Simultaneous Fine-grain Sleep Transistor Placement and Sizing for Leakage Optimization

Wang Yu, Lin Hai, Yang Huazhong, Luo Rong, Wang Hui
EE Department, Tsinghua University, Beijing, P.R. China
{wangyuu99, linhai99}@mails.tsinghua.edu.cn {yanghz, luorong, wangh}@tsinghua.edu.cn

Abstract*

With the growing scaling of technology, leakage power dissipation has become a critical issue of VLSI circuits and systems designs. Multi-threshold CMOS leads to about 10X leakage reduction in circuit standby mode. In this paper, we reduce leakage current through fine-grain sleep transistor (ST) insertion which makes it easier to guarantee circuit functionality at high speed and improves circuit noise margins [1]. We model the leakage current reduction problem as a mixed-integer linear programming (MLP) problem in order to simultaneously choose where to add the sleep transistors and the sleep transistors' sizes optimally. The model is solved with both continuous (MLP-C) and discrete (MLP-D) sleep transistor size constraints. Furthermore a method to speed up MLP-D model is introduced. Because of the better circuit slack utilization, our experimental results show that the MLP-C model can achieve 79.75%, 93.56%, 94.99% leakage saving when the circuit slow down is 0%, 3%, 5% respectively. The MLP-C model also achieves on average 74.79% less area penalty compared to the conventional fixed slowdown method when the circuit slowdown is 7%. The MLP-D model can achieve similar leakage saving compared to the MLP-C model. The MLP-CtoD method can speed up the MLP-D model 30X times with almost no difference in leakage reduction.

1. Introduction

With technology stepping into the submicron region, power issues have already reached a bottleneck in the design of portable and wireless electronic systems. The total power dissipation consists of dynamic power, short circuit power and leakage power, thus can be expressed as:

$$P_{total} = P_{dynamic} + P_{Leakage} + P_{shortcircuit}$$
$$= \sum_{i=1}^{N} (\frac{1}{2}\alpha_i f C_i V_{DD}^2 + I_{l,i} V_{DD} + \alpha_i f Q_{short,i} V_{DD})$$

Where, f is the operation frequency, V_{DD} is the supply voltage, and N is the number of gates. α_i, C_i, $I_{l,i}$, and $Q_{short,i}$ are the transition probability, load capacitance, leakage current, and short circuit charge of the i-th gate, respectively. The behavior of the short circuit power dissipation remains at around 10% of the total power dissipation [2]. With the development of the fabrication technology, leakage power dissipation has become comparable to switching power dissipation [3]. At the 90nm technology node, leakage power may make up 42% of total power [4].

Inevitably, techniques are necessary for reducing the increasing leakage power. These leakage control methods can be broadly categorized into two main categories: process level and circuit level techniques [5]. At the process level, leakage reduction can be achieved by controlling the dimensions (length, oxide thickness, junction depth, etc.) and doping profile in transistors. Here we talk about circuit design techniques, namely, adapt body bias [6], DVTS [7], input vector control [8], dual-V_t assignment [9] [10] and Multi-Threshold CMOS (ST insertion).

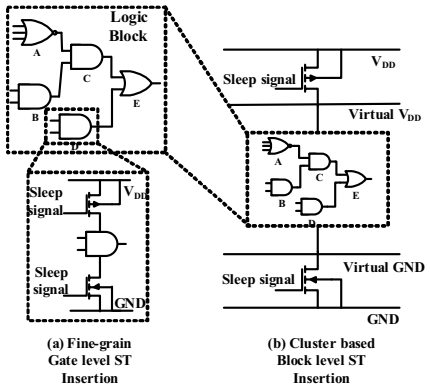

Figure 1. Fine-grain vs Cluster-based ST Insertion

Among these, Multi-Threshold CMOS (MTCMOS) is a valuable technique for reducing leakage power in the circuit standby mode. MTCMOS technique is essentially placing a sleep transistor between the gates and the power/ground (P/G) net in order to put them into sleep mode when the circuit is standby. The most popular MTCMOS technique is gating the power of sizable blocks using large sleep transistors which assumes that all gates have a fixed slowdown [11] [12] [13] [14] [15]. However,

* Project supported by National 863 project of China (No. 2004AA1Z10 50, No. 2005AA1Z1230) and National Natural Science Foundation of China (No.90207001, No. 60506010).

in recent years the use of sleep devices in the gate level [1] [16] (Figure 1. (a)), which has some advantages over the block level design (Figure 1. (b)), is raising people's concern.

