Subthreshold Leakage Modeling and Reduction Techniques

James Kao[1,3], Siva Narendra[2,3], Anantha Chandrakasan[3]
[1]*Silicon Labs* [2]*Intel Labs* [3]*Massachusetts Institute of Technology*
jkao@alum.mit.edu, siva.g.narendra@intel.com, anantha@mtl.mit.edu

Abstract

As technology scales, subthreshold leakage currents grow exponentially and become an increasingly large component of total power dissipation. CAD tools to help model and manage subthreshold leakage currents will be needed for developing ultra low power and high performance integrated circuits. This paper gives an overview of current research to control leakage currents, with an emphasis on areas where CAD improvements will be needed. The first part of the paper explores techniques to model subthreshold leakage currents at the device, circuit, and system levels. Next, circuit techniques such as source biasing, dual V_t partitioning, MTCMOS, and VTCMOS are described. These techniques reduce leakage currents during standby states and minimize power consumption. This paper also explores ways to reduce total active power by limiting leakage currents and optimally trading off between dynamic and leakage power components.

1. Background

Energy per operation, a key figure of merit in digital circuits, continually improves with process and supply voltage scaling. As a result, power supply voltage has scaled aggressively with each process generation. In order to sustain the traditional 30% improvement in gate delay for digital circuits each generation, MOSFET device threshold voltages (V_t) must scale aggressively as well. However, a reduction in V_t will cause an exponential increase in the device subthreshold leakage current (I_{off}). The drain current of a MOSFET in the subthreshold region can be expressed as (assuming V_{ds} is large compared to the thermal voltage),

$$I_{subthreshold} = I_0 10^{\frac{(V_{gs}-V_t)}{S}} \quad (1)$$

where I_o is the drain current with $V_{gs}= V_t$, and S is the subthreshold slope [1]. The subthreshold leakage current of a single MOS device (with gate and source grounded) is obtained by setting $V_{gs}=0$ in the above equation. Figure 1 shows that reducing V_t by about 85 mV increases the I_{off} by an order of magnitude. This relationship between I_{off} and V_t is captured by the subthreshold swing, S, which in this example is 85mV/decade. In the past, subthreshold leakage currents have been relatively small components of the total power consumed in digital circuits. However, as technology continues to scale, this exponentially increasing component can no longer be ignored.

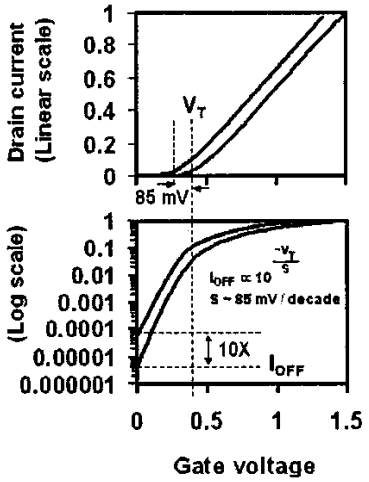

Figure 1. Relationship between threshold voltage (V_t) and subthreshold leakage current (I_{off}).

For example, projections show that in the 90nm process generation node, subthreshold leakage power can contribute as much as 42% of the total power as illustrated in Figure 2. For future technologies, it becomes essential to (*i*) predict the subthreshold leakage power using accurate models and (*ii*) enable design methodologies for known circuit techniques that reduce subthreshold leakage power. This tutorial paper presents an overview of current work in leakage estimation and control techniques.

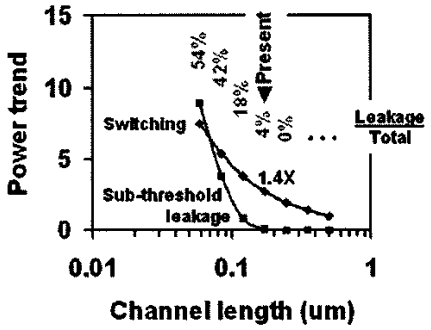

Figure 2. Trend in subthreshold leakage power and switching power with technology scaling.

2. Leakage Estimation

Subthreshold leakage currents are exponentially dependent on the threshold voltage, so any accurate model for leakage currents must take into account V_t variation effects. In particular, short channel effects can modulate the threshold voltage when the channel length of a MOSFET approaches the source-body and drain-body depletion widths. The charge in the channel due to these parasitic diodes become comparable to the depletion charge due to the MOSFET gate-body voltage [2], rendering the gate and body terminals less effective. As the band diagram illustrates in Figure 3, the finite depletion width of the parasitic diodes does not influence the energy barrier height that a long channel device must overcome to invert the channel.

However, as the channel length becomes shorter, both channel length and drain voltage reduce this barrier height. This two-dimensional short channel effect allows variations in channel length to change the barrier height. This condition manifests itself as threshold variation as shown in Figure 4.

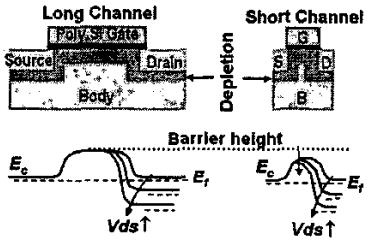

Figure 3. Barrier height lowering due to channel length reduction and drain voltage increase in an nMOS.

