Implications of Technology Scaling on Leakage Reduction Techniques*

Y-F. Tsai,
Department of CSE, Penn State University
ytsai@cse.psu.edu

D. Duarte
LTD, Intel Corporation
david.e.duarte@intel.com

N. Vijaykrishnan, M.J. Irwin
Department of CSE, Penn State University
{vijay, mji}@cse.psu.edu

ABSTRACT

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

Categories and Subject Descriptors
B.7.1 [**Types and Design Styles**]: VLSI, Advanced technologies;

General Terms
Design, Experimentation.

Keywords
Leakage reduction, technology scaling, 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. In order to prevent the negative effect on performance, the threshold voltage (V_{TH}) must be reduced proportionally with the supply voltage so that a sufficient gate overdrive is maintained. This reduction in the threshold voltage causes a 5x leakage current increase per generation, which in turn can increase the static power of the device to unacceptable levels. This clearly justifies the need for leakage reduction techniques, even for current technologies. 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 previous work discussing the effectiveness of leakage reduction techniques as technology scales. In [1], device measurements and a model predicting the scaling nature of the stacking effect were presented. The decreasing effectiveness of Body Bias Control (BBC) with scaling was shown in [2] using transistor and test chip leakage measurements. However, the influence of design style has not been considered in these works. In [3], the evaluation of run-time leakage reduction techniques applied to various functional units in datapath and memory structures designed using different design styles has been done. In this paper, we examine how the effectiveness of these techniques scale with technology.

The remainder of this paper is organized as follows. In Section 2, we briefly review the most commonly used leakage reduction techniques. In Section 3, the simulation framework is explained. Section 4 presents the results of our study and correlates them with some equations that can be used for early estimation of the scaling effects. Finally, some conclusions of the implications of technology scaling are given in Section 5.

DAC 2003, June 2-6, 2003, Anaheim, California, USA.

2. REVIEW OF RUN-TIME LEAKAGE REDUCTION TECHNIQUES

2.1 By Input Vector Control

Many researches have made evident the influence of input pattern on circuit leakage behavior, which is a consequence of the 'stacking effect' [4]. As the state of devices in a 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 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.

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. The basic idea is to shut down the power supply so that the leakage power of idle units is reduced. This can be done by inserting "sleep transistors" to cut the path from the power supply to the units or by controlling the supply voltage regulators.

* Acknowledgements: This work is supported in part by NSF 0082064, 0093085,and 0103583, NSR/RI 0202007, and MARCO 98-DF-600 GSRC.

3. EXPERIMENTAL SETUPS

From each technique described in Section 2, one particular run-time, controllable method was selected. To provide a comprehensive analysis of the effectiveness of each technique, all the major functional units and a 128-bit SRAM array were custom designed. For all designs and experiments, MicroMagic MAX is used for layout creation and HSPICE for circuit-level simulations under the conditions listed in Table 1. The possible leakage reduction is directly estimated from SPICE simulation and no further validation is possible for predictive technologies as no fabricated parts are available for testing.

Table 1: Summary of simulation conditions

Technology	V_{dd}	V_{th} (n/p)	Freq. (Hz)	Temp
0.25um	2.5V	470mV/-590mV	850M	85˚C
0.18um	1.8V	445mV/-447mV	1G	85˚C
0.07um	1.0V	200mV/-220mV	3G	85˚C

3.1 Input Vector Control (IVC)

In our evaluation, we start with the design of functional units that are front-ended by latches for use in pipelined datapaths. This latch is modified to support the input control logic as shown in Figure 1 In this design, when in sleep mode, the control_to_1 logic has two NMOS transistors in stack while the control_to_0 logic has two PMOS transistors in stack in the worst case. The use of stacking in the input control latch reduces the leakage power of the control logic tenfold.

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

In [5], 59 random input vectors were shown to achieve a 95% confidence of finding the input vector producing the least leakage current. The key to the proposed 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 measurements. Each input vector was simulated by HSPICE to find the input vector with the least leakage. The control unit was then added to the circuit, and the average obtainable leakage savings were found. Note that no validation of IVC for memory structures is performed since no savings will be gained due to the symmetric structure of the basic SRAM cell.

