Active Mode Leakage Reduction Using Fine-Grained Forward Body Biasing Strategy

Vishal Khandelwal and Ankur Srivastava
Department of Electrical and Computer Engineering,
University of Maryland at College Park.
{vishalk,ankurs}@glue.umd.edu

ABSTRACT

Leakage power minimization has become an important issue with technology scaling. Variable threshold voltage schemes have become popular for standby power reduction. In this work we look at another emerging aspect of this potent problem which is leakage power reduction in active mode of operation. In gate level circuits, a large number of gates are not switching in active mode at any given point in time but nevertheless are consuming leakage power. We propose a fine-grained Forward Body Biasing (FBB) Scheme for active mode leakage power reduction in gate level circuits without any delay penalty. Our results show that our optimal polynomial time FBB allocation scheme results in 70.2% *reduction in leakage currents. We also present a novel placement-driven FBB allocation algorithm that effectively reduces the area penalty using the post-placement area slack and results in* 39.7%*,* 64.7% *and* 67.1% *reduction in leakage currents for* 0%*,* 4% *and* 8% *area slack respectively.*

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 Power Optimization, Forward Body Biasing, Standard Cell Design

1. INTRODUCTION

In recent years, technology scaling has increased the role of leakage power in the overall power consumption of circuits. Supply voltage reduction is a widely accepted methodology for reducing dynamic power, but it has an adverse effect on circuit performance. To maintain high performance, the threshold voltage V_{th} must also be scaled down which causes an exponential increase in the sub-threshold leakage currents. Threshold voltage control through body biasing has been proposed in [6, 4, 12] as an effective technique to reduce leakage currents in deep-sub micron technologies.

In this paper, we propose a fine-grained Forward Body Biasing (FBB) scheme for leakage power reduction in the active mode for gate level circuits. Previous literature discuss the use of high V_{th} devices and the application of FBB to achieve high performance in active mode [6, 4, 12]. But no existing literature discusses the application of fine grained FBB methodology for leakage power minimization of gate level circuits in active mode.

We propose to fabricate all devices using the 2-D halo doping profile to obtain super high V_{th}. In active mode, high performance is obtained by applying FBB to these high V_{th} devices as opposed to using dual-V_{th} technology. We propose a polynomial-time optimal FBB allocation formulation. The key idea is to identify gates that are non-critical and increase their delay more than that of the critical gates (using appropriate FBB values) such that the overall leakage is minimized and the delay constraint is satisfied. In order to limit the number of distinct FBB values because of physical limitations, we also propose a clustering based FBB allocation algorithm. Furthermore, in order to minimize the area penalty associated with the scheme, we present a novel placement-driven fine-grained FBB allocation algorithm that utilizes the available area slack in each placement row for clustering.

The rest of the paper is organized as follows: section 2 discusses the preliminaries associated with this work, section 3 describes the fine-grained gate level FBB allocation scheme, section 4 describes the clustering based FBB allocation algorithm and section 5 presents our novel standard-cell based placement-driven FBB allocation scheme. Section 6 discusses our experimental results and section 7 contains the conclusions drawn from this work.

2. PRELIMINARIES

2.1 Device Equations

Let us understand the relevant gate level equations used in this approach. The delay d^i of a gate i can be expressed as

$$d_i = \frac{K_i C_{L_i} V_{dd}}{(V_{dd} - V_{th})^{\alpha}} \quad (1)$$

where K_i is the proportionality constant, C_{L_i} is the load capacitance at the gate output, V_{th} is the threshold voltage, $V_{dd} = 1.8$ V and α is the velocity saturation index (≈ 1.3 in 0.18-μm CMOS technology).

The sub-threshold leakage current I_{leak} of a gate is expressed as [11]

$$I_{leak} = \mu_n C_{ox} \frac{W_{eff}}{L_{eff}} e^{1.8} V_T^2 e^{\frac{V_{gs}-V_{th}}{nV_T}} (1 - e^{\frac{-V_{ds}}{V_T}}) \quad (2)$$

where μ_n is the N-mobility , C_{ox} is the oxide capacitance,

ISLPED'04, August 9–11, 2004, Newport Beach, California, USA.

V_{th} is the threshold voltage, V_T is the thermal voltage = 26mV and n is the sub-threshold swing parameter.

