Two-phase Fine-grain Sleep Transistor Insertion Technique in Leakage Critical Circuits

Yu Wang, Yongpan Liu, Rong Luo, Huazhong Yang, Hui Wang
E.E. Dept., Tsinghua University, Beijing, P.R.China
(8610)62772966
{wangyuu99, ypliu99}@mails.tsinghua.edu.cn

ABSTRACT
Multi-threshold CMOS is a valuable leakage reduction method in circuit standby mode. Reducing leakage current through fine-grain sleep transistor insertion (*FGSTI*) makes it easier to guarantee circuit functionality and improves circuit noise margins. In this paper, we first indicate the negligible dependence of ST size on the amount of leakage saving which makes the two-phase *FGSTI* reasonable based on our leakage current and delay models. Then we introduce a novel two-phase *FGSTI* technique: a) ST placement and b) ST sizing, which are formally modeled as two linear programming (LP) models respectively. Our experimental results show that the two-phase *FGSTI* technique can achieve 78.91%, 92.55%, 97.97% leakage saving when the circuit slowdown is 0%, 3%, 5% respectively. Comparing to the simultaneous ST placement and sizing method using mix integer linear programming (MLP) [1], our technique leads to on average 2% more leakage current reduction while at least 10X runtime saving since fewer variables and constraints with less approximation are used in the LP models. When the circuit slowdown is large enough to perform conventional fixed slowdown method, our technique can still achieve 75.48% ST area saving. Moreover, we show that when the circuit slowdown is 0%, it should be carefully considered to use *FGSTI* technique due to a large amount of leakage feedback gates.

Categories and Subject Descriptors:
J.6 [**Computer Aided Engineering**]: Computer aided design (CAD), B.6.3 [**Design Aids**]: Optimization

General Terms
Algorithms, Design

Keywords
Leakage current reduction, two-phase fine-grain sleep transistor insertion, mixed integer linear programming.

1. INTRODUCTION
With the development of the fabrication technology, leakage power dissipation has become comparable to switching power dissipation [2]. As we all know, the total power dissipation consists of dynamic power, short circuit power and leakage power. The behavior of the short circuit power dissipation remains at around 10% of the total power dissipation [3]. At the 90nm technology node, leakage power may make up 42% of total power [4]. Leakage power reduction techniques can be broadly categorized into two main categories [5]: process level and circuit level techniques. The circuit level techniques consist of adapt body bias [6], DVTS [7], input vector control [8], dual-V_t assignment [9-11] and Multi-Threshold CMOS (ST insertion) [1] [12-18]. Among these, Multi-Threshold CMOS (MTCMOS) technique is essentially placing a ST between the gates and the power/ground (P/G) net in a circuit in order to put it 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 is concluded as *block based ST insertion* (*BBSTI*) technique. In *BBSTI* techniques, all the gates in the block are assumed to have a fixed slowdown, so it is also called fixed slowdown method. The existing literatures on *BBSTI* techniques [12-16] present some details in clustering gates into blocks in order to optimize the leakage current and ST size. All these literatures focus on how to reduce the ST area penalty along with a remarkable leakage saving: [12] first gives out a mutual exclusion method; [13] [14] present several fast heuristic techniques for efficient gate clustering; [15] [16] propose a distributed sleep transistor network (DSTN) approach which assumes that all the sleep devices are connected to further reduce the area penalty.

Although *BBSTI* techniques greatly reduce the area penalty, they induce large ground bounce in the P/G network which has adverse effects on circuit speed and noise immunity [18]. What is more, ST size is determined by the worst case current of the clustering block which is quite difficult to determine without comprehensive simulation [12]. Thus it is harder to guarantee circuit functionality for large blocks with only one ST [17].

In recent years gate level ST insertion, which can be also called *fine-grain ST insertion* (*FGSTI*) technique [1] [17] [18] (as shown in figure 1 (a)) shows some advantages over the *BBSTI* technique (as shown in figure 1 (b)). 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 *FGSTI* technique leads to a smaller simultaneous switching current when the circuit changes between standby mode and active mode, thus improves circuit noise margins. Furthermore, better circuit slack utilization can be achieved as the slowdown of each gate is not fixed, and then leads to a further reduction of leakage and area. As shown in [18], *FGSTI* technique corresponds to an area penalty of roughly 5% using standard cell placement.

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

* This work is sponsored by the grants from National 863 project of China (No. 2005AA1Z1230) and NSFC (No. 60506010).

