Chapter 2

LEAKAGE DEPENDENCE ON INPUT VECTOR

Siva Narendra[§], Yibin Ye[¶], Shekar Borkar[¶], Vivek De[¶], and Anantha Chandrakasan*
[§]*Tyfone, Inc., USA,* [¶]*Intel Corp., USA, and* **Massachusetts Institute of Technology, USA*

2.1 INTRODUCTION

As described earlier to limit the energy and power increase in future CMOS technology generations, the supply voltage (V_{dd}) will have to continually scale. The amount of energy reduction depends on the magnitude of V_{dd} scaling. Along with V_{dd} scaling, the threshold voltage (V_t) of MOS transistors will have to scale to sustain the traditional 30% gate delay reduction. These V_{dd} and V_t scaling requirements pose several technology and circuit design challenges. In this chapter the term leakage refers to sub-threshold leakage, unless otherwise explicitly mentioned.

One of challenge with technology scaling is the rapid increase in sub-threshold leakage power due to V_t reduction. Should the present scaling trend continue it is expected that the sub-threshold leakage power will become a considerable constituent of the total dissipated power. In such a system it becomes crucial to identify techniques to reduce this leakage power component. It has been shown previously that the stacking of two off transistors has significantly reduced sub-threshold leakage compared to a single off transistor. The stack effect can therefore be used not jus for leakage reduction by forcing stacks, but also using natural stacks that existing in logic gates. Natural stacks can be realized by loading an appropriate primary input vector such that it propagates to maximize the total channel width of stacked transistors that are *OFF*.

In this chapter we present a model that predicts the stack effect factor, which is defined as the ratio of the leakage current in one off transistor to the leakage current in a stack of two off transistors [1]. Model derivation based on transistor fundamentals and verification of the model through statistical

transistor measurements from 0.18 μm and 0.13 μm technology generations are presented. The scaling nature of the stack effect leakage reduction factor is also discussed. The derived model for leakage reduction depends on fundamental transistor parameter. This makes the model viable to predict potential leakage savings using stack effect techniques in future transistors.

There are number of solutions including reverse body bias, power gating, and multi-performance transistors can be used to reduce power during standby mode. All of these will be discussed in detailed in the later chapters.

In this chapter after the introduction of stack effect, we will review a new standby leakage control scheme which exploits the large reduction in leakage current achievable by simultaneously turning *OFF* more than one transistor in NMOS or PMOS stacks. Usually, a large circuit block consists of a significant number of logic gates where transistor stacks already exist, such as the PMOS stack in NOR or NMOS stacks in NAND gates.

This first solution, using stack effect in natural stacks that already exists, enables effective leakage reduction during standby mode by installing a vector at the inputs of the circuit block so as to maximize the number of PMOS and NMOS stack with more than one *OFF* transistor. In contrast to the other leakage reduction techniques this scheme offers leakage reduction with minimal overheads in area, power, and technology requirements. Extensive circuit simulations of a sample circuit block to (*a*) elucidate the dynamics of leakage reduction using transistor stacks, (*b*) influence on overall leakage power reduction of the circuit block during both active and standby modes of operation, and (*c*) determine the standby leakage reductions due to the use of natural stacks will be discussed [2].

Another solution to the problem of ever-increasing leakage is to force a non-stack transistor to a stack of two transistors without affecting the input load. By ensuring iso-input load, the previous gate's delay and the switching power will remain unchanged. Logic gates after stack forcing will reduce leakage power, but incur a delay penalty, similar to replacing a low-V_t transistor with a high-V_t transistor in a dual-V_t design. In a dual-V_t design the low-V_t transistors are used in performance critical paths and the high-V_t transistors in the rest. Further details of dual-V_t design technique will be described in Chapter 8 under multi-performance transistors.

Usually a significant fraction of the transistors can be high-V_t or forced-stack since a large number of the paths are non-critical. This will reduce the overall leakage power of the chip without impacting operating clock frequency. In this chapter we discuss the stack forcing method to reduce leakage in paths that are not performance critical. This stack forcing technique can be either used in conjunction with dual-V_t or can be used to reduce the leakage in a single-V_t design.

Although it is not covered in depth in this chapter, it should be pointed out that vector dependent leakage behavior can not only be used to reduce standby sub-threshold current, but also total standby leakage current in the presence of tunneling sources. The current of transistor due to just the gate leakage is more when a transistor is *ON* compared to *OFF*, due to larger area. The gate leakage area of a transistor that is *OFF* is usually just the drain-gate overlap area, while in the case of a transistor that is *ON* it usually includes the drain-gate overlap, source-gate overlap, and channel areas. This is reverse of sub-threshold leakage current, therefore understanding of the relative contribution of the different leakage currents and proper methodology to identify the leakage minimizing input vector is critical [3]. Having said that, it is also necessary to realize under most conditions for logic circuits sub-threshold leakage will be a more dominant component.

