第27卷 第2期 半 导 体 学 报 Vol.27 No.2
2006年2月 CHINESE JOURNAL OF SEMICONDUCTORS Feb.,2006

Fine-Grain Sleep Transistor Insertion for Leakage Reduction*

Yang Huazhong†, Wang Yu, Lin Hai, Luo Rong, and Wang Hui

(Tsinghua University, Beijing 100084, China)

Abstract: A fine-grain sleep transistor insertion technique based on our simplified leakage current and delay models is proposed to reduce leakage current. The key idea is to model the leakage current reduction problem as a mixed-integer linear programming (MLP) problem in order to simultaneously place and size the sleep transistors optimally. Because of better circuit slack utilization, our experimental results show that the MLP model can save leakage by 79.75%, 93.56%, and 94.99% when the circuit slowdown is 0%, 3%, and 5%, respectively. The MLP model also achieves on average 74.79% less area penalty compared to the conventional fixed slowdown method when the circuit slowdown is 7%.

Key words: leakage current reduction; fine-grain; sleep transistor insertion; delay model; mixed-integer linear programming

EEACC: 1265A; 1130B

CLC number: TN406 **Document code:** A **Article ID:** 0253-4177(2006)02-0258-08

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, and can thus be expressed as

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

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 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].

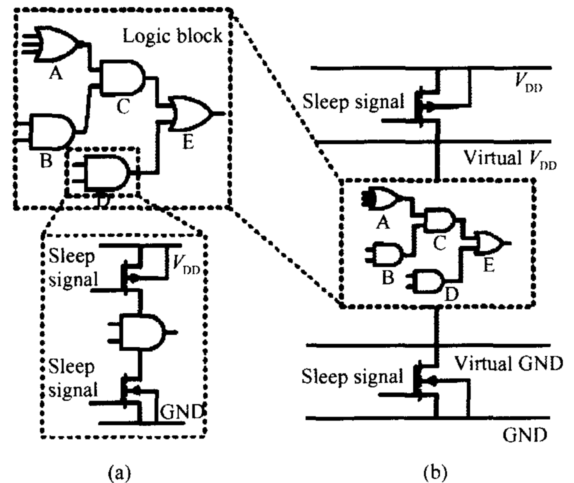

Fig.1 Fine-grain versus cluster-based ST insertion (a) Fine-grain gate level ST insertion; (b) Cluster based block level ST insertion

New techniques are necessary to reduce leakage power. 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

* Project supported by the National High Technology Research and Development Program of China (Nos. 2004AA1Z1050, 2005AA1Z1230) and the National Natural Science Foundation of China (Nos. 90207001, 60506010)

† Corresponding author. Email: yanghz@tsinghua.edu.cn

Received 26 October 2005 ©2006 Chinese Institute of Electronics

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).

Among these, multi-threshold CMOS (MTCMOS) is a valuable technique for reducing leakage power in the circuit standby mode. The MTCMOS technique consists essentially of 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 in 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[11~15]. However, in recent years the use of sleep devices at the gate level[1,16] (Fig. 1 (a)), which has some advantages over the block level design (Fig. 1 (b)), has raised some concern.

The existing literature on MTCMOS circuits[11~15] present cluster based methods for sleep transistor insertion and sizing. Reference [11] first gives out a mutual exclusion method to reduce the area penalty. References [12] and [13] present several heuristic techniques for efficient gate clustering and try to mitigate the ground problem by introducing an additional power penalty. In Refs. [14] and [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 a 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]. Therefore, it is harder to guarantee circuit functionality for large blocks with only one sleep transistor[1].

The fine-grain MTCMOS design methodology is discussed in Refs. [1] and [16]. In Ref. [1], a fine-grain MTCMOS design methodology and several design rules are proposed. The authors also make a comparison between local and global devices. Reference [16] presents a selective sleep transistor insertion methodology with better utilization of circuit slack. They first select where to put the sleep transistors with a heuristic method and then solve an LP model to optimize the sleep transistor size. The second step can give an optimal size, but 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.

(1) Our newly developed leakage current and delay models of a single gate are proposed, which are much simpler and more exact than the ones in traditional fine grain sleep transistor insertion strategies.

