Leakage in Nano-Scale Technologies: Mechanisms, Impact and Design Considerations

Amit Agarwal, Chris H. Kim, Saibal Mukhopadhyay, and Kaushik Roy
School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47906, USA
001-765-494-2361
<amita, hyungil, sm, kaushik@ecn.purdue.edu>

ABSTRACT
The high leakage current in nano-meter regimes is becoming a significant portion of power dissipation in CMOS circuits as threshold voltage, channel length, and gate oxide thickness are scaled. Consequently, the identification of different leakage components is very important for estimation and reduction of leakage. Moreover, the increasing statistical variation in the process parameters has led to significant variation in the transistor leakage current across and within different dies. Designing with the worst case leakage may cause excessive guard-banding, resulting in a lower performance. This paper explores various intrinsic leakage mechanisms including weak inversion, gate-oxide tunneling and junction leakage etc. Various circuit level techniques to reduce leakage energy and their design trade-off are discussed. We also explore process variation compensating techniques to reduce delay and leakage spread, while meeting power constraint and yield.

Categories and Subject Descriptors
B.3.7.1 [**Integrated Circuits**]: Types and Design Styles - Microprocessors and microcomputers, VLSI.

General Terms: Design, Performance, Experimentation.

Keywords: Leakage current, Circuit design, Process variation.

1. INTRODUCTION
CMOS devices have been scaled down aggressively in each technology generations to achieve higher integration density and performance. However, the leakage current has increased drastically with technology scaling and has become a major contributor to the total IC power. Moreover, the increasing statistical variation in the process parameters has emerged as a serious problem in the nano-scaled circuit design and can cause significant increase in the transistor leakage current. Designing with the worst case leakage may cause excessive guard-banding, resulting in a lower performance. Hence, accurate estimation of the total leakage current considering the effect of random variations in the process parameters is extremely important for designing CMOS circuits in the nano-meter regime.

DAC 2004, June 7-11, 2004, San Diego, California, USA

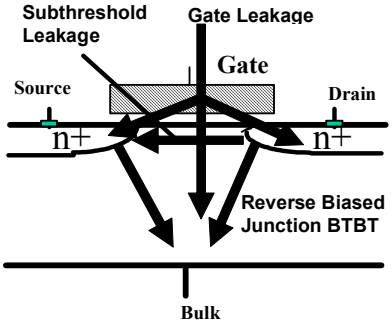

Figure 1. Major leakage mechanisms in a CMOS device.

Different leakage mechanisms contribute to the total leakage in a device. Among them, the three major ones can be identified as: Subthreshold leakage, Gate leakage and Reverse biased drain-substrate and source-substrate junction Band-To-Band-Tunneling leakage [1-2]. With technology scaling each of these leakage components increases drastically, resulting in increase in the total leakage current. Moreover, each component depends differently on the transistor geometry (gate length, Source-Drain extension length, oxide thickness, junction depth, width, the doping profile and "halo" doping concentration), the flat-band voltage, and the supply voltage [2]. Hence, statistical variation in each of these parameters results in a large variation in each of the leakage components, thereby, causing significant spread in leakage and delay. In the nano-meter regime, a significant portion of the total power consumption in high performance digital circuits is due to leakage currents. Because high performance systems are constrained to a predefined power budget, the leakage power reduces the available power, impacting performance. It also contributes to the power consumption during standby operation, reducing battery life. Hence, techniques are necessary to reduce leakage power while maintaining the high performance.

Process parameter variation, which is increasing as technology scales, impacts the frequency and leakage distribution of fabricated chips [3]. It can be observed that there is a correlation between the leakage power and the frequency of operation. Because of die-to-die and within die variations, many dies may not achieve the desired frequency target, while others may fail the maximum leakage power specification. Leakage spread poses stringent design trade-off in leakage sensitive circuits, e.g wide-OR domino gate, to achieve high performance while maintaining sufficient yield [4]. Process compensating techniques that reduce the delay and leakage spread, while meeting power constraint and high yield are indispensable in future design.

