Impact of process scaling on the efficacy of leakage reduction schemes

Yuh-Fang Tsai•, David Duarte§, N. Vijaykrishnan•, and Mary Jane Irwin•
§LTD, Intel Corporation, Hillsboro, OR; david.e.duarte@intel.com
•Dept. of CSC, Penn State University, University Park, PA; {ytsai, vijay, mji}@cse.psu.edu,

ABSTRACT

The effects of technology scaling on three run-time leakage reduction techniques (Input Vector Control, Body Bias Control and Power Supply Gating) are evaluated by determining their limits and benefits, in terms of the potential leakage reduction, performance penalty and area and power overhead in 0.25um, 0.18um, 0.07um and 0.065um technologies. HSPICE simulation results and estimations with various function units and memory structures are presented to support a comprehensive analysis.

Keywords
Leakage reduction, technology scaling, process variations, low power.

1. INTRODUCTION

As technology scales down, the supply voltage must be reduced such that dynamic power can be kept at reasonable levels and power delivery can still be performed within functional requirements. But in order to prevent the negative effect on performance incurred, the threshold voltage (V_{TH}) must be reduced at the same rate such that a sufficient gate overdrive is maintained. This reduction in the threshold voltage causes an increase in the leakage current of about 5 times per generation [1], which in turn can increase the static power of the device to unacceptable levels. Thus leakage reduction is predicted necessary in the future. Among the emerging leakage reduction techniques, some require modification of the process technology, achieving leakage reduction during the fabrication/design stage while others are based on circuit-level optimization schemes that require architectural support, and in some cases, technology support as well, but are applied at run-time (dynamically).

There is some work discussing the effectiveness of leakage reduction techniques as technology scales. In [2], a model and device measurements predicting the scaling nature of the stacking effect were presented. The decreasing effectiveness of BBC with scaling was shown in [3] using transistor and test chip leakage measurements. However, the influence of design style and some other issues brought by technology scaling have not been considered in these works. One of these issues is the sensitivity of leakage power to the process variations in gate length and threshold voltage. It has been shown that the 30mV variation in threshold leakage can result in 20x difference in leakage power in 0.18um technology [4]. The impact of process variations becomes even severe in scaled technologies and should be considered when evaluating the leakage reduction techniques. Moreover, it is expected that hi-K dielectric materials will be used in more aggressive technologies. While the maturity of hi-K dielectric materials is still under debate, the contribution of gate leakage to the total leakage remains indisputable. In this paper, however, we focus our study on run-time subthreshold leakage reduction techniques applied to different functional units in data path and memory structures designed using different design styles. Our goal is to examine the effectiveness of currently used leakage reduction techniques in future technologies, considering the scaling impact, not only on the leakage reduction effectiveness but also on the incurred overheads.

The remainder of this paper is organized as follows. In Section 2, we briefly review the most commonly used leakage reduction techniques. The simulation framework is explained in Section 3 while the results of our study and correlated equations that can be used in early estimation for scaling impacts are presented in Section 4. We then discuss the influence of process variations in Section 5 and finally, some conclusions of the implications of technology scaling are given in Section 6.

2. REVIEW OF RUN-TIME LEAKAGE REDUCTION TECHNIQUES

The run-time leakage reduction techniques are based on reducing the leakage by changing the bias conditions in the four terminals of a transistor. We can generalize them into three categories:

2.1 By Input Vector Control

Many researchers have used models and algorithms to estimate nominal [5] and minimum and maximum leakage of a given circuit [6]. This work has made evident the influence of the input pattern on the circuit leakage behavior, which is a consequence of the 'stacking effect' [7]. As the state of devices in the stack is determined by their corresponding inputs, which in turn are determined by the unit's input signals, the goal can be expressed as finding

Acknowledgements: This work is supported in part by GSRC, Career Award 0093085, 0202007 and 0130143 and Intel Corporation.

Table 1: Power characteristics of the units under evaluation.

Technology (um)	Average Leakage Power (nW) (% of dynamic power) 0.25	0.18	0.07	0.065	Average Dynamic Power (mW) 0.25	0.18	0.07	0.065
32-bit Carry Lookahead Adder	712(<0.01)	1550(0.01)	555300(3.43)	117754.3 (4.89)	19.29	13.1	1.61	2.41
16x16-bit Array Multiplier	1946.12 (<0.01)	2292.8(<0.01)	817467.8 (8.6)	1060000 (10.06)	165.42	88.65	9.51	10.54
32-bit Shifter	5700 (0.07)	8600 (0.26)	44000 (31.27)	153738.3 (9.37)	7.73	3.31	1.42	1.64
3-to-1 Multiplexer (9-bit)	16.3 (<0.01)	29.4 (<0.01)	5110 (4.26)	10965.65 (5.77)	1.51	0.96	0.12	0.19
32 2-input XOR (32-bit word)	14.7 (<0.01)	28.9 (0.02)	9560 (8.69)	9440.34 (6.94)	1.33	0.13	0.11	0.14
32 2-input NAND (32-bit word)	26.9 (0.01)	24.9 (0.06)	5250 (8.75)	14857.27 (9.29)	0.32	0.04	0.06	0.16
32 2-input AND (32-bit word)	31.2 (<0.01)	30.9 (0.03)	5060 (3.61)	9559.25 (3.98)	0.74	0.11	0.14	0.24
32 2-input NOR (32-bit word)	69.1 (0.02)	84.4 (0.11)	3350 (5.58)	6603.9 (5.5)	0.45	0.08	0.06	0.12
32 2-input OR (32-bit word)	73.6 (0.01)	92.2 (0.06)	8850 (8.05)	19414.94 (7.11)	0.92	0.15	0.11	0.27
128-bit SRAM Array	44.04(<0.01)	66.16(<0.01)	207200(21.14)	312000 (22.29)	33.28	11.68	0.98	1.4