3.2 Body Bias Control (BBC)

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. The substrate bias level is manually modified to the optimum level for BBC. It must be noted that for 0.07um technology, the Berkeley Predictive Transistor Model (BPTM) does not capture the degradation of threshold voltage caused by substrate bias due to SCE. For our simulations of 0.07um technology, in addition to the previous modification, we adjusted the threshold voltage value according to (1) manually in the netlists. Figure 2 shows the achievable increase in threshold voltage by changing the substrate bias. For 0.07um technology, the achievable threshold voltages, with and without considering SCE, are shown to illustrate the importance of capturing this effect.

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

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. Simulation results show that there are optimum V_{sb} values. The use of large values of V_{sb} (i.e., above 1V for 0.25um, 0.8V for 0.18um, and 0.4V for 0.07um) is not recommended since the large values of V_{sb} increase gate/junction leakage (as they are a function of the voltage difference between the gate/junction and the substrate). Unlike the 0.25um and 0.18um technology which have an optimum substrate bias level that balances the conflicting tradeoff between subthreshold and gate/junction leakage, for 0.07um technology, overall leakage keeps reducing with increasing substrate bias. This is because subthreshold leakage is always larger than gate/junction leakage at high temperature for this technology. We assume the use of high-k dielectric gate oxide technology to control gate leakage. However, because of reliability concerns, there is a limit to the magnitude of substrate bias. Specifically, the voltage gap between gate and substrate should be limited to the burn-in power supply, which is typically 1.4 times the V_{dd} level.

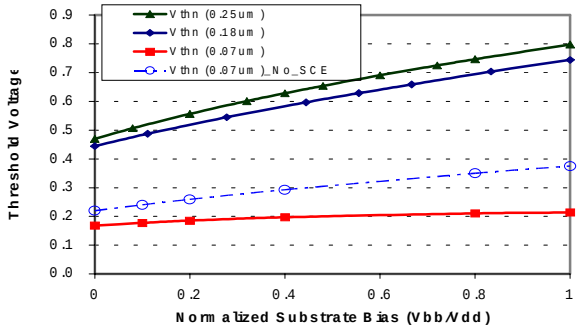

Figure 2: The achievable threshold voltage by BBC.

3.3 Power Supply Gating (PSG)

3.3.1 Datapath logic

In our approach for datapath components, a PLL circuit with a voltage follower (buffer) is used as a voltage regulator to set the supply level to GND level in sleep mode. Our implementation supports two levels of hierarchy in supply gating. 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 at the output buffers provides support for local supply gating of only the units being powered by a particular buffer.

3.3.2 Memory structure

A sleep transistor is inserted between the power supply and the cells to control the transition between active and sleep mode, as per the Gated-Vdd technique. 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.

Table 2: Various performance parameters for Input Vector Control (IVC).

Technology (um)	Leakage Reduction (%) .25	.18	.07	Area Overhead (%) .25/0.18/0.07	Min. idle time .25 (in us)	.18 (in us)	.07 (in ns)
32-bit Carry Lookahead Adder	29	30	28.5	1.84	43.91	12.39	13.60
16x16-bit Array Multiplier	9.67	11.66	6.34	0.26	1318.48	497.38	110.84
32-bit Shifter	73.2	78.22	76.53	0.54	1.17	0.29	19.53
3-to-1 Multiplexer (9-bit)	43.39	51.82	56.93	3.3	108.88	22.68	8.34
32 2-input XOR (32-bit word)	31.33	39.4	35.96	12.74	8.66	0.34	0.28
32 2-input NAND (32-bit word)	57.5	58.8	64.66	18.74	1.03	0.08	0.17
32 2-input AND (32-bit word)	48.03	48.7	33.9	16.44	6.42	0.58	2.33
32 2-input NOR (32-bit word)	53.8	62.2	71.64	13.74	0.48	0.05	0.24
32 2-input OR (32-bit word)	46.7	47.7	50.33	10.92	2.94	0.27	0.91

Table 3: Various performance parameters for Body Bias Control (BBC).