This paper is organized as follows. Different leakage mechanisms in nano-scaled CMOS devices are introduced in section 2. Section 3 describes various circuit techniques to reduce different leakage

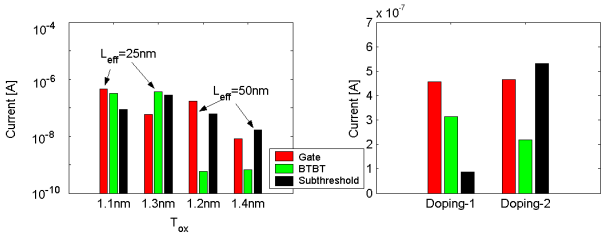

Figure 2. Variation of different leakage components with (a) oxide thickness; (b) doping profile. "Doping-1" has a stronger halo profile than "Doping-2". For NMOS devices 25nm and 50nm of effective channel length.

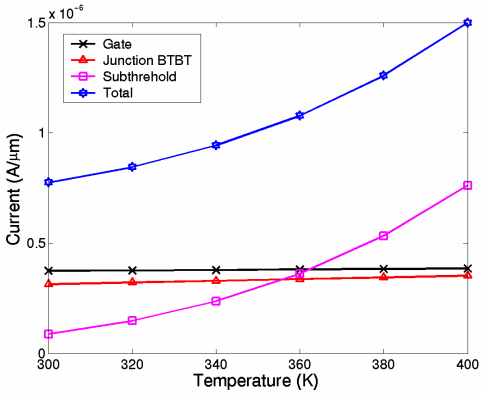

Figure 3. Simulation result for variation of different leakage components with temperature for NMOS device (L_{eff}=25nm).

components. Process variation and its effect on leakage are presented in section 4. Section 5 describes possible solution to compensate for the process variation. We conclude this paper in section 6.

2. LEAKAGE CURRENTS IN SCALED DEVICES

The leakage current in a nano-scale CMOS device is mainly due to (1) subthreshold conduction, (2) gate direct tunneling current, (3) junction tunneling leakage, (4) gate induced drain leakage (GIDL), (5) hot carrier injection current, (6) puncthrough current, etc. Among these, the three major components are subthreshold leakage, gate leakage and junction leakage (Figure 1). In this section, we will briefly discuss each of these three components.

2.1 Subthreshold Leakage

In the "off" state of a device ($V_{gs} < V_{th}$), diffusion of the minority carriers through the channel causes current to flow from the drain to the source of a transistor. This is known as subthreshold current. The subthreshold current depends exponentially on both gate-to-source voltage and V_{th}. A reverse bias applied at the substrate increases V_{th}, thereby reducing the subthreshold current (Body effect). In a long channel device, V_{th} does not depend on the drain bias nor on the channel length. However, in short channel devices source and drain depletion regions penetrate significantly into the channel and control the potential and the field inside the channel. This is known as the short channel effect (SCE). Due to the short channel effect, V_{th} reduces with (i) reduction in channel length (V_{th} roll off), and (ii) increase in the drain bias (Drain Induced Barrier Lowering (DIBL)) [1,2]. This results in large subthreshold current in the short channel devices. On the other hand, in nano-scaled devices with ultra-thin oxides, high electric field at the surface (E_s) and high substrate doping causes quantization of inversion-layer-electron energy, which in turns modulates V_{th}. Due to quantum mechanical behavior of the substrate electrons, more band bending is required to populate the lowest subband, which is at higher energy than the bottom of the conduction band. This increases V_{th}, thereby reducing the subthreshold current [2,5].

2.2 Gate Direct Tunneling Current

Gate direct tunneling current is due to the tunneling of electrons (or holes) from the bulk silicon and source/drain (S/D) overlap region through the gate oxide potential barrier into the gate (or vice-versa) [2]. The tunneling current increases exponentially with the decrease in the oxide thickness and the increase in the potential drop across oxide. Major components of gate tunneling in a scaled MOSFET device are [6]: (1) Gate to S/D overlap region current (Edge Direct Tunneling (EDT)) components (I_{gso} & I_{gdo}) ,(2) Gate to channel current (I_{gc}), part of which goes to source (I_{gcs}) and rest goes to drain (I_{gcd}), (3) Gate to substrate leakage current (I_{gb}). The overlap tunneling dominates the gate leakage in an 'off' (V_{gs}=0) transistor, whereas, gate-to-channel tunneling control the gate current in an 'on' device.

2.3 Junction Tunneling Current

A high electric field across a reverse biased p-n junction causes significant current to flow through the junction due to tunneling of electrons from the valence band of the p-region to the conduction band of the n-region (Band-To-Band-Tunneling (BTBT)) [2]. In an MOSFET when the drain-substrate and/or the source-substrate junction is reverse, biased at a potential higher than that of the substrate, a significant BTBT current flows through the junctions.