the input pattern that maximizes the number of disabled transistors in all stacks across the unit. Once this vector is found, we can switch the input vector to this minimum leakage input when the unit is idle for a period of time. The implementation of the input vector control technique requires minimal architectural support. The sleep signal that determines whether the device is active or not may be already implemented in most designs but we still need to determine the threshold of idleness beyond which the input vector control is beneficial as there is an overhead energy associated with the transition to sleep (low leakage) mode.

2.2 By Increasing the Threshold Voltage

This technique has different implementations, but all of them require some process technology support to change the threshold voltage of some (or all) transistors from the default defined for the technology. Some implementations in this category includes Multiple Threshold Voltage CMOS (MTCMOS), which assigns low threshold devices in the critical path while high threshold devices are used in non-critical path, Dynamic Threshold MOS (DTMOS), in which the body and gate of each transistor are tied together such that whenever the device is off, low leakage is achieved while when the device is on, higher current drives are possible, and Variable threshold CMOS (VTCMOS), which raises V_{TH} during standby mode by making the substrate voltage either higher than V_{dd} (P devices) or lower than ground (N devices).

2.3 By Gating the Supply Voltage

The last approach considered is power supply gating. There are many ways in which this technique can be implemented, but the basic idea remains: to shut down the power supply so that the idle units do not consume leakage power. This can be done by inserting "sleep transistors" to cut the path from the power supply to the units [8] or by controlling the supply voltage regulators. The latter can also support Dynamic Voltage Scaling (DVS), which is a popular technique for dynamic power management.

3. EXPERIMENTAL SETUP

From the techniques described in Section 2, one per category has been chosen, each of which is controllable at run-time. To obtain a comprehensive analysis of the effectiveness of each technique, the following major function units were custom designed with their power characterizations listed in Table 1: a 32-bit adder, a 16x16 multiplier, a 32-bit shifter, a 9-bit multiplexer, various 32-bit wide Boolean logic functions and a 128-bit SRAM array. The device models for 0.25um, 0.18um, and 0.07um use a BSIM3 model while for 0.065um, a BSIM4 model, which includes gate leakage, is used. The simulation results in 0.065um technology (with gate leakage) are compared against that in 0.07um technology (without gate leakage) to reflect the impact of technology improvement, especially the inclusion of hi-K dielectric materials. Note that the gate oxide material is assumed to be SiO_2. Due to the much higher energy required for hole tunneling in SiO2, gate leakage for a PMOS device is typically one order of magnitude smaller than an NMOS device with identical T_{ox}. For all designs and experiments, MicroMagic MAX is used for layout creations and HSPICE for circuit-level simulations on the conditions listed in Table 2. The possible leakage reduction is directly estimated from SPICE simulation. Take note that Short Channel Effect (SCE) and Drain Induced Barrel Lower (DIBL) have been considered when obtaining the leakage power.

Table 2: Summary of simulation conditions.

Technology	V_{dd}	V_{th} (n/p)	Temp
0.25um	2.5V	470mV/-590mV	85˚C
0.18um	1.8V	445mV / -447mV	85˚C
0.07um/0.065um	1.0V	200mV / -220 mV	85˚C

3.1 Input Vector Control

The assumption is that all designs are front-ended by latches, which is reasonable as most functional units are

normally used in pipelined datapaths. An implementation of the input control logic with reasonable area overhead is shown in Figure 1. In this design, when in sleep mode, the control_to_1 logic has two NMOS transistors in stack while control_to_0 logic has two PMOS transistors in stack in the worst case. This property reduces the leakage power of the control logic by ten folds and thus realizes its feasibility even when the leakage percentage elevates.

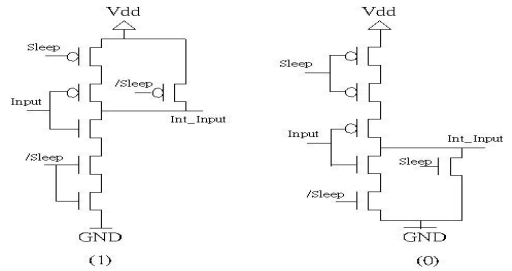

Figure 1: Modified low-leakage latches with optimum sleep values stored (1 left, 0 right).

