ gen</td>
<td>58.2</td>
<td>83.7</td>
</tr>
<tr>
<td>Domino PGZ merge4</td>
<td>134.7</td>
<td>195.1</td>
</tr>
<tr>
<td>Domino PGZ sum</td>
<td>24.8</td>
<td>41.3</td>
</tr>
<tr>
<td>24-bit Booth multiplier</td>
<td>57,242</td>
<td>90,441</td>
</tr>
</tbody>
</table>

4. Conclusion

This study demonstrated that DTMOS is a good circuit design choice for low-power systems targeting supply voltages around 0.5V designed in 0.13µm SOI. The best performance for DTMOS comes when driving large, fixed wire loads and/or for high fan-in gates. Simulation results show that DTMOS provides reduced delay over FB-CMOS of about 35% for large wire loads and 40% for high fan-in while reducing the energy dissipated by 20% at 20X widths and 0.5V $V_{DD}$. Table 2 summarizes how delay and PDP change over parameters for FO1 gates and FO4 gates at 0.5V $V_{DD}$. Of the various topologies of DTMOS available, using a near-static body bias around 0.5V provides the lowest-power and lowest-delay. An additional advantage of using near-static DTMOS biasing is the ability to reduce leakage currents during low circuit activity periods by setting the bias to zero volts.

5. Acknowledgements

The authors wish to thank Jeff Burns of IBM’s Austin Research Lab for his guidance and Noah Zamdmer, Melanie Sherony, Mohamed Talbi, Yue Tan, and Jean-Olivier Plouchard of IBM for access to their 0.13µm SOI design kit.

6. References


