Fine-grain Sleep Transistor Placement Considering Leakage Feedback Gate

Yu Wang, Hui Wang, Huazhong Yang
Circuit and System Division, E.E. Department
Tsinghua University
Beijing, P.R.China
wangyuu99@mails.tsinghua.edu.cn

***Abstract*—Fine-grain sleep transistor insertion (FGSTI) technique is easier to guarantee circuit functionality and improves circuit noise margins while achieves a considerable leakage saving when the circuit is standby. However, when the circuit slowdown is not enough to assign sleep transistors (ST) to each gate, a large amount of leakage feedback (LF) gates should be used to avoid floating states, and these additional buffers will induce large area and dynamic power penalty. In this paper, we propose a multi-object optimization method to simultaneously reduce the LF gate number and the leakage current. Our experimental results show that, when the circuit slowdown varies from 0% to 5%, comparing with method only considering the leakage current reduction, we can achieve on average 4X-9X LF gate number reduction while the leakage difference is only about 8% of original circuit leakage.**

I. INTRODUCTION

At the 90nm technology node, leakage power may make up 42% of total power [1]. Since the leakage issue will become more and more important in the future VLSI circuit design with the technology scaling, various techniques are proposed to reduce the leakage power from system level down to physical level [2]. In burst mode type circuits, where the system spends the majority of the time in an idle standby state, Multi-Threshold CMOS (MTCMOS) is proved to be a very effective technique for leakage current reduction during the standby mode [3-10].

The most popular MTCMOS technique is gating the power of sizable blocks using large ST's which assumes that all the gates in one block have a fixed slowdown [3-5], which is concluded as block based ST insertion (BBSTI) technique. The existing literatures on BBSTI techniques [3-5] focus on how to reduce the ST area penalty along with a remarkable leakage saving. Although BBSTI techniques greatly reduce the area penalty, they induce large ground bounce which has adverse effects on circuit speed and noise immunity [8]. ST size is determined by the worst case current which is quite difficult to determine without comprehensive simulation [3]. Thus it is harder to guarantee circuit functionality for large blocks with only one ST [6].

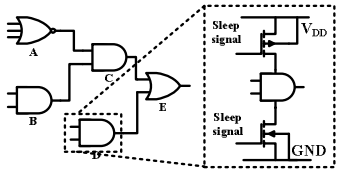

Figure 1. Fine-grain sleep transistor insertion technique

In recent years gate level ST insertion, which can be also called fine-grain ST insertion (FGSTI) technique [6-10] (shown in Figure 1) shows some advantages over the BBSTI technique. It is easier to guarantee circuit functionality in an FGSTI technique as ST sizes are not determined by the worst case current of large circuit blocks. And the FGSTI technique leads to a smaller simultaneous switching current when the circuit changes between standby and active mode, thus improves circuit noise margins. As shown in [8], FGSTI technique corresponds to an area penalty of roughly 5% using standard cell placement. However, when the circuit slowdown is not enough to assign ST to each gate, a large amount of leakage feedback (LF) gates may be used to avoid floating states [7]. The LF gates number may exceed as much as 80% of the gates with ST when the effect of LF gate is not considered in FGSTI technique. Thus the additional buffers in the LF gates will induce large area and dynamic power penalty [10].

In [6], a fine-grain MTCMOS design methodology and several design rules are proposed. The authors also make a comparison between local and global devices. Recently, in [8] a one-shot heuristic algorithm is used to determine where to put the sleep transistor in FGSTI design considering LF gates, but how to perform FGSTI technique when the circuit slowdown is 0% isn't addressed and the one-shot heuristic algorithm may easily fall into a local optimal result. Our previous work [9] presents a mixed integer programming (MLP) model for FGSTI technique to determine ST placement and sizing simultaneously without considering the influence of LF gate. In [10], we prove that FGSTI technique can be performed in a two phase manner: first, decide where to put the ST and achieve most of the leakage saving; and

Project supported by National 863 project of China (No. 2005AA1Z 1230) and National Natural Science Foundation of China (No.90207001, No. 60506010).

then resize the ST's to reduce the area overhead along with further leakage current reduction. The LF gate number is calculated and compared under different circuit slowdowns without optimization.