In [9], 59 random input vectors were shown to achieve a 95% confidence of finding the input vector producing the least leakage current. The key to their approach was the fitting of a Gaussian distribution to the leakage profile obtained by the selected input vectors. In our approach, 180 random input vectors were generated to fit a Gaussian distribution of leakage measurement. Each input vector was simulated by HSPICE to find the input vector with the least leakage and highest obtainable savings. Note that no validation of IVC for memory structures is performed since no savings will be gained due to the symmetric structure of SRAM structure. However, a technique called leakage-biased bitlines (LBB) [10], which mitigates the bitline leakage flowing through the access transistors, is based on a concept similar to that of IVC. Instead of forcing the bitlines of inactive subbanks with a sleep vector, it simply turns off the hi-V_{th} NMOS precharging transistors and lets the bitlines float. The leakage current from the bit cells automatically biases the bitlines to a mid-rail voltage that minimizes the bitline leakage current. We evaluate this technique as the IVC scheme for memory structure by delaying the precharge of bitlines in SRAM cell arrays.

3.2 Body Bias Control

VTCMOS is used as the sample technique for body bias control as it requires architectural support and does not rely completely on hardware design choices and placement, allowing it to be applied at runtime. This is a required feature for useful comparison against the other techniques studied. To provide the substrate bias, we modified the netlists generated from the layouts and manually adjusted the body voltages of P and N devices, which, by default, are wired to V_{dd} and ground, respectively. This method is based on the fact that the device models for these technologies are from TSMC mature processes and the Short Channel Effect (SCE), which has a strong impact on threshold voltage, is well captured. However, this is not applicable for the 0.07um model: the Berkeley Predictive Transistor Model (BPTM), which does not include the degradation of threshold voltage caused by substrate bias due to SCE. The amount of reduction in threshold voltage is expressed in Equation (1). For our simulations of 0.07um technology, in additional to the previous modification, we adjusted the threshold voltage value according to Equation (1) manually in the netlists. Figure 2 shows the achievable increase in threshold voltage by changing the substrate bias. For 0.07um technology, both achievable threshold voltages, with and without considering SCE, are shown to illustrate the importance of including SCE.

$$\Delta V_{th} = (4.8 * t_{ox} * (\Phi + V_{sb})) / L_{eff} \qquad (1)$$

Where t_{ox} is the oxide thickness, Φ is potential barrier, V_{sb} is the substrate bias, and L_{eff} is the effective device length.

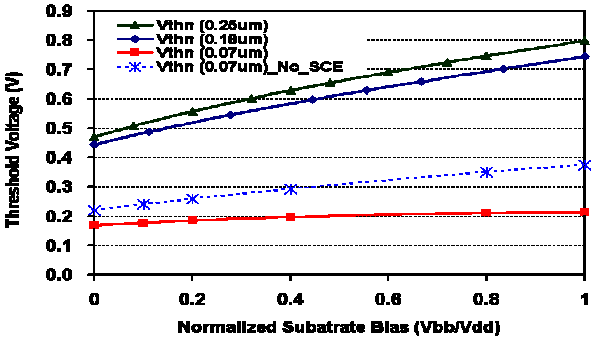

Figure 2: The achievable threshold voltage by biasing the substrate.

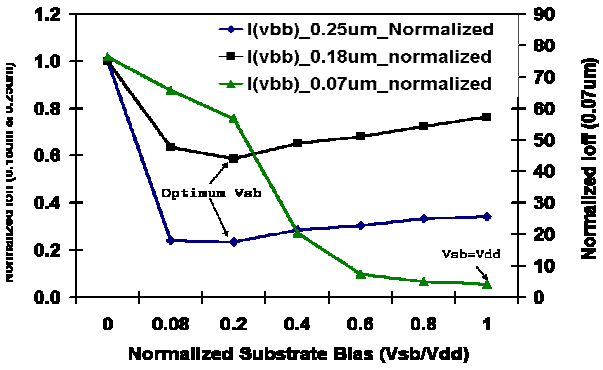

Figure 3: Leakage current vs. substrate bias showing the optimum substrate bias.

Figure 3 shows the simulated average leakage current and optimum V_{sb} value. The use of large values of V_{sb} (i.e., above 1V for 0.25um, 0.5V for 0.18um, and 0.4V for 0.07um and 0.065um) is not recommended since the large values of V_{sb} increase gate/junction leakage significantly as the electric field across the oxide is determined by the voltage difference between the gate junction and substrate.