For short channel MOS devices, the drain to source voltage and channel length affects V_t, which can be described below [3].

$$V_t = V_{fb} + |2\phi_p| + \frac{\lambda_b}{C_{ox}}\sqrt{2qN\varepsilon_s(|2\phi_p|+V_{sb})} - \lambda_d V_{ds}; \quad (2)$$

The above equation resembles that of a long channel device except for the terms for body effect factor[1], λ_b, [2] and DIBL, λ_d [4]. Note that the λ_d model in [4] is empirical and cannot be assumed to be universally true.

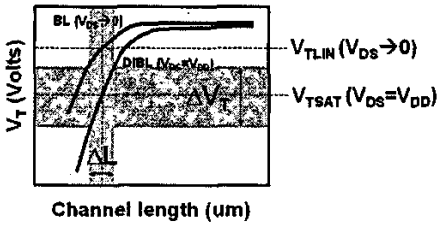

Figure 4. Barrier lowering (BL) resulting in V_t roll-off with channel length reduction. Drain induced barrier lowering (DIBL) reduces V_t for short channel devices and increases V_t roll-off. For short channel devices channel length variation (ΔL) translates to threshold voltage variation (ΔV_t).

Since these device level effects contribute exponentially to subthreshold leakage power, it is important to consider within-die threshold voltage variations to model subthreshold leakage power accurately. The rest of this section explores different system level leakage estimation techniques based on fundamental device and architectural parameters. Most of this discussion will focus on standby subthreshold leakage, which corresponds to the total power dissipated by MOS devices that are "off" when there is no system level activity. Some additional work is then presented to model active leakage currents, which correspond to the leakage components of the MOS devices that are switched off when the system is operational.

2.1 Standby leakage current bound estimation

Due to the wide variation expected in the die-to-die and within-die V_t of MOS devices during the lifetime of a process technology, present subthreshold leakage current prediction techniques provide lower and upper bounds on the subthreshold leakage current. In older technology generations, basing system design on the two subthreshold leakage current bounds was acceptable since subthreshold leakage power was a negligible component of the total power.

The lower bound subthreshold leakage ($I_{leak\text{-}l}$) prediction of a chip is given as follows,

$$I_{leak\text{-}l} = \frac{w_p}{k_p} I^o_p + \frac{w_n}{k_n} I^o_n \quad (3)$$

where w_p and w_n are the total PMOS and NMOS device widths in the chip; k_p and k_n are factors that determine percentage of PMOS and NMOS device widths that are in the "off" state; I^o_p and I^o_n are the nominally expected subthreshold leakage currents per unit width of PMOS and NMOS devices in a particular chip. The nominal subthreshold leakage current is obtained for devices with mean threshold voltage or channel length. The upper bound subthreshold leakage current ($I_{leak\text{-}u}$) prediction of a chip is related to the device subthreshold leakage as follows,

$$I_{leak\text{-}u} = \frac{w_p}{k_p} I^{3\sigma}_{off-p} + \frac{w_n}{k_n} I^{3\sigma}_{off-n} \quad (4)$$

where, $I^{3\sigma}_{off\text{-}p}$ and $I^{3\sigma}_{off\text{-}n}$ are the worst-case subthreshold leakage current per unit width of PMOS and NMOS devices. The worst case subthreshold leakage current is obtained for devices with threshold voltage or channel length 3σ lower than the mean subthreshold leakage currents per unit width of PMOS and NMOS devices in a particular chip.

In the next section a leakage current estimation model that predicts the leakage based on fundamental device and architectural parameters by including within-die variation will be described [5]. An analytical approach for leakage estimation including within-die variations is discussed in [6]. The benefit of an accurate model that is based on fundamental device and architectural parameters is its ability to predict leakage power of a design that has not been fabricated yet.

2.2 Standby leakage estimation including within-die variation

To include the impact of within-die threshold voltage or channel length variation, it is necessary to consider the entire range of subthreshold leakage currents, not just the mean subthreshold leakage or the worst case subthreshold leakage. Let us assume that the within-die threshold voltage or channel length variation follows a normal distribution with respect to transistor width, with μ being the mean and σ being the sigma of the distribution. Let I^o be the subthreshold leakage of the device with the mean threshold voltage or channel length. Then by performing the weighted sum of devices of different subthreshold leakage, we can predict the total subthreshold leakage of the chip. This is achieved by integrating the threshold voltage or channel length distribution multiplied by the subthreshold leakage, as shown below,

$$I_{leak} = \frac{I^o w}{k} \frac{1}{\sigma\sqrt{2\pi}} \int_{xmin}^{xmax} e^{\frac{-(x-\mu)^2}{2\sigma^2}} e^{\frac{(\mu-x)}{a}} dx \quad (5)$$

[1] The body effect factor captures Barrier Lowering (BL) in short channel MOS devices and the resulting reduction in the body terminal's influence on the channel charge.

In the above equation, the first exponent predicts the fraction of the total width for the device subthreshold leakage predicted by the second exponent. If the distribution considered within-die is threshold voltage variation then x in the above equation represents threshold voltage and a will be equal to $n\phi_t$, where ϕ_t is the thermal voltage and n is $1+(C_d/C_{ox})$ [1]. If the distribution considered is channel length then x in the above equation will represent channel length, l, and a will be equal to λ. λ can be predicted for a technology by measuring the relationship between channel length and device subthreshold leakage. In the rest of this section, we will assume that the distribution of interest is the channel length, since this parameter is used to characterize a technology. With this assumption the integral can be simplified [5] and replicated for NMOS and PMOS devices resulting in subthreshold leakage current of a chip as,

$$I_{leak\text{-}w} = \frac{I_p^o w_p}{k_p} e^{\frac{\sigma_p^2}{2\lambda_p^2}} + \frac{I_n^o w_n}{k_n} e^{\frac{\sigma_n^2}{2\lambda_n^2}} \qquad (6)$$

where, w_p and w_n are the total PMOS and NMOS device widths in the chip; k_p and k_n are factors that determine percentage of PMOS and NMOS device widths that are in "off" state; I^o_p and I^o_n are the expected mean subthreshold leakage currents per unit width of PMOS and NMOS devices in a particular chip; σ_p and σ_n are the standard deviation of channel length variation within a particular chip; λ_p and λ_n are constants that relate channel length of PMOS and NMOS devices to their corresponding subthreshold leakages.

