Low-power Fanout Optimization Using MTCMOS and Multi-Vt Techniques

Behnam Amelifard
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-9481
amelifar@usc.edu

Farzan Fallah
Fujitsu Laboratories of America
Sunnyvale, CA
(408) 530-4544
farzan@fla.fujitsu.com

Massoud Pedarm
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-4458
pedram@ceng.usc.edu

ABSTRACT
This paper addresses the problem of low-power fanout optimization. We show that due to neglecting short-circuit current, previous analytical techniques proposed to optimize the area of a fanout tree may result in excessive power consumption. This shows, to achieve a low-power fanout tree, an accurate power consumption model should be used as the objective function. Moreover, we propose an efficient method to minimize the total power consumption of a fanout tree by using MTCMOS and Multi-Vt techniques. Experimental results show that depending on the activity factor of the circuit, the proposed technique can reduce the power consumption of the fanout tree 18% to 45%.

Categories and Subject Descriptors
B.6.3 [**Design Aids**]: Automatic synthesis, Optimization

General Terms
Algorithms, Design, Performance

Keywords
Low-power design, fanout optimization, fanout tree, buffer chain

1. INTRODUCTION
Fanout optimization, an operation performed in logic synthesis, is the problem of building an inverter tree topology between a source and some sinks and sizing the inverters in this topology so that the driving capacitance at the source is less than an upper bound and the timing constraints at sinks are met [1][2]. Different objective functions have been considered for the fanout optimization problem such as minimizing area [2][3][4], minimizing power consumption [3][5], and minimizing load on the source [6]. In this paper we minimize the total power consumption. Since both dynamic and leakage power dissipation of an inverter chain are proportional to its area, it has been widely accepted that power minimization of the fanout tree is equivalent to its area optimization [3][5]. In this paper, however, we show that due to short-circuit power dissipation, minimizing area does not necessarily result in a minimized power dissipation solution and the solution obtained from an area optimization technique may dissipate excessive short-circuit power.

To reduce both the active power and the standby leakage power, we utilize multi-Vt and MTCMOS techniques in a fanout tree. For doing this, at the first step, we use high-threshold voltage inverters in the fanout tree to reduce the leakage power consumption in both active and standby modes. If due to a high delay penalty high-Vt inverters cannot be used in the chain, then by using the MTCMOS technique, we try to reduce the leakage power consumption in the standby mode.

The remainder of the paper is organized as follows. After presenting the preliminaries in Section 2, in Section 3 the problem of low-power fanout optimization with one sink, i.e., an inverter chain, is formulated. Section 4 shows how a low-power fanout tree can be designed from the power-optimized inverter chains. Simulation results are given in Section 5, while Section 6 concludes the paper.

ISLPED'06, October 4–6, 2006, Tegernsee, Germany.

2. PRELIMINARIES
2.1 Delay Model
In logical effort, the delay of a gate with input capacitance C_{in}, which drives the load capacitance C_L, is modeled as,

$$D = \tau_0 (p + gh) \quad (1)$$

where τ_0 is a technology-dependent parameter, g is the logical effort of the gate, $h=C_L/C_{in}$ is the electrical effort and p is the parasitic delay of the gate. τ_0 is a constant and without losing generality it can be assumed to be one. For an inverter, the value of logical effort g equals one and p is the ratio of diffusion to input capacitance of the template inverter, denoted by p_0, i.e., $p_0=C_{diff,T}/C_{in,T}$. Since both input and diffusion capacitances of an inverter are scaled linearly when changing the size of inverter, for a scaled inverter the ratio of diffusion-to-gate capacitance remains constant. i.e.,

$$p_0 = C_{diff} / C_{in} \quad (2)$$

where C_{diff} is the diffusion capacitance at the output.

In a multi-Vt technology, the values of the logical effort and parasitic delay change as follows [5][8],

$$g_h = \frac{(V_{dd} - V_{t,0})^{\alpha}}{(V_{dd} - V_{t,h})^{\alpha}}, \quad p_h = p_0 \frac{(V_{dd} - V_{t,0})^{\alpha}}{(V_{dd} - V_{t,h})^{\alpha}} \quad (3)$$

where g_h and p_h are the logical effort and parasitic delay for an arbitrary $V_{t,h}$ threshold voltage, $V_{t,0}$ is threshold voltage of the template inverter and V_{dd} is the supply voltage and α is a technology parameter which is around 1.3 for short-channel devices.

