

Skewed Flip-Flop Transformation for Minimizing Leakage in Sequential Circuits

Jun Seomun
Dept. of Electrical
Engineering, KAIST
Daejeon 305-701, Korea

Jaehyun Kim
Dept. of Electrical
Engineering, KAIST
Daejeon 305-701, Korea

Youngsoo Shin
Dept. of Electrical
Engineering, KAIST
Daejeon 305-701, Korea

ABSTRACT

Mixed V_t has been widely used to control leakage without affecting circuit performance. However, current approaches target the combinational circuits even though sequential elements, such as flip-flops, contribute an appreciable proportion of the total leakage. A skewed flip-flop (SFF) is obtained by slightly increasing the gate length of a subset of the transistors in a conventional flip-flop. The resulting SFF will exhibit very skewed characteristics in terms of leakage and delay, which depend on the transistors that are replaced. We present an algorithm that selectively substitutes SFFs for conventional flip-flops in sequential circuits, such that the timing constraint is still satisfied while the leakage from the flip-flops is reduced. When combined with the mixed V_t technique, an average leakage saving of 16% is achieved, compared to the use of mixed V_t alone.

Categories and Subject Descriptors: B.6.1 [Logic Design]: Design Styles—*Sequential circuits*; B.7.1 [Integrated Circuits]: Types and Design Styles—*VLSI*

General Terms: Algorithms, Design

Keywords: Flip-flop, sequential circuit, leakage current, mixed V_t

1. INTRODUCTION

Scaling down of transistors size has resulted in dramatic increase of leakage current. MOSFET threshold voltages are commonly scaled down to compensate for the reduced circuit performance at a low supply voltage, which leads to an exponential increase in sub-threshold leakage. Gate oxide is also scaled down for better control of MOSFET channel current, which is another reason for the ever-increasing gate leakage. The overall leakage current has now become a major contributor to total power consumption. In many technologies, it takes up to 50% of the overall power [1].

A mixed V_t circuit [2] utilizes low-threshold voltage (V_t) gates on timing-critical paths and high (and/or normal) V_t gates on paths which are not critical to timing. This autonomously reduces both active and standby leakage. As opposed to other circuit techniques such as power gating and reverse body bias [1], mixed V_t is a design-time technique that does not require any designer effort, although there is a small increase of manufacturing cost. Many algorithms for the use of mixed V_t have been proposed. However, all the algorithms proposed so far target the combinational portion of a circuit, although sequential elements such as flip-flops are respon-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DAC 2007, June 4–8, 2007, San Diego, California, USA.

Copyright 2007 ACM 978-1-59593-627-1/07/0006 ...\$5.00.

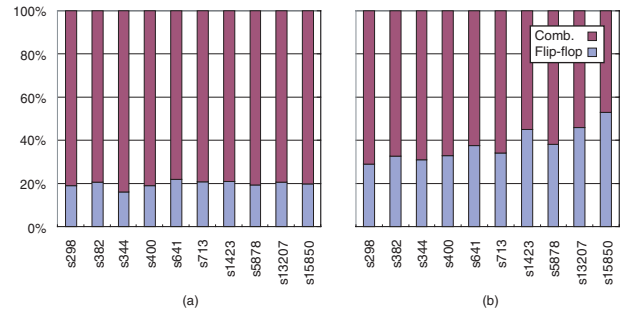


Figure 1: Distribution of leakage in combinational subcircuits and flip-flops of several ISCAS benchmarks: (a) before and (b) after the use of mixed V_t .

sible for an appreciable proportion of the total leakage. Figure 1(a) shows that the flip-flops contribute, on average, 20% to the total leakage of ISCAS benchmark circuits. But if mixed V_t is used in designing the combinational subcircuits [3], then the proportion of leakage in flip-flops goes up to 31% on average and can get as high as 54%.

In this paper, we propose the concept of *skewed flip-flops (SFFs)*. These flip-flops have very unequal leakage (and timing) characteristics for different present- and next-state combinations. This is made possible by increasing the gate lengths [4] of different combinations of transistors in a conventional flip-flop. We will go on to propose an algorithm that utilizes these skewed flip-flops to reduce the overall flip-flop leakage while maintaining the original cycle time of a circuit. Then, as before, we apply conventional mixed V_t design to the combinational portion of the circuit. The results show that we can reduce leakage by an additional 16% on average.