Table 3: Various performance parameters of IVC. The performance penalty is less than 1 cycle

Technology(um)	Leakage Reduction (%)				Area Overhead (%)	Min. idle time			
	0.25	0.18	0.07	0.065	0.25/0.18/0.07	0.25 (in us)	0.18 (in us)	0.07 (in ns)	0.065(in ns)
32-bit Carry Lookahead Adder	29	30	28.5	8.89	1.84	43.91	12.39	13.60	29.36
16x16-bit Array Multiplier	9.67	11.66	6.34	11.94	0.26	1318.48	497.38	110.84	49.33
32-bit Shifter	73.2	78.22	76.53	61.22	0.54	1.17	0.29	1.67	7.59
3-to-1 Multiplexer (9-bit)	43.39	51.82	56.93	28.68	3.3	108.88	22.68	8.34	11.64
32 2-input XOR (32-bit word)	31.33	39.4	35.96	61.35	12.74	8.66	0.34	0.28	0.22
32 2-input NAND (32-bit word)	57.5	58.8	64.66	92.25	18.74	1.03	0.08	0.17	0.12
32 2-input AND (32-bit word)	48.03	48.7	33.9	31.2	16.44	6.42	0.58	2.33	2.28
32 2-input NOR (32-bit word)	53.8	62.2	71.64	36.34	13.74	0.48	0.05	0.24	0.46
32 2-input OR (32-bit word)	46.7	47.7	50.33	57.54	10.92	2.94	0.27	0.91	0.92

Table 4: Various performance parameters of LBB.

	Leakage Reduction (%)				Transition Energy (fJ)				Minimum idle time			
Technology (um)	0.25	0.18	0.07	0.065	0.25	0.18	0.07	0.065	0.25	0.18	0.07	0.065
32x4-bit SRAM Array	50.1	30.06	24.35	25.48	68.5	16.7	1.21	1.61	13.8us	0.38us	0.73ns	0.93ns

For 0.07um and 0.065um technologies, there is no optimum substrate bias level since, at the range of substrate bias applied, subthreshold leakage is always larger than gate/junction leakage. However, because of reliability concern, the voltage gap between gate and substrate should be limited to the burn-in power supply level, which is commonly estimated to be 1.4 times of V_{dd} level.

3.3 Power Supply Gating

3.3.1 Datapath logic

In our approach for datapath logics, a PLL circuit with a voltage follower as shown in Figure 4 is use as the voltage regulator to set the supply level to GND level in sleep mode. In this figure, two situations are possible. The sleep signal provides a way to perform global leakage reduction by shutting down the PLL and consequently all supply voltages that depend on the reference voltage generated (V_c), while the enable signal in the buffer provides support for local supply gating of only the units being powered by that particular buffer.

3.3.2 Memory structures

For the memory structure, a sleep transistor is inserted between the supply and cells to control the transition between active and sleep mode. The main benefit for choosing this technique is that the data can be preserved by correctly sizing the sleep transistor. Due to the regular structure of the SRAM array, the sizing of the sleep transistor can be done efficiently.

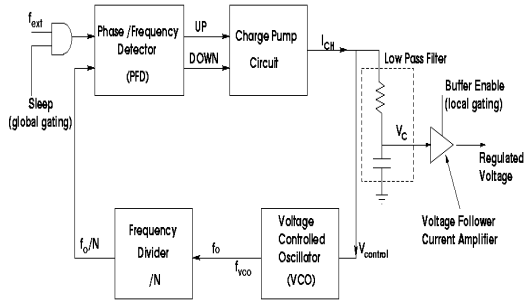

Figure 4: The PLL as a voltage regulator.

4. TECHNOLOGY SCALING IMPACT ANALYSIS

In this section, the overheads in terms of power penalty, area and performance incurred by each technique are analyzed and the simulated/estimated values are presented. The results shown in this section are acquired under the assumption of no process variations.

4.1 Input Vector Control

4.1.1 Datapath logic

It is predicted that the "stacking effect" will be more efficient for smaller technologies, which implies the improving effectiveness of IVC with technology scaling. The reason behind this is the increasing prominence of Drain Induced Barrier Lowering (DIBL). The HSPICE results in 0.25um and 0.18um technologies shown in Table 3 confirm this prediction.

In terms of power overhead, the only contribution comes from the transition from the state in which the unit was, to the minimum-leakage state once the unit enters the

sleep mode. Note that if the switching incurred in setting the input to the desired pattern causes the dynamic power consumption larger than that of the leakage at the current state for the given idle time, there will no savings. In other words, the amount of time that the unit remains idle must be long enough so that the dynamic power used in setting the low-leakage input is less than the consumed leakage power during the same time if no low-leakage input is set. In [10], this minimum idle time is formulated as:

$$t_{idle} > \frac{E_{tr1} + E_{tr2} + P_{leak\_avg} \cdot (t_{tr1} + t_{tr2})}{(P_{leak} - P_{leak\_n})} \approx \frac{E_{tr} + P_{leak\_avg} \cdot (2t_{tr})}{(P_{leak} - P_{leak\_n})}$$

While all the parameters are shown in Figure 5. The technology scaling impact on power overhead is evaluated by the measurements from HSPICE simulation in different technologies. Due to the increasing leakage reduction, the minimum idle time decreases with technology scaling.

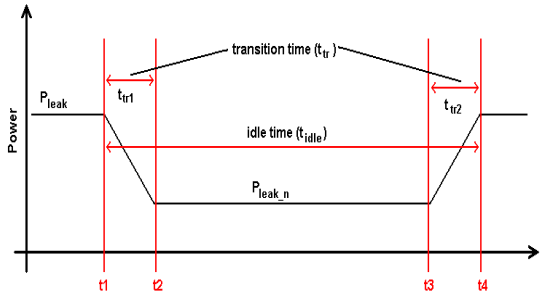

Figure 5: Representation of leakage power behavior during scheme application.