In an MTCMOS circuit when the sleep transistor is ON, it can be modeled as a resistor whose resistance is inversely proportional to its width; hence, for an inverter [7],

$$g^r = 1, \quad p^r = p_0 \quad (4)$$

and,

$$g^f = (1 + \kappa \frac{C_{in}}{w}), \quad p^f = (1 + \kappa \frac{C_{in}}{w}) p_0 \quad (5)$$

where g^r and p^r (g^f and p^f) are the logical effort and parasitic delay for the rise (fall) delay, w is the width of the sleep transistor and κ is a constant which depends on the technology and the threshold voltage of the sleep transistor.

2.2 Power Dissipation Model

The power dissipation of an inverter has three components: dynamic power, short circuit power, and leakage power. The dynamic power is equal to,

$$P_{dyn} = \chi f V_{dd}^2 C \qquad 6$$

where χ is the switching activity of the inverter, f is the frequency, and C is the sum of the input gate capacitance and output diffusion capacitance of the inverter, i.e., $C=C_{diff}+C_{in}$. By using (2), (6) can be re-written as,

$$P_{dyn} = \chi f V_{dd}^2 (1+p_0) C_{in} \qquad 7$$

The second source of power dissipation in digital circuit is due to the short-circuit current. Several techniques have been proposed to address the problem of short circuit power estimation [10], but due to their complexity, they may not be very useful during a gate-level optimization process. In this paper, by observing the fact that short-circuit power dissipation of an inverter is a linear function of its size and input transition time [10], and also the fact that input transition time itself can be approximated as a linear function of the electrical effort of the previous stage in the chain, the short-circuit power dissipation of the i^{th} inverter in a chain is modeled as,

$$P_{sc} = K_{sc} h_{i-1} C_{in} \qquad 8$$

where K_{sc} is a technology-dependent parameter, h_{i-1} is the electrical effort of the i-1^{th} inverter and C_{in} is the input capacitance of the i^{th} inverter. Transistor level SPICE simulations show this technique, despite its simplicity, is accurate enough to be used in a gate-level optimization technique.

The third source of power dissipation is leakage. In current technologies, the major components of leakage current are subthreshold and gate-tunneling currents [9]. The total leakage power dissipation of an inverter can be modeled as

$$P_{leak} = (K_{sub} + K_{ox}) C_{in} \qquad 9$$

where K_{sub} and K_{ox} are technology parameters which depend on the effective channel length, oxide thickness, temperature, and supply voltage. Moreover, K_{sub} is also a function of the threshold voltage.

Having had different components of the power consumption, the total power dissipation of inverter i in a chain can be expressed as,

$$P = P_{dyn} + P_{leak} + P_{sc} = C_i (K_{dyn} + K_{sub} + K_{ox} + K_{sc} h_{i-1}) \qquad 10$$

where C_i is the input capacitance of inverter i and $K_{dyn} = \chi f V_{dd}^2 (1+p_0)$.

3. LOW-POWER INVERTER CHAINS

In our approach, to construct the low-power fanout tree topology and size the inverters in the tree, the problem is decomposed into sub-problems in the forms of inverter chains, and each sub-problem is separately solved for each sink. The solutions to the sub-problems are then merged to find the solution to the main problem. So, in this section we formulate the problem of minimizing power dissipation of an inverter chain under timing and input capacitance constraints, i.e.,

$$\begin{cases} \min \quad Power \\ s.t \quad (1) \quad Delay \le T \\ \qquad (2) \quad C_1 \le C_{\max} \end{cases} \qquad 11$$

where T is the timing constraint on the sink, C_1 is the input capacitance of the first inverter, and $C_{\max}$ is the maximum tolerable load on the source.