This paper presents a novel ST placement method to simultaneously reduce the LF gate number and the leakage current based on the two-phase FGSTI technique [10]. The LF gate and normal ST gate are compared, to prove that a carefully sized LF gate can be used to substitute a normal ST gate without affecting the circuit performance. The multi-object ST placement problem is formulated to provide the designer the relationship between LF gate number and the leakage current reduction.

The paper is organized as follows. In Section II, detailed information of LF gate is introduced. In Section III, the leakage current and delay models are presented and analyzed to prove the rationality of the two-phase FGSTI technique. Our multi-object ST placement technique is proposed in Section IV. Experimental results are presented and analyzed in Section V. In Section VI, we conclude this paper.

II. LEAKAGE FEEDBACK GATE

A. Circuit Scheme

When the circuit slowdown is not large enough to change all the gates in the circuit, FGSTI technique can cause a gate with ST to drive a gate without ST. This will lead to a floating state at the output of the gate with ST and may cause large power dissipation due to the short circuit current in the gate without ST. As mentioned in [8], the LF gate structure [7] shown in figure 2 should be used in order to avoid the floating states. The important characteristic of LF gate is that

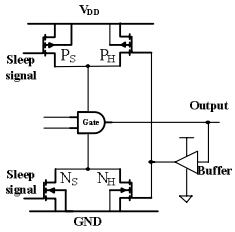

Figure 2. Leakage feedback gate

depending on the state of the latest output, one but not both, helper ST's (P_H or N_H) is turned on by the feedback buffer, thus the output state of the LF gate are set to "1" or "0".

B. Comparison with Normal ST Gate

During the standby state, both high V_t sleep devices P_S and N_S are turned off, only one of the helper sleep devices will be kept on to drive the output signal to the appropriate rail. On the other hand, when the circuit is active, both high V_t ST's P_S and N_S are turned on. One and only one of the helper ST's will be turned on to accelerate the circuit speed since the feedback buffer is sensitive to the change of the output signal. The signal propagation delay of an inverter with ST and an LF gate are compared under same load capacitance and shown in Figure 3. The sizes of helper ST's are the same as those of the original ST's. As we can see, the rise slope of a LF gate is steeper than that of a normal ST gate. Therefore, we can conclude that every gate with ST can be replaced by a carefully sized LF gate without affecting the circuit delay constraints.

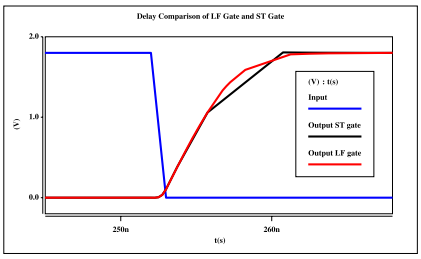

Figure 3. Delay comparison-- a LF gate and a normal gate with ST

III. LEAKAGE CURRENT AND DELAY MODELS

A combinational circuit is represented by a directed acyclic graph (DAG) $G = (V, E)$. 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. In this section, the leakage current and delay models are given out and briefly analyzed to show the rationality of two-phase FGSTI.

A. Leakage Current Model

The original leakage current of gate v is denoted as $I_{w/o}(v)$, while the leakage current of gate v assigned with high V_t ST is denoted as $I_w(v)$. Obviously, the leakage current of gate v with ST depends on the ST's size. We choose the largest ST size $(W/L)_v = 16$ during ST placement for simplicity, which leads to the minimum delay overhead as shown below. Due to the stacking effect, $I_{w/o}(v)$ is about two orders of magnitude larger than $I_w(v)$. Thus if more gates in the circuit can be assigned with ST's, more leakage saving can be achieved. Extensive HSPICE simulations are used to create two leakage current look up tables for all the gates in the circuits to represent these two values: $I_{w/o}(v)$ and $I_w(v)$. Here the leakage current of an LF gate and a normal ST gate are assumed to be the same value: $I_w(v)$.