2. PRELIMINARIES

2.1 Computation of Idle State Probabilities

We will exploit the idle state probabilities of flip-flops, i.e. the state probabilities when the circuit is in idle. For D flip-flop, the probabilities of D-input and Q-output are the same and equal to the state probability. However, if we only consider a sequence of idle intervals, which interleave with active intervals, the probabilities of D-input and Q-output are different. They can be derived as follows. We will suppose that the design starts from a state transition graph (STG) and we also assume that the (idle) input probability distribution is available. The probability of the present state (Q) can be obtained from eigenvector of the transition matrix corresponding to the unit eigenvalue [5]. The present state probabilities together with the input probabilities are propagated [6] through the combinational subcircuit to yield the probability distribution of the next states of the inputs D. If we have a structural description of a design, we can simulate the circuit with a sequence of (idle) input patterns, monitor the next and present states, and derive their probabilities.

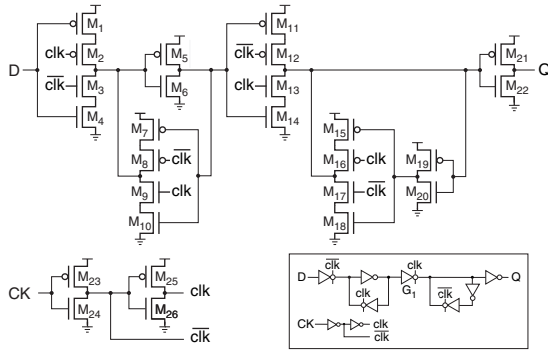


Figure 2: An example D flip-flop with inverter and tristate inverter implementation.

Table 1: Groups of transistors that make up leakage sources for a given D-input or Q-output

Group name	Transistors
D_0	M_4, M_5, M_{10}, M_{14}
D_1	M_1, M_6, M_7, M_{11}
Q_0	M_{18}, M_{19}, M_{21}
Q_1	M_{15}, M_{20}, M_{22}

2.2 Gate-Length Biasing

Gate-length biasing involves a small increase in the gate lengths of devices. In a 130-nm industrial process, it is reported [4] that an 8 nm increase in gate length yields 30% decrease in leakage with a 5% increase in delay for a minimum size inverter. This large decrease in leakage with just a small delay occurs because the nominal gate length of the technology is usually very close to the knee of the leakage versus gate length curve that is produced by short channel effects.

3. SKEWED FLIP-FLOPS

3.1 Design of an SFF

Figure 2 shows an example (positive edge-triggered) D flip-flop. The leakage of this flip-flop is determined by its D-input and Q-output¹. We define groups of transistors that are turned off (thus becoming the leakage source) for a given input or output of the flip-flop. Table 1 shows these groups. The transistors of group D_0 , for example, are those that are turned off when the D-input is low (refer to Figure 2).

Depending on the pair of groups (one for the D-input and another for the Q-output) that we take for gate-length biasing, we can design four different new flip-flops, which are all examples of skewed flip-flops (SFFs). The SFF SF_{00} , for example, has gate-length biased transistors belonging to groups D_0 and Q_0 . Since we are increasing the gate length of the transistors that are the leakage source when both the D-input and Q-output are low, the leakage of SF_{00} for that input-output combination can be made very low. Most flip-flops generate both phases of the clock signal internally through cascaded inverters, as shown in Figure 2. We therefore apply gate-length biasing to M_{24} and M_{25} , since they are turned off when the circuit is idle. Due to these transistors, even when both the D-input and the Q-output are high, the leakage of SF_{00} is still reduced.

The transistors that are driven by the clock need separate attention. The transistors that are turned on (M_2, M_3, M_{16}, M_{17}) when the circuit is idle do not need gate-length biasing. The transistors

¹We assume the clock input (CK) to be low when the circuit is idle. This is a reasonable assumption due to the widespread use of clock gating.