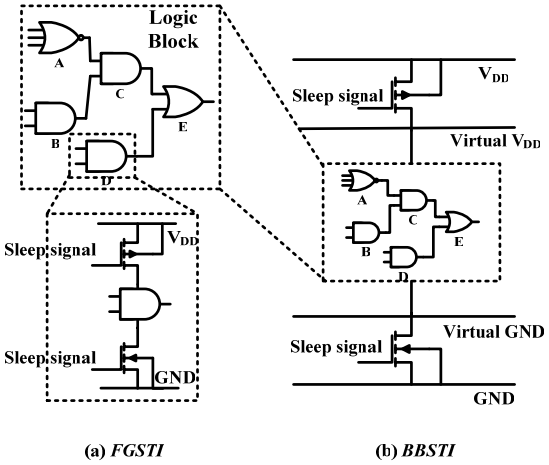

Figure. 1 *Fine-grain ST insertion* (*FGSTI*) vs *Block based ST insertion* (*BBSTI*)

In [17], a fine-grain MTCMOS design methodology and several design rules are proposed. The authors also make a comparison between local and global sleep devices. In [18], a selectively ST insertion methodology with better utilization of circuit slack is proposed in detail. They use a heuristic method to determine ST placement and sizing when circuit slowdown is 3%. When the circuit slowdown exceeds 5%, they solve an LP model to get optimal ST size. Although their method can give out an optimal sizing result, the heuristic step may lead to a local optimal point because ST placement is affected by ST sizing under their leakage current model assumption. In [1], a simultaneous fine-grain ST placement and sizing method using MLP is presented. The MLP model leads to an accurate result, but its computation time is considerably long.

This paper presents a novel two-phase *FGSTI* technique which has three contributions to leakage reduction:

(a) Simple leakage current and delay models of a single gate are proposed. Our model analysis is the first to provide the designer the negligible dependence of ST size on the amount of leakage saving which makes the two-phase *FGSTI* reasonable.

(b) The novel two-phase *FGSTI* technique: a) ST placement and b) ST sizing are modeled as two simple LP models respectively. Fewer variables and constraints with less approximation are used in our models, thus our two-phase *FGSTI* technique is more accurate and faster compared to simultaneous ST placement and sizing method [1]. Our ST placement can achieve an impressive leakage saving when the conventional fixed slowdown method can not be performed. Furthermore, if the circuit slowdown is large enough to use conventional fixed slowdown method, our ST sizing still leads to a much smaller total ST size.

(c) We show that when the circuit slowdown is 0%, it may be inappropriate to use *FGSTI* technique due to the usage of different type ST to avoid floating states. A large amount of buffers bring not only additional area, but also considerable dynamic power penalty.

The paper is organized as follows. In Section 2, our leakage current and delay models are given out and analyzed to prove the rationality of our two-phase *FGSTI* technique. The two-phase *FGSTI* technique 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

In this section, the leakage current and delay models used in our two-phase *FGSTI* technique are given out and analyzed to prove that a *FGSTI* design can be performed in two phases. ST is used with variable size which is decided by the process technology in our two-phase *FGSTI* design. 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.

2.1 Leakage current model

For the gates without ST, a leakage lookup table is created 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:

$$I_l^{w/o}(v) = \sum_{IN} I_l(v, IN) \times PB(v, IN) \quad (1)$$

Where $I_l(v, IN)$ and $PB(v, IN)$ are the leakage current and the probability of gate v under input pattern IN.

We simply use a linear model to represent leakage current $I_l^{ST}(v)$ based on HSPICE simulation results:

$$I_l^{ST}(v) = A(v) \times (W/L)_v \quad (2)$$

where $A(v)$ is constant and decided by the gate type. Here we assume all the input patterns have same probability and estimate every $A(v)$ for all the standard cells in the library. Consider two standard cells: NOR2XL and NAND4XL in the TSMC 0.18μm standard cell library, the largest error is about 52% as shown in table 1. The error of linear approximation may be neglected in *FGSTI* due to Law of large numbers [19] with the growing circuit size. As we will mention in Section 2.3, the influence of the linear model error on the *FGSTI* technique will be diminished by the large difference between leakage current of a gate with or without ST.