2.2 STACK EFFECT

To reiterate, should the present scaling trend continue it is expected that the sub-threshold leakage power will become as much as 50% of the total power in the 0.09 μm generation [4]. Under this scenario, it is not only important to be able to predict sub-threshold leakage power more accurately as discussed in the previous section, it becomes crucial to identify techniques to reduce this leakage power component. It has been shown previously that the stacking of two *OFF* transistors has significantly reduced sub-threshold leakage compared to a single *OFF* transistor [2, 5, 6]. This concept of stack effect is illustrated in Figure 2-1.

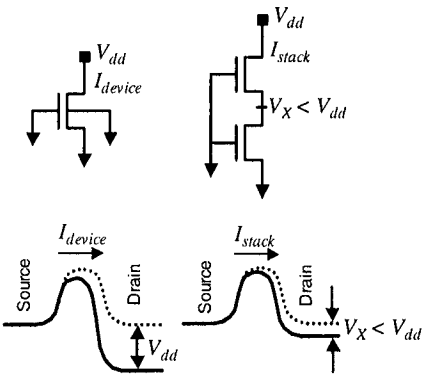

Figure 2-1. Leakage difference between a single *OFF* transistor and a stack of two *OFF* transistors. As illustrated by the energy band diagram, the barrier height is modulated to be higher for the two-stack due to smaller drain-to-source voltage resulting in reduced leakage.

In this section, a model is derived that predicts the stack effect factor, which is defined as the ratio of the leakage current in one *OFF* transistor to the leakage current in a stack of two *OFF* transistors. Model derivation based on transistor fundamentals and verification of the model through statistical transistor measurements from 0.18 μm and 0.13 μm technology generations are presented. The scaling nature of the stack effect leakage reduction factor is also discussed.

Let I_1 be the leakage of a single transistor of unit width in *OFF* state with its $V_{gs} = V_{bs} = 0$ V and $V_{ds} = V_{dd}$. If the gate-drive, body bias, and drain-to-source voltages reduce by ΔV_g, ΔV_b, and ΔV_d respectively from the above-mentioned conditions, the leakage will reduce to,

$$I'_1 = I_1 \quad 10^{-\frac{1}{S}\left[\Delta V_g + \lambda_d \Delta V_d + k_\gamma \Delta V_b\right]}$$

where S is the sub-threshold swing, λ_d is the drain-induced barrier lowering (DIBL) factor, and k_γ is the body effect coefficient. The above equation assumes that the resulting $V_{ds} > 3kT/q$ [7]. For a two-transistor stack shown in Figure 2-2 a steady state condition will be reached when the intermediate node voltage V_{int} approaches V_x such that the leakage currents in the upper and lower transistors are equal. Under this condition, the leakage currents in the upper and lower transistors can be expressed as,

$$I_{stack\text{-}u} = w_u I_1 \quad 10^{\frac{-(1+\lambda_d + k_\gamma)V_x}{S}}$$

$$I_{stack\text{-}l} = w_l I_1 \quad 10^{\frac{-\lambda_d (V_{dd} - V_x)}{S}}$$

and the intermediate node voltage by equating the two current can be derived to be,

$$V_x = \frac{\lambda_d V_{dd} + S \log \frac{w_u}{w_l}}{1 + k_\gamma + 2\lambda_d}$$

For short channel transistors the body terminal's control on the channel is negligible compared to gate and drain terminals, implying $k_\gamma << 1 + 2\lambda_d$.

Hence, the steady state value, V_x, of the intermediate node voltage can be approximated as,

$$V_x \approx \frac{\lambda_d V_{dd} + S \log \frac{w_u}{w_l}}{1+2\lambda_d}$$

Substituting V_x in either $I_{stack\text{-}u}$ or $I_{stack\text{-}l}$ will yield the leakage current in a two-stack given by,

$$I_{stack} = w_u^{\alpha} w_l^{1-\alpha} I_1 \; 10^{\frac{-\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$where \quad \alpha \approx \frac{\lambda_d}{1+2\lambda_d}$$

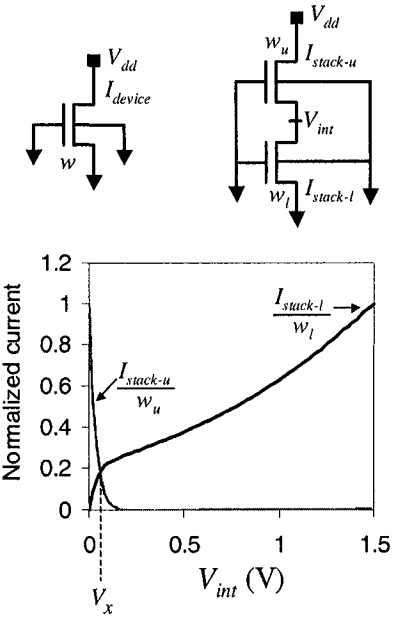

Figure 2-2. Load line analysis showing the leakage reduction in a two-stack.

The leakage reduction achievable in a two-stack comprising of transistors with widths w_u and w_l compared to a single transistor of width w is given by,

$$X = \frac{I_{device}}{I_{stack}} = \frac{w}{w_u^{\alpha} w_l^{1-\alpha}} 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)}$$

$$= 10^{\frac{\lambda_d V_{dd}}{S}(1-\alpha)} \quad when \quad w_u = w_l = w$$

The stack effect factor, when $w_u = w_l = w$, can be rewritten as,

$$X = 10^{\frac{\lambda_d V_{dd}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)} = 10^{U}$$

where U is the universal two-stack exponent which depends only on the process parameters, λ_d and S, and the design parameter, V_{dd}. Once these parameters are known, the reduction in leakage due to a two-stack can be determined from the above model. It is essential to point out that the model assumes the intermediate node voltage to be greater than $3kT/q$.