Table 2: Leakage of skewed flip-flops

FF	Leakage (nA)			
	DQ = 00	DQ = 01	DQ = 10	DQ = 11
Orig.	963.9	1119.1	1163.9	858.7
SF_{00}	203.8	654.5	521.5	533.9
SF_{01}	501.9	237.0	839.7	326.4
SF_{10}	340.3	792.4	250.2	410.8
SF_{11}	639.9	581.4	716.7	204.5

Table 3: Timing characteristics of skewed flip-flops

FF	Delay (ps)			
	Rising T_{su}	Falling T_{su}	Rising T_{c-q}	Falling T_{c-q}
Orig.	17.0	14.6	28.2	28.4
SF_{00}	19.3	11.2	34.2	30.0
SF_{01}	19.2	11.2	31.9	35.0
SF_{10}	15.9	17.2	35.5	29.9
SF_{11}	15.9	17.3	32.4	31.9

M_8 and M_9 , two out of three transistors in tristate inverter, are turned off. Since the leakage through tristate inverter is already small, M_8 and M_9 are not candidates for gate-length biasing. The same is true of M_{12} and M_{13} in SFFs SF_{00} and SF_{11} , because the input and output of the tristate inverter G_1 are different. However, in SF_{01} , both of the input and output of G_1 are low. The leakage through M_{13} and M_{14} is small, since both are turned off. However, M_{12} is between V_{dd} and the output, which is low, with M_{11} turned on, is thus a leakage source. Therefore, in SF_{01} , we subject M_{12} to gate-length biasing, as well as the transistors in groups D_0 and Q_1 . Similarly, in SF_{10} , we subject gate-length biasing to M_{13} , as well as to the transistors in groups D_1 and Q_0 .

In order to test our skewed flip-flops, we used a 45-nm predictive technology model [7]. For the conventional D flip-flop shown in Figure 2, we applied a gate-length biasing of 4 nm. Table 2 compares the leakage of a original flip-flop and the four SFFs. As expected, SF_{00} , exhibits the lowest leakage when both the D-input and the Q-output are low. Note that the leakage when both the D-input and the Q-output are high also drops due to the gate-length biasing of M_{24} and M_{25} .

Note that the leakage currents in Table 2 are idle-state ones, which are all lower than those of conventional flip-flop. However, since we increased the gate length of some transistors in SFFs, the increased switching current can outweigh the reduced leakage current in active state. However, the simulation results even under the worst case condition (100% switching activity of D-input and more than 1 GHz clock frequency) show that the sum of switching and leakage current of SFFs is still smaller than that of conventional flip-flop.

3.2 Timing Characteristics of SFF

The timing parameters of SFFs, namely the setup time T_{su} and the clock-to-Q delay T_{c-q} , are shown in Table 3 together with those of a conventional flip-flop. Because of the way we select a subset of the transistors for gate-length biasing, the SFFs exhibit a very asymmetric timing behavior.

Figure 3 shows the waveforms that explain the timing characteristics of SF_{00} . Note that the timing parameters of a flip-flop are measured with respect to its clock input (CK); they are affected by the late arrival of $c1k$ and $c1\bar{k}$, which are internally generated and thus lag behind CK (refer to Figure 2). The rising D-input (refer to Figure 2 and Figure 3) is captured in the master latch using M_4 , M_5 , and M_{10} , which are all slower in SF_{00} than in the conventional flip-flop, because we increased their gate lengths. However, a D-input can be captured only after $c1k$ arrives at the gate input of M_9 , and $c1k$ arrives later than it does in a conventional flip-flop (for the same rising CK), because the gate lengths of M_{24} and M_{25} have also been increased. Figure 3(a) explains the increased rising setup

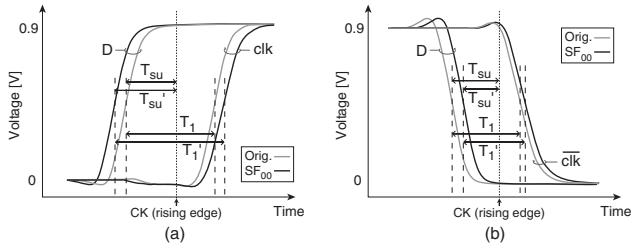


Figure 3: Comparison of SF_{00} and a conventional flip-flop: (a) rising T_{su} , and (b) falling T_{su} .

time of 2.3 ps ($T'_{su} - T_{su}$), although the increase in the delay from D-input to $\overline{\text{clk}}$ ($T'_1 - T_1$) is even larger. A falling D-input is not affected by gate-length-biased transistors, and $\overline{\text{clk}}$ is delayed due to M_{24} . That is why, as shown in Figure 3(b), the falling setup time is decreased rather than increased. The rising and falling clock-to-Q delays of SF_{00} and the timing parameters of the other SFFs can be understood similarly.