(2) A formal mixed-integer linear model of the leakage current reduction problem provides the designers with the relation between leakage current and circuit constraints, and makes it possible to simultaneously select and optimize the place to put the sleep transistors and the size of the sleep transistors. 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 affected, our model can save an impressive amount of leakage. Furthermore, if the conventional fixed slowdown method can be performed, our method still leads to better leakage saving and a much smaller total sleep transistor size.

(3) The model can be solved with a discrete sleep transistor size constraint which is more practical in real life.

2 Preliminaries

First we define leakage current and the delay model. A cell-based design flow with a given cell library is used. We assume that sleep transistors with variable sizes, which are determined 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.

$$P_{leakage}(G) = V_{DD} \times I(G) \quad (2)$$

The circuit average leakage current can be calculated as the sum of the individual gates' average leakage current. The leakage current of a CMOS gate is determined 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:

$$I_l(v) = \sum_{IN} I_l(v, IN) \times PB(v, IN) \quad (3)$$

where $I_l(v, IN)$ 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 into this gate or not. For the gates without sleep transistor, we create a leakage reference 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

$$I_l^{w/o}(v) = \sum_{IN} I_l(v, IN) \times PB(v, IN) \quad (4)$$

The subthreshold leakage currents with sleep transistors are given by Ref. [17]:

$$I_l^{ST}(v) = \mu_n C_{ox}(W/L)_v e^{1.8} V_T^2 e^{\frac{V_{gs} - V_{THhigh}}{nV_T}} (1 - e^{\frac{-V_{ds}}{V_T}}) \quad (5)$$

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. As we will explain below, V_{ds} is the voltage drop V_x which is decided by $(W/L)_v$, and thus the relationship between I_l^{ST} and $(W/L)_v$ is complicated. Here we present our simplified leakage current $I_l^{ST}(v)$ model:

$$I_l^{ST} = A(v) + B(v) \times (W/L)_v \quad (6)$$

where A(v), B(v) are constants that 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. Referring to our model, we can give 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 = 1	1.33615	1.3403	0.31%	1.695831	1.70132	0.31%
W/L = 1	1.3618	1.36286	0.08%	1.730075	1.7316	0.09%
W/L = 1	1.407875	1.40798	<0.01%	1.791681	1.79216	0.03%
W/L = 1	1.4988	1.49822	-0.04%	1.914263	1.91328	-0.05%

Notice $I_l^{ST}(v)$ is still sensitive to the input pattern. The data shown in Table 1 are the average leakage currents assuming 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 15 times larger than $I_l^{ST}(v)$. We estimate every A(v) and B(v) for all the standard cells and find that, on average, the B(v)'s are around 1% of A(v), and thus the variation range of $I_l^{ST}(v)$ is about 15% of A(v).

Thus we use a 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 de-

sign, we have to insert sleep transistors into the original gates in the given library. As shown in Ref. [18], the delay of the gate is affected by the sleep transistor insertion. The load dependent delay $D^{w/o}(v)$ of gate v without sleep transistors can be expressed as

$$D^{w/o}(v) = \frac{KC_L V_{DD}}{(V_{DD} - V_{THlow})^{\alpha}} \tag{7}$$

where C_L, V_{THlow}, α, and 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

$$D^{ST}(v) = \frac{KC_L V_{DD}}{(V_{DD} - 2V_x - V_{THlow})^{\alpha}} \tag{8}$$

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

$$\Delta D(v) = D^{ST}(v) - D^{w/o}(v) \tag{9}$$

Referring to Eqs. (6~8), we can get an approximate $\Delta D(v)$ with negligible difference using the Taylor series expansion:

$$\Delta D(v) = D^{ST}(v) - D^{w/o}(v) = \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}}\right)^{-\alpha} - 1\right) D^{w/o}(v) \overset{\text{Taylor}}{=} \left(\left(1 + \alpha\frac{2V_x}{V_{DD} - V_{THlow}} + \alpha(\alpha+1) \times \left(\frac{2V_x}{V_{DD} - V_{THlow}}\right)^2 + \cdots\right) - 1\right) D^{w/o}(v)$$
$$\approx \left(\alpha\frac{2V_x}{V_{DD} - V_{THlow}} + \alpha(\alpha+1)\left(\frac{2V_x}{V_{DD} - V_{THlow}}\right)^2\right) \times D^{w/o}(v) \triangleq \left(\Gamma V_x + \frac{\alpha+1}{\alpha}\Gamma^2 V_x^2\right) \times D^{w/o}(v) \tag{10}$$

