Leakage Power Analysis and Reduction for Nanoscale Circuits

Leakage current in the nanometer regime has become a significant portion of power dissipation in CMOS circuits as threshold voltage, channel length, and gate oxide thickness scale downward. Various techniques are available to reduce leakage power in high-performance systems.

Amit Agarwal
Intel Corp.

Saibal Mukhopadhyay
Arijit Raychowdhury
Kaushik Roy
Purdue University

Chris H. Kim
University of Minnesota

CMOS devices have scaled downward aggressively in each technology generation to achieve higher integration density and performance. However, leakage current has increased drastically with technology scaling and become a major contributor to the total IC power. Different leakage mechanisms contribute to the total leakage in a device. As Figure 1 shows, the three major types of leakage mechanisms are

- subthreshold,
- gate, and
- reverse-biased, drain- and source-substrate junction band-to-band-tunneling (BTBT).[1]

With technology scaling, each of these leakage components increases drastically, resulting in an increase in the total leakage current.

The increase in different leakage components with technology scaling has two major implications in leakage estimation and low-power logic design. First, these increases add up to a dramatic increase in total leakage. More importantly, each of the leakage components becomes equally important in nanoscaled devices.[2] Hence, the relative magnitudes of the leakage components play a major role in low-leakage logic design.[3]

In the nanometer regime, leakage currents make up a significant portion of the total power consumption in high-performance digital circuits. Because high-performance systems must work within a predefined power budget, leakage power reduces the available power, impacting performance. It also contributes to the power consumption during standby operation, reducing battery life. Hence, designers require techniques that reduce leakage power while maintaining high performance. Moreover, as different leakage components become more important with technology scaling, each leakage reduction technique needs reevaluation in scaled technologies where subthreshold conduction is not the only leakage mechanism. Designers will require new, low-power circuit techniques to reduce total leakage in high-performance nanoscale circuits.

Leakage components

In addition to the three major leakage com-

Published by the IEEE Computer Society

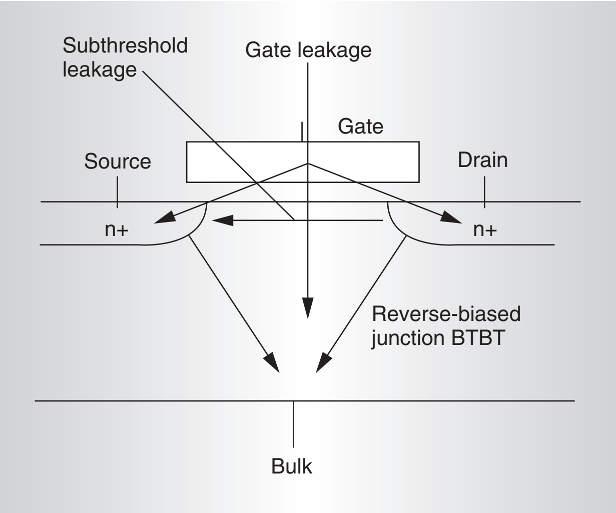

Figure 1. Major leakage components in a transistor.

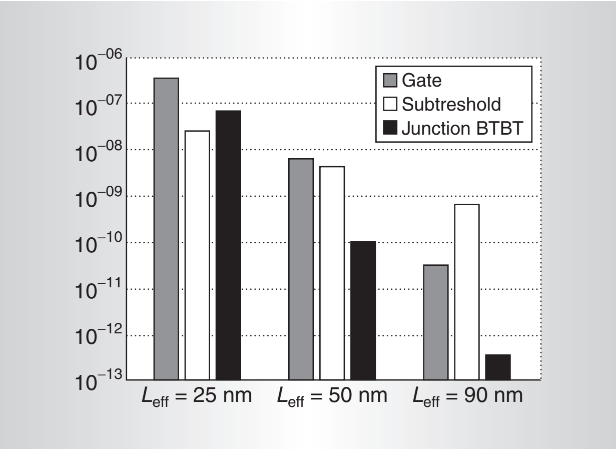

Figure 2. Contribution of different leakage components in NMOS devices[4] at different technology generations. The leakage values are extracted using device simulation in Medici. V_{DD} values are chosen following the ITRS guidelines (0.7 V at 25 nm, 0.9 V at 50 nm, and 1.2 V at 90 nm).

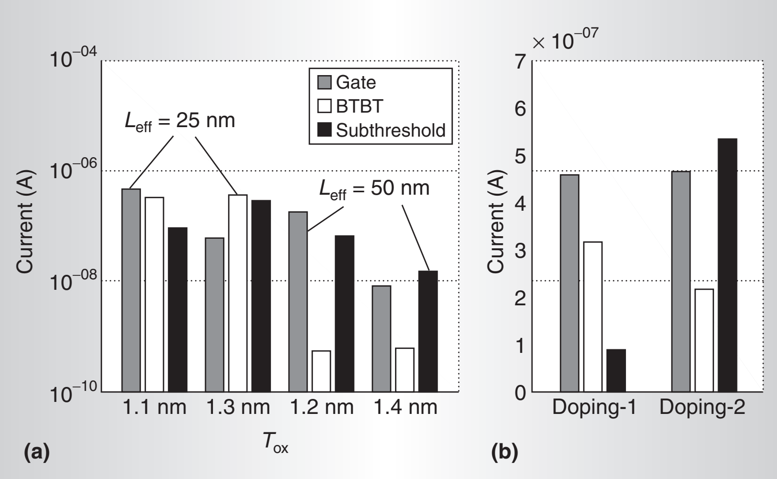

Figure 3. Variation of different leakage components with technology generation and oxide thickness (a), and doping profile (b). Doping-1 has a stronger halo profile than Doping-2. We extracted leakage values using device simulation in Medici. We chose V_{DD} as 0.7 V at 25 nm and 0.9 V at 50 nm.

ponents, there are others, such as gate-induced drain leakage (GIDL) and punch-through current. But those components are not very serious in normal modes of operations. GIDL will be of concern in cases where $V_{GD} < 0$ (V_{GD}, voltage across the gate and the drain of the transistor), and pass-gate logic is definitely a part of it. However, for the range of V_{DD} suggested by the *International Technology Roadmap for Semiconductors*, we observe that maximum negative $V_{GD} = -V_{DD}$, and it does not result in any significant GIDL.