4. SFF TRANSFORMATION (SFX)

4.1 Overview

The input to our design process is a netlist for a sequential circuit, which is obtained from conventional logic synthesis. We assume that all the gates including flip-flops are initially at low V_t . For each flip-flop in the netlist, we need to know the signal probabilities of its D-input ($p(D)$) and its Q-output ($p(Q)$) when the circuit is idle. This data can be obtained as described in Section 2. For each flip-flop, we can then compute the probabilistic leakage if it were to be replaced by an SFF:

$$L = (1 - p(D))(1 - p(Q))L_{00} + (1 - p(D))p(Q)L_{01} + p(D)(1 - p(Q))L_{10} + p(D)p(Q)L_{11}, \quad (1)$$

where L_{ij} indicates the leakage of the replacement SFF when the D-input is at logic i and the Q-output is at logic j (refer to Table 2). We now find the SFF that minimizes (1), and then substitute that SFF for the original flip-flop in the netlist.

Once all the original flip-flops have been replaced by SFFs, the leakage from the flip-flops is reduced, but there will be timing violations. Therefore, we now select the SFFs which are causing timing problems, and replace them with other SFFs with better timing parameters. However, this is not always possible, or not the best solution, due to a couple of problems. First, it may happen that there is no dominating SFF in terms of timing parameters. For example, SF_{10} has the worst rising T_{c-q} , but its falling T_{c-q} and rising T_{su} are the best. Second, as well as solving the timing problems, we want to leave adequate slacks for the gates in the combinational subcircuit, so that as many gates as possible can take advantage of high V_t , which is very hard, if not impossible, to predict at this point in the design.

The conversion of the original flip-flops to SFFs (and vice versa) is abrupt in terms of leakage and timing. To achieve a smoother transition, we will define a new pair of intermediate flip-flops, called *half-skewed flip-flops* (HSFs). HSF_0 has gate-length biasing on transistors M_{15} , M_{18} , M_{19} , M_{20} , M_{21} and M_{22} (refer to Figure 2), while M_1 , M_4 , M_5 and M_6 are gate-length biased in HSF_1 . The inverters to generate the clock signals are not included in the gate-length biasing. It is easy to see that HSF_0 does not affect the setup time, while the clock-to-Q delay remains the same in HSF_1 , as shown in Figure 4(a). The reduction in leakage, as shown in Figure 4(b), is not significant, but the HSFs can now guarantee either the setup time or the clock-to-Q delay.

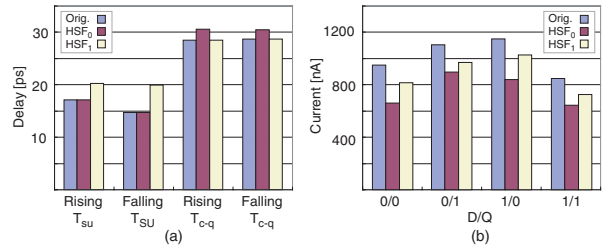


Figure 4: Half-skewed flip-flops: (a) timing parameters and (b) leakage current.

```

SFX:
L1   $p_c = \text{critical path}$ 
L2  if ( $d(p_c) > \text{delay constraint}$ ) {
L3     $f_s = \text{flip-flop at the leading end of } p_c$ 
L4     $f_t = \text{flip-flop at the trailing end of } p_c$ 
L5    if ( $f_s = \phi$ ) SUBSTITUTE( $f_t$ , trail, phase of  $p_c$  at trailing end)
L6    else if ( $f_t = \phi$ ) SUBSTITUTE( $f_s$ , lead, phase of  $p_c$  at leading end)
L7    else {
L8       $\delta_s = T_{c-q}(f_s) - T_{c-q}(FF_o)$ 
L9       $\delta_t = T_{su}(f_t) - T_{su}(FF_o)$ 
L10     if ( $\delta_s < \delta_t$ ) SUBSTITUTE( $f_t$ , trail, phase of  $p_c$  at trailing end)
L11     else SUBSTITUTE( $f_s$ , lead, phase of  $p_c$  at leading end)
L12   }
L13   go to L1
L14 }
L15 else return

SUBSTITUTE ( $f$ , end, phase):
L16 if ( $f \in \{SF_{00}, SF_{01}, SF_{10}, SF_{11}\}$ ) {
L17   case (end, phase):
L18     lead, rising: if ( $f \neq SF_{01}$ ) { $f = SF_{01}$ } else { $f = HSF_1$ }
L19     lead, falling: if ( $f \neq SF_{10}$ ) { $f = SF_{10}$ } else { $f = HSF_1$ }
L20     trail, rising: if ( $f \neq SF_{11}$ ) { $f = SF_{11}$ } else { $f = HSF_0$ }
L21     trail, falling: if ( $f \neq SF_{00}$ ) { $f = SF_{00}$ } else { $f = HSF_0$ }
L22 }
L23 else if ( $f \in \{HSF_0, HSF_1\}$ ) {
L24   case (end):
L25     lead: if ( $f \neq HSF_1$ ) { $f = HSF_1$ } else { $f = FF_o$ }
L26     sink: if ( $f \neq HSF_0$ ) { $f = HSF_0$ } else { $f = FF_o$ }
L27 }
L28 }
L29 return