Since both dynamic and leakage power dissipation of an inverter are proportional to its size, if short-circuit power consumption is ignored, the problem of finding the minimum power consumption inverter chain is the same as finding the minimum area inverter chain. In [2] the problem of minimizing the area of an inverter chain given a constraint on the delay of the chain and a constraint on the load of the source has been formulated using logical effort. By using Lagrangian relaxation technique [11], it can be shown that when the input capacitance constraint of the fanout chain is "loose", i.e., in the optimal solution $C_1<C_{\max}$, such a formulation results in a solution in which the following relation holds among the sequence of electrical efforts,

$$h_{i+1} = h_i (h_i - h_{i-1} + 1) \qquad 12$$

where h_0 is defined as 0 and h_1 can be found from solving this polynomial equation,

$$\sum_{i=1}^{n} h_i = T - n p_0 \qquad 13$$

It can be verified that in (12),

$$h_{i+1} > h_1^{2^i} \qquad 14$$

From (14), it is easy to see the value of h_i's grow exponentially and based on (8), the short circuit power dissipation of the inverters grows very fast.

Based on the above fact, we give a precise objective function for minimizing the total power dissipation of an inverter chain. In order to simplify the equation, without losing generality, we assume the driver and load of the chain are fixed-sized inverters. The driver is called 0^{th} inverter, while the load is called n+1^{th} inverter. Hence, the power dissipation of the inverter chain (i.e., the objective function of (13)) can be modeled as,

$$Power = \tilde{C}_L \sum_{i=1}^{n} \frac{1 + k_\varphi h_{i-1}}{\prod_{j=i}^{n} h_j} + \tilde{C}_L k_\varphi h_n \qquad 15$$

where $\tilde{C}_L = C_L (K_{dyn} + K_{sub} + K_{ox})$ and $k_\varphi = K_{sc} / (K_{dyn} + K_{sub} + K_{ox})$.

Since the size of the load is fixed, the dynamic and leakage power dissipation of the load inverter are constant; however, the short-circuit power consumption of this inverter is a function of the electrical effort of the last stage in the chain. Therefore, we have included the short-circuit power dissipation of the load into the objective function as the last term.

The constraints of (11) in logical effort notion are similar to those in [2], i.e., the delay constraint can be expressed as,

$$Delay = \sum_{i=1}^{n} (p_0 + h_i) \le T \qquad 16$$

while the input capacitance constraint can be written as,

$$C_1 = C_L / \prod_{i=1}^{n} h_i \le C_{\max} \qquad 17$$

Therefore, problem (11) is a minimization of a posynomial function with posynomial inequality constraints that can be easily solved in polynomial time [11]. Notice that to find the minimum inverter chain, the abovementioned mathematical program should be solved for different values of n. The upper and lower bounds of n are similar to those in [2] and [5]; however, based on the polarity of the sink node, only even or odd numbers of inverters between these bounds are considered when searching for the optimum solution [5].

Although by solving the above mathematical problem the total power consumption in the active mode is reduced, the standby leakage power consumption is weakly decreased. Many techniques have been proposed to reduce the standby leakage power, while maintaining high performance in the active mode. A combination of MTCMOS [12] and multi-Vt techniques has been shown to be very effective in reducing the standby leakage power dissipation [13]. In this scheme, by using high-Vt transistors in the non-critical paths their active and standby leakage power consumption is reduced. For the gates on the critical path low-Vt transistors are used to achieve the high performance but MTCMOS technique is applied to these gates to reduce the standby subthreshold current. We use a similar technique to suppress standby-mode leakage power consumption of our fanout trees.

Notice if the threshold voltage of all inverters in the inverter chain increases to $V_{t,h}$, (16) should be modified to,

$$Delay = g_h\left(\sum_{i=1}^{n} p_0 + h_i\right) \le T \qquad 18$$

where g_h is obtained from (3). Moreover, due to exponential reduction of the subthreshold current, $\tilde{C}_L$ in (15) should be changed to,

$$\tilde{C}_L = C_L(K_{dyn} + K_{sub}\exp(V_{t,0} - V_{t,h}) + K_{ox}) \qquad 19$$

In the following sub-section we show how to modify this mathematical program for the case that MTCMOS technique is used.

3.1 Low-power MTCMOS inverter chain

Figure 1 shows three different ways to build an MTCMOS inverter chain. Although the techniques shown in Figure 1.b and 1.c seem to be more area-efficient, they are not compatible with the merge transformations that we are going to use to build the fanout tree from the inverter chains. Hence, in the remainder of this paper we assume the structure of Figure 1.a for the MTCMOS inverter chain.