In nano-scale devices due to higher doping at the junctions this current becomes significant and can considerably increases the total leakage current [5]. The junction tunneling current depends exponentially on the junction doping (junction electric field) and the reverse bias across the junction . Hence, application of a reverse substrate bias results in large junction tunneling, whereas, use of forward body bias helps to reduce it.

2.4 Inter-Dependence of Different Leakage Components

In nano-scale devices the different leakage currents depend strongly on each other through the device geometry and the doping profile. Understanding of this inter-dependence is essential in studying and controlling the total leakage in a CMOS device and circuit. To reduce the subthreshold leakage in scaled devices the short channel effect has to be effectively controlled [1]. This requires the use of ultra-thin oxide (to increase the vertical electric field across the gate) and use of highly doped region (Halo implants) near the source-substrate and the drain-substrate junction. However, use of an ultra-thin oxide in the nano-scaled devices (T_{ox} < 2nm) results in a considerable gate direct tunneling current (Figure 2a) [7]. On the other hand, introduction of the "Halo" implants results in large junction field and hence large junction tunneling current (Figure 2b). Moreover, different leakage components show different temperature dependence. The subthreshold current increases exponentially with the temperature, whereas the gate leakage is almost insensitive to the temperature variation (Figure 3) [7]. The

Table 1. Circuit techniques to reduce leakage

Design time techniques	Run time techniques	
	Standby leakage reduction	Active leakage reduction
Dual-V_{th}	Natural Stacking Sleep Transistor FBB/RBB	DVTS

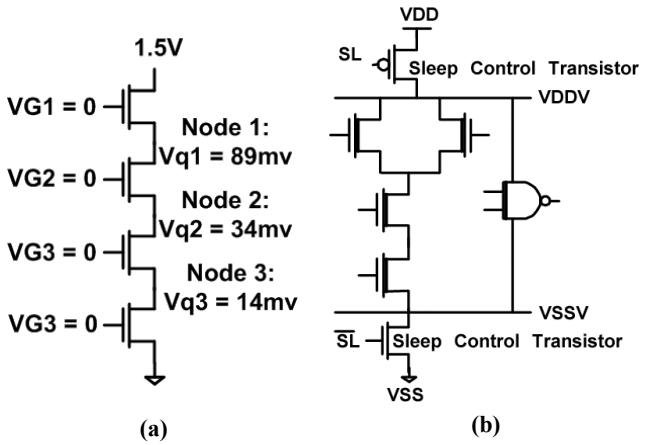

Figure 4 (a) Effect of transistor stacking on source voltage, (b) Schematic of MTCMOS circuit with low V_{th} device shaded.

junction tunneling leakage increases slowly with the temperature. Hence, the leakage component that dominates the total leakage depends on the device geometry, the doping profile and the operating temperature. Any leakage control technique in nano-scaled devices has to consider its impact on each of the leakage mechanisms along with the total leakage.

3. CIRCUIT TECHNIQUES TO REDUCE LEAKAGE

Since circuits are mostly designed for the highest performance — say to satisfy overall system cycle time requirements — they are composed of large gates, highly parallel architectures with logic duplication. As such, the leakage power consumption is substantial for such circuits. However, not every application requires a circuit to operate at the highest performance level all the time. Modules, in which computation is bursty in nature e.g. functional units in a microprocessor or sections of a cache, are often idle. It is of interest to conceive of methods that can reduce the leakage power consumed by these circuits. Different circuit techniques have been proposed to reduce leakage energy utilizing this slack without impacting performance. These techniques can be categorized based on when and how they utilize the available timing slack (Table 1).

3.1 Design Time Techniques

Design time techniques exploit the delay slack in non-critical paths to reduce leakage. These techniques are static; once it is fixed, it cannot be changed dynamically while the circuit is operating.

