Analysis and Optimization of Enhanced MTCMOS Scheme

Rahul M Rao
Department of EECS
Univ. of Michigan, MI, USA
(rmrao@eecs.umich.edu)

Jeffrey L Burns
IBM Research
Yorktown Heights, NY, USA
(jlburns@us.ibm.com)

Richard B Brown
Department of EECS
Univ. of Michigan, MI, USA
(brown@eecs.umich.edu)

Abstract

Stacking of "off" transistors has been shown to reduce sub-threshold leakage of the stack. This paper presents an analysis of optimal selection of device widths for such stacked configurations for total leakage reduction. We show that forced stacking always results in an increase in gate leakage if identical performance is desired. We further present an analysis of optimal width ratios for sub-threshold and gate leakage reduction, and derive bounds on the input occurrence probability that ensures total leakage reduction with forced stacking. We demonstrate that leakage is greatly minimized if the stack is optimized for gate leakage rather than for sub-threshold leakage. Finally, we investigate optimization for total leakage and show that as gate leakage becomes dominant, optimization for gate leakage will be identical to total leakage optimization.

1. Introduction

In this paper, we investigate the effect of transistor stacks on gate leakage and present optimal design guidelines for total leakage reduction using stacked sleep devices. The contribution of leakage power to the total power consumption of an integrated circuit has been increasing at an alarming rate in recent technology generations due to the aggressive scaling of device threshold voltages and oxide thicknesses. Several schemes have been proposed for leakage power reduction in standby mode such as MTCMOS [1], VTCMOS [2], SC-CMOS [3], etc. Among these, MTCMOS is most popular because of its ease of implementation. A conventional MTCMOS scheme implementation is shown in Fig. 1 (left) with a sleep device (T_a) added between the low V_{TH} circuitry and the ground node. In [4], the use of multiple *off* devices in a transistor stack was proposed to exploit the stacking effect. An analysis of the scaling of stack effect and the leakage savings obtained was presented in [5]. However, stacking of devices results in a performance degradation and requires the application of low-leakage sleep-state vectors to exploit the stacking effect. An enhanced MTCMOS scheme was presented in [6], where stacked sleep devices (or power switches) were used for leakage control in standby state. An implementation of the enhanced MTCMOS scheme is shown in Fig. 1 (right) with sleep transistors T_b and T_c added between the low V_{TH} circuitry and the ground node. This scheme eliminates the need for input vector control by using a single sleep-state signal S which is driven low in sleep state, thereby turning *off* both the sleep devices. However, stacking of sleep devices can result in a performance degradation in active mode and hence the stacked sleep devices T_b and T_c need to be optimally sized to obtain maximum leakage savings while minimizing the performance degradation. It was also shown that having two *off* devices is the optimal stack height for the sleep devices. In [7], an optimal sizing scheme for the widths of stacked sleep devices was presented. However, this analysis was based on sub-threshold leakage, with gate leakage being ignored, and only considers leakage in standby mode. With gate leakage increasing at a greater rate than sub-threshold leakage and expected to become the dominant leakage component in future technology generations [8], it is necessary to understand the effect of gate leakage on any leakage minimization scheme. In addition, it is also essential to consider the effect of any leakage reduction scheme on the leakage power drawn during active operation.

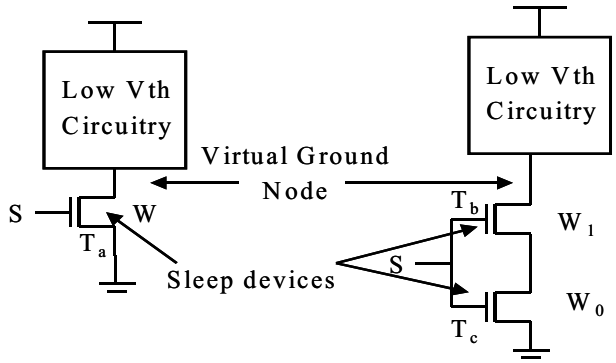

Fig. 1 Conventional MTCMOS scheme (left) and enhanced MTCMOS scheme with stacked sleep devices (right) for leakage reduction

Proceedings of the 17th International Conference on VLSI Design (VLSID'04)
1063-9667/04 $ 20.00 © 2004 IEEE

The objective of this work is to analyze the stacking effect while taking gate leakage of the sleep devices both during active and sleep state into consideration. Section 2 provides an overview of stacking effect and its application in an enhanced MTCMOS scheme. In Section 3, we analyze the effect of stack optimization with sub-threshold leakage as the objective function, and show that stacking always results in an increase in gate leakage. In Section 4, we present a device-width optimization scheme for gate leakage reduction, while an analysis of total leakage minimization is presented in Section 5. The findings of the paper are summarized in Section 6.