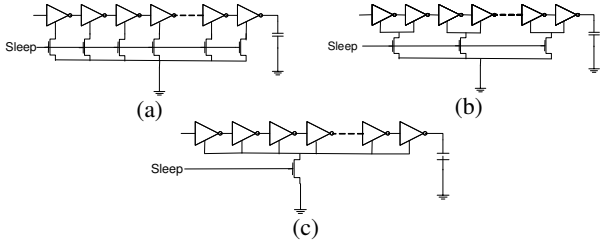

Figure 1. Different scenarios for implementing MTCMOS inverter chains

By using (3) and (4), the rise and fall delays of the inverter chain from sink to source, d^r and d^f, can be expressed as functions of the electrical efforts of the inverters in the chain. Since d^f and d^r are not equal, we define the delay of the inverter to be the maximum of the fall and rise delays. To minimize the total power consumption of the MTCMOS inverter chain, the delay constraint in (11) must be modified as,

$$Delay = \max\{d^f, d^r\} \le T \qquad 20$$

The objective function of (11) also needs to be modified to model the gate-tunneling current of the sleep transistors in the active mode. Thus, (15) should be modified as,

$$Power = \sum_{i=1}^{n}\left(\tilde{C}_L\frac{1+k_\phi h_{i-1}}{\prod_{j=i}^{n} h_j} + \tilde{K}_{ox}w_i\right) + \tilde{C}_L k_\phi h_n \qquad 21$$

On the other hand, in practice there is a budget for the total size of sleep transistors in the chain. So, in the mathematical program (11) a third constraint should be added to limit the total size of the sleep transistors. With these modifications, the low-power MTCMOS inverter chain optimization can be expressed as,

$$\begin{cases} \min \quad Power \\ s.t \quad (1) \quad Delay \le T \\ \qquad (2) \quad C_1 \le C_{\max} \\ \qquad (3) \quad Sleep\ Transistor\ Size \le W_0 \end{cases} \qquad 22$$

where W_0 is the budget on the size of the sleep transistor.

4. BUILDING A FANOUT TREE

In this section we show how to build a fanout tree with more than one sink. The typical fanout tree we want to build is shown in Figure 2, where the first m sinks are not on the critical path and hence the corresponding tree can be designed using high-Vt devices, while the next k sinks are on the critical path and hence

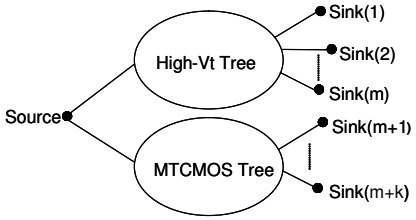

Figure 2. MTCMOS and High-Vt fanout trees

MTCMOS technique should be used for them. [6] introduced two transformations that could be performed on a fanout tree, namely merging and splitting and it showed that these transformations maintain the same area, delay, and input capacitance. We have extended the merging and splitting techniques, as shown in Figure 3, to handle MTCMOS fanout trees.

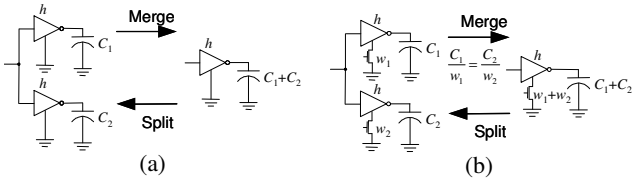

Figure 3: Split and merge transformations (a) original transformations (b) extended transformation for MTCMOS inverters

It can be shown that split/merge transformations and their extended versions applied to a fanout tree preserve the delay, input capacitance, and power dissipation values of the tree. By using these transformations, any fanout optimization problem with N sink nodes can be converted to N inverter chain optimization problems, whose respective power dissipation will be the same. To apply such transformations two issues should be addressed. The first issue is the input capacitance allocation to different chains in a decomposed fanout tree. It was shown in [2] that this problem is NP-complete and a heuristic, which we use in this paper, has been developed to allocate the input capacitance. The second issue is finding the inverter chains for which the high-Vt can be used and allocating the sleep transistor width among other chains. To address these problems, after the input capacitance allocation, we determine which inverter chains can work with the high-Vt. For an inverter chain whose allocated input capacitance is $C_{\max,i}$, the load capacitance is $C_{L,i}$, and the required time at the sink is T_i, this can be done by checking if the following relation holds for at least one even or odd n depending on the polarity of the sink,

$$S_{i,n} = T_i - \left(ng_h(C_{L,i}/C_{\max,i})^{1/n} + ng_h p_0\right) \ge 0 \qquad 23$$