To compare the power estimation accuracy of the various models standby subthreshold leakage power measurements on 960 samples of a 0.18 µm 32-bit microprocessor were carried out. The subthreshold leakage current (with V_{gs} = 0 V and $V_{ds} = V_{DD}$) and effective channel length measurements of test devices that accompany each microprocessor were measured to determine I^o_p, I^o_n, λ_p, and λ_n. Using these individual device measurements, with w_p and w_n obtained from the design, the subthreshold leakage power was calculated using the $I_{leak\text{-}l}$, $I_{leak\text{-}u}$, and $I_{leak\text{-}w}$ formulae.

The prediction accuracy of the formulae is summarized in Figure 5. As the figure indicates the subthreshold leakage power for most of the samples are under predicted by 6.5X if the lower bound technique is used and over predicted by 1.5X if the upper bound technique is used. The measured-to-calculated subthreshold leakage ratio for the majority of the device samples is 1.04 for the new technique described in this paper. The calculated subthreshold leakage is within ±20% of the measured subthreshold leakage for more than 50% of the samples if the new $I_{leak\text{-}w}$ technique is used. Only 11% and 0.2% of the samples fall into this range for the $I_{leak\text{-}u}$ and $I_{leak\text{-}l}$ techniques respectively.

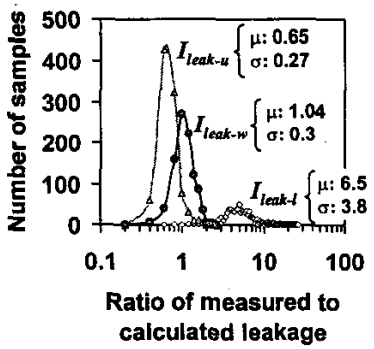

Figure 5. Ratio of measured to calculated subthreshold leakage current distribution for $I_{leak\text{-}u}$, $I_{leak\text{-}l}$, and $I_{leak\text{-}w}$ techniques (Sample size: 960).

2.3 CAD tool requirements for leakage estimation

The techniques mentioned above can also be used to estimate full-chip active leakage power by dividing the entire chip into multiple iso-temperature and iso-power supply regions. I^o_p, I^o_n, w_p, w_n, λ_p, and λ_n will have to be determined for each region. Since the die temperature and power supply voltage depends on the total power consumed in different regions, for active leakage power estimation it is therefore necessary to solve thermal and leakage estimation models simultaneously. Development of electro-thermal computer aided design tools that comprehend within-die device and environmental variation will be vital for sub-100 nm design.

If a chip can be experimentally measured, active leakage currents can be directly quantified. For example, [7] describes an accurate way to measure the active leakage power component by extrapolating from the slope of energy-versus-time curves for different operating frequencies in a digital circuit. For a fixed supply, the switching energy to perform an operation is independent of frequency, but the active leakage energy is proportional to the execution time. With this insight, it is possible to then decouple the total active power consumption into the switching component and the subthreshold leakage component. In actuality the linear term contains all DC current terms, of which subthreshold leakage is the most dominant.

The subthreshold leakage power estimation modeling described so far did not assume the use of any leakage management techniques. Some of these techniques will be described in the subsequent section of this tutorial. When these techniques are employed, the leakage estimation model will have to be modified to comprehend these changes. Computer aided design tools that can enable active and standby subthreshold leakage calculation with "what-if" benefit analysis of different leakage reduction techniques will also be of interest.

3. Circuit Techniques to Manage Subthreshold Leakage

An important area of research today is developing circuit techniques to reduce subthreshold leakage currents in both the active and standby periods to minimize total power consumption. Standby leakage currents are especially wasteful in burst mode systems (such as cell phones or pagers) where circuits spend a significant portion of the time in an idle mode where no computation takes place. A large number of circuit techniques have been developed to turn off these leakage currents when performance is not needed. As technology continues to scale, subthreshold leakage currents become so large that they must be balanced during the active state as well.

In general, there have been two main approaches to control subthreshold leakage currents: source biasing and direct V_t manipulation. In source biasing, the main idea is to bias the source terminal of an "off" transistor in order to exponentially reduce the leakage currents of that device. The other way to lower subthreshold leakage currents is to directly adjust the V_t of transistors within the circuit. This can be accomplished by using a multiple threshold voltage process where a combination of low and high V_t devices can be used to select between high performance and low leakage requirements. Alternatively, a variable threshold voltage technology (like a triple well process using body biasing) could also be used to explicitly alter the threshold voltage.

3.1 Source biasing

The source biasing principle is illustrated in Figure 6, where a positive bias is applied during the standby state to the source terminal of an "off" device.