The area overhead can be hidden if the unit is frontended with latches and thus is small for the function units as shown in Table 3; otherwise, the area overhead is proportional to the number of input nodes of the unit. Note that the area overhead is fixed across technologies.

IVC can mitigate both subthreshold leakage and gate leakage. However, the best-input vectors for these two leakage mechanisms are different. Thus, as can be seen from the data of 0.07um and of 0.065um technologies in Table 3, the savings depends on the percentages of these two mechanisms as well as the design styles.

4.1.2 Memory

The simulation results of applying Leakage-Biased Bitline (LBB) on a 32x4-bit SRAM cell array are shown in Table 4. Different from IVC applied on datapath logic, the efficacy decreases as technology scales.

The transition energy penalty happens when restoring the charge back to the bitlines before the memory cells can be used and wakeup latency is the precharging time, which is delayed until the subbank needs to be accessed. Since there is no extra hardware needed to implement this technique, there is no area overhead.

Since the bias condition of the SRAM 6-T cell in floating state incurs less gate leakage than that in the precharge state, BBL performs somewhat better when considering gate leakage.

4.2 Body Bias Control

Table 5 shows the various performance parameters for BBC. Our results confirm the decreasing effectiveness of BBC, which is illustrated by the curves for the 0.07um technology in Figure 2, which shows the reduced V_{th} control ability of substrate bias due to the previously mentioned increased effect of SCE and V_{th} roll off. While prior research [12] indicates that the optimum substrate bias level for leakage reduction depends only on the process technology, our results reveal that the design style plays an important role in selecting this value as well. The raised threshold voltage increases the leakage current of "on" NMOS's (PMOS's) which pass a weak 1(0) or transistors in pass-transistor logic while reduces that of "off" transistors. However, this situation diminishes in smaller technologies.

Table 5: Various performance parameters for Body Bias Control (BBC).

Technology (um)	Leakage Reduction (%) /<Vsb (V)>				Transition Energy (pJ)				Minimum idle time (us)			
	0.25	0.18	0.07	0.065	0.25	0.18	0.07	0.065	0.25	0.18	0.07	0.065
32-bit Carry Lookahead Adder	81.7<0.5>	60.55<0.5>	39.53	20.38	21.40	25.44	323.23	769.21	30.07	9.04	4.43	5.84
16x16-bit Array Multiplier	77.94<1.0>	85.65<0.8>	41.5	11.94	89.84	44.21	4691.60	7225.60	50.73	4.61	4.24	6.42
32-bit Shifter	91.83<1.0>	73.96<0.5>	46.1	18.39	136.62	124.01	2467.59	1018.53	18.42	5.79	3.89	6.02
3-to-1 Multiplexer (9-bit)	76.82<0.8>	61.36<0.5>	50.9	26.42	1.24	0.73	27.82	69.34	89.86	27.47	3.61	5.41
32 2-input XOR (32-bit word)	93.9<1.0>	92.6<0.8>	85.45	14.16	1.35	0.73	51.58	64.10	95.04	25.55	3.55	6.35
32 2-input NAND (32-bit word)	48.5<0.5>	50.9<0.5>	51.12	28.37	1.28	0.64	26.77	92.38	66.34	30.59	2.93	5.21
32 2-input AND (32-bit word)	50.1<0.5>	57.5<0.5>	59.88	10.07	1.43	0.74	26.19	65.99	61.71	25.98	3.08	6.59
32 2-input NOR (32-bit word)	66.7<1.0>	66.3<0.5>	57.91	10.62	2.24	1.50	18.22	45.74	33.54	8.62	3.56	6.63
32 2-input OR (32-bit word)	64.7<0.5>	69.6<0.5>	51.83	24.11	2.86	1.69	48.69	123.97	39.98	8.20	3.74	5.54
128-bit SRAM Array	85.96<1.0>	88.76<0.8>	48.8	7.16	5.63	2.24	1495.79	2170.97	143.8	38.54	7.22	6.71

The power overhead is represented by the circuitry in charge of adjusting the body bias voltage. The circuit presented in [13] uses a charge pump to change the substrate level to an optimum standby bias and a charge injector to perform the recovery to active mode in reasonable time while trying to keep the area overhead to a minimum. In this implementation, there is a portion of the circuit that continuously draws current from the supply, but its effect can be ignored due to small magnitude (around 1nA and can be kept small with careful design as technology scales down). The bulk of the power overhead is in the energy required to charge the substrate when the system is entering a sleep mode. Since the transition time for fully charging the substrate is comparatively longer, there are two cases to be considered. Independent of how fast the substrate is charged, the energy required to charge the substrate can be estimated as:

- for $t_{idle} >= t_{sleep}$ where the substrate is fully charged to the optimum substrate bias level:

$$E_{ch-subs} = (\Delta V_{ch})^2 C_{subs} r = (\Delta V_{ch})^2 (C_{sub/A} A)$$

- for $t_{idle} < t_{sleep}$ where the substrate is partially charged to a level less than the optimum substrate bias:

$$E_{ch-subs}(t_{tr1}) = \frac{t_{tr1}}{t_{sleep}} * (\Delta V_{ch})^2 C_{subs} r = \frac{t_{tr1}}{t_{sleep}} * (\Delta V_{ch})^2 (C_{sub/A} A)$$