As technology scales downward, the supply voltage must also scale down to reduce dynamic power and maintain reliability. However, this requires the scaling of V_{th} to maintain a reasonable gate overdrive. V_{th} scaling and reduction, because of short-channel effects (SCEs) such as drain-induced barrier-lowering (DIBL),[2] result in an exponential increase in subthreshold current. To control the SCE and to increase the transistor drive strength, oxide thickness must also become thinner in each technology generation. Aggressive scaling of oxide thickness results in a high direct-tunneling current through the transistor's gate insulator.[2] On the other hand, scaled devices require higher substrate doping densities and the application of "halo" profiles (implants of a high doping region near the source and drain junctions of the channel) to reduce the width of the depletion region for the source- and drain-substrate junctions.[2] A narrower depletion region width helps to control the short-channel effect. The high doping density near the source- and drain-substrate junctions causes a significantly large BTBT current through these junctions under high reversed bias.[2] We conclude that each of the leakage components increases with technology scaling, as Figure 2 shows.

Figure 3 shows the different leakage components of NMOS devices with physical gate lengths of 25 and 50 nm.[4] The plot also shows the results at different oxide thicknesses based

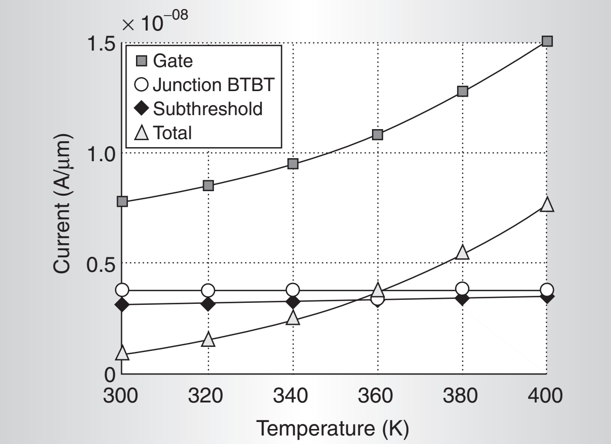

Figure 4. Simulation result for variation of different leakage components with temperature for NMOS device of L_{eff} = 25 nm.[4]

on device simulation. We varied only the oxide thickness in the simulations for a particular technology node (doping remained constant). The gate and subthreshold leakages correlate strongly with oxide thickness; a high oxide thickness results in low gate leakage. Although long-channel MOSFET theory maintains that higher oxide thickness helps to increase the threshold voltage, it worsens the short-channel effect.[2] If the short-channel effect is not very high (as in the 50 nm device in Figure 3a), increasing T_{ox} might reduce the subthreshold leakage. However, in a nanoscale device where SCE is extremely severe (in the 25 nm device in the present case), an increase in the oxide thickness will increase the subthreshold leakage (Figure 3a). Similarly, the subthreshold leakage and the junction BTBT are strongly coupled through the doping profile.

Figure 3b shows the different leakage components of a 25-nm device at different doping profiles (oxide thickness and V_{DD} remained constant). A strong “halo” doping reduces the subthreshold current but results in a high BTBT. Reducing the halo strength lowers the BTBT, but increases subthreshold current considerably. We conclude that the magnitude of leakage components and their relative dominance with respect to each other depend strongly on device geometry and doping profile.

The basic physical mechanisms governing the different leakage current components have different temperature dependences. Subthreshold current is governed by the carrier diffusion that increases with an increase of temperature. Since tunneling probability of an electron through a potential barrier does not depend directly on temperature, the gate and the junction band-to-band tunneling are less sensitive to temperature variations. However, increasing temperature reduces silicon’s band gap, which is the barrier height for tunneling in BTBT. Hence, the junction BTBT should increase with an increase in temperature.

Figure 4 shows the effect of temperature variation on individual leakage component of the previously mentioned 25-nm NMOS device based on the device simulation. In Figure 4, we observe that the subthreshold leakage increases exponentially with temperature, the junction BTBT increases slowly with temperature, and the gate leakage is almost independent of temperature variation. Figure 4 shows that for this particular NMOS device, at T = 300 K (a possible temperature in the standby mode) the gate leakage is the dominant leakage component. However, the subthreshold and BTBT leakages become dominant at T = 400 K (a possible temperature in active mode). Hence, it can be concluded that the individual leakage components and the total leakage depend strongly on temperature (or mode of operation).

It is evident that in nanoscaled devices all of the different leakage components become important and their magnitude depends strongly on the device structure, doping profile, and temperature.

Circuit techniques to reduce leakage in logic

Since circuits are mostly designed for the highest performance—to satisfy overall system cycle time requirements, for example—they typically consist of large gates, highly parallel architectures with logic duplication. The leakage power consumption is substantial for such circuits. However, not every application requires a fast circuit to operate at the highest performance level all the time. Modules in which computation is bursty (such as certain functional units or cache sections) are often idle. Thus, there is an opportunity to reduce the leakage power consumed by such circuits.

Researchers have proposed different circuit techniques to reduce leakage energy without impacting performance by using this slack. In Table 1, we categorize these techniques based on when and how they use the available timing slack. Dual V_{th} statically assigns high V_{th} to some transistors in the noncritical paths at design time to reduce leakage current. Techniques, which use the slack at runtime, fall into two groups, depending on whether they reduce standby or active leakage. Standby-leakage reduction techniques put the entire system in a low-leakage mode when computation is not required. Active-leakage reduction techniques slow down the system by dynamically changing V_{th} to reduce leakage when maximum performance is not needed. In active mode, the operating temperature increases because of the transistors' switching activities. This has an exponential effect on subthreshold leakage, making this the dominant leakage component during active mode and amplifying the leakage problem.

Table 1. Circuit techniques to reduce leakage.

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

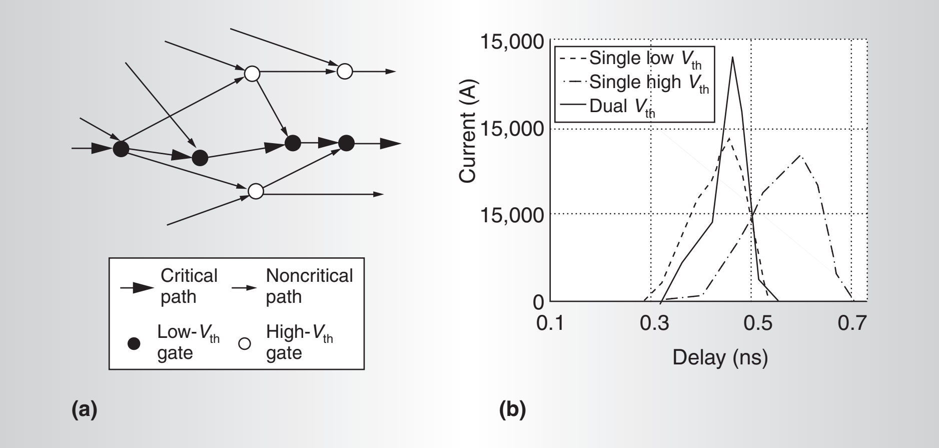

Figure 5. A dual V_{th} CMOS circuit (a), and a path distribution of dual V_{th} and single V_{th} CMOS (b).