To confirm the model's accuracy we performed transistor measurements on test structures fabricated in 0.18 μm and 0.13 μm process technologies. Results discussed in the rest of the section are from NMOS transistor measurements, but similar results hold true for PMOS transistors as well.

Figure 2-3 shows NMOS transistor measurements under different temperature, V_{dd}, body bias, and channel length conditions for 0.18-μm technology generations, which prove the accuracy of the theoretical model. It is important to note that the model discussed above doesn't include the impact of diode junction leakages that originate at the intermediate stack node. In Figure 2-3, the model's accuracy deviates the most under reverse body bias for nominal channel length transistors, where the ratio of diode junction leakage to sub-threshold leakage current increases.

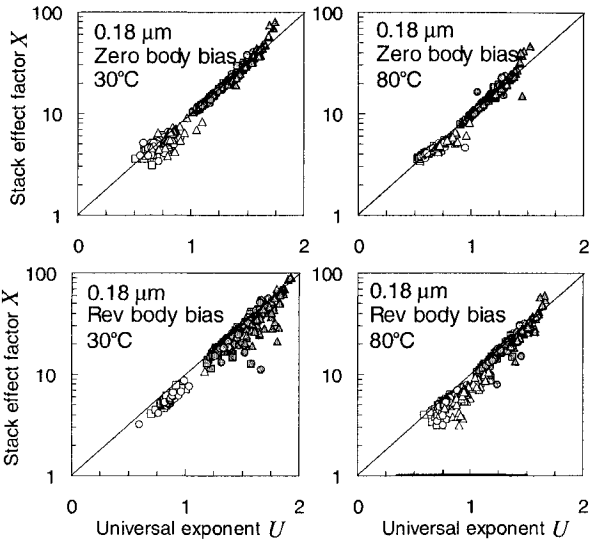

Figure 2-3. Measurement results showing the relationship between stack effect factor X for a two-stack to the universal exponent U. Lines indicate the relationship as per the analytical model and symbols are from measurement results. White symbols are for nominal channel transistors and gray symbols are for transistors smaller than the nominal channel length. Triangle, circle, and square symbols are for V_{dd} of 1.5, 1.2, and 1.1 V respectively. Zero body bias is when the body-to-source diode of the transistor closet to the power supply is zero biased and reverse body bias is when the diode is reverse biased by 0.5 V.

It is known that the stack effect factor strongly depends on λ_d as suggested by the model. In addition, a decrease in the channel length (L) will increase λ_d in a given technology [8]. So, any increase in the leakage of a single transistor due to decrease in L will not increase leakage of a two-stack at the same rate. This is illustrated in Figure 2-4 where increase in two-stack leakage is at a slower rate than that of a single transistor. Therefore, variation in L will result in smaller effective threshold voltage variation for a two-stack compared to a single transistor. Figure 2-5 illustrates the average stack effect factor for the nominal channel transistors in both 0.18 µm and 0.13 µm technology generations obtained from both the measurements and the model. The increase in stack effect factor at a given V_{dd} with technology scaling is attributed to increase in λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t transistor in 0.13 µm technology generation is due to the same effect.

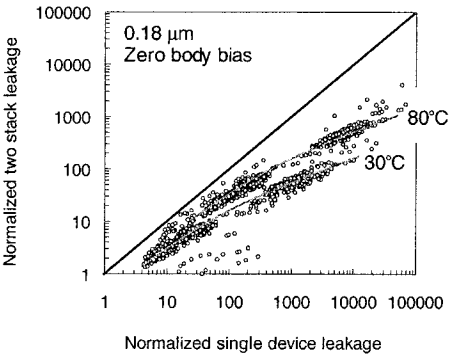

Figure 2-4. Measurement results indicate a slower rate of increase in leakage of two-stack compared to that of a single transistor. This should translate to reduction in the variation of effective threshold voltage.

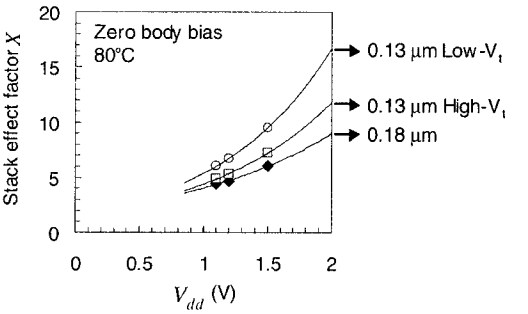

Figure 2-5. Nominal channel length transistor measurement results showing stack effect factor across two technology generations. The increase in stack effect factor is attributed to worsening of short channel effect, λ_d, which is predicted by the analytical model. The higher stack effect factor for the low-V_t transistor in 0.13 µm technology generation is attributed to the same reason. Lines are from analytical model and symbols are from measurement.