Where t_{tr1} is the period when the charge pump is charging the substrate, t_{sleep} is the time for the substrate to be fully charged to the desired substrate bias level, A is the area utilized and $C_{sub/A}$ is the capacitance per unit of area from the substrate to the active regions (P or N). We assume linear relationships between the transition time and transition energy and between the transition time and the obtained reduced leakage power. This assumption results in a pessimistic but safe since the deviation of the transition energy (a smaller transition energy is estimated) is less than that of the reduced leakage power (a larger reduced leakage power is estimated) and thus the estimated idle time is the worst-case number. The leakage power consumed during the time the scheme is applied, P_{lavg}, can be estimated as simply the average of P_{leak} and $P_{leak\_n}(t)$ for convenience. Note that since the leakage current of the substrate bias control circuitry is only 1nA, its leakage power can be neglected. The minimum idle time thus can be formulated as:

- for $t_{idle} >= t_{sleep}$
- $$t_{idle} > \frac{E_{tr} - P_{leak\_n} \cdot (t_{sleep} - t_{wakeup})}{(P_{leak} - P_{leak\_n})}$$
- for $t_{idle} < t_{sleep}$

$$t_{idle} \geq -\frac{t_{sleep} * P_{leak\_n}}{P_{leak} - \frac{2 * E_{ch-subs}}{t_{sleep}}}$$

Where t_{wakeup} is the transition time from sleep mode to active mode.

The performance overhead happens when changing the substrate level. The transition time of the charging circuits can be estimated as:

$$t_{delay} = (\Delta V_{sub} * C_{sub}) / W_{driving\_device} * I_{on}$$

Where ΔV_{sub} is the voltage difference of substrate to be charged, C_{sub} is the substrate capacitance, $W_{driving\_device}$ is the width of driving devices and I_{on} is the transistor saturation current. To satisfy the feasibility and to match the speed improvement of commercial products, we scale up the size of driving transistors in the charging circuit so that the delay is scaled by 0.7x per generation. The incurred area and power overhead across technologies can be estimated with the other parameters scaled using the scaling factors in [14].

The estimated data in Table 6 shows the increasing area overhead. However, despite of the decreasing effectiveness of BBC, the data in Table 5 shows the minimum idle time reduces. This is not only because of the assumption of scaled delay time but also of the larger percentage of leakage

Since threshold voltage has negligible impact on gate leakage, the efficacy of BBC, which reduces subthreshold leakage by controlling threshold voltage, is expected to be even significantly less when taking gate leakage into consideration.

4.3 Power Supply Gating

4.3.1 Datapath logic

Since the PSG technique reduced the power supply level to GND level in sleep mode, the leakage reduction is virtually 100% cross all technologies.

The power and area overhead come from the global PLL and local buffer circuitry. Since the estimation is at the granularity of the functional unit level, only the overhead of the local buffer is included giving the penalty caused by the global PLL is hidden when a whole system is considered.

Table 6: Performance penalty and delay time for transitions between sleep mode and active mode and area penalty for Body Bias Control (BBC).

Technology (um)	0.25	0.18	0.07/ 0.065
Active to sleep mode delay time (us)	30	21	7.2
Performance Penalty (wake up delay time) (ns)	30	21	7.2
Area Penalty (%)	1.12	1.44	2.76

Table 7: Various performance parameter of PLL-Based PSG. The leakage reduction is virtual 100% for all units.

	Area Overhead (%)			Buffer Enable Time (ns)			Buffer Nominal Power (uW)			Minimum Idle Time (us)		
	0.25	0.18	0.07	0.25	0.18	0.07	0.25	0.18	0.07	0.25	0.18	0.07
32-bit Carry Lookahead Adder	7.20	11.79	4.97	459.07	248.37	5.41	463.96	437.39	64.80	80.60	14.42	4.89
16-bit x 16-bit Multiplier	9.67	12.78	4.74	5511.46	2353.05	44.70	3971.08	2955.72	380.80	353.99	92.36	2.73
32-bit Shifter	2.19	2.28	3.09	183.96	62.76	4.77	186.52	111.05	57.20	4.03	0.66	5.42
3:1 Multiplexer (9-bit)	8.88	11.64	4.58	35.94	18.20	0.40	37.24	32.72	5.20	275.58	55.72	3.94
32 2-input XOR (32-bit word)	7.30	4.05	3.79	31.65	2.46	0.37	32.92	5.05	4.80	269.15	7.68	1.93
32 2-input NAND (32-bit word)	7.09	5.66	4.18	7.62	0.76	0.20	8.68	2.05	2.80	35.39	2.74	1.92
32 2-input AND (32-bit word)	8.48	6.01	7.05	17.61	2.09	0.47	18.76	4.39	6.00	70.56	6.07	4.64
32 2-input NOR (32-bit word)	6.04	4.80	3.26	10.71	1.52	0.20	11.80	3.39	2.80	19.37	1.62	3.01
32 2-input OR (32-bit word)	5.60	3.89	3.51	21.89	2.84	0.37	23.08	5.72	4.80	37.19	2.78	2.09

Table 8: Parameters of Gated-Vdd applied to a 128-bit SRAM array. P:PMOS, N:NMOS, C:CMOS sleep transistor.