We use a constant $\Gamma = 2\alpha/(V_{DD} - V_{THlow})$ to simplify Eq. (9) since V_{THlow}, α, V_{DD} are all technology-dependent constants. 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]

$$I_{ON}(v) = \mu_n C_{ox}(W/L)_v\left((V_{DD} - V_{THhigh})V_x - \frac{V_x^2}{2}\right)$$
$$\approx \mu_n C_{ox}(W/L)_v(V_{DD} - V_{THhigh})V_x \tag{11}$$

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

$$V_x = \frac{I_{ON}(v)}{\mu C_{ox}(V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_v} = \Psi(v) \times (W/L)_v^{-1} \tag{12}$$

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

$$\Delta D(v) = \left(\Gamma\Psi(v) \times (W/L)_v^{-1} + \frac{\alpha+1}{\alpha}\Gamma^2\Psi(v)^2(W/L)_v^{-2}\right) \times D^{w/o}(v) \tag{13}$$

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

3 MLP model construction

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

3.1 Objective function

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

$$I_l(v) = I_l^{w/o}(v) \times (1 - ST(v)) + I_l^{ST} \times ST(v) \tag{14}$$

Therefore we represent the total leakage current by

$$I(G) = \sum_{v \in V}\left(I_l^{w/o}(v) \times (1 - ST(v)) + I_l^{ST} \times ST(v)\right) \tag{15}$$

Referring to Eqs. (3) and (5), we can replace Equation (13) with

$$I(G) = \sum_{v \in V}\left\{\left(\sum_{IN} I_l(v, IN) \times PB(v, IN)\right) \times [1 - ST(v)] + [A(v) + B(v) \times (W/L)_v] \times ST(v)\right\} \tag{16}$$

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

3.2 Timing constraints

First we consider the primary input (PIs) and

output (POs) gates of the circuit. The arrival times t_a of all the PIs are set to zero, while the required times of all the POs are less than the overall circuit delay T_{req}.

$$t_a(m) = 0, \quad m \in PI \tag{17}$$

$$t_a(n) + D(n) \le T_{req}, \quad n \in PO \tag{18}$$

Then we notice that the sum of gate v's arrival time and its delay must be less than or equal to 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:

$$t_a(i) + D(i) \le t_a(j) \tag{19}$$

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

$$\begin{aligned} D(v) &= D^{w/o}(v) + \Delta D(v) \times ST(v) \\ &= D^{w/o}(v) + \Gamma\Psi(v) D^{w/o}(v) \times (W/L)_v^{-1} \times ST(v) \\ &+ \frac{\alpha+1}{\alpha}\Gamma^2 \Psi(v)^2 D^{w/o}(v) \times (W/L)_v^{-2} \times ST(v) \end{aligned} \tag{20}$$

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 Ref. [19] to linearize these exponential expressions with inequalities:

$$\begin{gathered} WL(v) \ge 2^k \times W(v) + (1 - k) \times 2^k, \\ k = 0, 1, \cdots, W_{max} \\ WLN(v) \ge -2^k \times W(v) + (1 - k) \times 2^k, \\ k = -W_{max}, -W_{max} + 1, \cdots, 0 \\ WLN2(v) \ge -2^k \times 2W(v) + (1 - k) \times 2^k, \\ k = -2W_{max}, -2(W_{max} + 1), \cdots, 0 \end{gathered}$$

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

$$\begin{gathered} WS(v) = (W/L)_v \times ST(v) = WL(v) \times ST(v) \\ WSN(v) = (W/L)_v^{-1} \times ST(v) = \\ WLN(v) \times ST(v) \\ WSN2(v) = (W/L)_v^{-2} \times ST(v) = \\ WLN2(v) \times ST(v) \end{gathered}$$

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

$$\begin{gathered} 0 \le C \le B \\ C \le MA \\ C \ge B - M(1 - A) \end{gathered}$$

Since $W(v) \in [0, W_{max}]$, $WL(v)$, $WLN(v)$, and $WLN2(v)$ all have upper bounds. This completes our MLP model for leakage minimization. The general form of our MLP model is given in Fig. 2.