In 0.13-µm generation, the low-V_t transistor will dominate chip leakage. Figure 2-6 shows the scaling of stack effect from a 0.18 µm transistor to a 0.13 µm low-V_t transistor based on transistor measurements under different V_{dd} scaling scenarios. Since λ_d is expected to increase due to worsening transistor aspect ratio and since V_{dd} scaling will slow down due to related challenges [9], stack effect leakage reduction factor is expected to increase with technology scaling. The predicted scaling of stack effect factor from 0.18 µm to 0.06 µm is depicted in Figure 2-7.

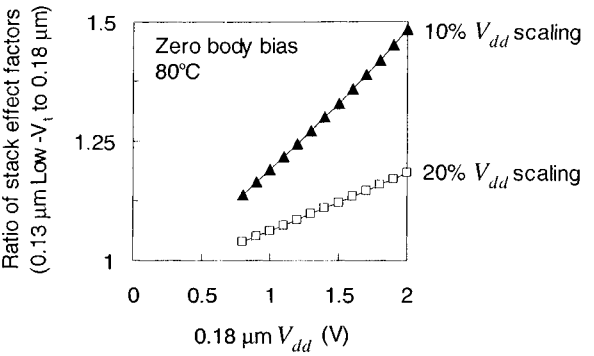

Figure 2-6. Nominal channel length transistor measurement results indicating the scaling of stack effect factor from 0.18 μm to 0.13 μm low-V_t under different V_{dd} scaling conditions. The low-V_t transistor will dominate leakage in 0.13 μm technology, so the comparison is made with the low-V_t transistor.

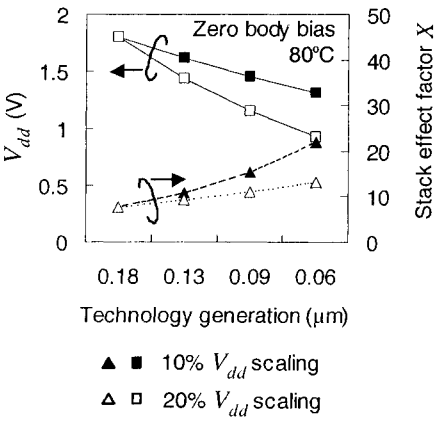

Figure 2-7. Prediction in the scaling of stack effect factor for two V_{dd} scaling scenarios in nominal channel length transistors. V_{dd} for 0.18 μm is assumed to be 1.8 V.

This scaling nature of stack effect factor makes it a powerful technique for leakage reduction in future technologies. In the next sections, we describe a circuit technique for taking advantage of stack effect to reduce leakage at a functional block level. In the first case, the natural stacks present in circuit blocks are used to reduce leakage in standby state, by loading appropriate input vectors to maximize amount of transistor width in stack mode. In the next case, forced stacks are used to minimize leakage of transistors in non-performance critical paths.

2.3 LEAKAGE REDUCTION USING NATURAL STACKS

Typically, a large circuit block contains a significant number of logic gates where transistor stacks are already present, like the PMOS stack in NOR or NMOS stack in NAND gates. The technique described here enables effective leakage reduction during standby mode by loading a vector at the primary inputs of the circuit block so as to maximize the number of PMOS and NMOS stack transistor widths with more than one *OFF* transistor. In contrast to techniques reported in the past [10, 11, 12], the proposed scheme offers leakage reduction with minimal overheads in area, power, and process technology change. In particular, this technique has the potential to replace the need for a high-V_t transistor for standby leakage.

Extensive results from circuit simulations of individual logic gates and a 32-bit static CMOS adder, designed in a 0.1 μm, is discussed to elucidate the dynamics of leakage reduction due to transistor stacks, examine its influence on the overall leakage power of the adder during both active and standby modes of operation, and determine the standby leakage reductions yielded by application of the new leakage control technique. Two different V_t values were considered throughout the analysis. The low-V_t is 100 mV smaller than the high-V_t.

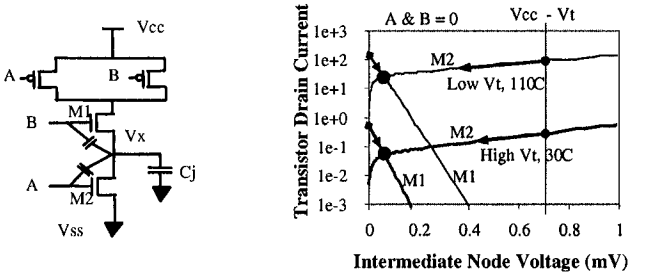

Figure 2-8. 2 NMOS stack in a NAND gate and DC solution for intermediate node voltage.

A 2-input NAND gate is used to illustrate the dynamics of leakage reduction in 2-transistor stacks with both transistors *OFF*, as shown in Figure 2-8. From the DC solution of NMOS sub-threshold current characteristics, shown in Figure 2-8, it is clear that the leakage current through a 2-transistor stack is approximately an order of magnitude smaller than the leakage of a single transistor. This reduction in leakage is can be viewed to come about due to negative gate-to-source biasing and body-effect induced V_t increase in M1, or reduced drain-to-source voltage in M2 which causes its V_t to increase, as the voltage V_x at the intermediate node converges

to ~100 mV. Thus, as shown in Figure 2-9, smaller amounts of leakage reduction are obtained at higher temperatures due to larger sub-threshold swing. For 3- or 4-transistor stacks, the leakage reduction is found to be 2-3X larger in both NMOS and PMOS, as illustrated in Figure 2-10.