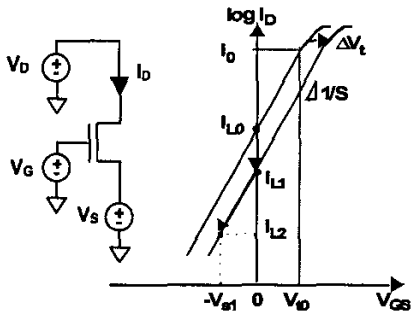

Figure 6. Source biasing effect with applied V_G.

The subthreshold leakage current is significantly reduced because the body effect raises the threshold voltage of the device. Additionally, the V_{gs} voltage becomes negative. The net effect is that the "off" device is turned off more strongly and leakage currents can be reduced during standby modes. This source biasing principle is the underlining mechanism for several different standby leakage reduction schemes. The switched source impedance concept [8] is a special case where a degenerating resistor is used to generate the biased source voltage. For high performance, the degenerating resistor is bypassed to ground, but during the standby state, the resistor is used to bias the source terminal of the "off" device. Another variation known as self reverse biasing [9,10] replaces the switched source impedance with another "off" transistor so that the equilibrium value is set through a series of "off" devices. This technique was first applied to decoded wordline driver circuits.

3.2 Stack effect

A final example of the source biasing principle is illustrated by using transistor stacks within the logic gates to control leakage [11,12]. In effect, two series-connected "off" transistors will have lower leakage currents compared to a single "off" device due to self-reverse biasing effects. Figure 7 shows a load line technique illustrating the difference between one "off" device versus two series "off" devices. For the case where the two series devices are both turned off, the leakage currents are smaller since the internal series node causes V_{gs} for the upper device to become negative. Another way to interpret the leakage is that V_{ds} of lower device will be V_x, which is much smaller (and thus can no longer be ignored in Equation 1) than that of the single "off" device whose V_{ds} will be V_{DD}.

Modeling of leakage reduction factor, X, due to a stack of two series "off" devices can be expressed based on fundamental technology parameters [13] as,

$$X = 10^{\frac{\lambda_d V_{dd}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)} = 10^{U} \qquad (7)$$

where U is the universal two-stack exponent which depends only on the process parameters, DIBL (λ_d) and subthreshold swing (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 $3kT/q$. In the above equation we assumed $w = w_u = w_l$. A more generic equation can be found in [13].

If an appropriate vector can be clocked into a logic block during the standby state, then leakage currents can be reduced by maximizing the number of series-connected "off" paths. In [14], this idea was extended to inserting extra series "off" devices into single stack paths. This provides moderate leakage reduction while using a standard single threshold voltage technology. One difficulty with this approach is developing the proper CAD tools to identify single stack candidates with enough slack such that inserting extra series devices will not adversely impact performance.

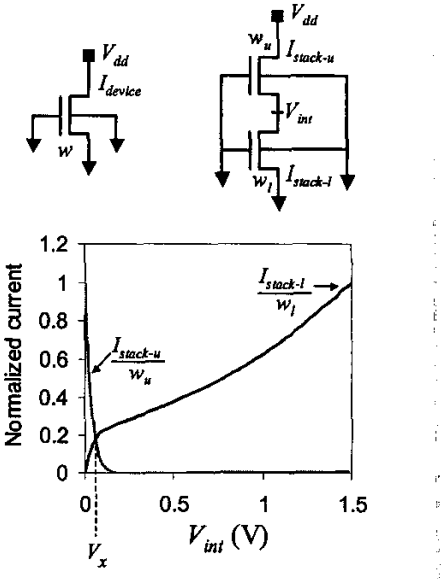

Figure 7. Leakage reduction due to stack effect for series "off" devices.

3.3 Dual V_t partitioning

Subthreshold leakage currents are exponentially dependent on device V_t, and can be changed by several orders of magnitude by switching between high and low threshold voltages. In many modern processes, multiple threshold voltage devices are readily available to the circuit designer. A dual V_t process requires only an extra mask layer to select between high and low threshold voltages, which provide the designer with transistors that are either fast (but high leakage) or slow (but low leakage).

A straightforward way to take advantage of these modern technologies is through a dual V_t partitioning procedure. A circuit can be partitioned into high and low threshold voltage gates or transistors, which will trade off between performance and reduced leakage currents. For instance, critical paths within a circuit should be implemented with low V_t to maximize performance, while non-critical paths should be implemented with high V_t devices to minimize leakage currents. By using fast, leaky devices only when necessary, leakage currents can be significantly reduced in both the standby and active modes compared to an all low V_t implementation.

Dual V_t partitioning is a popular leakage reduction technique because the circuit operation remains the same as for a single V_t implementation, yet critical parts of a circuit can use scaled V_t devices to maintain performance at low supply voltages. [15,16] are examples of research in low power processors that have utilized this technique.

There are practical limitations to the use of dual V_t partitioning to reduce leakage currents. In many optimized designs there are many critical delay paths. Therefore, a large fraction of all paths in the circuit must be implemented with low V_t devices, which reduces the effectiveness of this technique. Another limitation is that CAD tools must be developed and integrated into the design flow to help optimize the partitioning process. It is not straightforward to identify which gates can be made high and low V_t without changing the delay profiles of the circuit. For example, one partitioning scheme that can be applied to

random combinational logic is to first implement the circuit with all low V_t devices to ensure the highest possible performance, and then to selectively implant non-critical gates to be high V_t. However, not all non-critical gates can be converted to high V_t without degrading performance as illustrated in Figure 8.