3.1.1 Dual Threshold CMOS

In logic, a high V_{th} can be assigned to some transistors in the non-critical paths so as to reduce subthreshold leakage current, while the performance is not sacrificed by using low V_{th} transistors in the critical path(s) [8]. No additional circuitry is required, and both high performance and low leakage can be achieved simultaneously. Dual threshold CMOS is effective in reducing leakage power during both standby and active modes. Many design techniques have been proposed, which consider upsizing of high V_{th} transistor [9] in dual V_{th} design to improve performance, or upsizing additional low V_{th} transistor to create more delay slack and then converting it high V_{th} to reduce leakage power. Upsizing the transistor affects switching power and die area that can be traded off against using a low V_{th} transistor, which increases leakage power. Instead of changing the channel doping profile to change V_{th}, a higher t_{ox} can be used to obtain a high V_{th} device for dual threshold CMOS circuits. In order to suppress the SCE, the high t_{ox} device needs to have a longer channel length as compared to the low t_{ox} device. Multiple t_{ox} CMOS (MoxCMOS) [10] can optimize the power consumption due to subthreshold leakage, gate oxide tunneling leakage as well as switching power.

With the increase in V_{th} variation and supply voltage scaling, it is becoming difficult to maintain sufficient gap among low V_{th}, high V_{th} and supply voltage required for dual V_{th} design. Furthermore, dual V_{th} design increases the number of critical paths in a die. It has been shown in [3] that as the number of critical paths on a die increases, within-die delay variation causes both mean and standard deviation of the die frequency distribution to become smaller, resulting in reduced performance.

3.2 Run Time Techniques

3.2.1 Standby Leakage Reduction Techniques

A common architectural technique to keep the power of fast, hot circuits within bounds has been to freeze the circuits — place them in a standby state — any time when they are not needed. Standby leakage reduction techniques exploit this idea to place certain sections of the circuitry in standby mode (low leakage mode) when they are not required.

3.2.1.1 Natural Transistor Stacks

Leakage currents in NMOS or PMOS transistors depend exponentially on the voltage at the four terminals of transistor (section 2.1). Increasing the source voltage of NMOS transistor reduces subthreshold leakage current exponentially due to negative V_{gs}, lowered signal rail (V_{cc}-V_s), reduced DIBL and body effect. This effect is also called self-reverse biasing of transistor. The self-reverse bias effect can be achieved by turning off a stack of transistors [11]. Turning off more than one transistor in a stack raises the internal voltage (source voltage) of the stack, which acts as reverse biasing the source (Figure 4a). The voltages at the internal nodes depend on the input applied to the stack. Functional blocks such as NAND, NOR or other complex gates readily have a stack of transistors. Maximizing the number of off transistors in a natural stack by applying proper input vectors can reduce the standby leakage of a functional block. A model and heuristic is proposed in [12] to estimate leakage and to select the proper input vectors to minimize the leakage in logic blocks.

Since gate and junction leakage are also important in scaled technologies, the input vector control technique using a stack of transistors needs to be reinvestigated to effectively reduce the total leakage. It has been shown that with high gate leakage, the traditional way of using stacking fails to reduce leakage and in the worst case might increase the overall leakage [13]. The gate leakage depends on the voltage drop across different region of transistor. Applying "00" as the input to a two transistors stack only reduces subthreshold leakage and does not change the gate

leakage component. It has been shown that using "10" reduces the voltage drop across the terminals, where the gate leakage dominates, thereby lowering the gate leakage while offering marginal improvement in subthreshold leakage [13]. In scaled technologies where gate leakage dominates the total leakage, using "10" might produce more savings in leakage as compared to "00".

3.2.1.2 *Sleep Transistor (Forced Stacking)*

This technique inserts an extra series connected transistor (sleep transistor) in the pull-down/pull-up path of a gate and turns it 'off' in the standby mode of operation [14]. During regular mode of operation, the extra transistor is turned on. This provides substantial savings in leakage current during standby mode of operation. However due to the extra stacked transistor (sleep transistor), the drive current of forced-stack gate is lower resulting in increased delay. Hence, this technique can only be used for paths that are non-critical. If the V_{th} of the sleep transistor is high, extra leakage saving is possible. The circuit topology is known as MTCMOS (Figure 4b) [15].