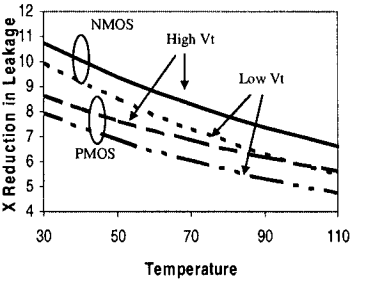

Figure 2-9. Leakage reduction in 2 NMOS and 2 PMOS stacks at different temperatures and different target threshold voltages, from simulations.

	High Vt	Low Vt
2 NMOS	10.7X	9.96X
3 NMOS	21.1X	18.8X
4 NMOS	31.5X	26.7X
2 PMOS	8.6X	7.9X
3 PMOS	16.1X	13.7X
4 PMOS	23.1X	18.7X

Figure 2-10. Leakage current reduction in multiple stacked transistors.

It is essential that we point out an anomaly – according to Figure 2-10, the simulation results show that low-V_t transistors have lower leakage reduction compared to high-V_t transistors. This is contradictory to the measurements and the model derived in the previous section. Low-V_t transistors have larger DIBL therefore should have larger leakage reduction due to stack effect as per the measurements and model. The simulation results due to the models used do not predict the expected behavior of leakage reduction due to stack effect when the V_t is lowered.

Generally speaking, this should be a note of caution to the reader, do not always believe the simulations without proper validation! Absolute values of measured results will probably be different from the simulation results described in this section. It is also important to keep in mind, that measured results will always have a statistical spread of values instead of a single value due to the impact of process variation on leakage, as shown in the previous section. Other than the mentioned threshold voltage related

anomaly the simulation result's ability to quantify the benefit of natural stacks for leakage reduction presented in this section holds.

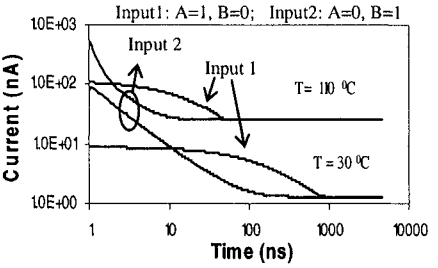

Figure 2-11. Transient behavior of leakage current convergence time constant in a 2 NMOS stack under different temperature and initial input conditions.

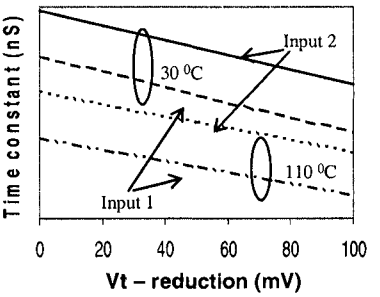

Figure 2-12. Dependence of leakage convergence time constant of stack leakage on threshold voltage, temperature, and initial input conditions.

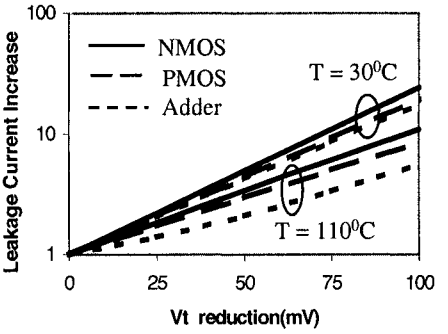

Figure 2-13. Leakage current increase with threshold voltage reduction at the transistor and adder block levels.

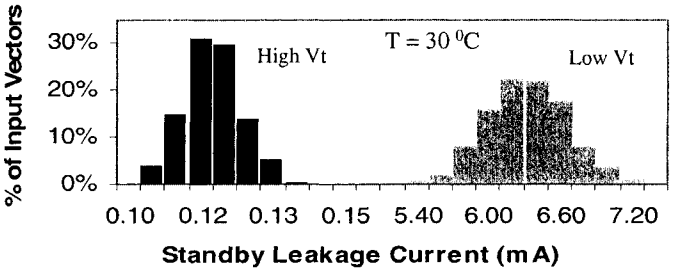

Figure 2-14. Distribution of standby leakage current in the 32-bit adder for a large number of random input vectors.

Back to the simulated data, the time required for the leakage current in transistor stacks to converge to its final value is dictated by the rate of charging or discharging of the capacitance at the intermediate node by the sub-threshold drain current of M1 or M2. This time constant as shown in Figure 2-11 is, therefore, determined by drain-body junction and gate-overlap capacitances per unit width, the input conditions immediately before the stack transistors are turned *OFF*, and transistor sub-threshold leakage current, which depends strongly on temperature and V_t. Therefore, the convergence rate of leakage current in transistor stacks increases rapidly with V_t reduction and temperature increase, as shown in Figure 2-11 and Figure 2-12. For Low-V_t transistors in the 0.1 μm technology, this time constant in 2-NMOS stacks at 110°C ranges from 5-50 ns depending on input conditions before both transistors are turned *OFF*.