The existing literature on MTCMOS circuits [11-15] present cluster based methods for sleep transistor insertion and sizing. [11] first gives out a mutual exclusion method to reduces the area penalty. [12] [13] present several heuristic techniques for efficient gate clustering and try to mitigate the ground problem by introducing additional power penalty. In [14] [15], a Distributed Sleep Transistor Network (DSTN) approach is proposed which connects all the sleep devices to reduce the area penalty.

Although cluster based methods reduce the area penalty, they induce large ground bounce in the P/G network which has adverse effects on circuit speed and noise immunity [16]. What is more, the sleep transistor's size is determined by the worst case current of the clustering block. However identifying the worst case is quite difficult without comprehensive simulation [11]. Thus it is harder to guarantee circuit functionality for large blocks with only one sleep transistors [1].

The fine-grain MTCMOS design methodology is discussed in [1] [16]. In [1], a fine-grain MTCMOS design methodology and several design rules are proposed. The authors also make a comparison between local and global devices. [16] presents a selectively sleep transistor insertion methodology with better utilization of circuit slack. They first select where to put the sleep transistors by a heuristic method and then solve an LP model to get optimal sleep transistor size. Although the second step can give out an optimal sizing result, the first step may lead to a local optimal point. Furthermore, in the second step they assume the sleep transistor size is continuous which is not the real case.

This paper presents three contributions to leakage reduction through fine grain sleep transistor insertion. First, we give out our newly developed leakage current model and delay model of a single gate, which are much simpler and more exact than the models in traditional fine grain sleep transistor insertion strategy. Secondly, a formal mixed-integer linear model of the leakage current reduction problem provides the designers the relations between leakage current and circuit constraints, and makes it possible to decide where to put the sleep transistors and the sizing of the sleep transistor simultaneously and optimally. The model can be solved when the circuit slowdown is not long enough to perform the conventional fixed slowdown based sleep transistor insertion. Even if the circuit performance is not influenced, our model can achieve an impressive leakage saving. Furthermore, if the conventional fixed slowdown method can be performed, our method still leads to a larger leakage saving and a much smaller total sleep transistor size. Finally, we show that the model can be solved with discrete sleep transistor size constraint which is much practical in the real life.

The paper is organized as follows. In Section 2, we give out our leakage current and delay model for a single gate. The detail of MLP model construction is proposed in Section 3. The implementation and experimental results are presented and analyzed in Section 4. In Section 5, we conclude this paper.

2. Preliminaries

First we give out our definition of leakage current and delay model. A cell-based design flow with a given cell library is used. We assumed sleep transistors with variable size which is decided by the process technology are used in our fine-grain sleep transistor insertion design. A combinational circuit is represented by a directed acyclic graph (DAG) $G = (V, G)$. A vertex $v \in V$ represents a CMOS gate from the given library, while an edge $(i, j) \in E$, $i, j \in V$ represents a connection from vertex i to vertex j. We define $I_l(v)$, $D(v)$ as the leakage current and delay of gate v respectively.

2.1 Leakage current model

The average leakage power dissipation $P_{leakage}(G)$ of the circuit can be expressed as the product of the average leakage current and power supply voltage.

$$P_{leakage}(G) = V_{DD} \times I(G) \tag{1}$$

The circuit average leakage current can be calculated as the sum of the individual gate's average leakage current. As we all know the leakage current of a CMOS gate is decided by its structure and input pattern. We define the probability of a gate v under input pattern IN as $PB(v, IN)$. Thus the leakage current of a gate v in the circuit can be expressed as:

$$I_l(v) = \sum_{IN} I_l(v, IN) \times PB(v, IN) \tag{2}$$

$I_l(v, IN)$ is the leakage current of gate v under input pattern IN.