```

Figure 5: Algorithm for skewed flip-flop transformation.

4.2 SFX Algorithm

Now we have three groups of flip-flops: the originals, the HSFs, and the SFFs. We start with all SFFs, as discussed in Section 4.1, and try to improve the timing: replace some SFFs with some other SFFs having better timing parameters. If we fail, we start to use HSFs, and try again. If that also fails, we return to the original unmodified flip-flop.

The procedure SFX, shown in Figure 5, is a sketch of the algorithm for iteratively identifying a critical path p_c (L1 and L11) and converting one of the flip-flops at the leading (L3) or at the trailing end (L4) of p_c , if the delay in the critical path $d(p_c)$ is larger than the delay constraint (L2). The procedure terminates when the delay of the (most) critical path is within the constraint.

Note that if p_c has only one flip-flop (L5 and L6), because it either starts from the primary input or ends at the primary output, the choice of flip-flop is obvious. If we have two flip-flops on the critical path p_c , then we select the one that has the largest increase in its timing parameters (L7, L8, L9, and L10), compared with the original flip-flop (FF_o).

In the SUBSTITUTE procedure, if the selected flip-flop f is an SFF (L12), then we try to substitute one of the SFFs with the smallest timing parameter (either T_{su} or T_{c-q} and either rising or falling) in its category for f . If the flip-flop f already has the best timing parameter in its category, then we choose an HSF with the smallest timing parameter in that category. If the best combination of HSF and timing parameter still fail, then we have to revert to the original

Table 4: Experimental results on ISCAS benchmark circuits

Benchmark			Mixed V_t			SFX + Mixed V_t			
Name	# Gates	# FFs	Comb. (μA)	SE (μA)	Total (μA)	Comb. (\times)	SE (\times)	Total (\times)	O/H/ S
s298	130	14	30	13	43	0.97	0.44	0.81	1/ 3/ 10
s344	144	15	31	15	46	0.99	0.54	0.85	0/ 3/ 12
s349	142	15	31	15	46	1.00	0.54	0.86	0/ 2/ 13
s382	185	21	38	19	57	1.06	0.38	0.84	0/ 3/ 18
s400	198	21	38	19	57	1.12	0.36	0.87	0/ 4/ 17
s444	199	21	49	19	68	1.12	0.36	0.91	0/ 2/ 19
s526	258	21	41	19	60	0.99	0.55	0.85	6/ 3/ 12
s641	206	19	30	18	48	0.99	0.45	0.79	0/ 1/ 18
s713	206	19	34	18	52	1.00	0.45	0.81	0/ 1/ 18
s838	416	32	70	30	100	1.03	0.37	0.83	0/ 7/ 25
s5378	1534	163	244	155	399	1.07	0.42	0.82	4/16/143
s9234	1457	135	280	121	401	1.03	0.36	0.83	0/15/120
Average						1.04	0.44	0.84	

flip-flop FF_o .

5. EXPERIMENTAL RESULTS

We performed experiments on a set of sequential circuits taken from the ISCAS benchmarks. Each circuit was synthesized with SIS [8] and mapped into a 45-nm gate library, which we built based on a predictive model [7]. The library consists of 23 cells: seven flip-flops, three inverters, three 2-input NOR gates, and 2-input, 3-input, and 4-input NAND gates each in three different sizes. Technology mapping was done using a weighted sum of area and delay as the cost function, and gate sizing was performed during technology mapping.

In Table 4, the three columns under the heading Mixed V_t respectively show the leakage current of the combinational subcircuit, the sequential elements (flip-flops), and the sum of the two; these figures are for mixed V_t in the combinational subcircuit. We implemented a mixed- V_t algorithm similar to that of [3], which enumerates all the possible V_t assignments in a topological order. Each circuit was simulated ten times with SPICE using ten different idle vectors for the primary inputs when the circuit is idle, and leakage current was taken as an average of them.