Table 1. Leakage current in NOR2XL and NAND4XL

	Leakage current in NOR2XL (fA) $A(v)$=1.60495			Leakage current in NAND4XL (fA) $A(v)$= 2.97335		
	Hspice	Our model	Error	Hspice	Our model	Error
w/o ST	14606.8	N/A	N/A	12261.3	N/A	N/A
(W/L) =2	5.3899	3.2099	-40.4%	10.19664	5.9467	-41.7%
(W/L) =4	5.9464	6.4198	7.96%	10.9738	11.8934	8.38%
(W/L) =8	8.5931	12.8396	49.4%	15.58464	23.7868	52.6%
(W/L) =16	27.6482	25.6792	-7.12%	51.37328	51.3733	-7.40%

2.2 Delay model

As shown in [20], gate delay is influenced by the ST insertion. The load dependent delay $d^{w/o}(v)$ of gate v without ST is given by:

$$d^{w/o}(v) = \frac{KC_L V_{DD}}{(V_{DD} - V_{THlow})^{\alpha}} \quad (3)$$

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 $d^{ST}(v)$ of gate v with ST can be expressed as:

$$d^{ST}(v) = \frac{KC_L V_{DD}}{(V_{DD} - 2V_x - V_{THlow})^{\alpha}} \quad (4)$$

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}. $\Delta D(v)$ is derived from the above equations:

$$\Delta d(v) = d^{ST}(v) - d^{w/o}(v) = \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d^{w/o}(v) \quad (5)$$

$I_{ON}(v)$ is the current flowing through ST in gate v during the active mode, which can be expressed as given by [18]:

$$I_{ON}(v) = \mu_n C_{ox}(W/L)_v((V_{DD} - V_{THhigh})V_x - \frac{V_x^2}{2}) \\ = \mu_n C_{ox}(W/L)_v(V_{DD} - V_{THhigh})V_x \quad (6)$$

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

$$V_x = \frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_v} \quad (7)$$

Refer to equation (3) and (4), V_x in gate v due to ST insertion can also be given out as:

$$V_x = \frac{1}{2}\left(1 - \left(\frac{d^{w/o}(v)}{d^{ST}(v)}\right)^{1/\alpha}\right)(V_{DD} - V_{THlow}) \quad (8)$$

2.3 Relationship between ST placement and sizing

From previous part, we find that a linear leakage current model may have an error as large as 50% comparing with the HSPICE simulation results. Refer to [18], the leakage current for a gate with ST is also modeled as a linear function from [21]:

$$I_l^{ST}(v) = \mu_n C_{ox}(W/L)_v e^{1.8} V_T^2 e^{\frac{V_{gs} - V_{THhigh}}{nV_T}}(1 - e^{\frac{-V_{ds}}{V_T}}) \quad (9)$$

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 ST size of gate v. Notice that their model is also linear by assuming parameters except $(W/L)_v$ are constant decided by process information and gate structure. Such a linear model will also consume comparative error as our leakage current model.

However, as we explore the leakage current model further, the leakage current of a gate without ST is much larger than that of a gate with ST as shown in Table 2, so that the error of the linear model can be neglected in the *FGSTI* procedure. In table 2, the leakage current of cells in the TSMC 0.18μm standard cell library under two different ST conditions: with ST or without ST are compared. As shown in table 1, the leakage current of a gate with ST become larger with a larger (W/L). Therefore, the largest leakage current of a gate with ST is derived by setting the (W/L) of a ST to 16 which is the maximum ratio of ST in our *FGSTI* technique.

Table 2. Leakage current comparison of standard cells (fA)

Cell Name	$I^{w/o}$	I^{ST}	$I^{w/o}/I^{ST}$	Cell Name	$I^{w/o}$	I^{ST}	$I^{w/o}/I^{ST}$
NAND2XL	14076.3	45.03	313	AND3X4	54900.3	53.4	1028
NAND2X4	84392.0	45.5	1854	BUFX4	80876.7	53.4	1513
INVXL	14213.2	36.9	388	NOR2X1	16261.1	27.6	516
NOR2XL	14606.8	27.6	528	CLKINVX4	38763.4	37.2	1043
XOR2XL	95853.9	53.4	1794	NAND4X4	72554.2	51.4	1411
NAND4XL	12261.2	51.4	239	AND2XL	26956.7	53.4	505
NAND3XL	14186.1	49.3	288	AND4XL	13768.6	53.4	258
AND2X4	60305.1	53.4	1129	OR4XL	33827.5	53.4	633
AND4X4	48899.1	53.4	915	CLKINV8	69175.0	37.7	1833

As shown in table 2, the leakage current difference is at least 238X. Referring to equation (5) and (7), the delay difference is less than 20% of the original gate delay under the same condition. 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 comparing with the original gate without ST. Also from table 1, the leakage current difference of a gate with different ST size is less than 1% of the original gate leakage. Hence, the leakage current variation range due to the change of ST size can be neglected since it is much smaller compare to the leakage saving of changing a gate's ST condition. In a word, although leakage current *is* reduced by sizing the ST, ST placement is not affected by ST sizing due to the large gap between their effects on leakage saving.