2. Forced Stacking Scheme

Stacking of devices has been shown to reduce sub-threshold leakage significantly [1][9]. The concept of forced stacking is illustrated in Fig. 1. In this case, a single device of width W is replaced by two devices of width W_1 and W_0 with their inputs tied together. In the enhanced MTCMOS configuration, this forced stacking is used for the sleep devices. Since the stacked sleep devices are driven by a conditional sleep signal S and not by the previous logic stage, this forced stacking does not increase the capacitive load of the previous stage, and the signal propagation times during active-mode operation can remain unaffected if the stacked sleep devices are appropriately sized. The stacking of sleep devices further reduces the sub-threshold leakage of the circuit in sleep state. Since the stacked sleep devices can be common to several stages of logic circuitry, the associated area penalty is minimal. In standby mode, the sleep devices are off and since their threshold voltage is higher than that of devices in the low V_{TH} circuitry, the virtual ground node is pulled to near V_{DD}. Hence, in this analysis, we consider the leakage of the circuit in standby mode to be given by the leakage of the sleep devices. Though this analysis only considers NMOS sleep devices, it can easily be applied to configurations that use PMOS sleep devices.

Gate leakage of a device is exponentially dependent on its gate-to-drain and gate-to source bias. The gate leakage of any device in the circuit can be determined on the basis of its bias state in the circuit [10]. Let *NG*0 and *NG*1 represent the gate leakage for a unit-width NMOS transistor in S(0) and S(1) bias states respectively as shown in Fig. 2, and let *I* represent the sub-threshold leakage of a unit-width device in S(0) state. In the S(2) state, the device has zero gate-to-drain and gate-to-source bias and hence the gate leakage in this state is negligible. Also, let $p1$ represent the probability of the circuit being in active mode, i.e., the probability of the input S being at logic high state.

In the conventional MTCMOS scheme, when the circuit is in active mode, the sleep device T_a is turned *on* and hence is in S(1) state. When the circuit is in standby mode, the sleep transistor is turned *off*. Since the sleep device usually has a higher threshold voltage than the active circuitry, the virtual ground node can be assumed to be nearly at the supply rail. Hence, the sleep device can be assumed to be in S(0) state.

The gate leakage for the conventional MTCMOS scheme (Fig. 1 left) can then be approximated as

$$Ig_{single} = W \cdot p1 \cdot NG1 + W \cdot (1 - p1) \cdot NG0$$

The total leakage is the sum of the sub-threshold and gate leakage of the sleep device and is given by

$$Itot_{single} = Ig_{single} + (1 - p1) \cdot NG0 \qquad (1)$$

For the enhanced MTCMOS scheme of Fig. 1, the sleep devices are *on* during active mode, and hence are in state S(1). In the sleep state, both the sleep devices are turned *off*. The virtual ground node can be assumed to be nearly at the supply rail as in the previous case. The intermediate node between the two stacked sleep devices will settle at a voltage very close to ground. Hence, the upper device can be assumed to be in S(0) state, while the lower device can be assumed to be in S(2) state.

The sub-threshold stack effect factor, which is defined as the ratio of the sub-threshold leakage current in one *off* device to the sub-threshold leakage current in a stack of two *off* devices [5] is given by

$$X = \frac{W}{W_1^{\alpha} \cdot W_0^{1-\alpha}} \cdot 10^{\frac{\lambda V_{dd}}{S} \cdot (1-\alpha)} \qquad (2)$$

Here, body effect has been neglected, λ is the DIBL factor, S represents sub-threshold swing, and α is given by $\alpha = \frac{\lambda}{1 + 2\lambda}$

The gate leakage for the enhanced MTCMOS Scheme (Fig. 1 right) can be approximated by

$$Ig_{stack} = (W_1 + W_0) \cdot p1 \cdot NG1 + W_1 \cdot (1 - p1) \cdot NG0$$

The gate leakage factor can be defined as the ratio of gate leakage current in conventional MTCMOS scheme to the gate leakage in enhanced MTCMOS and is given by