Design time techniques

Design time techniques exploit the delay slack in noncritical paths to reduce leakage. These techniques are static; once they become part of the design, there is no way to dynamically change them while the circuit is operating.

Dual-threshold CMOS

In logic, designers can assign a high V_{th} to some transistors in the noncritical paths to reduce subthreshold leakage current. This permits the use of low V_{th} transistors in the critical path(s), preserving performance.[5] This technique does not require additional circuitry, and can support both high performance and low leakage simultaneously.

Figure 5a illustrates the basic idea of a dual-V_{th} circuit. Figure 5b shows the path distribution of dual- and single-V_{th} standard CMOS for a 32-bit adder. Dual-V_{th} CMOS has the same critical delay as the single low-V_{th} CMOS circuit, but the transistors in the noncritical paths can use a high V_{th} to reduce leakage power. Dual-threshold CMOS is effective in reducing leakage power during both standby and active modes. Researchers have proposed many design techniques, some of which consider upsizing a high V_{th} transistor[6] in dual-V_{th} design to improve performance, or upsizing an additional low-V_{th} transistor to create more delay slack and then converting it to high V_{th} to reduce leakage power. Upsizing the transistor affects switching power and die area. Designers can trade off such effect against using a low-V_{th} transistor, which increases leakage power.

Domino logic can be susceptible to leakage—especially in wide or domino gates. Low-threshold evaluation logic reduces noise immunity. So, for scaled technologies, domino logic can require larger keeper transistors, which in turn can affect speed. Figure 6 shows a typical dual-V_{th} domino logic circuit for low-leakage noise-immune operations.[7] Because of the fixed transition directions in domino logic, you can easily assign low V_{th} to all transistors that switch during evaluate mode and

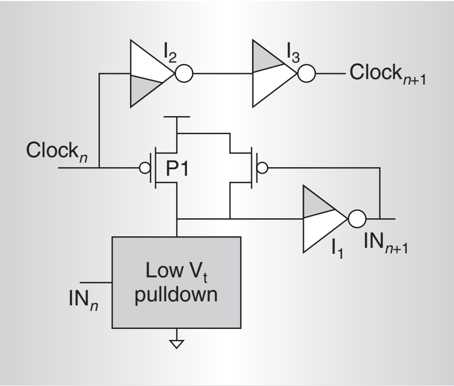

Figure 6. A dual V_{th} domino gate with low V_{th} devices shaded.[7]

high V_{th} to all transistors that switch during precharge modes. When a dual-V_{th} domino logic stage goes into standby mode, the domino clock must be high (evaluate) to shut off the high-V_{th} devices (that is, P_1, I_2 PMOS, and I_3 NMOS). Furthermore, to ensure that the internal node remains at solid logic *zero*, which turns off the high-V_{th} keeper and I_1 NMOS, the initial inputs into the domino gate must be set high.

The fabrication process can achieve a high-V_{th} device by varying different parameters, or changing the doping profile, using higher oxide thickness, and increasing the channel length. Each parameter has its own trade-off in terms of process cost, effect on different leakage components, and SCEs.

Doping profile

Increasing the channel doping densities[8] is a commonly used technique to achieve higher threshold voltages. It does require two additional masks, resulting in high process cost. However, the threshold voltage can vary because of the nonuniform distribution of the doping density, making it difficult to achieve dual-threshold voltages when the threshold voltages are very close to each other. High-V_{th} is also achievable by increasing the strength of halo, increasing the peak doping, moving the position of the lateral peak of the halo close to the channel's center, and moving the position of the halo's vertical peak away from the bottom junction and toward the surface. However, increasing the halo strength increases junction tunneling (Figure 3b), which might become severe in nanoscaled devices where junction tunneling is a significant portion of total leakage.

Oxide thickness

A higher oxide thickness (T_{ox}) can yield a high V_{th} device for dual-threshold CMOS circuits. Higher T_{ox} not only reduces the subthreshold leakage but reduces gate oxide tunneling, since the oxide tunneling current decreases exponentially with an increase in oxide thickness. Since higher oxide thickness reduces the gate capacitance, it is also beneficial for dynamic power reduction.[8] However, in a nanoscale device where SCEs are extremely severe (as in 25-nm devices), an increase in the oxide thickness will increase the subthreshold leakage (Figure 3a). To suppress SCEs, the high-T_{ox} device must have a longer channel length compared to low-T_{ox} devices.[8] Advanced process technology is required for fabricating multiple oxide thicknesses.

Channel length

For short-channel transistors, the threshold voltage increases with the increase in channel length (V_{th} roll-off). A multiple-channel-length design uses the conventional CMOS technology. However, for the transistors with feature size close to 0.1 μm, designs can use halo implants[2] to suppress SCE. This causes a very sharp V_{th} roll-off; and hence, it is nontrivial to control the threshold voltages near the minimum feature size for such technologies. The longer transistor lengths for the high-V_{th} transistors will increase the gate capacitance, which has a negative effect on the performance and power.

With the increase in V_{th} variation and supply voltage scaling, it is becoming difficult to maintain sufficient gap among the 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 that as the number of critical paths on a die increases, within-die delay variation causes both the mean and standard deviation of the die's frequency distribution to become smaller, resulting in reduced performance.[9]

Runtime 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 they are not needed. Standby-leakage reduction techniques exploit this idea to place certain sections of the circuit in standby mode (low-leakage mode) when they are not required.

Exploiting natural transistor stacks

Leakage currents in NMOS or PMOS transistors depend exponentially on the voltage at the four transistor terminals. Increasing the source voltage of the NMOS transistor exponentially reduces subthreshold leakage current because of negative V_{gs}, a lowered signal rail (V_{DD} – V_s), reduced DIBL, and body effect. This effect, also called *self-reverse biasing* of the transistor, is achievable by turning off a stack of transistors. Turning off more than one transistor in a stack raises the stack's internal voltage (source voltage), which acts as a source reverse.