Technology (um)	Leakage Reduction (%) /<Vsb (V)> .25	.18	.07	Transition Energy (pJ) .25	.18	.07	Minimum idle time (us) .25	.18	.07
32-bit Carry Lookahead Adder	81.7<0.5>	60.55<0.5>	39.53<0.4>	21.40	25.44	323.23	30.1	9.04	4.43
16x16-bit Array Multiplier	77.94<1.0>	85.65<0.8>	41.5<0.4>	89.84	44.21	4691.60	50.7	4.61	4.24
32-bit Shifter	91.83<1.0>	73.96<0.5>	46.1<0.4>	136.62	124.01	2467.59	18.4	5.79	3.89
3-to-1 Multiplexer (9-bit)	76.82<0.8>	61.36<0.5>	50.9<0.4>	1.24	0.73	27.82	89.9	27.47	3.61
32 2-input XOR (32-bit word)	93.9<1.0>	92.6<0.8>	85.45<0.4>	1.35	0.73	51.58	95.0	25.55	3.55
32 2-input NAND (32-bit word)	48.5<0.5>	50.9<0.5>	51.12<0.4>	1.28	0.64	26.77	66.3	30.59	2.93
32 2-input AND (32-bit word)	50.1<0.5>	57.5<0.5>	59.88<0.4>	1.43	0.74	26.19	61.7	25.98	3.08
32 2-input NOR (32-bit word)	66.7<1.0>	66.3<0.5>	57.91<0.4>	2.24	1.50	18.22	33.5	8.62	3.56
32 2-input OR (32-bit word)	64.7<0.5>	69.6<0.5>	51.83<0.4>	2.86	1.69	48.69	40.0	8.20	3.74
128-bit SRAM Array	85.96<1.0>	88.76<0.8>	48.8<0.4>	5.63	2.24	1495.79	143.8	38.54	7.22

4. TECHNOLOGY SCALING IMPACT ANALYSIS

4.1 Input Vector Control

It is predicted that the "stacking effect" will be more efficient for smaller technologies, which implies effectiveness improving 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. However, the leakage reduction data in 0.07um is underestimated since the BPTM does not capture the increasing DIBL factor and fails to present the increasing effectiveness.

The only penalty is the transition from the current state to the minimum-leakage state as the unit enters the sleep mode. This minimum idle time in order to gain savings is estimated by plugging our simulation results in the formula presented in [3]. Due to the increasing leakage reduction, the minimum idle time decreases with technology scaling.

If the unit is front-ended with latches, the area overhead is small as shown in Table 3.

4.2 Body Bias Control

Table 3 shows the various performance parameters for BBC. Our results confirm that BBC will be less effective with technology scaling, which can be further supported by the curves for 0.07um in Figure 2. The power overhead is represented by the circuitry in charge of adjusting the body bias voltage. Our implementation 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 its small magnitude (around 1nA for 0.25um and can be kept small with careful design as technology scales down). The bulk of the power overhead is in the energy required to change the substrate when the system is entering the sleep mode. Independent of how fast the substrate is charged, the energy required to charge the substrate can be estimated as:

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

Where ΔV_{ch} is the substrate bias level, A is the area utilized and $C_{sub/A}$ is the capacitance per unit of area between the substrate and the active regions (P or N). If the idle time is less than the time needed to fully charge the substrate, the incurred transition energy is proportional linearly to the idle time. Note that since the leakage current of the substrate bias control circuitry is small (1nA for 0.25um), its leakage power can be neglected. The minimum idle time thus can be estimated as shown in Table 3. The performance overhead is represented by the wakeup time needed to discharge the substrate, which can be estimated as:

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

Where ΔV_{sub} is the voltage difference at the substrate to be discharged, 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 circuit 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 [6].

Table 5: Various performance parameter of PLL-Based PSG. The leakage reduction is virtually 100% for all units.