Minimize:

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

Subject to:

{Timing constraints}

$t_a(m) = 0, \quad m \in PI$

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

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

$D(v) = D^{w/o}(v) + \Gamma\Psi(v) D^{w/o}(v) \times WSN(v)$

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

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

{Variable bounds}

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

$ST(v)$ are binary variables

Fig. 2 MLP model for leakage minimization

3.4 MLP model with discrete size constraint

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

4 Implementation and experimental results

We use ISCAS85 benchmark circuits to evaluate our MLP model. The netlists are synthesized using the synopsys design compiler and a TSMC 0.18μm standard cell library. The leakage current reference table is generated by HSPICE with a TSMC 0.18μm CMOS process and a 1.8V supply condition. The values of various transistor param-

eters have been taken from the TSMC library. For all the gates in the circuit, V_{THhigh} = 500mV, V_{THlow} = 300 mV, I_{ON} = 200μA. The experiments are set up with a specialized static timing analysis (STA) tool[10] to automatically generate the timing information. 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 slowdowns, we cannot get a valid solution through the conventional fixed slowdown method. On the other hand our MLP model can save leakage current by an impressive amount. When the performance slowdown is 7% and 9%, the conventional fixed slowdown method is implemented with a larger area penalty and less leakage current is saved compared with our MLP-C model.

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

As shown in Table 2, the MLP-C model can save leakage by 79.75% without affecting the circuit performance is not. When the circuit slowdown is 3% and 5%, then 93.56%, 94.99% of the leakage is saved respectively through our MLP-C model. As we can see, our MLP-C model can save more leakage in the 5% circuit slowdown condition than the fixed slowdown method can with a 7% or 9% circuit slowdown. However, the difference of the saved leakage between our model and the conventional fixed slowdown method is not as large as that mentioned in Ref. [16]. In our experimental results, the difference of leakage saved between our MLP-C model and the fixed slowdown method under the same circuit slowdown condition is within 11%. That is caused by the different leakage current models. When the performance slowdown is larger than 6%, our MLP-C model can get an 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 the MLP-C model and the fixed slowdown method. As we mentioned above, the difference in leakage saved is not very large. However, our MLP-C model can achieve a much smaller sleep transistor area penalty. With a 7% circuit slowdown, our MLP-C model saves sleep transistor area by 74.79% compared to the fixed slowdown method.

When the circuit slowdown is below 6%, not all the gates in the circuit can use the sleep transistor scheme, and 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 Ref. [16] show that the area penalty is just around 5% through a standard cell placement methodology.

5 Conclusion

We have presented a mixed integer linear programming model to simultaneously place and size the sleep transistor in our fine-grain sleep transistor design to minimize the leakage current. Novel leakage current and delay models of the fine-grain sleep transistor design are presented in order to build up the MLP model. Our MLP model can minimize the leakage current to about 79.75% without affecting the circuit performance. Our experimental results show that the MLP-C model can achieve save leakage by 93.56% and 94.99% when the circuit slowdown is 3% and 5%, respectively. The MLP-C model also achieve on average an area penalty 74.79% less than the conventional fixed slowdown method when the circuit slowdown is 7%.

References

[1] Calhoun B H, Honoré F A, Chandrakasan A P. A leakage reduction methodology for distributed MTCMOS. IEEE J Solid-State Circuits, 2004, 39(5): 818

[2] Duarte D, Vijaykrishnan N, Irwin M J, et al. Formulation and validation of an energy dissipation model for the clock generation circuitry and distribution networks. Proc of VLSI Design, 2001: 248

[3] Moore G. No exponential is forever: but forever can be delayed. IEEE ISSCC Dig Tech Papers, 2003: 20

[4] Kao J, Narendra S, Chandrakasan A. Subthreshold leakage modeling and reduction techniques. Proc of ICCAD, 2002: 141