Note the value inside the parentheses is the minimum delay of a high-Vt inverter chain with n inverters. In this case, we say the sink is *non-critical* and the corresponding chain can work with high-Vt devices without violating the timing constraint of the sink. On the other hand, if (23) does not hold for any n, the sink is called *critical*. In this case an MTCMOS inverter chain should be used in the corresponding chain. However, it should be noticed the circuit at the end of the MTCMOS chain needs to be in standby mode whenever the chain is in standby mode; otherwise, very high short circuit current flows through the circuit. If the critical sink does not drive an MTCMOS gate, only low-Vt inverters (without sleep transistors) are used in the corresponding chain.

To allocate the sleep transistor width to different MTCMOS inverter chains, we use the following heuristic. From the set of constraints of (22) it can be seen that in an MTCMOS inverter chain with n inverters, the power cost is a decreasing function of

Table 1: Comparison between *MinPowerFO* and *MinAreaFO* for a few inverter chains

Circuit	Circuit Specification				Fanout Type	ΔA (%)	ΔP (%)				
	C_{in}	C_{out}	T	P			$\acute{\alpha}$=10%	$\acute{\alpha}$=30%	$\acute{\alpha}$=50%	$\acute{\alpha}$=70%	$\acute{\alpha}$=90%
C1	1	100	23	+	high-Vt	54.11	53.78	29.55	19.75	14.44	11.12
C2	2	135	20	+	high-Vt	154.81	44.57	30.20	25.73	23.56	22.27
C3	2	100	21	-	high-Vt	152.11	43.34	33.97	31.47	30.32	29.65
C4	2	100	17	-	MTCMOS	196.81	57.59	34.05	24.82	19.88	16.81
C5	2	70	15	+	MTCMOS	166.49	48.14	26.07	18.67	14.95	12.72
C6	4	550	20	-	MTCMOS	204.94	49.18	27.37	20.01	16.32	14.09
Average						**154.87**	**49.43**	**30.20**	**23.41**	**19.91**	**17.78**

Table 2: Comparison between *MinPowerFO* and *MinAreaFO* for a few fanout optimization problems

Circuit	Circuit Specification						ΔA (%)	ΔP (%)				
	C_{in}	$C_{out,max}$	T_{min}	T_{max}	P+	P-		$\acute{\alpha}$=10%	$\acute{\alpha}$=30%	$\acute{\alpha}$=50%	$\acute{\alpha}$=70%	$\acute{\alpha}$=90%
T1	25	550	15	23	2	3	87.33	48.35	38.79	29.23	19.67	10.12
T2	20	1100	14	50	3	3	189.34	39.95	34.52	29.09	23.66	18.23
T3	17	135	40	90	4	1	163.44	52.34	46.38	40.43	34.47	28.52
T4	14	550	9	32	1	5	209.43	34.55	30.74	26.94	23.14	19.34
T5	10	70	12	52	2	6	121.22	57.77	47.18	36.60	26.01	15.43
T6	14	100	12	21	7	3	178.91	39.44	34.88	30.32	25.76	21.21
Average							**158.27**	**45.40**	**38.74**	**32.10**	**25.45**	**18.80**

the available slack defined as (23). Since using sleep transistors in the chain incurs delay overhead and reduces the available slack, we allocate the sleep transistor budget in a way that a larger transistor width is assigned to a chain with less slack, i.e.,

$$W_i = \frac{1/S_i}{\sum_{j=1}^{k_0} 1/S_j} W_{tot} \qquad 24$$

where $k_0 \le k$ is the number of MTCMOS chains, W_i ($1 \le i \le k_0$) is the width of the sleep transistor allocated to the i^{th} chain, $S_i = \max_n\{S_{i,n}\}$ is the slack of the i^{th} chain, and W_{tot} is the total budget for sleep transistor width.