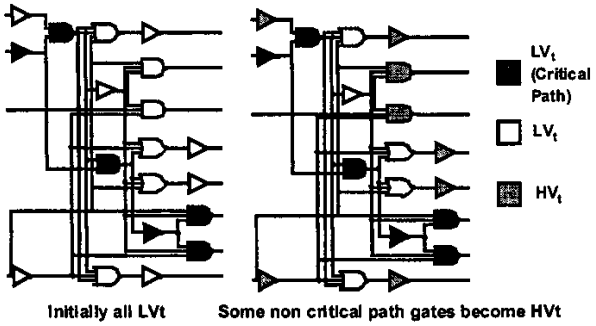

Figure 8. Dual V_t partioning scheme where only some non-critical path gates can be made high V_t.

[17] describes a cell-by-cell method for assigning dual threshold voltage values while [18] describes a breadth-first search based algorithm for selecting and assigning optimal high V_t gates to ensure that timing constraints are maintained. Better leakage reduction can be achieved if individual transistors within gates themselves are optimized to have multiple threshold voltage options. [19] proposes an initial methodology for synthesizing dual V_t circuitry using readily available tools. A recent research direction has been to explore dual V_t partitioning techniques in conjunction with transistor sizing algorithms to provide even better performance and leakage reduction [20-22].

Significant research is still required to improve dual V_t partitioning algorithms, and to develop effective tools that will integrate with existing design techniques. However, there exists a natural limit where dual V_t partitioning may not reduce standby leakage currents enough for ultra low power, high performance applications. As a result, several other leakage reduction techniques are important.

3.4 Multi-threshold CMOS (MTCMOS)

MTCMOS (multi-threshold CMOS) is a dual V_t technique that is very effective at reducing standby leakage currents [23]. The basic principle is illustrated in Figure 9 where a low V_t computation block is gated with high V_t power switches (sleep devices).

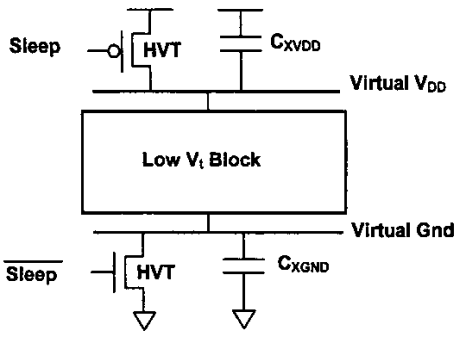

Figure 9. MTCMOS principle showing both polarity sleep devices and parasitic virtual power/ground capacitances. Capacitances help filter out transient ground bounce but not DC components.

The high V_t switches are used to disconnect the power supplies during the standby state, resulting in very low leakage currents set by the high threshold voltage of the series transistor. During the active state, the high V_t switches are turned on, and the internal logic transitions through fast low V_t devices. Although both PMOS and NMOS gating transistors are shown in Figure 9, only one polarity sleep device is actually required to reduce leakage currents if the logic block is purely combinational.

3.4.1 Sleep transistor sizing

MTCMOS circuits require proper sizing of the sleep transistor. This device needs to be large enough to maintain performance, yet still be area efficient and incur minimum energy overhead when switching between modes. Because sleep transistors can be very large in an MTCMOS circuit, it is important to develop CAD tools to efficiently size these devices to ensure functionality and to minimize costs [24].

The optimal way to size sleep transistors is to identify the worst case MTCMOS critical path to size the sleep transistor. Unfortunately, identifying the worst case path is not straightforward (requiring an exhaustive search for all possible vector transitions) because the delay not only depends on a signal propagating through a datapath, but also on how all other non-critical gates switch and contribute to the virtual ground or power bounce. Figure 10 illustrates how two different vectors that exercise the same critical paths in a CMOS multiplier cell will behave differently if implemented as an MTCMOS block (using an NMOS sleep transistor). The second vector exercises many extra transitions throughout the cell and requires a much larger sleep transistor to maintain performance.

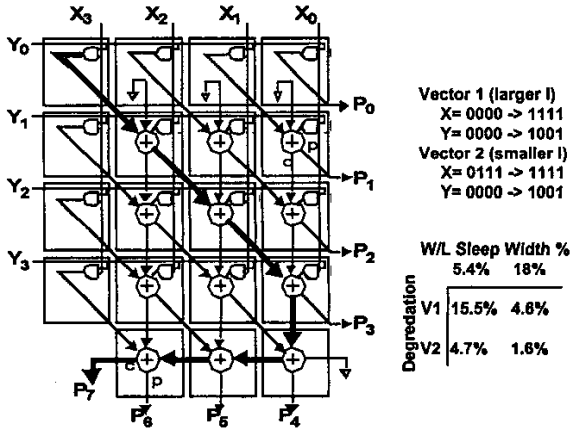

Figure 10. MTCMOS multipler showing dependency on sleep transistor sizing.

It is important to develop CAD tools to size MTCMOS sleep transistors and to trade off between minimizing silicon area overhead and increasing design complexity. [25] describes a preliminary sizing algorithm that ensures that the MTCMOS circuit will always meet performance constraints without performing an exhaustive search of all possible input vectors. This technique relies on dividing an MTCMOS block into smaller mutually exclusive pieces that can be simulated more easily, and then merged together to determine the required sleep transistor width.