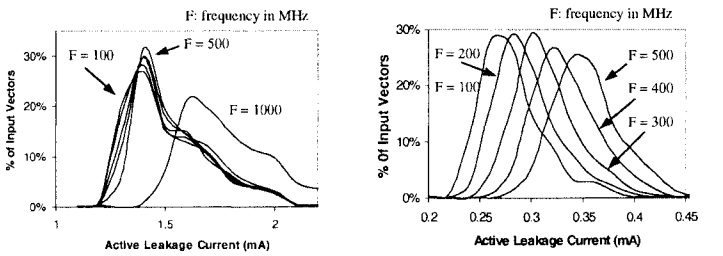

Figure 2-15. Distribution of active leakage current in the 32-bit adder with low-Vt transistors (left) and high-Vt transistors (right) at different frequencies.

Increase in the active and standby leakage of the 32-bit static CMOS Kogge-Stone adder with V_t-reduction, as shown in Figure 2-13, is smaller than that in individual transistors, due to the presence of a significant number of transistor stacks in the design. The standby leakage power varies by 30%-40%, depending on the input vector, as shown in Figure 2-14, which determines the number of transistor stacks in the design with more than one

OFF transistor. Figure 2-15 shows that the adder leakage during active operation is dictated by the sequence of input vectors as well as the operating clock frequency. Magnitude of the stack leakage time constant at elevated temperatures relative to the time interval between consecutive switching events determines the extent of convergence of the leakage to steady-state value. As a result, the active leakage corresponding to each input vector becomes higher as the clock frequency increases from 100 to 1000 MHz resulting in larger average leakage power at higher frequencies.

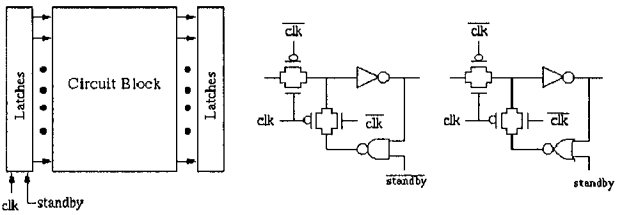

Figure 2-16. Implementation of the standby leakage control using natural stacks through input vector activation.

Figure 2-16 shows an implementation of the new leakage reduction technique where a standby control signal, derived from the clock gating signal, is used to generate and store a predetermined vector in the static input latches of the adder during standby mode so as to maximize the number of NMOS and PMOS stacks with more than one OFF transistor. Since the desired input vector for leakage minimization is encoded by using a NAND or NOR gate in the feedback loop of the static latch, minimal penalty is incurred in adder performance. As shown in Figure 2-17, up to 2X reduction in standby leakage can be achieved by this technique. In order that the additional switching energy dissipated by the adder and latches, during entry into and exit from "standby mode", be less than 10% of the total leakage energy saved by this technique during standby, the adder must remain in standby mode for at least 5 μs, as summarized in Figure 2-18.

A standby leakage control technique, which exploits the leakage reduction offered by natural transistor stacks, was presented. Based on simulation results that showed up to 10X leakage reduction at gate level resulted in up to 2X reduction in standby leakage power. By using natural stacks this can be achieved with minimal overheads in area, power, and process technology change. We also elucidated the dynamics of leakage reduction due to transistor stacks, and its influence on overall leakage power of large circuits. Since with technology scaling the leakage reduction due to stack effect is expected to increase as described in the previous section, this technique will become more effective. Additionally, the time constant for

leakage convergence depends on the sub-threshold leakage current itself, so with scaling this time constant will reduce rapidly due to exponential increase in sub-threshold leakage.

High Vt	Avg. Worst	% Reduction 35.4% 60.7%
Low Vt	Avg. Worst	33.3% 56.5%

Figure 2-17. Adder leakage reduction using the best input vector activation compared to the average and worst case standby leakage causing input vectors.

	High Vt	Low Vt
Savings	2.2uA	0.0384mA
Overhead	1.64 nJ	1.84 nJ
Min. time in standby	84 uS	5.4 uS

Figure 2-18. Standby leakage power savings and the minimum time required in standby mode.

2.4 LEAKAGE REDUCTION USING FORCED STACKS

As shown earlier, stacking of two transistors that are *OFF* has significantly reduced leakage compared to a single *OFF* transistor. However due to the iso-input load requirement and due to stacking of transistors, the drive current of a forced-stack gate will be lower resulting in increased delay. So, stack forcing can be used only for paths that are non-critical, just like using high-V_t transistors in a dual-V_t design [13, 14]. Forced-stack gates will have slower output edge rate similar to gates with high-V_t transistors. Figure 9 illustrates the use of techniques that provide delay-leakage trade-off. As demonstrated in the figure, paths that are faster than required can be slowed down which will result in leakage savings. Such trade-offs are valid only if the resulting path still meets the target delay. Figure 2-19 shows the delay-leakage trade-off due to n-stack forcing of an inverter with fan-out of 1 under iso-input load conditions in a dual-V_t 0.13 μm technology [15].