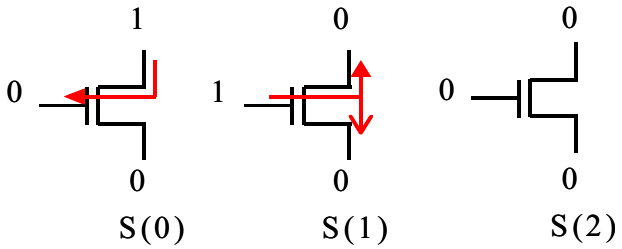

Fig. 2 Device bias states characterized for gate leakage approximation.

$$Y = \frac{W \cdot (p1 \cdot NG1 + (1-p1) \cdot NG0)}{(W_1 + W_0) \cdot p1 \cdot NG1 + W_1 \cdot (1-p1) \cdot NG0} \quad (3)$$

Thus, the total leakage of the stacked sleep devices is

$$Itot_{stack} = Ig_{stack} + I((1-p1)/X) \quad (4)$$

Stacking of devices can result in a performance penalty [5], unless the devices are appropriately sized to maintain identical conductances of the pull-down tree. It was shown in [7] that to ensure identical performance, the stacked sleep devices should be sized such that

$$W_1 = \frac{W \cdot W_0}{W_0 - W} \quad (5)$$

with the constraint that $W_0 > W$. These devices can be sized to satisfy the above constraint and optimized to minimize the leakage of the stack. The subsequent sections present an analysis of device width optimizations for the enhanced MTCMOS scheme to minimize sub-threshold, gate and total leakage while satisfying eqn. (5) to ensure performance equivalent to that of the conventional MTCMOS scheme.

3. Sub-threshold Leakage Optimization

In [7], it was derived that the optimal width ratio for maximum sub-threshold leakage reduction while maintaining equivalent performance is given by

$$\frac{W_1}{W_0} = \frac{1}{\alpha} - 1 \quad (6)$$

For this condition, it can also be shown that

$$W_0 = \frac{W}{(1-\alpha)}, and, W_1 = \frac{W}{\alpha} \quad (7)$$

The maximum sub-threshold stack factor is given by using eqn. (7) in eqn. (2)

$$X_{max} = \frac{10^{\frac{\lambda Vdd}{S} \cdot (1-\alpha)}}{\left(\frac{1}{\alpha}\right)^{\alpha} \cdot \left(\frac{1}{1-\alpha}\right)^{1-\alpha}}$$

Thus, for increased sub-threshold leakage savings the lower device T_c should be smaller than the upper device T_b [7]. This is because DIBL effect is reduced for the lower device due to its lower drain bias, and the fact that body effect (which is negligible in SOI) has been neglected. Using eqn.(6) in eqn.(3), the gate leakage factor can be re-written as

$$Y = \frac{W \cdot (p1 \cdot NG1 + (1-p1) \cdot NG0)}{W_0 \cdot \left(\frac{p1 \cdot NG1}{\alpha} + \frac{(1-\alpha)}{\alpha} \cdot (1-p1) \cdot NG0\right)}$$

To obtain savings in gate leakage, Y should be greater than 1 in the above equation. This results in an upper bound on the input occurrence probability, i.e., the percentage of time the circuit is in active mode, beyond which gate leakage increases with stacking:

$$p1 < \frac{\left\{1 - \left(\frac{W_0}{W} \cdot \frac{(1-\alpha)}{\alpha}\right)\right\}}{\frac{NG1}{NG0} \cdot \left(\frac{W_0}{W \cdot \alpha} - 1\right) + 1 - \left(\frac{(1-\alpha)}{\alpha} \cdot \frac{W_0}{W}\right)} \quad (8)$$

This can be simplified as

$$p1 < \frac{1 - \frac{1}{\alpha}}{\frac{NG1}{NG0} \cdot \left(\frac{1}{\alpha \cdot (1-\alpha)} - 1\right) + 1 - \frac{1}{\alpha}} \quad (9)$$

For an advanced industry-standard sub 0.1µm SOI technology, the above constraint results in $p1 < 0$, irrespective of the value of λ, i.e., the circuit should always be in sleep state. Thus, if the devices are sized to obtain maximum sub-threshold leakage savings under identical performance constraint, there is always an increase in gate leakage current.