The dependence of threshold voltage on the FBB voltage for a transistor can be expressed as [3]

$$V_{th} = V_{th_0} + \gamma[(2\phi_f + V_{SB})^{1/2} - 2{\phi_f}^{1/2}] \quad (3)$$

where V_{th_0} is the threshold voltage at zero reverse bias voltage. V_{SB} is the reverse bias voltage between the source and the substrate (body) of the transistor, γ is a process parameter and ϕ_f is a physical parameter.

Equation 1 establishes a relation between delay of a gate d_i and V_{th}. By replacing V_{th} in equation 2 in terms of d_i (using equation 1), we get a dependence between the gate delay and gate leakage. Thus, a range of gate delays would correspond to a range of gate leakage values. The final relation between leakage and delay can be expressed as

$$I_{leak} = \mu_n C_{ox}(\frac{W_{eff}}{L_{eff}})e^{1.8}V_T^2(1-e^{\frac{-V_{ds}}{V_T}})e^{\frac{V_{gs}}{nV_T}}e^{\frac{KC_L{V_{dd}}^{1/\alpha}}{nV_T{d_i}^{1/\alpha}} - \frac{V_{dd}}{nV_T}} \quad (4)$$

Equation 3 establishes the relation between the FBB and the threshold voltage V_{th} for a gate. We can utilize this relation and fabricate all devices at a high threshold voltage. During active mode of operation, FBB is applied to reduce their threshold voltage to meet the performance constraints. Hence, the devices are inherently in low leakage state and during active operation, their leakage is increased to the minimal possible level while meeting the performance constraints. This can be achieved by controlling the forward body bias of the corresponding delay critical gates.

2.2 Body Biasing Schemes and Device Considerations

There are two popular body biasing schemes, namely Forward Body Biasing (FBB) and Reverse Body Biasing (RBB). RBB schemes have been used to reduce sub-threshold leakage through body-effect while meeting the performance constraints in active mode by switching to Zero Body Bias (ZBB) [2, 9, 5]. In this work we have considered the FBB scheme since it has been seen that the BTBT component of leakage increases due to decrease in the body coefficient in scaled technologies with shrinking dimensions when using the RBB scheme. In [4], the authors show that FBB can be used to improve performance and robustness of leakage-sensitive circuits. V_t roll-off, Drain Induced Barrier Lowering (DIBL) as well as short channel effects which are more prominent in low V_{th} devices, are countered by using high V_{th} devices with FBB [4]. Previous work [4, 12] has shown that a high V_{th} device can achieve high drive current in active mode using a FBB.

In [6], various device optimization considerations have been discussed. Device engineering and FBB can be used to simultaneously achieve low leakage power and high drive currents. We propose to use 2-D super-halo doping profile devices [14] for fabricating high V_{th} devices in our circuits. There are three main components of leakage current, namely sub-threshold leakage, gate leakage and BTBT leakage. As a result of increasing the peak halo doping, there is an exponential increase in the BTBT leakage currents. [6] shows that gate work function engineering can be used to obtain super high V_{th} devices without affecting the BTBT leakage. We will consider only sub-threshold leakage as our optimization objective as [6] shows that gate leakage forms only a small part of the total leakage in this case. We have also assumed that for every gate in the design, both N-MOS and P-MOS transistors operate at the same threshold voltage after applying the assigned FBB values.

2.3 Motivation

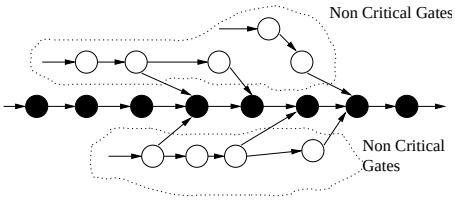

Figure 1: Motivational Example

It is known that in active mode, most of the gates in any circuit are not switching for a significant period of time. These gates are continuously leaking power even in active mode. Hence, in order to minimize the power consumption of circuits in active mode we also need to address the problem of leakage power minimization. In this work we try to address this problem by using super high V_{th} devices during circuit fabrication and then applying appropriate FBB values to meet the active mode performance constraints.