Figure 7a depicts a simple pull-down network of a four-input NAND gate. This pull-down network forms a stack of four transistors. If you turn off some transistors for a long time, the circuit reaches a steady state where leakage through each transistor is equal, and the voltage across each transistor settles to a steady-state value. In a case where only one NMOS device is off, the voltage at the source node of the off transistor would be virtually zero because all the other transistors that remain on will act as a short circuit. Hence, there is no self-reverse biasing effect, and the leakage across the off transistor is large. If more than one transistor is off, the source voltages of the off transistor—except the one connected to ground by the on transistors—will be greater than zero. The most negatively self-reverse biased transistor (since subthreshold leakage is an exponential function of gate-source voltage) will be the main determiner of overall leakage. The voltages at the internal nodes depend on the input applied to the stack.

Figure 7a shows the internal voltages when all four transistors are off. These internal voltages make the off transistors self-reverse biased. The reverse bias makes the leakage across the transistor that is off very small. Figure 7b shows the subthreshold leakage current versus the number of off transistors in a stack. There is a large difference in leakage current between one-and two-transistor-off cases. Turning off three transistors improves subthreshold leakage, but provides a diminishing return.

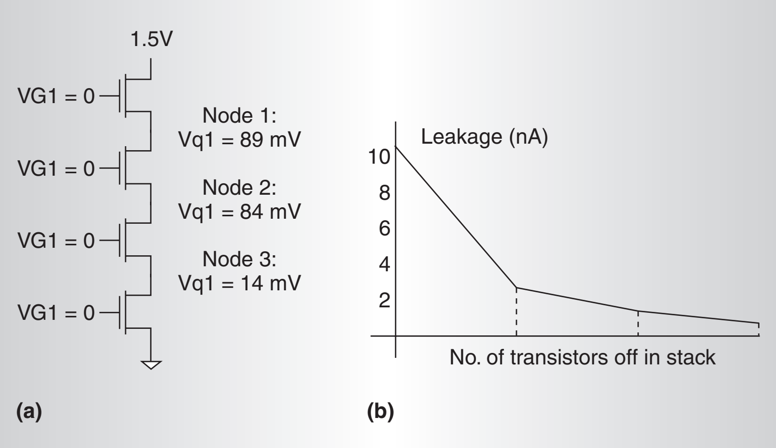

Figure 7. Effect of transistor stacking on source voltage (a), and leakage current versus number of transistors off in stack (b).

Table 2. Input vector control.

Circuit	Input vector	I_{DDQ} (nA)	Type of case
Four-input NAND	ABCD = 0000	0.60	Best
	ABCD = 1111	24.1	Worst
Three-input NOR	ABC = 111	0.13	Best
	ABC = 000	29.5	Worst
Full adder	ABCi = 111	7.8	Best
	ABCi = 001	62.3	Worst
4 bit ripple adder	ABC = 111	91.3	Best
	A = B = 1111, Ci = 1	94.0	Best
	A = B = 0101, Ci = 1	282.9	Worst

The voltages at the internal nodes depend on the input applied to the stack. Functional blocks such as NAND, NOR, or other complex gates have a ready 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. Lee et al.[10] propose a model and algorithm to estimate leakage and to select the proper input vectors to minimize the leakage in logic blocks. Table 2 shows the quiescent current (I_{DDQ}) flowing into different functional blocks for the best- and worst-case input vectors. All results are based on HSpice simulation using 0.18-µm technology with V_{DD} = 1.5 V. Results show that applying the proper input vector can efficiently reduce the total subthreshold leakage in the standby mode of operation.

Gate and junction leakage are also important in scaled technologies, and can be a significant portion of total leakage. So we must

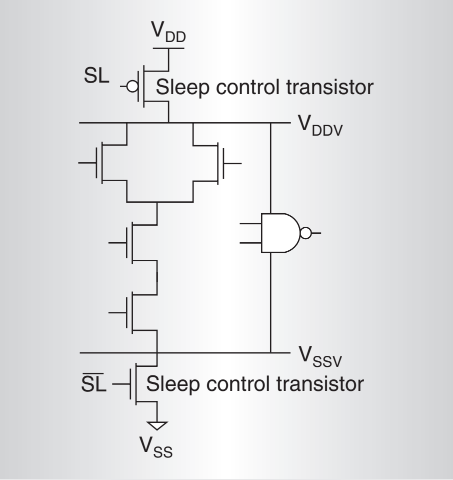

Figure 8. Schematic of an MTCMOS circuit with low V_{th} device shaded.[11]

reinvestigate the stack-of-transistors technique for these types of leakage.

Researchers have shown that with high gate leakage, the traditional way of using stacked transistors fails to reduce leakage and in the worst case might increase overall leakage.[3] Gate leakage depends on the voltage drop across different regions of a transistor. Applying 00 as the input to a two-transistor stack only reduces subthreshold leakage and does not change the gate leakage component. 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.[3] Therefore, in scaled technologies where gate leakage dominates the total leakage, using 10 might produce more savings in leakage, compared to 00. The source-substrate and drain-substrate junction BTBT leakage is a weak function of input voltage and hence we can neglect it in this analysis.

Forced-stack (sleep) transistor

This technique inserts an extra series-connected (sleep) transistor in the pull-down/pull-up path of a gate and turns it off in standby mode. The extra transistor is on during normal operation. This provides a substantial savings in leakage current during standby mode. However, the extra stacked transistor makes the drive current of the forced-stack gates lower, resulting in increased delay. Hence, this technique is only usable for noncritical paths. If the sleep transistor's V_{th} is high, extra leakage savings are possible. Such a circuit topology—one that incorporates and uses a high-Vt sleep transistor—is known as multithreshold CMOS or MTCMOS. Figure 8 shows an example.[11]

In fact, only one type (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 lower than that for a PMOS at the same width. The NMOS can thus be smaller than a corresponding PMOS.[12] However, MTCMOS can only reduce leakage power in standby mode, and the large inserted sleep transistors can increase the area and delay. Moreover, if the circuit must retain data in standby mode, it will require an extra high-V_{th} memory circuit to maintain the data.

Instead of using high-V_{th} sleep transistors, a super cutoff CMOS (SCCMOS) circuit uses low-V_{th} transistors with an inserted gate bias generator.[13] In standby mode, the gate is applied to V_{DD} + 0.4 V for PMOS (V_{SS} – 0.4 V for NMOS) by using the internal gate bias generator to fully cut off the leakage current. Compared to MTCMOS—in which 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.