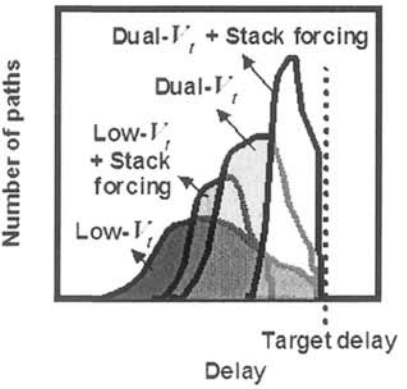

Figure 2-19. Stack forcing and dual-V_t can reduce leakage of gates in paths that are faster than required.

By properly employing forced-stack one can reduce standby and active leakage of non-critical paths even if a dual-V_t process is not available. This method can also be used in conjunction with dual-V_t. Stack forcing provides wider coverage in the delay-leakage trade-off space as illustrated in Figure 2-20.

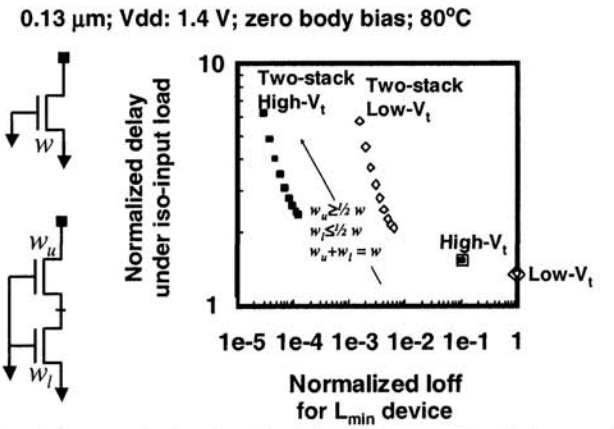

Figure 2-20. Simulation result showing the delay-leakage trade-off that can be achieved by stack forcing technique under iso-input load conditions. Iso-input load is achieved by making the gate area after stack forcing identical to before stack forcing. Several such conditions are possible, which enhances delay-leakage trade-off possible by stack forcing. The two-stack condition for a given V_t with the least delay is for $w_u = w_l = ½w$. This trade-off can be used with or without high-V_t transistors. The simulation anomaly described in Section 2.3 for Figure 2-10 is evident here as well.

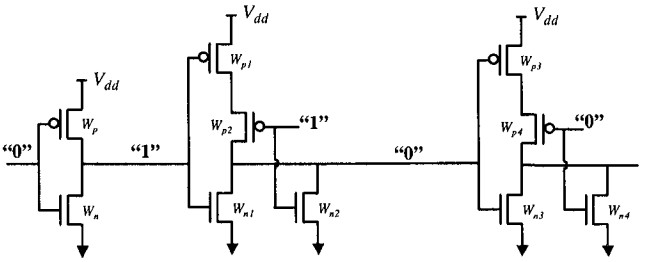

Figure 2-21. A sample path where natural stack is used to reduce standby leakage by applying a predetermined vector during standby. No delay penalty is incurred with this technique.

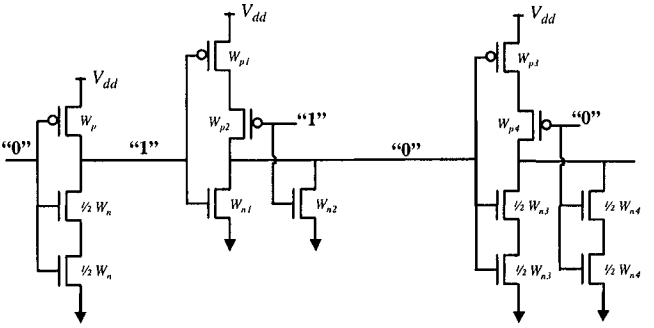

Figure 2-22. Using stack-forcing technique the number of logic gates in stack mode can be increased. This will enable further leakage reduction in standby mode. Increase in delay under normal mode of operation will be incurred.

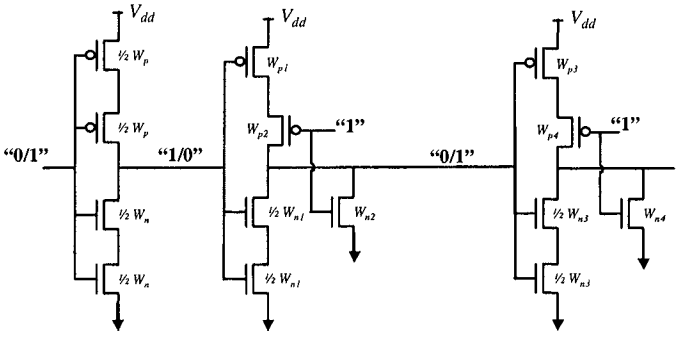

Figure 2-23. If a gate can have its input as either "0" or "1" and still force stack effect then that gate will have reduced active leakage. The more the number of inputs that can be either "0" or "1" the higher the probability that stack effect will reduce active leakage.