	Leakage Reduction (%)			Area Overhead (%)		Normalized Access Time		Minimum idle time (ns)		
	0.18	0.07	0.065	0.18	0.07/0.065	0.18	0.07/0.065	0.18	0.07	0.065
P	64.8	87.8	43.47	1.8	2.5	1	1	169.6	0.2	89.64
N	83.3	96.1	38.68	0.6	0.34	1.02	1.02	177	4.54	112.88
C	92.8	98.3	43.32	0.6	1.7	1.03	1.07	169.8	4.24	88.61

Table 9: Role of process variations (pv) in leakage reduction scheme efficiency.

	Leakage reduction (%) – IVC			Leakage reduction (%) - BBC		
	w/o pv	w/ pv	Efficiency impact	w/o pv	w/ pv	Efficiency impact
32-bit Carry Lookahead Adder	8.89	5.76	0.65	20.38	25.35	1.24
16-bit x 16-bit Multiplier	11.94	0.97	0.08	11.94	12.21	1.02
32-bit Shifter	61.22	53.35	0.87	18.39	19.85	1.08
3:1 Multiplexer (9-bit)	28.68	20.40	0.71	26.42	28.17	1.07
32 2-input XOR (32-bit word)	61.35	61.20	1.00	14.16	14.26	1.01
32 2-input NAND (32-bit word)	92.25	74.53	0.81	28.37	21.55	0.76
32 2-input AND (32-bit word)	31.2	31.11	1.00	10.07	10.12	1.00
32 2-input NOR (32-bit word)	36.34	36.16	1.00	10.62	10.68	1.01
32 2-input OR (32-bit word)	57.54	19.02	0.33	24.11	24.94	1.03
128-bit SRAM Array	N.A.	N.A.	N.A.	7.16	8.7	1.22

In contrast to what was done earlier with buffers for BBC, the driver is not sized for a constant delay overhead but to meet the corresponding unit's average current requirements during normal operation. Due to this reason, the results in Table 7 show that the area overhead and buffer enable time (performance penalty) depend on the unit the scheme is applied to. However, the incurred performance and power penalty decrease with technology scaling. Thus the minimum idle time decreases.

4.3.2 Memory structures

Gated-Vdd is used for the implementation of our PSG for memory structures. Simulation results in Table 8 show that the effectiveness improves and the minimum idle time decreases as expected. The sleep transistor is sized to preserve the data in sleep mode, both the area and performance penalty increase for smaller technologies. The noise analysis is critical when implementing this technique due to the low V_{dd} level for preserving data in cells.

From the results in Table 8, we can see that the efficiency of Gated-Vdd decreases when including gate leakage in evaluation and when gate leakage is comparable to subthreshold leakage. Gated-V_{dd} reduces both subthreshold leakage and gate leakage. However, by analyzing the bias conditions in sleep mode, we can find it to be less efficient in reducing gate leakage than

reducing subthreshold leakage. A reason for the drop in efficiency is the extra gate leakage incurred by the sleep transistors implementing Gated-V_{dd}. Another difference is the efficiency of different types of sleep transistors. When only subthreshold leakage is present, NMOS sleep transistor is superior than PMOS sleep transistor since NMOS can reduce the subthreshold leakage of the access transistors in a 6-T SRAM cell while the PMOS effect is less. However, when gate leakage is comparable to subthreshold leakage, PMOS sleep transistor performs better since it reduces the gate leakage of the pass transistors while an NMOS sleep device is not capable of doing so.

5. Process Variations Impact

To assess the impact of process variations on the efficacy of the various leakage reduction schemes, Monte Carlo analysis in Hspice was performed. We assumed 10% 3-sigma variations in both threshold voltage and gate length which are the dominate parameters influencing leakage current. The magnitude of the variations considered is extreme but are chosen to maximize the effect and facilitate observations.

Results in Table 9 show how the process variations affect the efficiency of IVC and BBC. The parameter "efficiency impact" is defined as the savings gained when assuming there is no process variations divided by the saving when assuming 3-sigma process variations. We can see that when process variations are included in the evaluation, the efficiency of IVC reduces while that of BBC increases. We see this trend as follows. For IVC, the "stack effect" is not as effective when gate length decreases. The leakage gained by a process-induced gate length reduction is more than the leakage eliminated with transistor stacking, for a given input vector combination. However, the overall impact depends on distribution of the variations and the resulting efficiency varies for different designs. We argue that with the inclusion of process variations, the minimum leakage vector would need to be recalculated as the leakage profile of the unit changes. For BBC, due to the exponential relation between the growth of leakage current and threshold voltage, the higher the threshold voltage, the less the increase in leakage current caused by variation in threshold voltage. BBC actually shifts the mean threshold voltage higher and thus the mean leakage current is further reduced when considering process variations.