Heo and Asanovic[14] proposed a sleep transistor technique to save leakage in domino gates. Figure 9 shows two small sleep transistors added to a conventional CMOS domino gate.[14] In standby mode, the clock is high, and the sleep signal is asserted. If the data input was high, node 1 would have been discharged. If the data input was low, node 1 would be high but leakage through the NMOS dynamic pull-down stack would slowly discharge the node to ground. The NMOS sleep transistor prevents any short-circuit current in the static output logic while the dynamic node discharges to ground. Node 2 would rise as static pull-up turns on, which would cause the NMOS transistors in the pull-down stacks of the following domino gates to turn on. This would accelerate the discharge of their internal dynamic nodes. Since sleep transistors are not in the critical path (the evaluation path), this technique incurs a minimal performance loss.

Forward or reverse body biasing

Variable-threshold CMOS (VTCMOS) is a body-biasing design technique.[15] Figure 10a shows the VTCMOS scheme. To achieve different threshold voltages, this scheme uses a self-substrate-bias circuit to control the body bias. In active mode, VTCMOS technique applies a zero body bias (ZBB). In standby mode, it applies a deep reverse body bias (RBB) 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.[16] However, more recent data shows that RBB's effectiveness in lowering I_{off} decreases as technology scales, because of the exponential increase in source-substrate and drain-substrate band-to-band tunneling leakage at the source-substrate and drain-substrate p-n junctions (because of halo doping in scaled devices).[16] Moreover, the shorter channel lengths as technology scales and the lower channel doping (to reduce V_{th}) worsen SCE and diminish the body effect. This in turns weakens RBB's V_{th} modulation capability.

For scaled technologies, researchers have recently proposed using forward body biasing (FBB) to achieve better current drive with fewer SCEs.[17] This circuit uses high-V_{th} transistors (high channel doping), reducing leakage in standby mode, while employing FBB in active mode to achieve high performance. Both high channel doping and FBB reduce SCEs, relaxing the scalability limitations on channel length imposed by V_{th} roll-off and DIBL. This results in higher I_{on} compared to low-V_{th} designs for same worst-case I_{off}, improving performance.

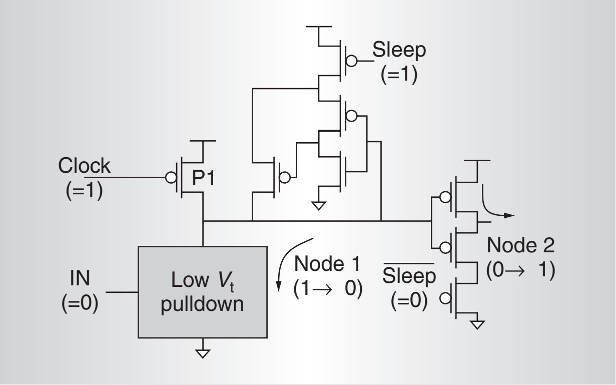

Figure 9. Domino gate with sleep transistor.[14]

A design can use RBB in standby mode together with FBB to further reduce the leakage current. Researchers have shown that FBB and high V_{th} along with RBB provides a 20 × reduction, as opposed to 3× for RBB and low V_{th}. FBB devices, however, have larger junction capacitance and body effect, which reduces the delay improvement, especially in stacked circuits. It is also possible to combine FBB with a lower V_{DD} to reduce the switching and standby leakage power yet achieve the same performance as for a high V_{DD}.

Raising the NMOS source voltage while tying the NMOS body to ground can produce the same effect as RBB. Applying a negative source voltage with respect to the body, which is tied to ground, can also provide FBB. Figure 10b illustrates the circuit diagram for these techniques.[18] The main advantage is that it then eliminates the need for a deep N- or triple-well process since the target system and the control circuitry can share the same substrate.

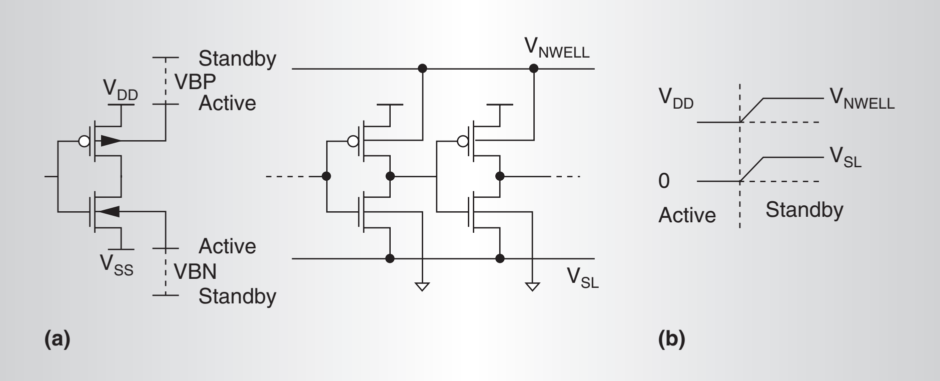

Figure 10. Variable threshold CMOS15 (a), and realizing body biasing by changing the source voltage with respect to body voltage, which is grounded[18] (b).

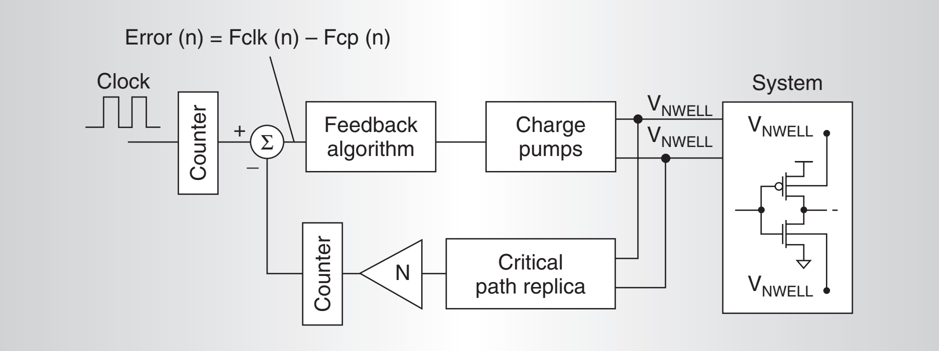

Figure 11. Dynamic V_{th} scaling system.

Active-leakage reduction techniques

During active mode, circuits work at high temperatures. Figure 4 shows that the subthreshold leakage increases exponentially with temperature, the junction BTBT increases slowly with temperature, and the gate leakage is almost independent of temperature variation. Because of the exponential increase in leakage, the active-leakage power in sub-100-nm generations accounts for a large fraction of the total power consumption, even during runtime. However, not every application requires a fast circuit to operate at the highest performance level all the time. Active-leakage reduction 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.