In our fine-grain sleep transistor insertion design the leakage of a gate in the circuit is also determined by whether the sleep transistor is inserted to this gate or not. For the gates without sleep transistor, we create a leakage lookup table for $I_l(v, IN)$ by simulating all the gates in the standard cell library under all possible input patterns. Thus the leakage current $I_l^{w/o}(v)$ can be expressed as:

$$I_l^{w/o}(v) = \sum_{IN} I_l(v, IN) \times PB(v, IN) \tag{3}$$

On the other hand, the subthreshold leakage current $I_l^{ST}(v)$ with sleep transistors are given by [17]:

$$I_l^{ST}(v) = \mu_n C_{ox} (W/L)_v e^{1.8} V_T^2 e^{\frac{V_{gs} - V_{THhigh}}{nV_T}} (1 - e^{\frac{-V_{ds}}{V_T}}) \tag{4}$$

where μ_n is the N-mobility, C_{ox} is the oxide capacitance, V_{THhigh} is the high threshold voltage, V_T is the

thermal voltage, n is the sub-threshold swing parameter, $(W/L)_v$ represents the size of the sleep transistor inserted to gate v. V_{ds} is decided by $(W/L)_v$, thus the relationship between $I_l^{ST}(v)$ and $(W/L)_v$ is complicated. Here we present our simplified leakage current $I_l^{ST}(v)$ model:

$$I_l^{ST} = A(v) + B(v) \times (W/L)_v \quad (5)$$

where $A(v)$, $B(v)$ are constants and are decided by the gate type.

Consider two standard cells: a two-input NAND and a four-input AND with fixed structure and size in the given library. We add high threshold voltage sleep transistor to the gates, and compare the leakage current of the gates with different sleep transistor sizes. Refer to our model, we can give out the $A(v)$, $B(v)$ of the NAND2 and AND4 respectively: 1.31774, 0.01128; 1.67104, 0.01514.

Table 1 Leakage current with different sleep transistor sizes in NAND2 and AND4

	Leakage current in NAND2 (pA)			Leakage current in AND4 (pA)		
	Hspice	Our model	Error	Hspice	Our model	Error
w/o ST	18.8938	N/A	N/A	22.67189	N/A	N/A
(W/L) =1	1.333825	1.32902	-0.36%	1.692819	1.68618	-0.39%
(W/L) =2	1.33615	1.3403	0.31%	1.695831	1.70132	0.31%
(W/L) =4	1.3618	1.36286	0.08%	1.730075	1.7316	0.09%
(W/L) =8	1.407875	1.40798	<0.01%	1.791681	1.79216	0.03%
(W/L) =16	1.4988	1.49822	-0.04%	1.914263	1.91328	-0.05%

Notice $I_l^{ST}(v)$ is still sensitive to the input patterns, the data shown in Table 1 is the average leakage current for which we assume all the input patterns have same probability. As shown in Table 1, the error is less than 0.39% and the original leakage current without sleep transistor is at least 15X larger than $I_l^{ST}(v)$. We estimate every $A(v)$, $B(v)$ for all the standard cells and find out, on average, $B(v)$'s are around 1% of $A(v)$, thus the variation range of $I_l^{ST}(v)$ is about 15% of A(v).

Thus we use lookup table to model the leakage current of gates with no sleep transistor, and linear equations to model the leakage current of gates with sleep transistors. As we can see, our leakage current model for a single gate is very simple and accurate.

2.2 Delay model

In our fine-grain sleep transistor insertion design, we have to insert sleep transistors to the original gates in the given library. As shown in [18], the delay of the gate is influenced by the sleep transistor insertion. The load dependent delay $D^{w/o}(v)$ of gate v without sleep transistors can be expressed as:

$$D^{w/o}(v) = \frac{KC_L V_{DD}}{(V_{DD} - V_{THlow})^{\alpha}} \quad (6)$$

where C_L, V_{THlow}, α, K are the load capacitance at the gate output, the low threshold voltage, the velocity saturation index and the proportionality constant respectively. The propagation delay $D^{ST}(v)$ with the presence of sleep transistors of gate v can be expressed as:

$$D^{ST}(v) = \frac{KC_L V_{DD}}{(V_{DD} - 2V_x - V_{THlow})^{\alpha}} \quad (7)$$

where V_x is the V_{ds} of the sleep transistor, that is to say the voltage drop from V_{DD} to the virtual V_{DD} as shown in Figure 1. We define the difference of $D^{w/o}(v)$ and $D^{ST}(v)$ as $\Delta D(v)$:

$$\Delta D(v) = D^{ST}(v) - D^{w/o}(v) \quad (8)$$

Refer to equation (6) (7) (8), we can get an approximate $\Delta D(v)$ with neglectable difference using Taylor series expansion:

$$\begin{aligned}
\Delta D(v) &= D^{ST}(v) - D^{w/o}(v) = \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}}\right)^{-\alpha} - 1\right) D^{w/o}(v) \\
&\overset{Taylor}{=} \left(\left(1 + \alpha\frac{2V_x}{V_{DD} - V_{THlow}} + \alpha(\alpha+1)\left(\frac{2V_x}{V_{DD} - V_{THlow}}\right)^2 + \cdots\right) - 1\right) D^{w/o}(v) \\
&= \left(\alpha\frac{2V_x}{V_{DD} - V_{THlow}} + \alpha(\alpha+1)\left(\frac{2V_x}{V_{DD} - V_{THlow}}\right)^2\right) \times D^{w/o}(v) \\
&= (\Gamma \times V_x + \frac{\alpha+1}{\alpha}\Gamma^2 \times V_x^2) \times D^{w/o}(v)
\end{aligned} \quad (9)$$