6. Conclusion

Table 10 shows trends of parameters while technology scales, based on the assumption of a scaling factor of 0.7x per generation for the delay time. It should be noted that the efficacy of BBC would reduce as technology scales while that of others increase. The effectiveness decrease of BBC is due to V_{TH} roll-off and elevating SCE. Note that the declining leakage reduction causes undesirable idle time needed to gain power saving. To solve this problem, larger driving devices are recommended at the expense of the area overhead. Our results show that even though the effectiveness of BBC decreases, the reduction will be significant (>50% in average) and the minimum idle time can be tuned to a desirable value with reasonable area overhead down to 0.07um technology.

The column 4 in Table 10 shows decreasing minimum idle time for all the techniques evaluated regardless the trends of effectiveness. This is due to the increasing percentage of the leakage power. The decreasing ratio shown is the ratio of the minimum idle time in cycles in 0.18um technology to that in 0.07um technology. A 0.7x (per generation) scaling factor of the cycle time is assumed. The scaling factor of minimum idle time is smaller than that of the cycle time. This promises the feasibility of these techniques even when there are less slacks of idleness resulted by the increasing operating speed.

All the evaluated techniques cause one-time delay penalty when waking up the units. However, Gated-Vdd is the only technique incurs run-time performance penalty due to the sleep transistor being always present. Because of the increasing driving current requirements, even though the wake up time decreases, the simulation result shows that the run-time performance penalty increases.

Table 10: Impact comparison of technology scaling. The decreasing ratio is the ratio of the minimum idle time in cycles in 0.18um to that in 0.07um. Cycle time scaled by 0.7x per generation.

Method	Leakage Reduction	Area Overhead	MIT (decreasing ratio)
Input Vector Control	Increase	Fixed	Decrease (x0.001)
IVC-based BBL	Decrease	Fixed	Decrease (x0.0025)
Body Bias Control	Decrease	Increase	Decrease (x0.58)
Supply Gating (local)	Increase	Depend	Decrease (x0.34)
Gated-Vdd	Increase	Increase	Decrease (x0.05)

The comparison of subthreshold leakage reduction techniques, when including gate leakage, suggests that gate leakage should be considered when choosing a leakage reduction scheme. Moreover, process variations are crucial to leakage power in advanced technologies. Initial observations indicate that the efficiency of BBC is more consistent than that of IVC

when considering process variations. Further study is required to understand better the mechanisms behind this behavior.

With the availability of functional units in datapath and memory structures, our analysis provides a comprehensive prediction and validation for the implications of technology scaling to the run-time leakage reduction techniques.

7. REFERENCES

[1] Borkar, S., "Design Challenges of Technology Scaling", IEEE MICRO, July-August 1999.

[2] Narendra, S., et al, "Scaling of stack effect and its application for leakage reduction," Low Power Electronics and Design, International Symposium on, pp. 195-200, '01.

[3] Keshavarzi, et al, "Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs," Low Power Electronics and Design, International Symposium on, pp 207-212, 2001

[4] S. Bokar, et al., "Parameter Variations and Impact on Circuits and Microarchitecture", Design Automation Conference, pp.338-342 , June 2003

[5] Cheng, Z., Johnson, M., Wei, L. and Roy, K., "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks", ISLPED 98, pp. 239-244.

[6] Johnson, M., Somasekhar, D. and Roy, K., "Models and Algorithms for Bounds in CMOS Circuits", IEEE Transactions on CAD of Integrated Circuits and Systems, Vol. 18, No. 6, June 1999, pp. 714-725.

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

[8] Mutoh, S, et al, "1-V Power Supply High-Speed Digital Circuit Technology with Multi-threshold Voltage CMOS", IEEE Journal of Solid-state Circuits, pp. 847-854, August 1995.

[9] Halter J., and Najm, F., "A Gate-level Leakage Power Reduction Method for Ultra Low Power CMOS Circuits, IEEE Custom Integrated Circuits Conference, pp. 475-478, 1997.

[10] Heo, S., et al, "Dynamic Fine-Grain Leakage Reduction Using Leakage-Biased Bitlines", International Symposium for Computer Architectures, pp. 137-147, May 2002

[11] Duarte, D. et al., "Evaluating Run-time Techniques for Leakage Reduction", 7th ASPDAC/15th International conference on VLSI Design, pp.31-38, Jan, 2002.

[12] Keshavarzi, A, et al, "Technology Scaling Behavior of Optimum Reverse Body Bias for Standby Leakage Power Reduction in CMOS IC's", Intl. Symp. Low Power Electronics and Design, pp. 252-254, Aug. 1999

[13] Kuroda, T., et al, "A 0.9V 150MHz 10mW 4mm2 2-D discrete cosine transform core processor with variable threshold-voltage (VT) scheme," IEEE Journal of Solid-State Circuits, pp. 1770-1779, November 1996.

[14] Duarte, D., "Clock Network and Phase-Locked Loop Power Estimation and Experimentation," PhD Thesis, Department of Computer Science and Engineering, Penn State University.