Dynamic V_{th} scaling (DVTS)

The DVTS scheme uses body biasing to adaptively change the V_{th} based on the performance demand. The circuit delivers the lowest V_{th} via ZBB, when the highest performance is required. When the performance demand is low, this scheme reduces the clock frequency, and raises V_{th} via RBB to reduce the runtime leakage power dissipation. In cases when there is no workload at all, the circuit can increase V_{th} to its upper limit to significantly reduce the standby leakage power. This scheme delivers just enough throughput for the current workload by tracking the optimal V_{th}. It considerably reduces leakage power by intermittently slowing down the circuit.

The literature documents the proposal of several DVTS system implementations.[19-20] Figure 11 shows DVTS hardware that uses continuous body bias control to track the optimal V_{th} for a given workload. A clock speed scheduler embedded in the operating system determines the (reference) clock frequency at runtime. 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—the frequency difference between the reference clock and the oscillator—goes into the feedback controller. The continuous feedback loop can also compensate for process, supply voltage, and temperature variations.

A simpler method, called the V_{th} hopping scheme, dynamically switches between low and high V_{th}, depending on the performance demand.[20] Figure 12 shows the schematic diagram for this scheme. Compared to the continuous body bias control in Figure 11, the discrete control has two levels of V_{th}. If control signal VTHlow_Enable is asserted, the transistors in the target system are forward body biased, and V_{th} is low. When the system can trade off performance for lower power consumption, VTHhigh_Enable is asserted, and V_{th} is high. The operating frequency of the target system is f_{CLK} when V_{th} is low and f_{CLK2} when the V_{th} is high. Researchers have verified an algorithm that adaptively changes V_{th} depending on the workload, applying it to an MPEG4 video encoding system. As mentioned in the previous section, the RBB's effectiveness is reducing because of the worsening SCE and increasing BTBT leakage at the source-substrate and drain-substrate junctions. Applying FBB together with RBB can achieve a better performance-leakage tradeoff for DVTS systems.

Circuit techniques to reduce leakage in cache memories

Figure 13a shows the 7 available terminals in a conventional six-transistor SRAM cell: V_{SL}, V_{PWELL}, V_{NWELL}, V_{DL}, V_{WL}, V_{BL}, and V_{BLB}. Researchers have proposed various SRAM cell architectures that control one or more of the seven terminal voltages during standby mode, to reduce the leakage components in Figure 13b. Each technique exploits the fact that the active portion of a cache is very small, which gives the opportunity to put the large idle portion in a low-leakage sleep mode. We evaluate the effectiveness and overhead of each technique based on the following criteria:

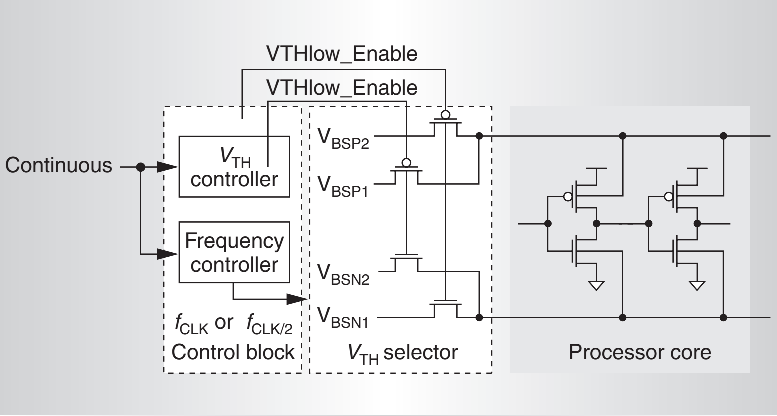

Figure 12. V_{th} hopping scheme.

- *Leakage reduction.* Although subthreshold leakage still continues to dominate the I_{off} at high temperatures, ultrathin oxides and high doping concentrations have led to a rapid increase in direct-tunneling gate leakage and BTBT leakage at the source and drain junctions in the nanometer regime. Each leakage reduction technique needs reevaluation in scaled technologies where subthreshold conduction is not the only leakage mechanism.
- *Performance.* Some techniques, such as source biasing, lengthen the delay of reading or writing to memory due to additional circuit in critical path. This negative effects needs assessment along side a given technique's benefits.
- *Mode transition overhead.* Although creating alternate mode can save power, most systems have a limited time and energy budget for mode transition. An assessment should consider the overhead in terms of transition latency and energy.
- *Stability.* The leakage reduction technique should not have a noticeable impact on SRAM cell stability or soft error rate.

Table 3 shows our assessment of the various techniques, based on these criteria.

The source biasing scheme raises the source line voltage (V_{SL}) in sleep mode,[21,22] which reduces subthreshold leakage because of the three effects described earlier. It also reduces the gate leakage in the cell because of the

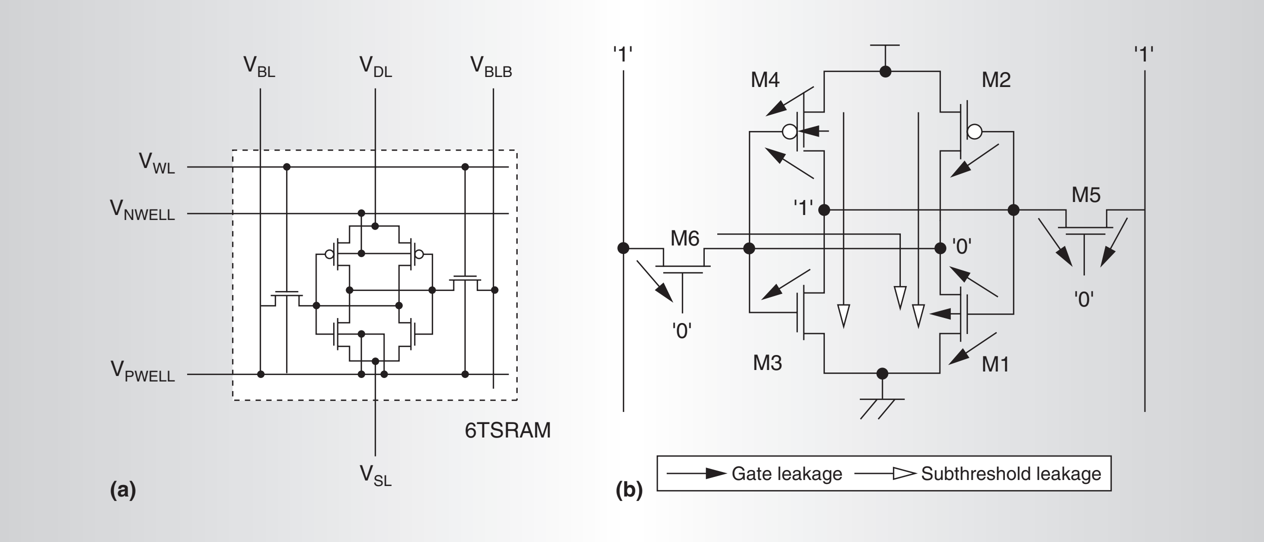

Figure 13. Seven terminals of the 6T SRAM cell (a), and dominant leakage components in a 6T SRAM (b).