Let us consider a motivational example as shown in figure 1. The nodes represent the gates in the circuit. The delay-critical nodes are represented by solid circles. We can see from figure 1, that there are a large number of non-critical nodes which can bear more delay penalty while still satisfying the required time constraint. Therefore, we can use this extra slack at non-critical nodes during FBB allocation and increase the savings in leakage power. Our strategy proposes that all devices are fabricated using super high V_{th} doping profiles. These devices have low leakage but higher delays. Depending on the delay criticality of each gate, we can allocate appropriate FBB values to lower its threshold voltage and therefore its delay. Hence, we can allocate appropriate FBB values to each gate such that the delay constraints are met while keeping the penalty in leakage power as low as possible.

3. FINE-GRAINED GATE LEVEL FBB SCHEME

In this section we present our fine-grained gate level FBB scheme.

3.1 Modeling

Let us first consider the gate level delay and leakage models used in this work.

3.1.1 Gate Delay Model

The propagation delay and output transition formula derived from the short channel MOSFET model [13] for a CMOS inverter is accurate in predicting the circuit behavior of deep sub-micron designs. It has been found that N series-connected MOSFET (SCMS) would show less than N times the delay of a single MOSFET for deep sub-micron designs. We can represent this as:

$$\frac{delay(SCMS)}{delay(Invertor)} = 1 + \zeta(N-1) \quad (5)$$

where ζ is a technology dependent parameter ($0 < \zeta < 1$ for deep sub-micron technologies). We use this model as our gate delay model. We take the 0.18 micron inverter as the

base case and use equation 5 to calculate the delays of all other gates in the technology library.

3.1.2 Sub-threshold Leakage Model

The standby current of a CMOS network can be expressed as a function of the current of a single CMOS transistor [10]. The ratio of the standby currents for single stacked transistor I_{s1}, double stacked transistor I_{s2} and triple stacked transistor I_{s3} can be expressed as:

$$I_{s1} : I_{s2} : I_{s3} = 1.8exp(\eta V_{dd}/nV_T) : 1.8 : 1 \quad (6)$$

where $0.04 \leq \eta \leq 1$, $1.4 \leq n \leq 1.5$, $V_{dd} = 1.8$ V and $V_T = 26$ mV. Putting these values in equation 6, we get

$$I_{s1} : I_{s2} : I_{s3} = 1 : 0.234 : 0.13 \quad (7)$$

We take the 0.18 micron inverter as the base case and use the ratio from equation 7 to calculate the leakage current of all other gates in the technology library. This defines our gate sub-threshold leakage estimation model.

3.2 Optimal Fine-Grained FBB Allocation

We address the problem of fine-grained FBB allocation by presenting a novel polynomial time optimal algorithm that tries to maximize the utilization of the existing slack in the circuit for savings in leakage power in active mode. The FBB allocation problem can be defined as follows:

Given a gate level circuit, the arrival time at each primary input and a required time constraint at each of the primary outputs, the problem is to optimally allocate FBB values to each gate for minimal leakage in active mode while satisfying the overall delay constraints on the circuit.

We have a Linear-Programming formulation that performs the optimal FBB allocation in polynomial time. We exploit the dependence of gate delay and gate leakage in equations 1 and 2 on the threshold voltage of the gate. Essentially, we budget the delay of each gate such that the total leakage is minimized and the delay constraint is satisfied. The dependence of FBB and the effective threshold voltage at each gate has been discussed in sub-section 2.1 as given by equation 3. A range of possible FBB values for each gate in the circuit would impose a range on the threshold voltage as well as the delay of each gate. Hence allocating a particular delay value to a gate implies a threshold voltage assignment which in turn implies a FBB allocation to that gate. We have assumed that it is possible for us to assign FBB values to each gate independently. We optimally assign delay budgets to each gate in the circuit such that our objective function, which is the sum of the leakage power over all gates is minimized.

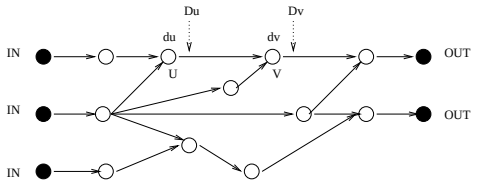

Figure 2: DAG representation

We represent our gate level circuit as a DAG, $G(V, E)$ as shown in figure 2. Each node in the DAG represents a gate in the circuit. We add a dummy IN node before each of the primary inputs which are shown as the black nodes marked IN in figure 2. We also add a similar dummy node OUT after each of the primary outputs which are shown as the black nodes marked OUT in figure 2. For each node u, we associate a variable d_u which represents the delay of that node. We also associate another variable D_u with each node which represents the arrival time at the output of node u. Now we consider two nodes u and v as shown in figure 2. Their corresponding variables have also been shown in the figure. The timing constraints on G(V,E) can be modeled as