[5] Roy K, Mukhopadhay S, Mahmoodi-Meimand H. Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. Proceedings of the IEEE, 2003, 91(2): 305

[6] Narendra S, Keshavarzi A, B Bloechel A, et al. Forward body bias for microprocessors in 130-nm technology generation and beyond. IEEE J Solid-State Circuits, 2003, 38(5): 696

[7] Kim C H, Roy K. Dynamic VTH scaling scheme for active leakage power reduction. Proc of DATE, 2002: 163

[8] Mukhopadhyay S, Neau C, Cakici R T, et al. Gate leakage reduction for scaled devices using transistor stacking. IEEE Trans Very Large Scale Integration Syst, 2003, 11(4): 716

[9] Wei L, Chen Z, Roy K, et al. Design and optimization of dual-threshold circuits for low-voltage low-power applications. IEEE Trans Very Large Scale Integration Syst, 1999, 7(1): 16

[10] Wang Yu, Yang Huazhong, Wang Hui. Signal-path level assignment for dual-V_t technique. Proceedings of IEEE PRIME, 2005: 52

[11] Kao J, Narendra S, Chandrakasan A. MTCMOS hierarchical sizing based on mutual exclusive discharge patterns. Proc of DAC, 1998: 495

[12] Anis M, Areibi S, Elmasry M. Dynamic and leakage power reduction in MTCMOS circuits using an automated efficient gate clustering technique. Proc of DAC, 2002: 480

[13] Wang Wenxin, Anis M, Areibi S. Fast techniques for standby leakage reduction in MTCMOS circuits. Proc of IEEE SOC, 2004: 21

[14] Long Changbo, He Lei. Distributed sleep transistors network for power reduction. Proc of DAC, 2003: 181

[15] Long Changbo, He Lei. Distributed sleep transistor network for power reduction. IEEE Trans Very Large Scale Integration Syst, 2004, 12(9): 937

[16] Khandelwal V, Srivastava A. Leakage control through fine-grained placement and sizing of sleep transistors. Proc of ICCAD, 2004: 533

[17] Mukhopadhyay S, Roy K. Modeling and estimation of total leakage current in cano-scaled CMOS devices considering the effect of parameter variation. Proc of ISLPED, 2003

[18] Mutoh S, Douski T, Matsuya Y, et al. 1-V power supply high speed digital circuit technology with multithreshold voltage CMOS. IEEE J Solid-State Circuits, 1995, 30(8): 847

[19] Feng G, Hayes John P. Gate sizing and V_t assignment for active-mode leakage power reduction. Proc of IEEE ICCD, 2004

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

[21] Kao J, Chandrakasan A. MTCMOS sequential circuits. Proc of ESSDERC, 2003

降低泄漏电流的细粒度休眠晶体管插入法 *

杨华中† 汪 玉 林 海 罗 嵘 汪 蕙

(清华大学电子工程系, 北京 100084)

摘要: 首先给出一种泄漏电流和延时的简化模型,并且在此基础上提出了一种降低泄漏电流的细粒度休眠晶体管插入法. 该方法的核心是利用混合整数线性规划方法同时确定插入细粒度休眠晶体管的位置和尺寸. 从实验结果可以发现,由于这种方法更好地利用了电路中的延时余量,所以在电路性能不受影响的情况下可以减小 79.75 %的泄漏电流;并且在一定范围内放宽电路的延时约束可以更大幅度地降低泄漏电流. 与传统的固定放宽延时约束的方法相比较,当延时约束放宽 7 %时,这种方法可以节约 74.79 %的面积.

关键词: 泄漏电流; 细粒度; 休眠晶体管; 延时模型; 混和整数线性规划

EEACC: 1265A; 1130B

中图分类号: TN406　**文献标识码**: A　**文章编号**: 0253-4177(2006)02-0258-08

* 国家高技术研究发展计划(批准号:2004AA1Z1050,2005AA1Z1230)和国家自然科学基金(批准号:90207001, 60506010)资助项目

† 通信作者. Email: yanghz@tsinghua.edu.cn

2005-10-26 收到

c 2006 中国电子学会