With technology scaling down, the leakage current difference may be smaller under different ST condition, but it will still be very large due to high V_{th} ST and stacking effect. Hence we can draw a conclusion that *the leakage reduction depends on where to insert ST and the leakage difference of each gate under different ST condition; while the area penalty is decided by the ST sizing procedure.*

We further assume that ST placement and sizing are *independent* in a *FGSTI* design. Therefore, we develop a two-phase *FGSTI* technique: first, ST placement can be performed to decide where to put the ST and achieve most of the leakage saving; and then ST sizing can be used to reduce the area overhead along with further leakage current reduction.

3. TWO-PHASE *FGSTI* TECHNIQUE

In this section two-phase *FGSTI* technique is modeled using linear programming. First ST placement phase shows how to place the ST as many as possible in order to reduce the total leakage, and then an optimal sizing method is given out for ST sizing phase to reduce the area overhead based on the ST placement information. At the end of this section, the simultaneous placement and sizing method [1] is briefly reviewed for comparison.

3.1 ST placement

A novel ST placement method is proposed that tries to maximize the leakage saving in the circuits through mixed integer linear programming (MLP). First, the object function for the total leakage current is constructed as below:

$$I(G) = \sum_{v \in V}\left(I_l^{w/o}(v) \times (1 - ST(v)) + I_l^{ST}(v) \times ST(v)\right) \quad (9)$$

where $ST(v)$ 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. As ST size is not considered, we choose the largest ST size $(W/L)_{\max}$ in equation (2) to obtain the minimum delay overhead. The leakage current of gate v with ST is given by:

$$I_l^{ST}(v) = A(v) \times (W/L)_{\max} \quad (10)$$

It can be derived that $I_l^{ST}(v)$ is a constant for each gate to simplify the MLP model further.

The timing constraints of $G(V, E)$ can be expressed as:

$$t_a(m) = 0 \qquad m \in PI \quad (11)$$

$$t_a(n) + d(n) \le T_{req} \qquad n \in PO \quad (12)$$

$$t_a(i) + d(i) \le t_a(j) \qquad \forall (i,j) \in E,\ i, j \in V \quad (13)$$

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 (5) and (7):

$$d(v) = d^{w/o}(v) + \Delta d(v) \times ST(v) \\ = d^{w/o}(v) + \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}}\right)^{-\alpha} - 1\right) d^{w/o}(v) \times ST(v) \quad (14) \\ = d^{w/o}(v) + \left(\left(1 - \frac{2\frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \frac{1}{(W/L)_{\max}}}{V_{DD} - V_{THlow}}\right)^{-\alpha} - 1\right) d^{w/o}(v) \times ST(v) \\ = d^{w/o}(v) + \Gamma d^{w/o}(v) \times ST(v)$$

where $d^{w/o}(v)$ and Γ are constant for each gate v. Similarly we choose the largest ST size $(W/L)_{\max}$ to get the minimum delay overhead.

Minimize:

$$I(G) = \sum_{v \in V} \left(I_l^{w/o}(v) \times (1 - ST(v)) + (A(v) \times (W/L)_{\max}) \times ST(v) \right)$$

Subject to:

{Timing constraints}

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

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

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

$$d(v) = d^{w/o}(v) + \Gamma d^{w/o}(v) \times ST(v) \qquad v \in V$$

{Variable bounds}

$ST(v)$ are binary variables

Figure 2. MLP model for leakage minimization through ST placement

The general form of our MLP model for ST placement is shown in figure 2. ST placement is similar as dual V_{th} assignment with fixed high and low V_{th} values, thereby it can also be solved by sensitive based heuristic algorithms which are previously dealing with dual Vth assignment [9-11].