Since the objective is to reduce the total leakage, a constraint can be derived for $p1$ to ensure that stacking results in total leakage reduction i.e., $Itot_{single}/Itot_{stack} > 1$ using eqns. (1) and (4) and is given by

$$p1 < \frac{Is + \left(W - \left(W_0 \cdot \frac{(1-\alpha)}{\alpha}\right)\right) \cdot NG0}{\left(\frac{W_0}{\alpha} - W\right) \cdot NG1 + \left(\left(W - \left(W_0 \cdot \frac{(1-\alpha)}{\alpha}\right)\right) \cdot NG0\right) + Is}$$

$$Is = W \cdot I \cdot \left(1 - \frac{1}{X}\right)$$

Using eqn. (6), this can be rewritten as

$$p1 < \frac{\frac{I}{NG0} \cdot \left(1 - \frac{1}{X}\right) + 1 - \frac{1}{\alpha}}{\frac{NG1}{NG0} \cdot \left(\frac{1}{\alpha \cdot (1-\alpha)} - 1\right) + 1 - \frac{1}{\alpha} + \frac{I}{NG0} \cdot \left(1 - \frac{1}{X}\right)} \quad (10)$$

Fig.3 shows the maximum input occurrence probability against DIBL factor λ, while Fig. 4 shows the total leakage savings factor as a function of $p1$ for λ = 0.1, which is a fairly typical value. As λ increases, the sub-threshold leakage factor X increases, resulting in a higher bound on the input occurrence probability. However, as the input occurrence probability increases, the increased gate leakage when both the devices in the stack are *on* results in a lower total leakage savings. It can be seen from Fig. 4 that for λ = 0.1, the total leakage savings factor for a sub-threshold leakage optimized stack is always less than 1, indicating that it is better to use a single sleep device than to use stacked sleep devices.

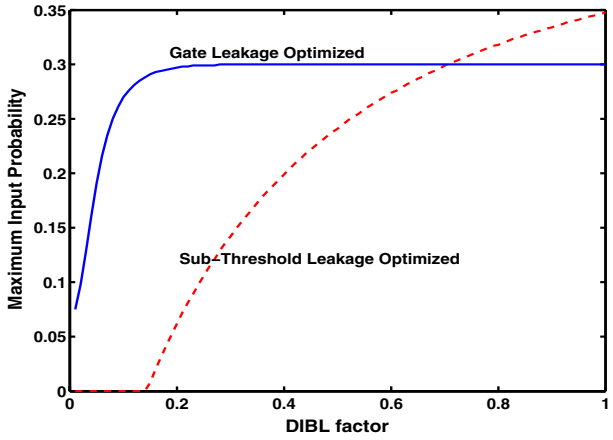

Fig. 3 Bound on input occurrence probability to obtain leakage savings using Enhanced MTCMOS Scheme

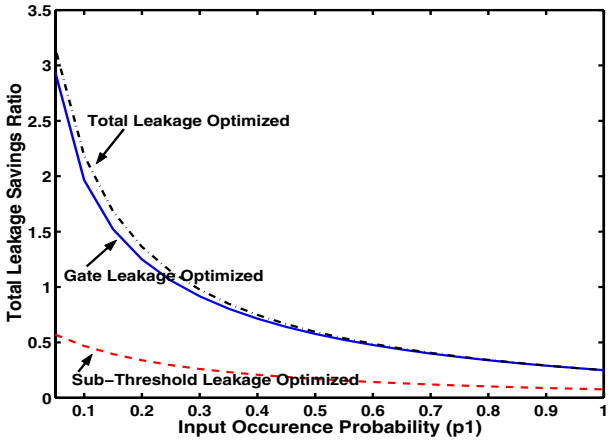

Fig. 4 Total leakage savings factors for different stack optimizations.

4. Gate Leakage Optimization

The gate leakage factor Y can be re-written using eqn. (5) in eqn. (3) as

$$Y = \frac{W \cdot (p1 \cdot NG1 + (1-p1) \cdot NG0)}{\left(\frac{W_0 \cdot W}{W_0 - W} + W_0\right) \cdot p1 \cdot NG1 + W_1 \cdot (1-p1) \cdot NG0} \quad (11)$$

In order to maximize gate leakage savings, the denominator of the above equation is minimized

$$\frac{\partial}{\partial W_0} Ig_{new} = (W_0^2 - 2WW_0) \cdot p1 \cdot NG1 - ((1-p1) \cdot W^2 \cdot NG0)$$

Setting the above expression to zero yields

$$W_0 = W \cdot (1+\beta), \text{ and, } W_1 = W \cdot \frac{1+\beta}{\beta} \quad (12)$$

$$\beta = \sqrt{1 + \frac{NG0}{NG1} \cdot \frac{(1-p1)}{p1}} \quad (13)$$

(The derivation details have been omitted).