We use a constant $\Gamma = 2\alpha/(V_{DD} - V_{THlow})$ to simplify the equation (9) since V_{THlow}, α, V_{DD} are all technology dependent constant. We suppose $I_{ON}(v)$ is the current flowing through the sleep transistor in the gate v during the active mode, and can be expressed as [16]:

$$\begin{aligned}
I_{ON}(v) &= \mu_n C_{ox}(W/L)_v\left((V_{DD} - V_{THhigh})V_x - \frac{V_x^2}{2}\right) \\
&= \mu_n C_{ox}(W/L)_v(V_{DD} - V_{THhigh})V_x
\end{aligned} \quad (10)$$

Thus the voltage drop V_x in gate v due to sleep transistor insertion can be expressed as:

$$\begin{aligned}
V_x &= \frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_v} \\
&= \Psi(v) \times (W/L)_v^{-1}
\end{aligned} \quad (11)$$

Here we use $\Psi(v)$ to simplify the equation. From above we can get $\Delta D(v)$ as:

$$\Delta D(v) = \left(\Gamma\Psi(v) \times (W/L)_v^{-1} + \frac{\alpha+1}{\alpha}\Gamma^2\Psi(v)^2 \times (W/L)_v^{-2}\right) \times D^{w/o}(v) \quad (12)$$

From equation (10), we can see V_x is slightly larger than the actual value, thus $\Delta D(v)$ is a little bit larger than the actual value which make our model more feasible to maintain the timing constraints of the circuit.

3. MLP model construction

We now construct an MLP model for simultaneous placement and sizing of sleep transistor. There are only two conditions to each gate v: with or without sleep transistor. We therefore define a binary variable $ST(v)$ to represent gate v's sleep transistor condition, where $ST(v)$ = 1 for gate v with sleep transistor inserted and $ST(v)$ = 0 for gate v without sleep transistor.

3.1 Objective function

We use equation (3) as basis to construct the objective function. Note that the leakage current of gate v $I_l(v)$ can be written as:

$$I_l(v) = I_l^{w/o}(v) \times (1 - ST(v)) + I_l^{ST} \times ST(v) \quad (13)$$

Therefore we represent the total leakage current by:

$$I(G) = \sum_{v \in V} \left(I_l^{w/o}(v) \times (1 - ST(v)) + I_l^{ST} \times ST(v) \right) \quad (14)$$

Refer to equation (3) and (5), we can hence replace equation (13) with:

$$I(G) = \sum_{v \in V} \left(\begin{array}{l} \left(\sum_{IN} I_l(v, IN) \times PB(v, IN) \right) \times (1 - ST(v)) \\ + \left(A(v) + B(v) \times (W/L)_v \right) \times ST(v) \end{array} \right) \quad (15)$$

where $ST(v)$ and $(W/L)_v$ are variables which decide where to put sleep transistor and how to size the sleep transistor respectively.

3.2 Timing constraints

First we consider the primary input (*PI*s) and output (*PO*s) gates of the circuit. The arrival time t_a of all the PIs are set to zero, while the required time of all the *PO*s are less than the overall circuit delay T_{req}.

$$t_a(m) = 0 \qquad m \in PI \quad (16)$$

$$t_a(n) + D(n) \le T_{req} \qquad n \in PO \quad (17)$$

Then we notice that the sum of gate v's arrival time and its delay must be smaller than the arrival time of gate v's fanout gates. That is to say, $\forall (i, j) \in E$, $i, j \in V$, we can derive the constraint as:

$$t_a(i) + D(i) \le t_a(j) \quad (18)$$

As we have already induced the definition of $ST(v)$, we can rewrite the delay of gate v as:

$$\begin{array}{l} D(v) = D^{w/o}(v) + \Delta D(v) \times ST(v) \\ = D^{w/o}(v) + \Gamma \Psi(v) D^{w/o}(v) \times (W/L)_v^{-1} \times ST(v) \\ + \dfrac{\alpha + 1}{\alpha} \Gamma^2 \Psi(v)^2 D^{w/o}(v) \times (W/L)_v^{-2} \times ST(v) \end{array} \quad (19)$$