B. Delay Model

The load dependent delay $d_{w/o}(v)$ of gate v without ST is given by:

$$d_{w/o}(v) = \frac{KC_L V_{DD}}{(V_{DD} - V_{THlow})^{\alpha}} \quad (1)$$

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 $d_w(v)$ of gate v with ST can be expressed as:

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

where V_x is the V_{ds} of the ST, that is to say the voltage drop from V_{DD} to the virtual V_{DD}. $\Delta d(v)$ from the above equations:

$$\Delta d(v) = d_w(v) - d_{w/o}(v) = \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d_{w/o}(v) \quad (3)$$

$I_{ON}(v)$ is the current flowing through ST in gate v during the active mode, and can be expressed as given by [8]:

$$\begin{aligned} I_{ON}(v) &= \mu_n C_{ox} (W/L)_v ((V_{DD} - V_{THhigh}) V_x - \frac{V_x^2}{2}) \\ &= \mu_n C_{ox} (W/L)_v (V_{DD} - V_{THhigh}) V_x \end{aligned} \quad (4)$$

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

$$V_x = \frac{I_{ON}(v)}{\mu_n C_{ox} (V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_v} \quad (5)$$

Combining equation (3) and (5), the propagation delay $d_w(v)$ of gate v with ST can be rewrite as:

$$\begin{aligned} d_w(v) &= = d_{w/o}(v) + \left(\left(1 - \frac{2 \frac{I_{ON}(v)}{\mu_n C_{ox} (V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_v}}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d_{w/o}(v) \\ &= d_{w/o}(v) + \varphi((W/L)_v) d_{w/o}(v) \end{aligned} \quad (6)$$

where $d_{w/o}(v)$ is constant which can be extracted from the technology library. Referring to equation (5), a larger $(W/L)_v$ leads to a smaller delay overhead. Here the largest ST size $(W/L)_v = 16$ is chosen which makes $\varphi((W/L)_v)$ a constant.

C. Rationality of Two-phase FGSTI

The leakage current difference of a gate is about 100X under different ST condition: with or without ST insertion. Referring to equation (3) and (5), the delay difference is less than 20% of the original gate delay under different ST condition (the ST size (W/L) is set to 16). However, the delay difference of a gate with different ST size is much larger; for example, setting the (W/L) of a ST to 1 will lead to about 140% additional delay compared to the original gate without ST. The leakage current variation range due to the change of ST size can be neglected because it is much smaller compared with the leakage saving of changing a gate's ST condition. Hence FGSTI technique can be performed in a two phase manner [10]: first, ST placement to achieve most of the leakage saving; and then ST sizing to reduce the area overhead along with further leakage current reduction.

IV. ST Placement Problem Fomulation

We propose a novel ST placement method that tries to maximize the leakage saving in the circuits and minimize the LF gate number simultaneously through mixed integer linear programming (MLP) models.

First, we construct the multi-object function as below:

$$\begin{aligned} & I_{leak} + \gamma N_{LF} \\ &= \sum_{v \in V} \left(I_{w/o}(v) \times (1 - ST(v)) + I_w(v) \times ST(v) \right) + \gamma \sum_{v \in V} (LF(v)) \end{aligned} \quad (7)$$

where I_{leak} is the total leakage current, N_{LF} is the LF gate number in the circuit; ST(v) is a binary variable to represent gate v's ST condition, $ST(v) = 1$ means gate v has ST inserted and $ST(v) = 0$ means gate v is without ST; $LF(v)$ is also a binary variable to represent gate v's LF gate condition, $LF(v) = 1$ means gate v is an LF gate and $LF(v) = 0$ means gate v is not an LF gate; γ is a weight value which can be modified by the circuit designer. The timing constraints of $G(V, E)$ can be expressed as:

$$t_a(m) = 0 \qquad m \in PI \quad (8)$$

$$t_a(n) + d(n) \le T_{req} \qquad n \in PO \quad (9)$$

$$t_a(i) + d(i) \le t_a(j) \qquad \forall (i,j) \in E, i, j \in V \quad (10)$$

where PI and PO refer to the primary input and primary output gates of the circuit; $t_a(v)$ represents the arrival time of gate v, T_{req} is the overall circuit delay; $d(v)$ represents the gate delay which can be expressed as using equation (6):

$$d(v) = d_{w/o}(v) + \varphi((W/L)_v)\big|_{(W/L)_v=16} \times d_{w/o}(v) \times ST(v) \quad (11)$$

where $d_{w/o}(v)$ and $\varphi((W/L)_v)\big|_{(W/L)_v=16}$ are constants for each gate.