Table 3. Low-leakage SRAM cell techniques.

Criteria	Source biasing (V_{SL})	RBB/FBB (V_{PWELL}, V_{NWELL})	Dynamic V_{DD} (V_{DL})	Leakage biased (V_{BL}, V_{BLB})	Negative word line (V_{WL})
Leakage reduction	Subthreshold, gate: ↓↓	Subthreshold: ↓↓ BTBT: ↑	Subthreshold, gate: ↓ Bit line leakage: ↑	Subthreshold, gate: ↓	Subthreshold: ↓, gate: ↑
Performance	Delay increase	No delay increase	No delay increase	No delay increase	No delay increase
Overhead	Medium transition overhead	Large transition overhead	Large transition overhead	Precharge latency overhead	Low charge pump efficiency
Stability	Increases soft-error rate	No impact on soft-error rate	Worst soft-error rate	No impact on soft-error rate	No impact on soft-error rate, high voltage stress

relaxed signal rail, $V_{DD} – V_{SL}$.[22] Such a scheme requires an extra NMOS to be series connected in the pull-down path to cut off the source line from ground during sleep mode; this in turn imposes an extra access delay. The reduced signal charge in sleep mode also causes the soft-error rate to rise, requiring additional error correction code circuits.

RBB the NMOS or PMOS transistors can reduce subthreshold leakage via body effect, but does not affect access time by switching to ZBB in active mode.[23] The body bias transition does impose a large latency or energy overhead because of the large V_{BB} swing and substrate capacitance. This scheme becomes less attractive in scaled technologies because the body coefficient decreases with smaller dimensions, and RBB also increases the source and drain junction BTBT leakage. For scaled technologies, a recently proposed design uses FBB to reduce subthreshold leakage and to achieve better current drive while maintaining reasonable junction BTBT.[24] A new high-V_{th} device optimized for FBB changes the doping profile by adjusting the peak halo doping (channel engineering) or uses gate material with a higher work function (a technique called *work function engineering*).[24] This scheme resolves the drawback associated with RBB SRAM and suggests a viable solution for reducing leakage in nanoscale memories.

A dynamic V_{DD} SRAM (DVSRAM) lowers supply voltage,[25] which in turn reduces the subthreshold, gate, and BTBT leakage. This scheme requires a smaller signal rail (V_{DL} – V_{GND}) compared to the SBSRAM for equivalent leakage savings. Although there is no impact on delay in active mode, the large V_{DD} swing between sleep and active mode imposes a larger latency or energy transition overhead than does SBSRAM. Moreover, the greatest drawback of the DVSRAM is that it increases the bit line leakage in sleep mode since the voltage level in the stored node also drops for lower V_{DD}. Therefore, this scheme is not suitable for dual-V_{th} designs where the speed-critical access transistors may already be using low-V_{th} devices with high leakage levels.

Researchers have also proposed a technique that biases the bitlines to an intermediate level to reduce the access transistor leakage via the DIBL effect.[26] Since only access transistors benefit from the leakage reduction, the overall leakage savings is moderate. Unlike the three previously mentioned techniques, it is necessary to apply this scheme to the entire subarray because different cache lines share the bit line. The main limitation comes from the fact that there is a precharge latency whenever a new subarray is accessed. This would require an architectural modification to resolve the multiple hit times in case the precharge instant is not known ahead of time.

The negative word line scheme[27] pulls V_{WL} down to a negative voltage during standby to avoid subthreshold leakage through the access transistors. However, it causes increased gate leakage and higher voltage stress in the access transistors. Although this technique has no impact on performance or soft-error rate, it causes a loss of power because of generating the negative bias using charge pumps. This becomes more serious as the supply voltage scales.

In each more advanced technology generation, semiconductor devices scale downward to achieve high integration density. At the same time, supply voltage also scales downward to achieve a lower switching energy per device. However, high performance also requires a commensurate scaling of the transistor threshold voltage, which in turn causes an exponential increase in subthreshold leakage current. So aggressive device scaling into the nanometer regime not only increases the subthreshold leakage but also has other negative impacts, such as increased drain-induced barrier-lowering, V_{th} roll-off, reduced on- to off-current ratio, and increased source-drain resistance. Avoiding these SCEs requires the incorporation of oxide thickness scaling and higher nonuniform doping, which results in an exponential increase in gate and junction BTBT leakage. Collectively, these factors lead to an increase in total leakage, making leakage current a major component of total power consumption. Hence, leakage reduction techniques are becoming indispensable to future designs. MICRO

Acknowledgments

The research was funded in part by Semiconductor Research Corp. (SRC 1078.001), the Gigascale System Research Center, Intel, and IBM.