The seventh, eighth, and ninth columns show the leakage (as factors of the data under the heading Mixed V_t), when we apply the SFX procedure to the technology-mapped netlist, followed by applying mixed V_t to the combinational subcircuit. The leakage of the flip-flops is cut by 56% on average, which is an understandable consequence of the distribution of flip-flops shown in the last column of Table 4 (the number of original flip-flops, HSFs, and SFFs, in the order from left to right). Since the overall delay overhead of SFFs is not significant, many flip-flops are converted to SFFs. The leakage in the combinational logic remains largely unchanged. For some benchmarks, we can see the decrease rather than increase of leakage implying that some paths have increased slack. This is partly due to the reduced setup times of some SFFs (refer to Table 3) and partly achieved by the heuristic pruning used in the mixed- V_t algorithm [3].

We also implemented a simple heuristic algorithm for applying mixed V_t both to the flip-flops and the combinational subcircuit. Figure 6 compares the total leakage from this approach with that from SFX followed by mixed V_t . Both are normalized to the results for mixed V_t applied to the combinational subcircuits alone. Our approach outperforms the heuristic just described most of the time, except for three benchmarks (s641, s5378, and s9234). These circuits have very few critical paths, and the remaining timing paths have large positive slacks that cannot be absorbed even by high V_t flip-flops and high V_t combinational gates.

6. CONCLUSION

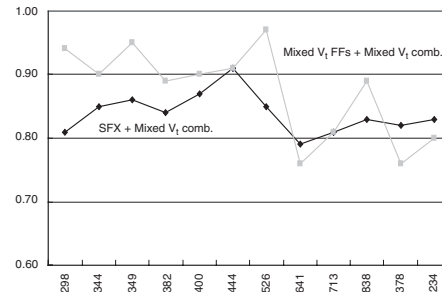


Figure 6: Comparison of mixed V_t applied to both flip-flops and the combinational subcircuit with our proposed approach.

Although in widespread use, the value of mixed V_t is limited, since it considers only the combinational portion of a circuit, even though the sequential elements contribute a non-negligible, and sometimes a significant, portion of the total leakage. We have proposed skewed flip-flops (SFFs) that exhibit very unequal leakage and timing characteristics. This concept is general and any kind of conventional flip-flop can be transformed to an SFF. We have presented a heuristic that substitutes SFFs for conventional flip-flops. An average saving of 16% of leakage was observed when this approach was compared to the use of mixed V_t alone.

Acknowledgment

This work was supported by Samsung Electronics.

References

- [1] S. G. Narendra and A. Chandrakasan, Eds., *Leakage in Nanometer CMOS Technologies*, Springer, 2005.
- [2] L. Wei, Z. Chen, M. Johnson, K. Roy, and V. De, "Design and optimization of low voltage high performance dual threshold CMOS circuits," in *Proc. Design Automation Conf.*, June 1998, pp. 489-494.
- [3] M. Ketkar and S. S. Sapatnekar, "Standby power optimization via transistor sizing and dual threshold voltage assignment," in *Proc. Int'l Conf. on Computer Aided Design*, Nov. 2002, pp. 375-378.
- [4] P. Gupta, A. B. Kahng, P. Sharma, and D. Sylvester, "Gate-length biasing for runtime-leakage control," *IEEE Trans. on Computer-Aided-Design*, vol. 25, no. 8, pp. 1475-1485, Aug. 2006.
- [5] L. Benini and G. De Micheli, "State assignment for low power dissipation," *IEEE Journal of Solide-State Circuits*, vol. 30, no. 3, pp. 258-268, Mar. 1995.
- [6] S. Ercolani, M. Favalli, M. Damiani, P. Olivo, and B. Ricc6, "Estimate of signal probability in combinational logic networks," in *Proc. European Test Conf.*, Apr. 1989, pp. 132-138.
- [7] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proc. Int'l Symp. on Quality Electronic Design*, Mar. 2006, pp. 585-590.
- [8] E. M. Sentovich, K. J. Singh, L. Lavagno, C. Moon, R. Murai, A. Saldanha, H. Savoj, P. R. Stephan, R. K. Brayton, and A. Sangiovanni-Vincentelli, "SIS: a system for sequential circuit synthesis," Tech. Rep., UCB/ERL M92/41, U. C. Berkeley, May 1992.