A gate v must be changed into LF gate if $ST(v) = 1$ and one of its fan-out gate is a gate without ST. Thus the binary variable $LF(v)$ should satisfy the following constraint:

$$LF(i) \ge ST(i) - ST(j) \qquad \forall (i,j) \in E, i, j \in V \quad (12)$$

The general form of our MLP model for ST placement is shown in figure 4.

Minimize:

$$I_{leak} + \gamma N_{LF} = \sum_{v \in V} \left(I_{w/o}(v) \times (1 - ST(v)) + I_w(v) \times ST(v) \right) + \gamma \sum_{v \in V} (LF(v))$$

Subject to:

{Timing constraints}

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

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

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

$$d(v) = d_{w/o}(v) + \varphi((W/L)_v)\big|_{(W/L)_v=16} \times d_{w/o}(v) \times ST(v) \qquad v \in V$$

{Variable constraints}

$ST(v)$ and $LF(v)$ are binary variables

$$LF(i) \ge ST(i) - ST(j) \qquad \forall (i,j) \in E, i, j \in V$$

Figure 4. MLP model for multi-object ST placement

V. Implementation and Experimental Results

ISCAS85 benchmark netlists are synthesized using Synopsys Design Compiler and a TSMC 0.18μm standard cell library. Two leakage current look up tables for all the standard cells are generated using HSPICE. The values of various transistor parameters are taken from the TSMC 0.18μm process library, i.e. V_{DD}=1.8V, V_{THhigh}=500mV, V_{THlow}= 300mV, and I_{ON}= 200μA for all the gates in the circuit. The timing constraints are set up with a static timing analysis (STA) tool [11], and the MLP models for ST placement are automatically generated. We use an LP solver named *lp_solve* [12] to solve the models.

We assume $(W/L)_v$= 16, corresponding to a delay variance of 6% if we assign ST's to all the gates in the circuit [9]. Thus when the circuit slowdown varies in the range of 6% circuit original delay, ST's can not be assigned to every gate in the circuit. The LF gate should be used when a gate with ST is driving a gate without ST. We first compare results of our multi-object ST placement (M-STP) with the ST placement without considering the LF gate (STP-WO) [10], which are shown in table I. The weight value γ is assumed to be 100.

In table I, if the LF gate is not considered during ST placement, about 37.1% of the gates with ST should be changed into LF structure if the circuit slowdown is 0%. When circuit slowdown is 3% and 5%, about 19.8% and 9.9% of the gate with ST should be changed into LF gate respectively. When the circuit slowdown is 0%, some of the benchmarks, such as C499, C1355, need to change 80.4% and 66.4% of normal ST gates into LF gates. This will lead to a large area increasing due to large number of high V_t feedback buffers and helper ST's. As the LF gate is considered during the multi-object ST placement, the LF gate number is about 9.3%, 3.3% and 1.1% of the total gates with ST when the circuit slowdown is 0%, 3% and 5% respectively. Meanwhile, the difference of leakage reduction rate is only 7.9%, 4.3% and 2.8%. For the two typical benchmarks we mentioned above: C499 and C1355, the LF gate becomes 35.2% and 16.1% of the gates with ST when the circuit slowdown is 0%.

TABLE I. DIFFERENT WEIGHT VALUE γ FOR C880

C880	0% circuit slowdown			3% circuit slowdown			5% circuit slowdown		
	I_{leak}(pA)	N_{ST}	N_{LF}	I_{leak}(pA)	N_{ST}	N_{LF}	I_{leak}(pA)	N_{ST}	N_{LF}
γ=0	619.2	352	105	232.2	370	68	126.1	375	46
γ=10	630.7	352	25	252.2	369	14	157.5	375	7
γ=50	723.6	350	21	365.6	366	10	199.7	373	4
γ=100	1292.3	315	11	541.4	359	6	255.7	370	3
γ=200	2415.9	263	2	1034.1	330	2	479.5	357	1

Furthermore, the weight value γ can be used to control the trade-off between leakage reduction rate and the LF gate number. Four different weight values: 10, 50, 100, 200 are used in our MLP model for C880. Table II shows the leakage current and LF number under different weight value. As in table II, when the circuit slowdown is 0%, a larger weight value γ should be chosen to reduce the LF gate number; when the circuit slowdown is becoming larger, the original LF gate number without any optimization reduces to a low level, so a smaller weight value γ can be used to get a larger leakage reduction rate with an acceptable LF gate number.