References

1. S. Borkar, "Design Challenges of Technology Scaling,'' *IEEE Micro,* vol. 19, no. 4, July-Aug. 1999, pp. 23-29.
2. Y. Taur and T.H. Ning, *Fundamentals of Modern VLSI Devices,* Cambridge Univ. Press, 1998.
3. S. Mukhopadhyay et al., "Gate Leakage Reduction for Scaled Devices Using Transistor Stacking," *IEEE Trans. VLSI Systems,* vol. 11, no. 4, Aug. 2003, pp. 716-730.
4. D.A. Antoniadis et al., "'Well-Tempered' Bulk-Si NMOSFET Device Home Page," http://www-mtl.mit.edu/researchgroups/Well/.
5. M. Ketkar et al., "Standby Power Optimization via Transistor Sizing and Dual Threshold Voltage Assignment," *Proc. Int'l Conf. Computer-Aided Design* (ICCAD 02), IEEE Press, 2002, pp. 375-378.
6. T. Karnik et al., "Total Power Optimization by Simultaneous Dual-V_t Allocation and Device Sizing in High-Performance Microprocessors," *Proc. 39th Design Automation Conf.* (DAC 02), IEEE Press, 2002, pp. 486-491.
7. J.T. Kao et al., "Dual-Threshold Voltage Techniques for Low-Power Digital Circuits," *IEEE J. Solid-State Circuits,* vol. 35, no. 7, July 2000, pp. 1,009-1,018.
8. N. Sirisantana, L. Wei, and K. Roy, "High-Performance Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide Thickness," *Proc. Int'l Conf. Computer Design* (ICCD '00), IEEE CS Press, 2000, pp. 227-234.
9. K.A. Bowman et al., "Impact of Die-to-Die and Within Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," *IEEE J. Solid-State Circuits,* vol. 37, no. 2, Feb 2002, pp. 183-190.
10. Z. Lee, et al., "Two-Dimensional Doping Profile Characterization of MOSFETs by Inverse Modeling Using Characteristics in the Subthreshold Region," *IEEE Trans. Electron Devices,* vol. 46, no. 8, Aug. 1999, pp. 1,640–1,649.
11. S. Mutoh et al., "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold Voltage CMOS," *IEEE J. Solid-State Circuits,* vol. 30, no. 8, Aug. 1995, pp. 847-854.
12. J. Kao, A. Chandrakasan, and D. Antoniadis, "Transistor Sizing Issues and Tool for Multi-threshold CMOS Technology," *Proc. ACM/IEEE Design Automation Conference* (DAC 34), ACM Press, pp. 409-414, June 1997.
13. H. Kawaguchi, K. Nose, and T. Sakurai, "A CMOS Scheme for 0.5V Supply Voltage with Pico-Ampere Standby Current," *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE Press, 1998, pp. 111-116.
14. S. Heo and K. Asanovic, "Leakage-Biased Domino Circuits for Dynamic Fine-Grain Leakage Reduction," *Proc. Symp. VLSI Circuits,* IEEE Press, June 2002, pp. 316-319.
15. T. Kuroda et al., "A 0.9V 150 MHz 10 mW 4 mm 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage Scheme," *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE Press, 1996, pp. 166-167.
16. A. Keshavarzi et al., "Effectiveness of Reverse Body Bias for Low Power CMOS Circuits," *Proc. 8th NASA Symp. VLSI Design,* IEEE Press, 1999, pp. 231-239.
17. S. Narendra et al. "Forward Body Bias for Microprocessors in 130-nm Technology Generation and Beyond," *IEEE J. Solid State Circuits,* IEEE Press, May 2003, pp. 696-701.
18. H. Mizuno et al., "An 18-μA Standby Current 1.8-V, 200-MHz Microprocessor with Self-

Substrate-Biased Data-Retention Mode," *IEEE J. Solid-State Circuits,* vol. 34, no. 11, Nov. 1999, pp. 1,492-1,500.

19. C.H. Kim and K. Roy, "Dynamic V_{th} Scaling Scheme for Active Leakage Power Reduction," *Proc. Design, Automation, and Test in Europe,* IEEE CS Press, 2002, pp. 163-167.
20. K. Nose et al., "V_{th}-Hopping Scheme for 82 Percent Power Saving in Low-Voltage Processors," *Proc. IEEE Custom Integ. Circ. Conf.*, IEEE Press, 2001, pp. 93-96.
21. A. Agarwal, H. Li, and K. Roy, "A Single-V_t Low-Leakage Gated-Ground Cache for Deep Submicron," *IEEE J. Solid-State Circuits,* IEEE Press, 2003, pp. 319-328.
22. A. Agarwal and K. Roy, "Noise Tolerant Cache Design to Reduce Gate and Subthreshold Leakage in Nanometer Regime," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 03), 2003, pp. 18-21.
23. C.H. Kim and K. Roy, "Dynamic V_t SRAM: A Leakage Tolerant Cache Memory for Low Voltage Microprocessors," *Proc. Int'l Symp. Low Power Electronics. and Design* (ISLPED 02), ACM Press, 2002, pp. 251-254.
24. C.H. Kim et al., "A Forward Body-Biased Low-Leakage SRAM Cache: Device and Architecture Considerations," *Proc. Int'l Symp. Low Power Electronics and Design* (ISLPED 03), ACM Press, 2003, pp. 6-9.
25. K. Flautner, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," *Proc. 29th Ann. Int'l Symp. Comp. Architecture* (ISCA-29), IEEE CS Press, 2002, pp. 148-157.
26. S. Heo et al., "Dynamic Fine-Grain Leakage Reduction Using Leakage-Biased Bitlines," *Proc. Int'l Symp. Comp. Architecture* (ISCA-29), IEEE CS Press, 2002, pp. 137-147.
27. K. Itoh et al., "A Deep Sub-V, Single Power-Supply SRAM Cell with Multi-V_t, Boosted Storage Node and Dynamic Load," *Proc. Symp. VLSI Circuits Digest of Technical Papers,* IEEE Press, 1996, pp. 132-133.

Amit Agarwal is a research engineer in circuit research lab at Intel. His research interests include low-power, high-performance, process-tolerant cache and register file design; low-power integrated device, circuit, architecture design; and reconfigurable architecture design with unreliable components for yield improvement. Agarwal has an MS and PhD in electrical and computer engineering from Purdue University and a BTech in electrical engineering from the Indian Institute of Technology, Kanpur, India. He is a member of the IEEE.

Saibal Mukhopadhyay is a PhD candidate in electrical and computer engineering at Purdue University. His research interests include the analysis and design of low-power and robust circuits using nanoscale CMOS and circuit design using double-gate transistors. Mukhopadhyay has a BE in electronics and telecommunication electrical engineering from Jadavpur University, Calcutta, India. He is a student member of the IEEE.

Arijit Raychowdhury is a PhD candidate in electrical and computer engineering at Purdue University. His research interests include device and circuit design for scaled silicon and nonsilicon devices. Raychowdhury has a BE in electronics and telecommunication engineering from Jadavpur University, Calcutta, India.

Kaushik Roy is the Roscoe H. George professor of electrical and computer engineering at Purdue University. His research interests include VLSI design and CAD for nanoscale silicon and nonsilicon technologies; low-power electronics for portable computing and wireless communications; circuit-level system integration testing and verification; and reconfigurable computing. Roy has a PhD in electrical and computer engineering from the University of Illinois at Urbana-Champaign and a BTech degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India. He is a Fellow of the IEEE.

Chris H. Kim is an assistant professor in the electrical and computer engineering department at the University of Minnesota. His research interests include theoretical and experimental aspects of VLSI system design in nanoscale technologies. Kim has a PhD in electrical and computer engineering from Purdue University, and an MS in biomedical engineering and BS in electrical engineering from Seoul National University, Korea. He is a member of IEEE.

Direct questions and comments about this article to Amit Agarwal, Mailstop: JF2-04, 2111 NE 25th Ave., Hillsboro, OR 97124; amit1.agarwal@intel.com.