The maximum gate leakage savings factor is given by

$$Y_{max} = \frac{(p1 \cdot NG1 + (1-p1) \cdot NG0)}{\frac{(1+\beta)^2}{\beta} \cdot p1 \cdot NG1 + \frac{1+\beta}{\beta} \cdot (1-p1) \cdot NG0} \quad (14)$$

Since β is always greater than 1, Y shall always be less than 1, i.e., stacking always results in an increase in gate leakage (if identical performance is desired). Thus, for maximum gate leakage savings, the upper device T_b must be smaller than the lower device T_c, which is contrary to the condition for sub-threshold leakage optimization, since the upper device exhibits gate leakage even when the device is in the *off* state.

For this optimal gate-leakage device width ratio, the sub-threshold stack effect factor is given by

$$X = \frac{\beta^{\beta}}{(1+\beta)} \cdot 10^{\frac{\lambda Vdd}{S} \cdot (1-\alpha)}$$

We can formulate a constraint for the input occurrence probability $p1$ such that stacking results in a reduction in total leakage. Fig 3 showed the maximum input occurrence probability that results in leakage savings for different values for DIBL factor λ, while Fig. 4 showed the total leakage savings as a function of $p1$ for λ = 0.1. It can be seen that the trend in the maximum input occurrence probability and total leakage savings factor are similar to those obtained for a sub-threshold leakage optimized stack. It is obviously better to optimize for gate leakage, since this results in larger savings in total leakage as compared to optimizing the device widths for sub-threshold leakage. For a circuit that is in active mode 10% of the time, the input occurrence probability is 0.1 and the total leakage savings factor for a gate-leakage optimized stack is nearly four times that of a sub-threshold leakage optimized stack. Also, for a given value of λ, the range of possible input occurrence probability for which leakage savings are obtained is greater for a gate leakage optimized stack.

5. Total Leakage Optimization

The optimal width of the two devices for maximum total leakage savings is strongly dependent on the ratio of sub-threshold leakage to gate leakage. There are contradictory requirements for the width ratio of the devices in the stack for minimizing sub-threshold and gate leakage, as derived in the previous two sections. For maximum sub-threshold leakage reduction, the lower device should be smaller than the upper device, while gate leakage savings is maximized when the lower device is larger than the upper device. Using eqns. (1), (4) and (5) a constraint for maximum total leakage savings was derived which was solved

numerically. Fig. 4 showed the total leakage savings ratio as a function of the input occurrence probability (for a λ of 0.1). It is seen that for a circuit that is in active mode for a very small percentage of time (i.e., has a low input occurrence probability), the maximum obtainable leakage savings is slightly greater than that obtained for a gate-leakage optimized stack. However, as the input occurrence probability increases, gate leakage begins to dominate and the maximum leakage savings ratio is identical to that obtained if the stack were optimized only for gate leakage.

In order to validate these findings, extensive simulations were performed on an 8-bit Brent Khung adder in a sub 0.1μm advanced SOI process at a supply voltage of 1 V and temperature of 85C. Figs. 5 and 6 plot the sub-threshold and gate leakage savings ratio as functions of different input occurrence probabilities and width-ratios. As the width ratio increases, the lower device T_c becomes smaller in comparison with the upper device T_b (in Fig. 1b) to satisfy the identical performance constraint of eqn. (5). The input occurrence probability represents the percentage of time that the circuit is in active mode. With an increase in the width ratio, i.e. as the lower device becomes narrower, sub-threshold leakage savings increases, which validates the analysis presented in Section 2. It is seen that significant reduction in sub-threshold leakage can be obtained by appropriately sizing the stacked sleep devices. Also, the sub-threshold leakage savings ratio is independent of the input occurrence probability.