VI. CONCLUSIONS

In this paper, we present a novel multi-object ST placement method during a two-phase FGSTI technique to reduce the leakage current and the LF gate number simultaneously. Our experimental results show that the multi-object ST placement can achieve about 4X, 5X and 9X LF gate number reduction when the circuit slow down is 0%, 3% and 5% respectively, while the difference of the leakage reduction rate is about 7.9%, 4.3% and 2.8%. The weight value γ can be used to get a good trade-off between LF gate number and leakage reduction rate.

REFERENCES

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

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

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

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

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

[6] 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.

[7] J. Kao, A. Chandrakasan, " MTCMOS Sequential Circuits," in Proc. of ESSDERC, Sept 2003.

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

[9] Y. Wang, H. Lin, R. Luo, H.Z. Yang, H. Wang, "Simultaneous Fine-grain Sleep Transistor Placement and Sizing for Leakage Optimization", in Proc. of ISQED 2006, pp.723-728.

[10] Y. Wang, Y.P. Liu, R. Luo, H.Z. Yang, H. Wang, "Two-phase Fine-grain Sleep Transistor Insertion Technique in Leakage Critical Circuits", accepted by ISLPED06.

[11] Y. Wang, H.Z. Yang, H. Wang, "Signal-path Level Dual-Vt Assignment for Leakage Power Reduction," in JCSC Vol. 15, No. 2 (2006), pp 197-216.

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

TABLE II. COMPARISON BETWEEN MULTI-OBJECT ST PLACEMENT (M-STP) AND ST PLACEMENT WITHOUT CONSIDERING THE LF GATE (STP-WO) [10]

ISCAS85 benchmark circuits	Original leakage current (pA)	Total gate number	0% circuit slowdown				3% circuit slowdown				5% circuit slowdown			
			I_{leak} (pA)		N_{LF}/N_{ST}		I_{leak} (pA)		N_{LF}/N_{ST}		I_{leak} (pA)		N_{LF}/N_{ST}	
			M-STP	STP-WO	M-STP	STP-WO	M-STP	STP-WO	M-STP	STP-WO	M-STP	STP-WO	M-STP	STP-WO
C432	4609.417	169	2954.6	1759.3	10/59	37/130	1243.37	463.7	4/111	31/151	586.6	205.5	3/132	19/157
C499	21374.953	204	15340.9	14479.8	**25/71**	**78/97**	1437.9	805.9	12/159	48/189	636.4	105.3	6/182	32/200
C880	9261.315	383	1292.3	619.2	11/315	105/352	541.4	232.2	6/359	68/370	255.7	126.1	3/370	46/375
C1355	11874.533	548	7827.5	6417.5	**40/248**	**206/308**	5604.5	5099.4	51/381	134/402	1570.5	945.2	8/512	64/535
C1908	23418.219	911	4535.9	2498.8	24/782	261/830	1040.2	590.0	8/862	171/878	357.6	224.5	5/895	83/900
C2670	35191.285	1279	1900.8	1356.6	22/1214	293/1235	833.1	364.6	3/1248	193/1264	328.0	161.8	2/1265	98/1274
C3540	40369.652	1699	2680.0	2060.4	41/1588	368/1617	1686.9	1020.4	11/1628	246/1658	521.8	270.4	4/1683	122/1690
C5315	56292.203	2329	7122.1	1660.9	14/2193	428/2253	5191.1	788.6	9/2270	294/2293	4284.5	433.8	7/2302	148/2312
C6288	40968.834	2447	10371.5	7427.8	170/1752	800/1948	3709.7	2545.6	81/2203	485/2282	1598.1	977.7	29/2350	213/2385
C7552	85523.934	3566	3877.0	3012.4	120/3378	937/3415	2067.8	1320.2	37/3461	577/3504	945.6	682.4	21/3524	256/3539
Average	N/A	N/A	**71.0%**	**78.9%**	**9.3%**	**37.1%**	**88.2%**	**92.5%**	**3.3%**	**19.8%**	**95.2%**	**98.0%**	**1.1%**	**9.9%**