3.2 Optimal ST sizing

After the ST condition for each gate v is decided, we use linear programming to get the optimal ST size. First the object function for optimal ST sizing is given out as below:

$$Area(ST) = \sum_{v \in V} \left((W/L)_v \times ST(v) \right) \tag{15}$$

where $ST(v)$ is a binary value given out in the ST placement phase; $(W/L)_v$ is a continuous variable. Because the transistor length for ST is assumed to be a constant: the minimum length, $(W/L)_v$ is used instead of $W \times L$ to represent the ST area. Moreover, the expression for $(W/L)_v$ can be derived from equation (7) and (8) as below:

$$\begin{aligned} (W/L)_v &= \frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \frac{1}{V_x} \\ &= \frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \left(\frac{1}{2} \left(1 - \left(\frac{d^{w/o}(v)}{d^{ST}(v)} \right)^{1/\alpha} \right) (V_{DD} - V_{THlow}) \right)^{-1} \end{aligned} \tag{16}$$

The timing constraints can also be expressed as equation (11), (12) and (13). The propagation delay $d^{ST}(v)$ of gate v with ST can be rewrite using equation (5) and (7) as:

$$\begin{aligned} d^{ST}(v) &= d^{w/o}(v) + \Delta d(v) \\ &= d^{w/o}(v) + \left(\left(1 - \frac{2V_x}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d^{w/o}(v) \\ &= d^{w/o}(v) + \left(\left(1 - \frac{2 \dfrac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \dfrac{1}{(W/L)_v}}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d^{w/o}(v) \end{aligned} \tag{17}$$

With a given boundary of $(W/L)_v$: $[(W/L)_{\min}, (W/L)_{\max}]$, the boundary of $d^{ST}(v)$ can be easily gained: $[\, d^{ST}_{\min}(v), d^{ST}_{\max}(v) \,]$ using equation (17). Consequently, the general form of our LP model for ST sizing is show in figure 3.

3.3 Simultaneous ST placement and sizing

The object function of simultaneous ST placement and sizing is very similar to ST placement as shown in equation (9):

$$I(G) = \sum_{v \in V} \left(I_l^{w/o}(v) \times (1 - ST(v)) + A(v) \times (W/L)_v \times ST(v) \right) \tag{18}$$

Minimize:

$$Area(ST) = \sum_{v \in V} \left(\frac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \left(\frac{1}{2} \left(1 - \left(\frac{d^{w/o}(v)}{d^{ST}(v)} \right)^{1/\alpha} \right) (V_{DD} - V_{THlow}) \right)^{-1} \times ST(v) \right)$$

Subject to:

{Timing constraints}

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

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

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

$$d(v) = d^{w/o}(v) + (d^{ST}(v) - d^{w/o}(v)) \times ST(v) \qquad v \in V$$

{Variable bounds}

$$d^{ST}_{\min}(v) \le d^{ST}(v) \le d^{ST}_{\max}(v)$$

Figure 3. LP model for optimal ST sizing

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

The timing constraints also follow equation (11), (12) and (13). Refer to equation (14), gate delay $d(v)$ for gate v can be derived as:

$$\begin{aligned} d(v) &= d^{w/o}(v) + \left(\left(1 - \frac{2 \dfrac{I_{ON}(v)}{\mu_n C_{ox}(V_{DD} - V_{THhigh})} \times \dfrac{1}{(W/L)_v}}{V_{DD} - V_{THlow}} \right)^{-\alpha} - 1 \right) d^{w/o}(v) \times ST(v) \\ &= d^{w/o}(v) + d^{w/o}(v) \times \Phi((W/L)_v) \times ST(v) \end{aligned} \tag{19}$$

As we can see from equation (18) and (19), this problem is actually a non-linear programming model. In [1], Taylor series expansion and piecewise linear approximation technique are used to get a mixed integer linear programming model. Some dummy variables are needed for linear approximation and more linearization constraints are added in the MLP model for each dummy variable. Unfortunately, the model size becomes extremely large with the increasing gate number in the circuit.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

4.1 Implementation

All ISCAS85 benchmark circuit netlists are synthesized using Synopsys Design Compiler and a TSMC 0.18μm standard cell library. A leakage current look up table of all the standard cells without ST is generated using HSPICE. In addition, every $A(v)$ in equation (2) for all the standard cells is estimated using HSPICE simulation results under different $(W/L)_v$. The values of various transistor parameters are taken from the TSMC 0.18μm process library. 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 specialized static timing analysis (STA) tool [10], and the MLP and LP models for ST placement and sizing are automatically generated. We use an LP solver named *lp_solve* [22] to solve the models.