3.3 Linearization constraints

First we define variable $W(v)$ for each gate, where $WL(v) = (W/L)_v = 2^{W(v)}$, $WLN(v) = (W/L)_v^{-1} = 2^{-W(v)}$, $WLN2(v) = (W/L)_v^{-2} = 2^{-2W(v)}$, and $W(v) \in [0, W_{\max}]$. We use a similar piecewise linear approximation technique in [19] to linearize these exponential expressions with inequalities:

$WL(v) \ge 2^k W(v) + (1 - k) \times 2^k$, where $k = 0, 1, \ldots, W_{\max}$

$WLN(v) \ge -2^k W(v) + (1 - k) \times 2^k$, where $k = -W_{\max}, -W_{\max}+1, \ldots, 0$

$WLN2(v) \ge -2^k \times 2W(v) + (1 - k) \times 2^k$, where $k = -2W_{\max}, -2(W_{\max}+1), \ldots, 0$

Secondly, in equation (15) and (19), a set of items to be linearized is :

$$WS(v) = (W/L)_v \times ST(v) = WL(v) \times ST(v)$$

$$WSN(v) = (W/L)_v^{-1} \times ST(v) = WLN(v) \times ST(v)$$

$$WSN2(v) = (W/L)_v^{-2} \times ST(v) = WLN2(v) \times ST(v)$$

where $WL(v)$, $WLN(v)$, $WLN2(v)$ are real variables while $ST(v)$ is binary. As in [19], $C = B \times A$, where A is a binary variable and M is an upper bound of B, is linearized as follows:

$$0 \le C \le B$$
$$C \le M \times A$$
$$C \ge B - M(1 - A)$$

Since $W(v) \in [0, W_{\max}]$, $WL(v)$, $WLN(v)$ and $WLN2(v)$ all have their upper bound. Hereto, we end up our MLP model for leakage minimization. And the general form of our MLP model is given out in Figure 2.

Minimize:

$$I(G) = \sum_{v \in V} \left(\begin{array}{l} \left(\sum_{IN} I_l(v, IN) \times PB(v, IN) \right) \times (1 - ST(v)) \\ + A(v) \times ST(v) + B(v) \times WS(v) \end{array} \right)$$

Subject to:

{Timing constraints}

$$t_a(m) = 0 \qquad m \in PI$$

$$t_a(n) + D(n) \le T_{req} \qquad n \in PO$$

$$t_a(i) + D(i) \le t_a(j), \qquad \forall (i, j) \in E,\ i, j \in V$$

$$D(v) = D^{w/o}(v) + \Gamma \Psi(v) D^{w/o}(v) \times WSN(v) + \frac{\alpha + 1}{\alpha} \Gamma^2 \Psi(v)^2 D^{w/o}(v) \times WSN2(v)$$

{Linearization constraints for $WL(v)$, $WLN(v)$, $WLN2(v)$, $WS(v)$, $WSN(v)$, $WSN2(v)$}

{Variable bounds}

$$0 \le W(v) \le W_{\max}, \qquad v \in V$$

$ST(v)$ are binary variables

Figure 2 Our MLP model for leakage minimization

3.4 MLP model with discrete size constraint

In our MLP model presented in Figure 2, $W(v)$ is a continuous real variable which is not the real case. Thus we add a constraint that the $W(v)$'s are integers, which means the sizes of the sleep transistors are powers of two. It is clear we can change the constraints to fit other discrete conditions of sleep transistors' sizes. We name the MLP model with continuous size constraints as MLP-C, the MLP model with integer size constraints as MLP-D.