3.4.2 MTCMOS sequential circuits

MTCMOS circuit techniques are effective for controlling leakage currents in combinational logic, but a drawback is that it can cause internal nodes to float and corrupt stored data in standard memory designs. As a result, much research has explored MTCMOS latch

designs that can eliminate leakage currents yet maintain state during standby modes. [23,26] proposed a simple MTCMOS latch that uses extra high V_t CMOS gates to provide an always powered re-circulation path during the sleep state. However, this circuit requires both polarity sleep devices with local power switches (separate from the shared virtual power and virtual ground lines) in order to eliminate leakage paths. [27] proposed a "balloon" circuit that decouples the high V_t storage nodes from the MTCMOS nodes so shared virtual power and ground lines could still be used, but adds unnecessary complications to the flip flop operation when entering and leaving the sleep state.

Work in [28] develops a more detailed analysis of sneak leakage paths in sequential MTCMOS circuits, and proposes several novel structures that have better performance than the standard MTCMOS latch and have simpler operation compared to the balloon circuits. Figure 11 shows an improved MTCMOS flip flop that avoids leakage paths yet enables one to share virtual power and ground lines.

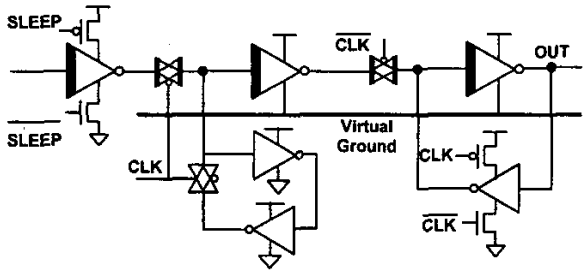

Figure 11. Improved MTCMOS Flip Flop

Figure 12 shows how "leakage feedback" gates [28] can be used to implement dynamic flip flops that can retain state during standby. The leakage feedback gate uses subthreshold leakage currents to hold the output of the gate to a solid logic level, yet fully eliminates subthreshold leakage paths from V_{DD} to ground. This ability to provide a constant output logic level during standby states makes the leakage feedback gate ideal for interfacing between MTCMOS blocks and CMOS logic blocks as well.

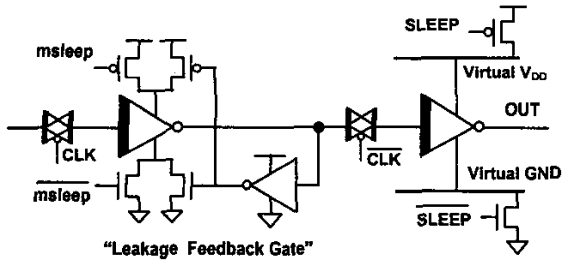

Figure 12. Dynamic flip flop using leakage feedback gate.

3.5 Dual threshold voltage domino logic

In order to cut off subthreshold leakage currents in CMOS gates, it is only necessary to strongly turn off either the PMOS pullup path or the NMOS pulldown path. For example, a single high V_t switch can be used like in MTCMOS, or high V_t transistors of the same polarity can be imbedded directly into the pullup or pulldown path of a gate. In both cases, leakage currents can be reduced with only one transition direction being degraded.

Dual threshold voltage domino logic is a leakage reduction technique that takes advantage of the known transition directions in domino logic [24]. By using either PMOS or NMOS high V_t "off" transistors imbedded within the gates themselves, it is possible to reduce the leakage currents without requiring external sleep devices or compromising performance. Domino logic is configured into a known state during the precharge phase, and is allowed to transition only in a fixed direction during the evaluate phase. Dual V_t domino logic implements all devices in the critical charge/discharge path with low V_t transistors and implements all devices that switch during precharge with high V_t transistors as illustrated in Figure 13.

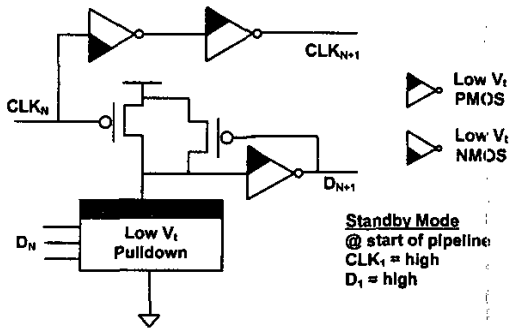

Figure 13. Dual V_t domino logic gate showing Nth stage of a pipeline.

By stalling the circuit during the active state, all high V_t devices are strongly turned off, which reduces standby leakage currents. In effect, this structure trades off increased precharge time (which is not in the critical path) with reduced leakage during the standby period.

Similar dual V_t methodologies can be applied to more general CMOS circuits as well. If the circuit has transitions that are only critical in one direction, or if the circuit is in a known configuration during the standby state, then an intelligent partitioning of dual V_t devices can yield improved performance and more efficient leakage reduction.

3.6 Variable threshold CMOS

Variable threshold CMOS, or VTCMOS, is another technique that has been developed to reduce standby leakage currents. Rather than employ multiple threshold voltage process options, VTCMOS relies on a triple well process where the device V_t is dynamically adjusted by biasing the body terminal. By applying maximum reverse biasing during the standby mode, the threshold voltage is shifted higher and subthreshold leakage current reduced. An example of the VTCMOS principle can be seen in [29] and analytical models describing the effectiveness of VTCMOS can be found in [30].

The VTCMOS approach provides the additional advantage that threshold voltage can be tuned during active mode to optimize performance. [31] uses VTCMOS to dynamically adjust V_t based on workload requirements to reduce active leakage currents. In [32,33], adaptive body biasing approaches were developed where circuit threshold voltages are tuned in a feedback loop so that critical paths operate only as fast as necessary. By allowing both forward and reverse body biasing directions, one can improve process yields and reduce leakage currents resulting from unexpected process variations. Finally VTCMOS provides circuit designers complete flexibility to set both V_{DD} and V_t during active modes to optimally balance between performance, dynamic power, and leakage power.