We assume $1 \le (W/L)_v \le 16$, corresponding to a least delay variance of 6% if ST is assigned to every gate in the circuit. We perform our two-phase *FGSTI* technique by first using the MLP model to get $ST(v)$ for all the gates in the circuit and then solve the LP model to get the optimal $(W/L)_v$ based on the results of $ST(v)$. The MLP model to simultaneously determine ST placement and sizing are also solved using the same LP solver under a same set of parameters for comparison with the two-phase *FGSTI* technique.

4.2 Results for two-phase *FGSTI* technique

For 0%, 3%, 5% circuit slowdown, we can not get a valid solution from conventional fixed slowdown method [12]. Thus the

leakage current saving for 0%, 3%, 5% circuit slowdown are compared between our two-phase *FGSTI* technique and MLP method [1]. As shown in table 3, our two-phase *FGSTI* technique can achieve 78.91% leakage saving even when the circuit slowdown is 0%. When the circuit slowdown is 3%, 5%, the leakage saving of our two-phase *FGSTI* technique is 92.55%, 97.97% respectively. Because of less approximation in ST placement phase, more ST's can be assigned to different gates and additional leakage saving is achieved. The leakage saving is about on average 2% more than the MLP method [1].

In table 4, we show that our two-phase *FGSTI* technique can achieve a very impressive runtime saving. As the LP model for ST sizing only need seconds to solve, the runtime saving is largely due to two reasons: one is the two-phase procedure of *FGSTI* technique and the other is less variables and constraints used in MLP model for ST placement. For circuit C432, there are only 271 constraints and 338 variables in our MLP model for ST placement; however, in [1] there are 2975 constraints and 1183 variables. Although the MLP problem still need a long time to solve, as we can see from some of the benchmark, especially the small ones, our two-phase *FGSTI* technique can achieve at least 10X runtime saving. We only list the results of 4 benchmarks, because other benchmarks take hours to get the optimal results. The stopping time criteria is set to 4 hours for larger circuits. Heuristic algorithms can get near optimal results with a very fast speed, but as we all know the heuristic may lead to local optimal and can not guarantee the optimality of the result; thus we can use the results of LP models as a reference.

When the circuit slowdown is larger than 6%, ST can be assigned to all the gates in the circuits, the two-phase procedure of *FGSTI* technique is changed to one: ST sizing, while ignoring the ST placement. Our LP model for ST sizing leads to a same result as optimal sizing method in [16]. We compare the area penalty with the fixed slowdown method and the MLP method in table 5. With 7% circuit slowdown, our ST sizing LP model causes 75.48% ST area saving compared to fixed slowdown method and the result is almost the same with MLP method. In table 5, ST area is calculated using equation (15), just summing up all the $(W/L)_v$, since the length of ST is a constant.

Table 5. ST sizing results comparison with MLP and fixed slowdown method

ISCAS85 benchmark circuits	7% circuit slowdown			9% circuit slowdown		
	ST sizing	MLP	Fixed slowdown	ST sizing	MLP	Fixed slowdown
C432	714	714	2317.72	596	597	1802.67
C499	1146	1146	2797.72	959	959	2176.01
C880	876	876	5252.58	780	780	4085.35
C1355	3365	3364.	7515.44	2719	2720	5845.35
C1908	2355	2355	12493.73	2081	2081	9717.36
C2670	2087	2088	17540.59	1937	1937	13642.71
C3540	3371	3371	23300.60	3160	3160	18122.72
C5315	4292	4293	31940.60	3917	3918	24842.74
C6288	11733	11733	33558.89	9610	9611	26101.41
C7552	8980	8981	48905.19	8197	8197	38037.45
Area saving	75.48%	N/A	N/A	73.0%	N/A	N/A

4.3 ST type Consideration

From table 3, when the circuit slowdown is below 6%, not all the gates in the circuit can be assigned with ST. FGSTI technique can cause a gate with ST to drive a gate without ST which leads to a floating state at the output of the gate with ST and large power dissipation in the gate without ST. As mentioned in [18], leakage feedback gate structure [23] shown in figure 4 is used in order to avoid the floating states. We assume that the leakage feedback structure can achieve the same delay as the normal ST insertion with a larger area and dynamic power consumption penalty. We examine all the gates with ST in the circuits and find out how many gates with ST should be changed into leakage feedback structure.