In fact, only one type (i.e. either PMOS or NMOS) of high V_{th} transistor is sufficient for leakage reduction. The NMOS insertion scheme is preferable, since the NMOS on-resistance is smaller at the same width and hence it can be sized smaller than a corresponding PMOS [16]. However, MTCMOS can only reduce leakage power in standby mode and the large inserted sleep transistors can increase the area and delay. Moreover, if data retention is required in standby mode, an extra high V_{th} memory circuit is needed to maintain the data. Instead of using high V_{th} sleep transistors, super cut-off CMOS (SCCMOS) circuit uses low V_{th} transistors with an inserted gate bias generator [17]. In standby mode, the gate is applied to V_{cc}+0.4V for PMOS (V_{SS}-0.4V for NMOS) by using the internal gate bias generator to fully cut off the leakage current. Compared to MTCMOS where it becomes difficult to turn on the high V_{th} sleep transistor at very low supply voltages, SCCMOS circuits can operate at very low supply voltages. The forced stacking technique can also be used in large cache memories to reduce leakage current [1].

3.2.1.3 Forward/Reverse Body Biasing

Variable threshold CMOS (VTCMOS) is a body biasing design technique [18]. Figure 5a shows the VTCMOS scheme. In order to achieve different threshold voltages, a self-substrate bias circuit is used to control the body bias. In the active mode, a zero body bias (ZBB) is applied. While in standby mode, a deep reverse body bias (RBB) is applied to increase the threshold voltage and to cut off the leakage current. Providing the body bias voltage requires routing a body bias grid and this adds to the overall chip area. Keshavarzi *et. al.* reported that RBB lowers IC leakage by three orders of magnitude in a 0.35 µm technology [19]. However, more recent data shows that the effectiveness of RBB to lower I_{off} decreases as technology scales due to the exponential increase in band-to-band tunneling leakage at the source/substrate and drain/substrate p-n junctions due to halo doping in scaled devices [19]. Moreover, smaller channel length with technology scaling and lower channel doping to reduce V_{th} worsen the short channel effect and diminish the body effect. This in turns weakens the V_{th} modulation capability of RBB.

For scaled technologies, recent design [20] has been proposed using forward body biasing (FBB) to achieve better current drive with less short channel effect. Circuit is designed using high V_{th} transistor (high channel doping) reducing leakage in standby mode, while FBB is applied in active mode to achieve high performance. Both high channel doping and FBB reduce short channel effect relaxing the scalability limit of channel length due to V_{th} roll off and DIBL. This results in higher I_{on} compared to low V_{th} design for same worst case I_{off}, improving performance. RBB can also be applied in standby mode together with FBB to further reduce the leakage current. It has been shown that FBB/Hight-V_{th} along with RBB reduces leakage by 20X, as opposed to 3X for the RBB/low-V_{th}. FBB devices however has larger junction capacitance and body effect, which reduces the delay improvement especially in stacked circuits. A new high V_{th} device optimized for FBB is proposed which changes the doping profile by adjusting the peak halo doping (channel engineering) or uses gate material with a higher work function (work function engineering) [21]. FBB can also be combined with lowering the V_{cc} to achieve same performance as high V_{cc}, while reducing the switching and standby leakage power.

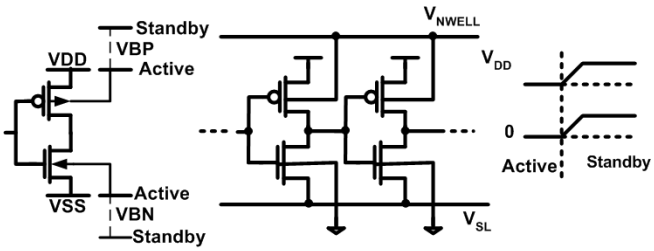

Figure 5. (a) Variable threshold CMOS, (b) Realizing body biasing by changing the source voltage with respect to body voltage, which is grounded.

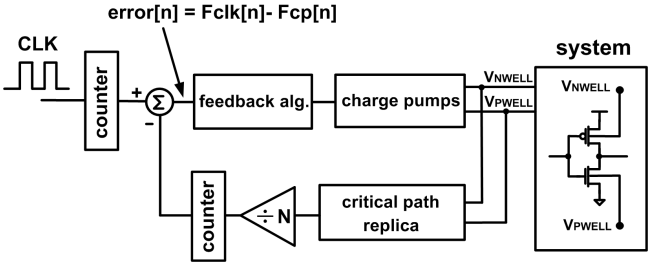

Figure 6. Dynamic V_{th} sclaing system.