Fig. 6 shows the gate leakage savings factor to be always lower than 1 as derived in Section 3. This re-emphasizes that forced stacking always results in an increase in gate leakage if identical performance is desired. At low input occurrence probability, the stacked sleep devices are *off* for a greater percentage of time and hence the gate leakage is dominated by the gate leakage through the upper device T_b. Gate leakage savings is thereby increased by having a narrower upper device as compared to the lower device, i.e., by having a smaller W/W_0 ratio. As the input occurrence probability increases, the gate leakage savings increases with increasing width ratio until W/W_0 is roughly 0.5, after which it begins to decrease. This is because the gate leakage in active mode is much higher than the gate leakage in standby mode and is dependent on the sum of the width of the two devices, since both the devices are in S(1) state. At a width ratio of 0.5, the two devices in the stack are of equal size and the sum of the device sizes, i.e., (W_1+W_0), which determines the gate leakage when the devices are *on*, is minimum.

The total leakage savings ratio of an enhanced MTCMOS scheme over a conventional MTCMOS scheme is shown in Fig. 7. At low input occurrence probabilities, the total leakage is substantially contributed to by both sub-threshold and gate leakage current. The contradictory optimal width conditions for sub-threshold and gate leakage minimization result in the total leakage savings ratio being maximized at a width ratio determined by the ratio of sub-threshold to gate leakage. Fig. 8 shows the best width ratio (W/W_0) for devices in an enhanced MTCMOS scheme for sub-threshold, gate and total leakage optimizations as functions of input occurrence probability. The best width ratio for a sub-threshold leakage optimized stack considers only sleep-state and hence is independent of the input occurrence probability. However, such a width ratio would result in increased gate leakage and hence is clearly sub-optimal while considering the total leakage of the circuit. The best width ratio for a gate leakage optimized stack increases with the input occurrence probability and saturates at a ratio of 0.5. It can be seen that the best width ratio for total leakage (i.e., sum of sub-threshold

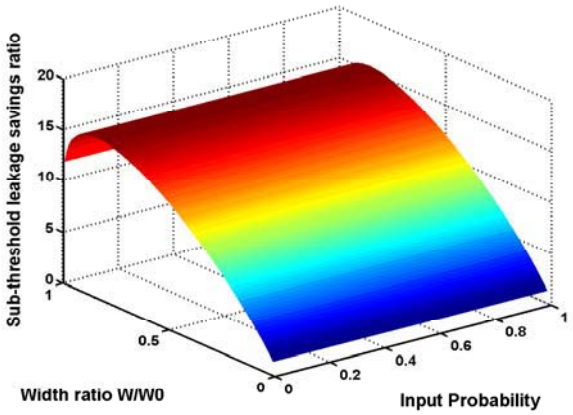

Fig. 5 Sub-threshold leakage savings ratio for different device widths and input occurrence probabilities.

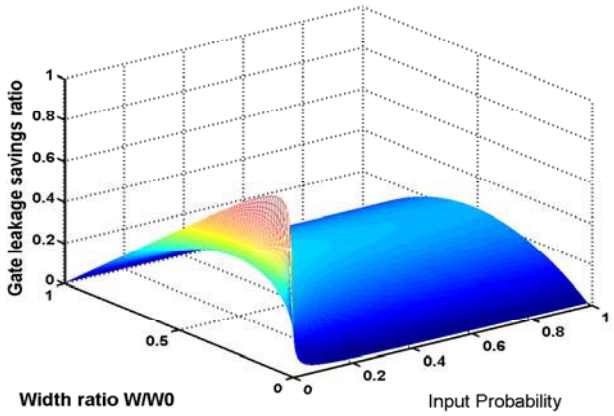

Fig. 6 Gate leakage savings ratio for different device widths and input occurrence probabilities.

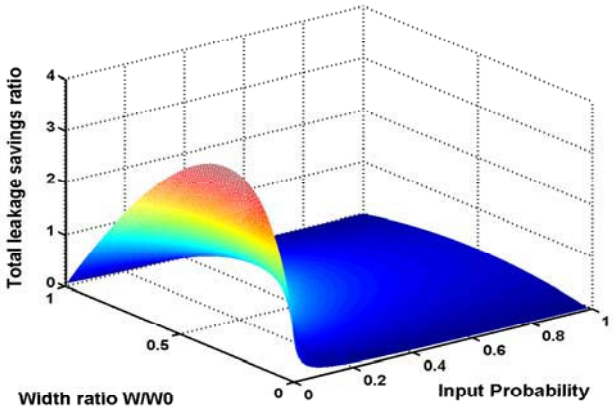

Fig. 7 Total leakage savings ratio for different device widths and input occurrence probabilities.