In table 6, when the circuit slowdown is 0%, about 82.75% of the total gates can change into gate with ST, and about 37.10% of the gates with ST should change into leakage feed back structure. When circuit slowdown is 3% and 5%, about 93.47% and 98.03% of the total gates can be changed into gates with ST, and only 19.78% and 9.90% of them should be changed into leakage feedback structure respectively. When the circuit slowdown is 0%, some of the benchmarks, such as C499, C1355, need to change 80.41% and 66.41% of original ST into leakage feedback struct-

Table 3. Leakage current comparison between two-phase *FGSTI* and MLP method

ISCAS85 benchmark circuits	Original I_{leak} (pA)	Total gate Num.	0% circuit slowdown				3% circuit slowdown				5% circuit slowdown			
			Two-phase FGSTI		MLP		Two-phase FGSTI		MLP		Two-phase FGSTI		MLP	
			I_{leak} (pA)	ST gate Num.	I_{leak} (pA)	ST gate Num.	I_{leak} (pA)	ST gate Num.	I_{leak} (pA)	ST gate Num.	I_{leak} (pA)	ST gate Num.	I_{leak} (pA)	ST gate Num.
C432	4609.417	169	1759.279	130	1964.911	127	463.719	151	463.719	151	205.459	157	205.459	157
C499	21374.953	204	14479.83	97	14587.294	101	805.901	189	1451.785	164	105.253	200	757.367	189
C880	9261.315	383	619.241	352	619.241	352	232.245	370	364.903	365	126.149	375	126.149	375
C1355	11874.533	548	6417.456	308	6712.258	287	5099.370	402	5220.438	386	945.191	535	4382.772	417
C1908	23418.219	911	2498.797	830	3177.773	831	590.002	878	1296.372	882	224.510	900	258.150	900
C2670	35191.285	1279	1356.567	1235	1382.020	1235	364.575	1264	667.634	1260	161.849	1274	269.792	1270
C3540	40369.652	1699	2060.401	1617	2251.211	1612	1020.377	1658	1558.656	1637	270.415	1690	611.115	1675
C5315	56292.203	2329	1660.947	2253	1841.916	2254	788.566	2293	1025.466	2283	433.763	2312	593.592	2305
C6288	40968.834	2447	7427.753	1948	8083.815	1903	2545.587	2282	3042.084	2248	977.670	2385	1088.476	2382
C7552	85523.934	3566	3012.412	3415	4190.660	3385	1320.156	3504	2004.873	3471	682.362	3539	975.917	3519
Leakage saving	N/A	N/A	**78.91%**	N/A	77.49%	N/A	**92.55%**	N/A	91.24%	N/A	**97.97%**	N/A	94.55%	N/A
Additional Leakage saving (MLP-two-phase)/MLP			6.31%				14.95%				62.75%			

Table 4. Runtime comparison between two-phase *FGSTI* and MLP method (Time in s)

ISCAS85 benchmark circuits	0% circuit slowdown				3% circuit slowdown				5% circuit slowdown			
	Two-phase FGSTI			MLP	Two-phase FGSTI			MLP	Two-phase FGSTI			MLP
	ST placement	ST sizing	Total		ST placement	ST sizing	Total		ST placement	ST sizing	Total	
C432	0.491	1.234	1.725	32.047	1.502	1.593	3.095	1905.094	0.551	1.625	2.176	635.016
C499	3.856	1.454	5.31	75.0	2400.802	2.390	2403.19	154825.17	23.303	2.938	26.241	99610.59
C880	0.351	7.619	7.97	134.109	0.501	6.578	7.079	2973.734	0.321	6.344	6.665	958.766
C2670	43.61	107.797	151.407	22121.922	42.532	221.734	264.266	27438.11	4.25	176.453	180.703	57462.64

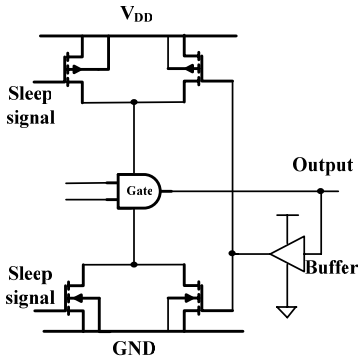

Figure 4. Leakage feedback structure

ure. This will lead to a large area increasing due to large number of high V_{th} feedback buffers. Therefore, if *FGSTI* technique will be used when there is no circuit slowdown, the penalty of using leakage feedback structure should be carefully examined because of the extra area and dynamic power consumption.