3.7 Optimal V_{DD}/V_t operating points

During the standby state, it is always beneficial to minimize the subthreshold leakage by making the effective "off" transistor V_t as large as possible. However, it is not so straightforward to effectively reduce subthreshold leakage currents during the active state. As described

earlier, a first step to reducing total active power is to implement non-critical blocks with high V_t devices wherever possible. However, for the rest of the circuit, there is a direct trade-off between reducing active leakage currents and improving performance.

With supply voltage scaling, dynamic power consumption is reduced quadratically, but performance can be maintained if threshold voltages scale at the expense of higher active leakage currents. For a particular operating condition, a circuit will exhibit a minimum in the total active power consumption corresponding to the balance point where an incremental decrease in dynamic switching power due to V_{DD} scaling is offset by an increase in the subthreshold leakage power due to V_t scaling. As a result, it may be more optimal to operate with higher active leakage currents if the voltage supply is scaled aggressively. On the other hand, for very leaky technologies it may be more optimal to operate at higher supply voltages if the leakage power can be reduced.

[34] explores in more detail how a variable threshold CMOS technology can be used to dynamically tune both V_{DD} and V_t as operating conditions change. Figure 14 shows theoretical curves for a processor showing power versus V_{DD} scaling (assuming V_t's are implicitly chosen). For different workload conditions, there exist different optimal combinations of V_{DD} and V_t that minimize the total power consumption. These operating points can change significantly as operating conditions change.

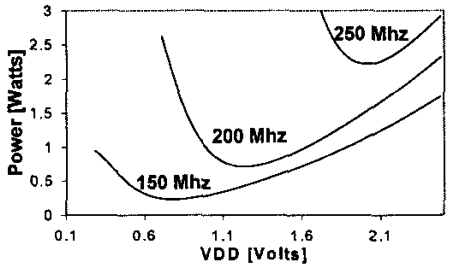

Figure 14. Impact of operating frequency on power vs. V_{DD} curves (with V_t implicitly set).

Figure 15 shows an architecture where a processor core's supply voltage and body bias can be automatically controlled to minimize total power consumption for a given throughput requirement. This two-dimensional control problem can be decoupled into two independent control loops where the V_{DD} controller searches for the lowest power supply point, while the body bias controller automatically adjusts the V_t to maintain performance at each operating point.

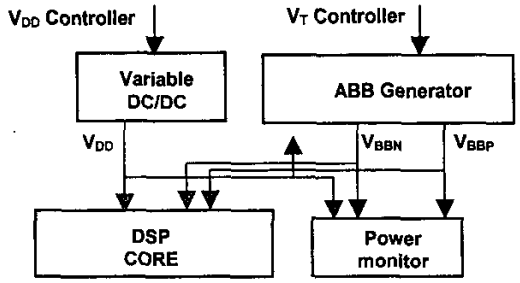

Figure 15. Architecture for minimum active power operation.

4.0 Conclusion

Technology scaling and Moore's law are driving forces behind the semiconductor industry. An unavoidable consequence is that subthreshold leakage currents for new technologies will become increasingly large. It is important to quantify the impact of leakage currents on overall power consumption and to develop circuit techniques for reducing their impact while maintaining performance. This tutorial describes many active research topics for quantifying and reducing leakage currents in the standby and active states. It is important to develop new CAD tools that will account for subthreshold leakage currents. Not only must leakage considerations be integrated into the design flow, but breakthroughs in efficiently implementing the leakage reduction techniques described in this tutorial will be necessary.

Acknowledgements

We would like to thank Ben Calhoun for providing valuable comments on this paper.

References

[1] Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, 1998.

[2] H.C. Poon, L.D. Yau, R.L. Johnston, D. Beecham, "DC Model for Short-Channel IGFET's," IEEE Intl. Electron Devices Meeting, pp. 156-159, December 1973.

[3] S. Narendra, D. Antoniadis, and V. De, "Impact of Using Adaptive Body Bias to Compensate Die-to-die V_t Variation on Within-die Vt variation," Intl. Symp. Low Power Electronics and Design, August 1999.

[4] K.K. Ng, S.A. Eshraghi, and T.D. Stanik, "An improved generalized guide for MOSFET scaling," IEEE Transactions on Electron Devices, vol. 40, pp. 1895-1897, October 1993.

[5] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A. Chandrakasan, "Full-chip Sub-threshold Leakage Power Prediction Model for sub-0.18 μm CMOS," Intl. Symp. Low Power Electronics and Design, pp. 19-23, August 2002.

[6] A. Srivastava, R. Bai, D. Blaauw, and D. Sylvester, "Modeling and Analysis of Leakage Power Considering Within-Die Process Variations," Intl. Symp. Low Power Electronics and Design, pp. 64-67, August 2002.

[7] Amit Sinha and Anantha Chandrakasan, "JouleTrack - A Web Based Tool For Software Energy Profiling", Proc. 38th Design Automation Conference, pp. 220-225, June 2001.

[8] M. Horiguchi, T. Sakata, K. Itoh, "Switched-Source-Impedance CMOS Circuit For Low Standby Sub-threshold Current Giga-Scale LSI's," IEEE JSSC, vol. 28, no. 11, pp. 1131-1135, November 1993.