Raising the NMOS source voltage while tying the NMOS body to ground can produce the same effect as RBB. Forward body biasing can also be realized by applying a negative source voltage with respect to the body, which is tied to ground. Figure 5b illustrates the circuit diagram of this technique [22]. The main advantage is that it eliminates the need for a deep N-well or triple well process since substrate of the target system and the control circuitry can be shared.

3.2.2 Active Leakage Reduction Techniques

Not every application requires a fast circuit to operate at the highest performance level all the time. Active leakage techniques exploit this idea to intermittently slow down the fast circuitry and reduce the leakage power consumption as well as the dynamic power consumption when maximum performance is not required.

3.2.2.1 Dynamic V_{th} Scaling (DVTS)

DVTS scheme uses body biasing to adaptively change the V_{th} based on the performance demand. The lowest V_{th} is delivered via ZBB, if the highest performance is required. When performance

demand is low, clock frequency is lowered and V_{th} is raised via RBB to reduce the run-time leakage power dissipation. In cases when there is no workload at all, the V_{th} can be increased to its upper limit to significantly reduce the standby leakage power. "Just enough" throughput is delivered for the current workload by tracking the optimal V_{th} while leakage power is considerably reduced by intermittently slowing down the circuit.

Several different DVTS system implementations have been proposed in literature [23,24]. Figure 6 shows a DVTS hardware that uses continuous body bias control to track the optimal V_{th} for a given workload. A clock speed scheduler, which is embedded in the operating system, determines the (reference) clock frequency at run-time. The DVTS controller adjusts the PMOS and NMOS body bias so that the oscillator frequency of the critical path replica tracks the given reference clock frequency. The error signal, which is the difference between the reference clock frequency and the oscillator frequency, is fed into the feedback controller. The continuous feedback loop can also compensate for process, supply voltage, and temperature variations. A simpler method called "V_{th} hopping scheme", which dynamically switches between low V_{th} and high V_{th} depending on the performance demand, is proposed in [24]. As mentioned in previous section, the effectiveness of RBB is expected to be low due to the worsening short channel effect and increasing band-to-band tunneling leakage at the source/substrate and drain/substrate junctions. FBB can be applied together with RBB to achieve a better performance-leakage tradeoff for DVTS systems.

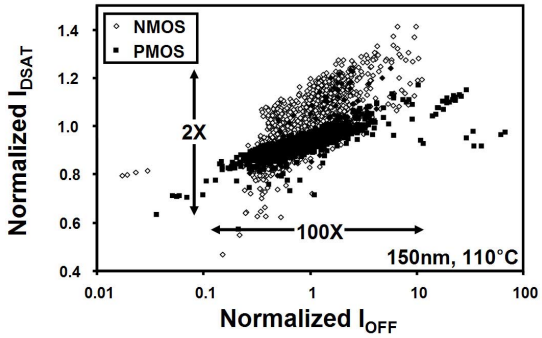

Figure 7. I_{DSAT} and I_{OFF} variation measured (150nm process).

4. PROCESS VARIATION AND LEAKAGE

Increasing inter-die and intra-die parameter variation in the nano-meter regime can result in a large variability in performance and power [25,26]. Due to aggravating short channel effect in nano-scale devices, variation in the channel length results in a large variation in threshold voltage (due to Vt-roll off) and hence subthreshold current. The variations in transistor width, oxide thickness or flat-band voltage can also be translated into the variation in threshold voltage. Moreover, in the nano-scaled devices (particularly for transistors with small width used in memory) the random placement of dopants also causes variation in device threshold voltage [2]. The principal cause of the variation in the gate leakage is the variation in oxide thickness as it exponentially depends on the oxide thickness. The variation in junction tunneling is principally caused by the variation in the doping profile. Among the three major leakage components, the junction tunneling current is least susceptible to parameter variation whereas the subthreshold leakage is the most sensitive component. Figure 7 shows the measured on-off current distribution for a 150nm CMOS process. A large spread in the leakage current can be observed and this makes it harder to achieve the target frequency while meeting the power constraints. Parameter variation can also reduce the robustness of dynamic circuits resulting in a large number of failing dies [29]. Hence, variation tolerant circuit design strategies are indispensable for improving the performance and robustness of nano-scale systems.