	Area Overhead (%)			Buffer Enable Time (ns)			Buffer Nominal Power (µW)			Minimum Idle Time (µs)		
	.25	.18	.07	.25	.18	.07	.25	.18	.07	.25	.18	.07
32-bit Carry Lookahead Adder	7.20	11.79	4.97	459.07	248.37	5.41	463.96	437.39	64.80	80.6	14.4	4.9
16-bit x 16-bit Multiplier	9.67	12.78	4.74	5511.46	2353.05	44.70	3971.08	2955.72	380.80	354.0	92.4	2.7
32-bit Shifter	2.19	2.28	3.09	183.96	62.76	4.77	186.52	111.05	57.20	4.0	0.7	5.4
3:1 Multiplexer (9-bit)	8.88	11.64	4.58	35.94	18.20	0.40	37.24	32.72	5.20	275.6	55.7	3.9
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	269.2	7.7	1.9
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	35.4	2.7	1.9
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	70.6	6.1	4.6
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	19.4	1.6	3.0
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	37.2	2.8	2.1

Table 6: 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)	
	.18	.07	.18	.07	.18	.07	.18	.07
P	64.8	87.8	1.8	2.5	1	1	170	0.2
N	83.3	96.1	0.6	0.34	1.02	1.02	177	4.5
C	92.8	98.3	0.6	1.7	1.03	1.07	170	4.2

4.2 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% across all technologies. The power and area overhead come from the top-level 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. Note that the overhead of the global PLL is hidden when the whole system is considered.

In contrast to what was done earlier with buffers for BBC, the driver is not sized for a constant delay overhead but instead to meet the corresponding unit's average current requirements during normal operation. Due to this reason, the results in Table 5 show that the area overhead and buffer enable time (performance penalty) depend on the unit whose supply rail is gated. We also observe that the minimum idle times decrease as incurred performance and power penalty decrease with technology scaling.

4.3.2 Memory structure

Gated-Vdd is used for the implementation of our PSG for memory structure. Simulation results in Table 6 show that the effectiveness improves and the minimum idle time decreases as expected. However, as the sleep transistor is sized to provide data-preserving in sleep mode, both the area and performance penalty increase for smaller technologies.

Table 7: Comparisons of impacts of technology scaling.

	Leakage Reduction	Area Overhead	Minimum Idle Time (decreasing ration)
IVC	Increase	Fixed	Decrease (x0.001)
BBC	Decrease	Increase	Decrease (x0.58)
PSG (local)	Increase	Depend	Decrease (x0.34)
Gated-Vdd	Increase	Increase	Decrease (x0.05)

5. CONCLUSION

Table 7 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 effectiveness of BBC reduces as technology scales while that of others increase. Our results show that even though the effectiveness of BBC decreases, the reduction will still be significant for 0.07um (>50% in average) and the minimum idle time can be tuned to a desirable value with reasonable area overhead. However, BBC is intrinsically more problematic for reliability as the high voltage across oxide decreases the lifetime of the devices.

The column 4 in Table 7 shows decreasing minimum idle time for all the techniques evaluated regardless of the trends for 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 trend indicates that there will be more opportunities in future technologies for applying the leakage mitigation techniques for even shorter duration of functional unit/memory idleness.

Our analysis provides a summary of the implications of technology scaling to the run-time leakage reduction techniques. This can be very useful for computer architects and system designers in deciding the power budget for the different components in the presence of leakage mitigation techniques.

6. REFERENCE

[1] Narendra, S., et al, “Scaling of stack effect and its application for leakage reduction,” ISLPED, pp. 195-200, ’01.

[2] Keshavarzi, et al, “Effectiveness of reverse body bias for leakage control in scaled dual Vt CMOS ICs,” ISLPED, pp 207-212, 2001

[3] Duarte, D, et al, “Evaluating Run-Time Techniques for Leakage Reduction”, ASPDAC, pp. 31-38, 2002

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

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

[6] Duarte, D, “Clock Network and Phase-Locked Loop Power Estimation and Experimentation”, PhD Thesis, Penn State University, May. 2002