[9] T. Kawahara, M. Horiguchi, Y. Kawajiri, G. Kitsukawa, T. Kure, "Sub-threshold Current Reduction for Decoded-Driver by Self-Reverse Biasing," IEEE JSSC, vol. 28, no. 11, pp. 1136-1144, November 1993.

[10] T. Sakata, K. Itoh, H. Horiguchi, M. Aoki, "Sub-threshold-Current Reduction Circuits for Multi-Gigabit DRAM's," IEEE JSSC, vol. 29, no.7, pp.761-769, July 1994.

[11] Z.Chen, L. Wei, K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," Proceedings of the International Symposium on Low Power Electronics and Design, pp. 239-244, 1998.

[12] Y. Ye, S. Borkar, V. De, "A New Technique for Standby Leakage Reduction in High-Performance Circuits," 1998 Symposium on VLSI Circuits, pp. 40-41, June 1998.

[13] 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 Electronics and Design, pp. 195-200, August 2001.

[14] M. Johnson, D. Somasekhar, L. Chiou, K. Roy, "Leakage Control with Efficient Use of Transistor Stacks in Single threshold CMOS," IEEE Transactions on VLSI Systems, vol.10, no. 1, pp. 1-5, February 2002.

[15] W. Lee, et al., "A 1V DSP for Wireless Communications," ISSCC, pp. 92-93, February, 1997.

[16] T. Yamashita, et al, "A 450 Mhz 64b RISC Processor Using Multiple Threshold Voltage CMOS," ISSCC Digest of Technical Papers, pp. 414-415, February 2000.

[16] N. Kato, et al., "Random Modulation: Multi-Threshold-Voltage Design Methodology in Sub-2V Power Supply CMOS," IEICE Transactions on Electronics, vol. E83-C, no.11, pp. 1747-1754, November 2000.

[18] L. Wei, Z. Chen, K. Roy, M. Johnson, Y. Ye, V. De, "Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications," IEEE Transactions on VLSI Systems, Vol. 7, no. 1, pp. 16-24, March 1999.

[19] M. Hiarabayashi, K. Nose, T. Sakurai, "Design Methodology and Optimization Strategy for Dual V_t Scheme using Commercially Available Tools," ISLPED, pp. 283-286, 2001.

[20] S. Sirichotiyakul, T. Edwards, C. Oh, J. Zuo, A. Dharchoudhury, R. Panda, D. Blaauw, "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing," IEEE Design Automation Conference, pp. 436-441, June 1999.

[21] J. Tschanz, et al., "Design Optimizations of a High Performance Microprocessor Using Combinations of Dual-V_t Allocation and Transistor Sizing," Proceedings VLSI Symposium, pp. 218-219, June 2002.

[22] S. Augsburger, B. Nikolic, "Combing Dual-Supply, Dual-Threshold and Transistor sizing for Power Reduction," to be presented at IEEE International Conference on Computer Design, September 2002.

[23] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," IEEE JSSC, vol. 30, no. 8, pp. 847-854, August 1995.

[24] J. Kao, A. Chandrakasan, "Dual V_t Techniques for Low Power Digital Circuits," IEEE JSSC, vol.35, no.7, pp. 1009-1018, July 2000.

[25] J. Kao, S. Narendra, A. Chandrakasan, "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns," 35th Design Automation Conference, pp. 495-500, June 1998.

[26] S. Mutoh, S. Shigematsu, Y. Matsuya, H. Fukada, T. Kaneko, J. Yamada, "1V Multithreshold- Voltage CMOS Digital Signal Processor for Mobile Phone Application", IEEE JSSC, vol. 31, no. 11, pp. 1795-1802, November 1996.

[27] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, J. Yamada, "A 1-V High-Speed MTCMOS Circuit Scheme for Power-Down Application Circuits," IEEE JSSC, vol. 32, no. 6, pp.861-869, June 1997.

[28] J. Kao, A. Chandrakasan "MTCMOS Sequential Circuits," 27th European Solid State Circuits Conference, pp. 332-335, September 2001.

[29] T. Kuroda, T. Fujita, et al, "A 0.9V, 150MHz, 10mW, 4mm2, 2-DCT Core Processor with Variable VT Scheme," IEEE JSSC, vol. 31, no. 11, pp. 1770-1778, November 1996.

[30] I. Hyunsik, T. Inukai, H. Gomyo, T. Hiramoto, T. Sakurai, "VTCMOS Characteristics and its Optimum Conditions Predicted by a Compact Analytical Model," ISLPED, pp. 123-128, 2001.

[31] K. Nose, M. Hirabayashi, H. Kawaguchi, S. Lee, T. Sakurai, "V_{th} Hopping Scheme to Reduce Subthreshold Leakage for Low-Power Processors," IEEE JSSC, vol. 37, no. 3, pp. 413-419, March 2002.

[32] M. Miyazaki, et al., "A Delay Distribution Squeezing Scheme with Speed-Adaptive Threshold-Voltage CMOS for Low Voltage LSIs," ISLPED, pp. 49-53, 1998.

[33] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," ISSCC Digest of Technical Papers, pp. 422-423, February 2002.

[34] M. Miyazaki, J. Kao, A. Chandraksan, "A 175mV Multiply-Accumulate Unit using an Adaptive Supply Voltage and Body Bias (ASB) Architecture," ISSCC Digest of Technical Papers, pp. 58-59, February 2002.