5. CIRCUIT TECHNIQUES FOR COMPENSATING PROCESS VARIATION

Post silicon tuning has become an attractive method to compensate the die-to-die and within-die parameter variation. Two main circuit components are required to implement a post silicon tuning technique; a leakage sensor or delay line that will detect the process variation, and a circuit technique such as adaptive body biasing to tune the performance/leakage of devices. The rest of this section describes post silicon tuning techniques that reduce the impact of parameter variation on performance and leakage of VLSI circuits.

5.1 Adaptive Body Biasing for Process Compensation

Due to the worsening parameter fluctuations some dies may not meet the target frequency, while others exceed the leakage power constraints. The slow dies which fail to meet the desired frequency can be forward body biased to improve performance while paying more leakage power. On the other hand, excess leakage dies can be reverse body biased to meet the leakage power specifications [27]. Measurements based on a 150nm CMOS test-chip demonstrates that adaptive body bias reduces the spread of the die frequency distribution by 7X, compared to a conventional zero body bias. This technique can also be expanded to compensate the within-die variation by applying different body biases for different parts of a chip [27]. It has also been shown in [28] that using adaptive V_{cc} in conjunction with adaptive body bias is more effective in maximizing the number of dies in the highest frequency bin.

5.2 Process Variation Compensation in Dynamic Circuits

Increasing I_{off} with process scaling has forced designers to upsize the keeper in dynamic circuits to obtain an acceptable robustness for worst-case leakage corner dies [29]. However, Figure 7 showing 100x+ variation in die-to-die NMOS I_{off} indicates that

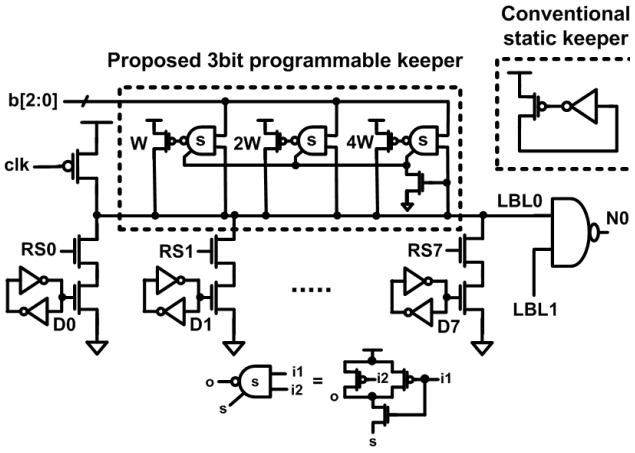

Figure 8. 8 way dynamic local bitline with process compensating dynamic (PCD) circuit technique.

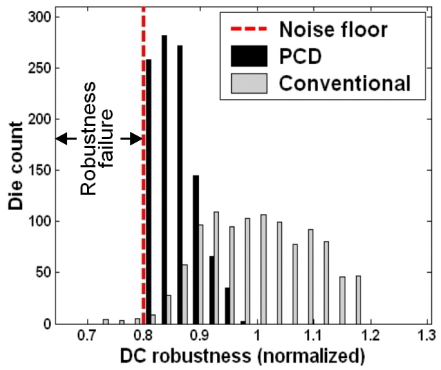

Figure 9. Robustness dist. of PCD vs conv. dynamic circuit.

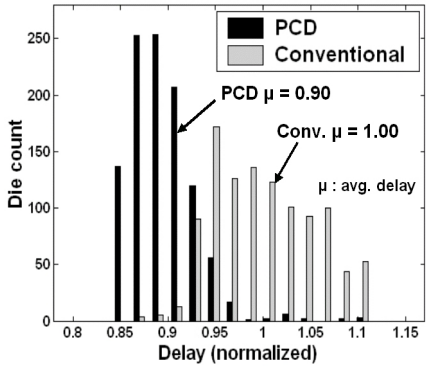

Figure 10. Delay dist of PCD vs conv. dynamic circuit.

(i) a large number of low leakage dies suffer from the performance loss due to the contention current with the unnecessarily strong keeper, while (ii) the excess leakage dies still cannot meet the robustness requirements even with a keeper sized for the fast corner leakage. Rather than using a fixed-strength keeper that will sacrifice the performance of low leakage dies, a post-silicon keeper technique that optimally programs the keeper strength based on the measured die leakage improves robustness and delay variation spread.