4. Implementation and experiment results

We use ISCAS85 benchmark circuits to evaluate our MLP model. The netlists are synthesized using Synopsys Design Compiler and a TSMC 0.18μm standard cell library. The leakage current look up table is generated by HSPICE with TSMC 0.18μm CMOS process and a 1.8v supply condition. The values of various transistor parame-

Table 2 Leakage current saving through MLP-C Model and Fixed-Slowdown Method

ISCAS85 benchmark circuits	Original I_{leak} (pA)	0% MLP-C (pA)	3% MLP-C (pA)	5% MLP-C (pA)	7% MLP-C (pA)	7% Fixed-Slowdown (pA)	9% MLP-C (pA)	9% Fixed-Slowdown (pA)
C432	5874.30	2177.01	541.24	302.50	251.97	284.04	249.617	273.74
C499	24680.41	10295.4	698.04	376.29	367.28	400.314	363.54	387.88
C880	11636.60	1237.92	765.195	633.96	591.67	679.20	589.75	655.85
C1355	14793.67	5625.89	1149.33	856.96	834.85	917.86	821.95	884.46
C1908	28369.39	3199.31	1558.53	1344.22	1334.86	1537.64	1329.39	1482.11
C2670	43212.81	3382.23	2124.93	2000.58	1995.78	2304.83	1992.74	2226.86
C3540	51098.54	4326.21	3078.78	2627.25	2619.62	3018.22	2613.9	2913.15
C5315	71369.01	5142.03	4127.72	3759.77	3633.8	4186.75	3626.29	4044.78
C6288	53758.63	10760	5011.99	3957.93	3606.19	4042.71	3563.75	3893.56
Leakage saving	N/A	79.75%	93.56%	94.99%	95.24%	94.61%	95.28%	94.80%

Table 3 Comparison between MLP-C and Fixed-slowdown

ISCAS85 benchmark circuits	7% MLP-C		7% Fixed- slowdown		9% MLP-C		9% Fixed- slowdown	
	I_{leak} (pA)	ST area (W/L)	I_{leak} (pA)	ST area (W/L)	I_{leak} (pA)	ST area (W/L)	I_{leak} (pA)	ST area (W/L)
C432	251.97	714.27	284.04	2317.714	249.617	596.4515	273.74	1802.67
C499	367.28	1146.2072	400.314	2797.714	363.54	959.1344	387.88	2176
C880	591.67	876.1366	679.20	5252.571	589.75	780.1343	655.85	4085.333
C1355	834.85	3364.689	917.86	7515.429	821.95	2719.648	884.46	5845.333
C1908	1334.86	2354.592	1537.64	12493.71	1329.39	2081.361	1482.11	9717.333
C2670	1995.78	2088.2674	2304.83	17540.57	1992.74	1936.51	2226.86	13642.67
C3540	2619.62	3370.65	3018.22	23300.57	2613.9	3160.092	2913.15	18122.67
C5315	3633.8	4293.24	4186.75	31940.57	3626.29	3917.95	4044.78	24842.67
C6288	3606.19	11732.626	4042.71	33558.86	3563.75	9610.65	3893.56	26101.33
Average saving	95.24%	74.79%	94.61%	N/A	95.28%	72.40%	94.80%	N/A

ters have been taken from the TSMC library. For all the gates in the circuit, we take V_{THhigh}=500mv, V_{THlow}= 300mv, I_{ON}= 200μA. The experiments are set up with a specialized static timing analysis (STA) tool [10] to automatically generate the timing information and furthermore our MLP models. The MLP models can be solved by various LP solvers, here we use an LP solver named *lp_solve* [20]. We assume $W_{\max}$= 4, that is to say: $1 \le (W/L)_v \le 16$, corresponding to a least delay variance of 6%. Thus for 0%, 3%, and 5% circuit slowdown, we can not get a valid solution through conventional fixed slowdown method. On the other hand our MLP model can lead to an impressive leakage current saving. When the performance slowdown is 7% and 9%, the conventional fixed slowdown method is implemented with a larger area penalty and a smaller leakage current saving compared with our MLP-C model.

As shown in Table 2, the MLP-C model can achieve 79.75% leakage saving even if the circuit performance is not influenced. When the circuit slowdown is 3% and 5%, the leakage saving is 93.56%, 94.99% respectively through our MLP-C model. As we can see, our MLP-C model can achieve more leakage saving in the 5% circuit slowdown condition than fixed slowdown method in the 7% or 9% circuit slowdown condition. However, the difference of the leakage saving between our model and conventional fixed slowdown method is not as large as that mentioned in [16]. In our experimental results, the difference of leakage saving between our MLP-C model and fixed slowdown method in the same circuit slowdown condition is within 11%. That is caused by the difference leakage current model. When the performance slowdown is larger than 6%, our MLP-C model can get a optimal result with all the $ST(v)$=1, which leads to the same result as optimal sizing with sleep transistors placed everywhere [16].