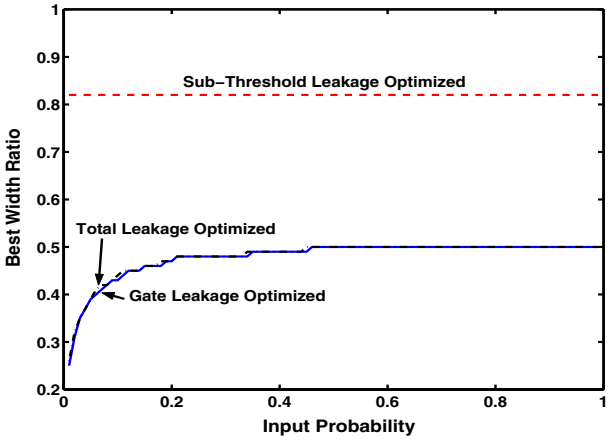

Fig. 8 Best width ratios (W/W_0) for gate, sub-threshold and total leakage optimizations.

and gate leakage) optimization is nearly equal to that for gate leakage optimization. Thus, optimizing the stacked-sleep devices in an enhanced MTCMOS scheme for gate leakage rather than for sub-threshold leakage results in almost maximum possible leakage savings at any input occurrence probability. With the increasing magnitude and dominance of gate leakage current, this can be expected to be true for future technology generations.

6. Conclusions

The increasing magnitude of sub-threshold and gate leakage currents has made leakage power an important factor in the design of integrated circuits. Most of the leakage mitigation techniques have focussed only on sub-threshold leakage with schemes such as stacking of *off* devices being proposed for leakage reduction. In this paper, we investigate the effect of the forced stacking scheme on gate leakage by analyzing an enhanced MTCMOS scheme that utilizes stacked sleep devices. We analyze device width optimizations with sub-threshold leakage and gate leakage as the objective functions and demonstrate that the total leakage savings obtained for a gate leakage optimized stack is higher than that obtained for a sub-threshold leakage optimized stack. We also derive bounds on the input occurrence probability for the application of stacking schemes and illustrate that optimizing the stacked sleep devices for gate leakage leads to nearly equivalent results as obtained by optimizing for total leakage.

7. References

[1] S. Mutoh, et.al, "1-V Power Supply High Speed Digital Circuit Technology with Multi-Threshold Voltage CMOS," *IEEE Journal on Solid State Circuits*, vol. 30, no. 8, pp. 847-854, Aug. 1995.

[2] T. Kurado, et.al, "A 0.9-V, 150MHz, 10-mW, 4mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold (Vt) Scheme," *IEEE Journal on Solid State Circuits*, vol. 31, no. 11, pp. 1770-1779, Nov. 1996.

[3] H. Kawaguchi, et.al, "A Super Cut-Off CMOS (SCCMOS) Scheme for 0.5V Supply Voltage with pico-Ampere Standby Current," *IEEE Journal on Solid State Circuits*, vol. 35, no. 10, pp. 1498-1501, Oct. 2000.

[4] Y. Ye, S. Borkar, V. De, "A New Technique for Standby Leakage Reduction in High-Performance Circuits," *Digest of Technical Papers, Symposium on VLSI Circuits*, pp. 40-41, Jun. 1998.

[5] S. Narnenda, et.al, "Scaling of Stack Effect and its Application to Leakage Reduction," *Proc. ISLPED,* pp. 195-200, Aug. 2001

[6] K. Das and R. Brown, "Novel Ultra Low-Leakage Power Circuit Techniques and Design Algorithms in PD-SOI for Sub-1V Applications," *Proc. International SOI Conference*, pp. 88 - 90, Oct. 2002.

[7] K. Das, et.al, "New Optimal Design Strategies and Analysis of Ultra-Low Leakage Circuits for Nano-Scale SOI Technology," *Proc. ISLPED,* pp. 168-171, Aug. 2003.

[8] International Technology Roadmap for Semiconductors, *http://public.itrs.net/Files/2001ITRS/Home.htm,* 2001 Edition

[9] M. Johnson, D. Somasekhar, K. Roy, "Leakage Control With Efficient Use of Transistor Stacks in Single Threshold CMOS," *Proc. DAC*, pp. 442-445, Jun. 1999.

[10] R. Rao, et.al, "Efficient Techniques for Gate Leakage Estimation," *Proc. ISLPED,* pp. 100-103, Aug. 2003.