5. SIMULATION RESULTS

The proposed technique in Sections 3 and 4, which we call *MinPowerFO*, has been developed in MATLAB optimization toolbox. To study the efficiency of our technique in reducing the power consumption of the fanout trees, we performed a set of experiments and compared the results of *MinPowerFO* with the results of *MinAreaFO*, which minimizes the area of the fanout tree [2]. The technology parameters we used in these sets of experiments are based on a 65nm technology node [14] and have been obtained by transistor level simulation of devices. In this technology, the supply voltage is 1.0V and the values of low and high threshold voltages are 0.2V and 0.3V, respectively. Moreover, the oxide thickness of both NMOS and PMOS transistors is $17A^0$. Simulation results for a few random problems, in the form of inverter chains, are shown in Table 1. In this table, C_{in} is the maximum allowed capacitance at the input of the inverter chain, C_{out} is the sink load, T is the required time at the sink, and P is the polarity of the sink. In each case, the constraint on the size of the sleep transistor has been assumed to be half of the total size of inverters in the minimum area solution. The power dissipation of circuits using these techniques has been compared for different activity factors $\acute{\alpha}$ (i.e., the percentage of the time the circuit is in the active mode). In this table, ΔA is the area increase of the *MinPowerFO* compared to that of *MinAreaFO*, while ΔP is the power reduction of the *MinPowerFO* technique compared to that of *MinAreaFO*.

In the second set of experiments, the fanout optimization problem is solved for a group of arbitrary problems. Each problem states one source and multiple sinks with capacitive load, required time, and polarity constraints specified for each sink. The specification of each circuit, including the maximum input capacitance (C_{in}), the number of sinks with positive and negative polarities (p+ and p-), the maximum and minimum required times of all sinks (T_{max} and T_{min}), and the maximum sink capacitances ($C_{out,\max}$), are shown in Table 2. From the table, one can see depending on the activity factor of the fanout circuit, the average power reduction ranges from 18% to 45%.

6. CONCLUSION

In this paper we showed that the fanout optimization with area and power objective functions are not the same and a fanout tree optimized for area may dissipate excessive short-circuit power. By modeling all components of power dissipation, we formulated the fanout optimization problem as a geometric program for a circuit with one sink. To reduce standby power consumption, we proposed using multi-Vt and MTCMOS fanout trees, where high-Vt fanout tree is constructed for the sinks on the non-critical paths, while the MTCMOS fanout tree is constructed for the sinks on the critical paths. Experimental results show the proposed technique is very effective in reducing the total power consumption of fanout trees for various activity factors.

7. REFERENCES

[1] Salek, A., *et al.* Hierarchical buffered routing tree generation. *IEEE Trans. on CAD,* 21, (May 2002), 554-567.

[2] Rezvani, P., *et al.* A fanout optimization algorithm based on the effort delay model. *IEEE Trans. on CAD,* 22, (Dec. 2003), 1671-1677.

[3] Zhou, D., Liu, X. Minimization of chip size and power consumption of high-speed VLSI buffers. In *Proc. ISPD*, 1997, 186-191.

[4] Singh, K. J., *et al.* A heuristic algorithm for the fanout problem. In *Proc. DAC*, 1990, 357–360.

[5] Amelifard, B., *et al*., Low-power fanout optimization using multiple threshold voltage inverters. In *Proc. ISLPED,* 2005, 95-98.

[6] Kung, D. S. A fast fanout optimization algorithm for near-continuous buffer libraries. In *Proc. DAC*., 1998, 352-355.

[7] Sutherland, I., *et al*. *Logical Effort: Designing Fast CMOS Circuits*. Morgan Kaufmann, San Fransisco, CA, 1999.

[8] Sakurai, T., *et al.* A simple MOSFET model for circuit analysis. *IEEE Trans. Electron Device*, 38 (Apr. 1991), 887-894.

[9] De, V., *et al.* Techniques for leakage power reduction. in *Design of High-Performance Microprocessor Circuit, Circuits*, Chandrakasan, A., *et al.*, IEEE , Piscataway , NJ, 2001.

[10] Pedram, M. Power minimization in IC design: Principles and applications. *ACM Trans. on Design Automation of Electronic Systems*, 1,1 (Jan. 1996), 3-56.

[11] Gill, P. E., *et al. Practical Optimization*, Academic Press, New York, 1981.

[12] Anis, M., *et al.* Dynamic and leakage power reduction in MTCMOS circuits using an automated efficient gate clustering Technique. In *Proc. DAC*, 2002, 480-485.

[13] Usami, K., *et al.* Automated selective multi-threshold design for ultra-low standby applications. In *Proc. ISLPED*, 2002, 202-206.

[14] http://www.eas.asu.edu/~ptm/