In Table 3, we compare the area penalty between MLP-C model and fixed slowdown method. As we mentioned above, the difference of leakage saving is not very large. However, our MLP-C model can achieve a much less sleep transistor area penalty. With 7% circuit slowdown, our MLP-C model leads to 74.79% sleep transistor area saving compared to fixed slowdown method.

Obviously, an MLP-D model is very time-consuming because of the integer constraints and the increasing circuit size. Thus we derived a fast method (MLP-CtoD) to solve it based on the MLP-C model. The circuit produced by this continuous sleep transistor size scenario provides us a set of $W(v)|_C$ and $ST(v)|_C$. We directly choose the sleep transistor size $W(v)|_D$ =Ceiling($W(v)|_C$), and $ST(v)|_D$= $ST(v)|_C$ as the result of MLP-D. Table 4 shows the comparison of MLP-C, MLP-CtoD and MLP-D. As we can see, the MLP-C and MLP-D model can get the same ST gate number under the same circuit slowdown condition. Therefore the difference of leakage saving through MLP-C and MLP-D is determined by the ST area difference. Refer to our leakage current model in Section 2 and data in Table 4, the difference of leakage saving is very small, about 0.1%. Meanwhile, as shown in Table 4, our MLP-CtoD method is a very good approximation to MLP-D model: the difference in leakage reduction is within 0.01% and the MLP-CtoD method is about 30X times faster than MLP-D model.

From Table 4, when the circuit slowdown is below 6%, not all the gate in the circuit can use the sleep transistor scheme, thus a MTCOMS gate may drive a traditional CMOS gate, which can put the output of the MTCMOS into a floating gate. We also use a leakage feedback gate structure [21] in order to avoid floating states. Meanwhile the results for the area penalty imposed by the fine-grain sleep transistor in [16] show that the area

penalty is just around 5% through a standard cell placement methodology.

5. Conclusions

We have presented a mixed integer linear programming method to simultaneously place and size the sleep transistor in our fine-grain sleep transistor design to minimize the leakage current. A novel leakage current and delay model of the fine-grain sleep transistor design is presented in order to build up the MLP model. Our MLP model can minimize the leakage current to about 79.75% even though the circuit performance is not influenced. Two MLP model: MLP-C and MLP-D with different sleep transistor size constraints are presented and compared. The MLP-D uses a discrete sleep transistor size constraint which is more practical. An MLP-CtoD method is introduced to speed up MLP-D model and approximate the MLP-D model very well. Our experimental results show that the MLP-C model can achieve 93.56%, 94.99% leakage saving when the circuit slowdown is 3%, 5% respectively. The MLP-C model also achieve on average 74.79% less area penalty compared to the conventional fixed slowdown method when the circuit slowdown is 7%. The MLP-D model can also achieve just 0.1% less leakage saving compared to the MLP-C model. The MLP-CtoD method can speed up the MLP-D model 30X times within almost no difference in leakage reduction.

6. Reference

[1] B. H. Calhoun, F. A. Honoré, and A. P. Chandrakasan, "A Leakage Reduction Methodology for Distributed MTCMOS," *IEEE JSSC* Vol. 39, No. 5, May 2004, pp. 818 - 826.

[2] D. Duarte, N. Vijaykrishnan, M. J. Irwin, and M. Kandemir, "Formulation and validation of an energy dissipation model for the clock generation circuitry and distribution networks," in *Procs. of VLSI Design*,2001, pp. 248 - 253.

[3] G. Moore, "No exponential is forever: But forever can be delayed," in *IEEE ISSCC* Dig. Tech. Papers, 2003, pp. 20 - 23.

[4] Kao J., Narendra S., Chandrakasan A., "Subthreshold Leakage modeling and reduction techniques", in *Procs. of ICCAD*, 2002, pp 141 – 149

[5] K. Roy, S. Mukhopadhay, H. Mahmoodi-Meimand,"Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits", in *Proceedings of the IEEE*, Vol. 91, No. 2, Februray 2003 pp 305 – 327

[6] Narendra, S.; Keshavarzi, A.; Bloechel, B.A.; Borkar, S.; De, V.; "Forward body bias for microprocessors in 130-nm technology generation and beyond", *IEEE JSSC*, Vol. 38 , No. 5 ,May 2003 pp. 696 - 701.

[7] Kim, C.H.; Roy, K.; "Dynamic VTH scaling scheme for active leakage power reduction", in *Procs of DATE* 2002 pp.163 - 167.