Functional blocks have naturally stacked gates such as NAND, NOR, or other complex gates. By maximizing the number of natural stacks in *OFF* state during standby by setting proper input vectors, the standby leakage of

functional block can be reduced, as was explained in the last section. Since it is not possible to force all natural stacks in the functional block to be in *OFF* state the overall leakage reduction at a block level will be far less than the stack effect leakage reduction possible at a single logic gate level [2]. With stack forcing the potential for leakage reduction will be higher. Figure 2-21 and Figure 2-22 illustrates such an example.

Forcing a stack in both n- and p-networks of a gate will guarantee leakage reduction due to stacking, independent of the input logic level. Such an example is shown in Figure 2-23. To reiterate, stack forcing can be applied to paths only if increase in delay due to stacking does not violate timing requirements. Gates that can force stack effect independent of its input vectors will automatically go into leakage reduction mode when the intermediate node of the stack reaches the steady state voltage. This will boost standby and active leakage reduction since no specific input vector needs to be applied.

2.5 SUMMARY

We presented a model based on transistor fundamentals that predicted the scaling nature of stack effect based leakage reduction. Transistor measurements verified the model's accuracy across different temperature, channel length, body bias, supply voltage, and process technology.

A standby leakage control technique, which exploits the leakage reduction offered by natural transistor stacks, was presented. Based on simulation results that showed up to 10X leakage reduction at gate level resulted in up to 2X reduction in standby leakage power. By using natural stacks this can be achieved with minimal overheads in area, power, and process technology change. Modes for using stack forcing to reduce standby and active leakage components were discussed.

Since with technology scaling the leakage reduction due to stack effect is expected to increase as described in the previous section, this technique will become more effective. Additionally, the time constant for leakage convergence depends on the sub-threshold leakage current itself, so with scaling this time constant will reduce rapidly due to exponential increase in sub-threshold leakage. These reasons make the stack effect based leakage reduction techniques attractive in nanoscale CMOS circuits.

REFERENCES

[1] S. Narendra, S. Borkar, V. De, D. Antoniadis, and A. Chandrakasan, "Scaling of stack effect and its application for leakage reduction," *Intl. Symp. Low Power Electronic and Design,* pp. 195-200, 2001.

[2] Y. Ye, S. Borkar, and V. De, "A Technique for Standby Leakage Reduction in High-Performance Circuits," *Symp. of VLSI Circuits*, pp. 40-41, 1998.

[3] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Simultaneous sub-threshold and Gate-Oxide Tunneling Leakage Current in Nanometer CMOS Design," *Intl. Symp. Low Power Electronic and Design*, pp. 287-292, 2003.

[4] A. Grove, http://www.intel.com/pressroom/archive/speeches/grove_ 20021210.pdf, *IEDM 2002 Keynote Luncheon Speech.*

[5] J.P. Halter and F. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits," *Custom Integrated Circuits Conf.*, pp. 475-478, 1997.

[6] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," *Intl. Symp. Low Power Electronics and Design,* pp. 239-244, 1998.

[7] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*, IEEE Press, pp. 46-47, 2000.

[8] Z. Liu, C. Hu, J. Huang, T. Chan, M. Jeng, P. Ko, and Y. Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Transactions on Electron Transistors*, vol. 40, no. 1, pp. 86-95, January 1993.

[9] Y. Taur, "CMOS Scaling beyond 0.1μm: how far can it go?" *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 6-9, 1999.

[10] S. Thompson et. al., *Symp. VLSI Tech.*, pp. 69-70, 1999.

[11] S. Mutoh et. al, *IEEE JSSC*, pp. 847-854, Aug. 1995.

[12] T. Kuroda et. al, *IEEE JSSC*, pp. 1770-1779, Nov. 1996.

[13] L. Su, R. Schulz, J. Adkisson, K. Beyer, G. Biery, W. Cote, E. Crabbe, D. Edelstein, J. Ellis-Monaghan, E. Eld, D. Foster, R. Gehres, R. Goldblatt, N. Greco, C. Guenther, J. Heidenreich, J. Herman, D. Kiesling, L. Lin, S-H. Lo, McKenn, "A high-performance sub-0.25 μm CMOS technology with multiple thresholds and copper interconnects," *Intl. Symp. on VLSI Technology, Systems, and Applications*, pp. 18-19, 1998.

[14] D. T. Blaauw, A. Dharchoudhury, R. Panda, S. Sirichotiyakul, C. Oh, and T. Edwards "Emerging power management tools for processor design," *Intl. Symp. Low Power Electronics and Design,* pp. 143-148, 1998.

[15] S. Tyagi, M. Alavi, R. Bigwood, T. Bramblett, J. Bradenburg, W. Chen, B. Crew, M. Hussein, P. Jacob, C. Kenyon, C. Lo, B. Mcintyre, Z. Ma, P. Moon, P. Nguyen, L. Rumaner, R. Schweinfurth, S. Sivakumar, M. Stettler, S. Thompson, B. Tufts, J. Xu, S. Yang, and M. Bohr, "A 130 nm Generation Logic Technology Featuring 70 nm Transistors, Dual Vt Transistors and 6 layers of Cu Interconnects," *Intl. Elec. Transistors Meeting*, pp. 567-570, December 2000.