Figure 8 shows the Process Compensating Dynamic (PCD) circuit scheme with a digitally programmable 3-bit keeper applied on an 8-way dynamic circuit [4]. Each of the three binary-weighted keepers with respective widths W, 2W, and 4W can be activated or deactivated by asserting appropriate globally routed signals b[2:0]. A desired effective keeper width can be chosen among {0, W, 2W,…, 7W}. Excess dynamic node capacitance of PCD scheme due to the keeper circuitry causes a 7% delay penalty. However, this penalty is offset by the opportunistic speedup achieved by keeper downsizing on low leakage dies, resulting in an overall performance improvement. Optimal keeper width is one-time programmed via fuses based on die leakage measurements. The PCD technique can be further improved to compensate within-die variation by locally generating the control bits b[2:0] using a self-contained on-die leakage sensor [18].

Statistical studies were carried out based on measured NMOS I_{off} data to show the robustness and delay distribution of the PCD technique vs. conventional static keeper scheme. Simulation results show a 5x reduction in the number of robustness failing dies (Figure 9) while simultaneously achieving 10% improvement in average performance (Figure 10) by selective keeper downsizing. Variation spread (σ/μ) of the robustness and delay distribution is reduced by 55% and 35%, respectively, resulting in a squeezed distribution for the PCD design. 92% of the dies can benefit from speedup via keeper downsizing, while the robustness of remaining 8% of the dies are recovered by an upsized keeper.

6. SUMMARY AND CONCLUSIONS

Continuous scaling of CMOS devices causes the leakage current to become a major component of total power consumption. In current deep sub-micron devices with low threshold voltage and thin oxide, sub-threshold, gate and junction leakage have become the dominant sources of leakage power. Each component is expected to increase with technology scaling. This paper explains different leakage mechanisms and explores circuit techniques to reduce leakage power in high performance systems. We have also discussed intrinsic parameter variations and its impact on performance and leakage. Finally, various post silicon tuning techniques to compensate process variation have been presented.

7. ACKNOWLEDGEMENT

This research was funded in part by SRC, DARPA-PACC, GSRC MARCO center, Intel, and IBM Corporation.

8. REFERENCES

[1] K. Roy et al. Proceeding of IEEE, Feb, 2003.

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

[3] K. A. Bowman et al. IEEE J. Solid State Circuits, Feb 2002.

[4] C. H. Kim et al. Symposium on VLSI Circuits, 2003.

[5] S. Mukhopadhyay et al. Proceedings of DAC, 2003.

[6] http://www-device.eecs.berkeley.edu/~bsim3/

[7] "Well-Tempered" Bulk-Si NMOSFET Device Home Page", Available: http://www-mtl.mit.edu/Well/

[8] L. Wei et al. IEEE Transactions on VLSI Systems, 16,1999.

[9] T. Karnik et al. ACM/IEEE DAC, 2002.

[10] N. Sirisantana et al. Int. Conf. on Computer Design, 2000.

[11] Y. Ye et al. IEEE Symposium on VLSI Circuits, 1998.

[12] Z. Chen et al. IEEE Int. Conf. on Comp. Aided Design, 1998.

[13] S. Mukhopadhyay et al. IEEE Tran. on VLSI Systems, 2003.

[14] M. C. Johnson et al. ACM/IEEE DAC, 1999.

[15] S. Mutoh et al. IEEE J. Solid State Circuits, 30, 1995.

[16] J. Kao et al. ACM/IEEE DAC, 1997.

[17] H. Kawaguchi et al. ISSCC, 1998.

[18] T. Kuroda et al. ISSCC 1996.

[19] A. Keshavarzi et al. NASA Symp. on VLSI Design, 1999.

[20] S. Narendra et al. IEEE J. Solid State Circuits, May 2003.

[21] C. H. Kim et al. ISLPED 2003.

[22] H. Mizuno et al. IEEE J. Solid-State Circuits., 34, 1999.

[23] C. H. Kim et al. ACM/IEEE DATE, 2002.

[24] K. Nose et al. IEEE Custom Integrated Circ. Conf., 93, 2001.

[25] R. Rao, et al. IEEE Transaction on VLSI, Feb. 2004.

[26] S. Mukhopadhyay, et al. ISLPED, 2003.

[27] J. Tschanz et al. IEEE J. Solid-State Circuits, Nov. 2002.

[28] J. Tschanz et al. IEEE J. Solid-State Circuits, May 2003.

[29] A. Alvandpour et al. IEEE J. Solid-State Circuits, May 2002.