[8] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim, and K. Roy, "Gate Leakage Reduction for Scaled Devices Using Transistor Stacking", *IEEE TVLSI*, Vol. 11, No. 4, Aug. 2003, pp. 716 - 729.

[9] L. Wei, Z. Chen, and K. Roy, "Design and Optimization of Dual Threshold Circuits for Low Voltage, Low Power Applications", *IEEE TVLSI*, Vol.2. 17, NO. 1, 1999, pp. 16-24.

[10] Yu Wang, Huazhong Yang, Hui Wang, "Signal-path level dual-Vt assignment for leakage power reduction", be appeared in JCSC Vol. 15, No. 2 (April 2006).

[11] J. Kao, S. Narendra, and A. Chandrakasan, "MTCMOS hierarchical sizing based on mutual exclusive discharge patterns," in *Procs. of DAC*, 1998, pp. 495–500.

[12] M. Anis, S. Areibi, and M. Elmasry, "Dynamic and leakage power reduction in MTCMOS circuits using an automated efficient gate clustering technique," in *Procs of DAC*, 2002, pp. 480–485.

[13] Wenxin Wang; Anis, M.; Areibi, S.; "Fast techniques for standby leakage reduction in MTCMOS circuits" in *Procs. of IEEE SOC*, 12-15 Sept. 2004 pp 21 – 24.

[14] Changbo Long; Lei He; "Distributed sleep transistors network for power reduction" in *Procs. of DAC*, 2-6 June 2003 pp. 181 – 186.

[15] Changbo Long; Lei He; "Distributed sleep transistor network for power reduction" in *IEEE TVLSI*, Volume: 12 , Issue: 9 , Sept. 2004 pp. 937 – 946

[16] V. Khandelwal, A. Srivastava; "Leakage Control Through Fine-Grained Placement and Sizing of Sleep Transistors ," in *Procs. of ICCAD* 2004, pp 533 - 536.

[17] S. Mukhopadhyay and K. Roy, " Modeling and Estimation of Total Leakage Current in Nano-scaled CMOS Devices Considering the Effect of Parameter Variation," in *Procs of ISLPED* Aug. 2003.

[18] S. Mutoh et al. "1-V Power Supply High Speed Digital Circuit Technology with Multithreshold Voltage CMOS," in *IEEE JSSC*, Vol. 30, No. 8 August 1995.

[19] Feng Gao and John P. Hayes; "Gate Sizing and Vt Assignment for Active-Mode Leakage Power Reduction," in *Procs. of IEEE ICCD*'04.

[20] http://groups.yahoo.com/group/lp_solve/

[21] J. Kao, A. Chandrakasan, " MTCMOS Sequential Circuits," in *Procs. of ESSDERC*, Sept 2003.

Table 4 Comparison of MLP-C, MLP-CtoD and MLP-D

Circuit slowdown	C432									
	MLP-C			MLP-CtoD			MLP-D			
	I_{leak} (pA)	ST Area (W/L)	ST Gate	I_{leak} (pA)	ST Area (W/L)	Time (Sec)	I_{leak} (pA)	ST Area (W/L)	ST Gate	Time (Sec)
0%	2177.01	489.56	137/169	2178	539	42.08	2178	539	137/169	1640.12
3%	541.24	721.741	157/169	542.71	795	1656.69	542.7	795	157/169	11398.14
5%	302.5	791.28	166/169	304.2	876	40.05	304.2	876	166/169	1057.3
7%	251.97	714.27	169/169	252.64	748	3.81	252.64	748	169/169	247.14
9%	249.62	596.45	169/169	250.61	646	3.88	250.61	646	169/169	206.53
Circuit slowdown	C880									
	MLP-C			MLP-CtoD			MLP-D			
	I_{leak} (pA)	ST Area (W/L)	ST Gate	I_{leak} (pA)	ST Area (W/L)	Time (Sec)	I_{leak} (pA)	ST Area (W/L)	ST Gate	Time (Sec)
0%	1237.92	678.33	356/383	1238.63	714	31.81	1238.62	714	356/383	3753.14
3%	765.2	831.91	373/383	765.92	868	1491.45	765.92	868	373/383	11398.14
5%	633.96	924.424	381/383	634.71	962	119.41	634.71	962	381/383	7258.91
7%	591.67	876.14	383/383	592.37	911	10.72	592.37	911	383/383	3256.92
9%	589.75	780.13	383/383	590.31	808	10.3	590.31	808	383/383	96.141