Table 6. Consideration of different ST type

ISCAS85 benchmark circuits	Total gate number	0% circuit slowdown		3% circuit slowdown		5% circuit slowdown	
		Total ST number	Leakage feed-back	Total ST number	Leakage feed-back	Total ST number	Leakage feed-back
C432	169	130	37	151	31	157	19
C499	204	97	78	189	48	200	32
C880	383	352	105	370	68	375	46
C1355	548	308	206	402	134	535	64
C1908	911	830	261	878	171	900	83
C2670	1279	1235	293	1264	193	1274	98
C3540	1699	1617	368	1658	246	1690	122
C5315	2329	2253	428	2293	294	2312	148
C6288	2447	1948	800	2282	485	2385	213
C7552	3566	3415	937	3504	577	3539	256
Average	N/A	82.75%	37.10%	93.47%	19.78%	98.03%	9.90%

5. Conclusions

In this paper, we present a novel two-phase *FGSTI* technique to reduce the leakage current using MTCMOS scheme. Simple leakage current and delay models for our two-phase *FGSTI* technique are proposed and analyzed to prove the rationality of our method. ST placement and sizing are modeled using MLP and LP models respectively. Our experimental results show that the two-phase *FGSTI* technique can achieve 78.91%, 92.55%, 97.97% leakage saving when the circuit slow down is 0%, 3%, 5% respectively. Moreover, two-phase *FGSTI* technique leads to 2% more leakage saving and at least 10X runtime saving comparing with simultaneous ST placement and sizing method using MLP. When the circuit slowdown is larger than 6%, the two-phase *FGSTI* can achieve 75.48% ST area saving comparing with fixed slowdown method. In conclusion, two-phase *FGSTI* technique is reasonable from our results. However, we show that the penalty of using leakage feedback structure during *FGSTI* technique should be carefully examined when the circuit slowdown is below 6%.

There are still some unsolved problems in *FGSTI* technique as our future work. Fast heuristic algorithms are needed for ST placement phase because the MLP model is very time consuming and can not handle large circuits. Furthermore, the detailed comparison between *FGSTI* and *BBSTI* techniques should be carefully examined in the physical level, such as place and routing penalty.

6. Reference

[1] Y. Wang, H. Lin, H.Z. Yang, R. Luo, H. Wang, "Simultaneous Fine-grain Sleep Transistor Placement and Sizing for Leakage Optimization," in *Proc. of ISQED'06*, 2006, pp. 723-728.

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

[3] 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 *Proc. of VLSI Design*, 2001, pp. 248 - 253.

[4] J. Kao, S. Narendra, A. Chandrakasan, "Subthreshold Leakage modeling and reduction techniques", in *Proc. 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 *Proc. of the IEEE*, Vol. 91, No. 2, Februray 2003 pp 305 – 327.

[6] S. Narendra et.al, "Forward body bias for microprocessors in 130-nm technology generation and beyond", in *IEEE JSSC*, Vol. 38 , No. 5 ,May 2003 pp. 696 - 701.

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

[8] S. Mukhopadhyay et. al., "Gate Leakage Reduction for Scaled Devices Using Transistor Stacking", in *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", in *IEEE TVLSI*, Vol.2. 17, NO. 1, 1999, pp. 16-24.

[10] Y. Wang, H.Z. Yang, H. Wang, "Signal-path Level Dual-Vt Assignment for Leakage Power Reduction," in Journal of Circuits, System and Computers, 2006, Vol. 15, No. 2, pp:197-216.

[11] Qi Wang; Vrudhula, S.B.K.; "Algorithms for minimizing standby power in deep submicrometer, dual-Vt CMOS circuits," in *IEEE TCAD*, Vol: 21 ,Issue: 3 ,March 2002 pp. 306 - 318 .

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

[13] 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.

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

[15] C. Long; L. He; "Distributed sleep transistors network for power reduction" in *Proc. of DAC*, 2-6 June 2003 pp. 181 – 186.

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

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

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

[19] W. Feller, "The Strong Law of Large Numbers," in An Introduction to Probability Theory and Its Applications, Vol.1, 3rd ed. New York: Wiley, pp.243-245,1968.

[20] 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.

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

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

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