$$\begin{aligned}
d_v - D_v + D_u &\leq 0 && \forall\ e(u,v) \in E && (8)\\
d^i_{min} \leq\ d_i &\leq d^i_{max} && \forall\ vertex\ i \in V, i \in S && (9)\\
D^i_{IN} &= T^i_{arrival} && \forall\ vertex\ i \in IN && (10)\\
D^i_{OUT} &\leq T^i_{con} && \forall\ vertex\ i \in OUT && (11)\\
d^i_{IN} &= 0 && \forall\ vertex\ i \in IN && (12)\\
d^i_{OUT} &= 0 && \forall\ vertex\ i \in OUT && (13)
\end{aligned}$$

For all the IN nodes, the D_{IN} values have been set to the corresponding arrival time values $T_{arrival}$ for the signals. The delay of the IN nodes denoted by d_{IN} have been set to zero. Similarly for the OUT nodes, their delay d_{OUT} values have been set to zero and the corresponding D_{OUT} values have been set to be less than or equal to the required time constraint T_{con} at the corresponding primary output node. The above LP formulation assigns delay budgets to all the gates of the circuit such that the utilization of the available slack is maximum. The range of possible FBB values that can be assigned is imposed as the corresponding range of delay budgets for each gate, which is denoted by $[d_{min}, d_{max}]$. The objective of optimal FBB allocation is to minimize the total leakage power of the circuit in active mode which can be represented as

$$min(\Sigma_{i \in V} P_i) = min(\Sigma_{i \in V} V_{dd} I^i_{leak}) \quad (14)$$

As illustrated before the dependence between gate leakage and gate delay is given as follows

$$I_{leak} = \mu_n C_{ox} (\frac{W_{eff}}{L_{eff}}) e^{1.8} V_T^2 (1 - e^{\frac{-V_{ds}}{V_T}}) e^{\frac{V_{gs}}{nV_T}} e^{\frac{KC_L {V_{dd}}^{1/\alpha}}{nV_T {d_i}^{1/\alpha}} - \frac{V_{dd}}{nV_T}} \quad (15)$$

Theorem: Optimal FBB Allocation Algorithm is polynomial time solvable

Proof: A LP formulation is polynomial time solvable if the objective function is a separable convex function under a set of linear constraints [7]. Let us consider the minimization objective function in equation 14. We can see that it is a separable function since each term P_i depends only on the variable d_i as seen from equation 15. Hence, grouping together all the other constant symbols into constants K_1 and K_2, we can represent P_i from equations 14 and 15 as

$$P_i = K_1 e^{\frac{K_2}{{d_i}^{1/\alpha}}} \quad (16)$$

where K_1 and K_2 are positive constants. We now try to prove the convexity of the objective function P_i. Since $1/\alpha = 0.77$, ${d_i}^{1/\alpha}$ is a concave function. Also, the delay d_i is positive by definition, hence ${d_i}^{1/\alpha}$ is a positive concave function. Therefore, $1/{d_i}^{1/\alpha}$ is a convex function (inverse of a positive concave function). Since K_2 is a positive constant, $K_2/{d_i}^{1/\alpha}$ is also a positive convex function. This implies that $e^{K_2/{d_i}^{1/\alpha}}$ is also a convex function. Since K_1 is

a positive constant, $P_i = K_1 e^{K_2/{d_i}^{1/\alpha}}$ is a convex function. Thereby we have shown that the objective function is convex separable. Hence, according to the result from [7], Optimal FBB Allocation Algorithm is polynomial time solvable.

4. OPTIMAL CLUSTER-BASED FBB ALLOCATION

We now consider another variation of the fine-grained FBB allocation problem. In section 3.2, we assumed that we could assign each gate an independent FBB value. From a fabrication point of view, routing these large number of voltage supply lines to the substrate of the transistors might not be feasible. Let us suppose that there is a limit to the maximum number of distinct FBB values that can be allocated to the gate in the circuit. The problem can be defined as follows:

Given a gate level circuit, a clustering of gates such that each gate within the same cluster has the same FBB value, the arrival time at each primary input and a required time constraint at each of the primary outputs, the problem is to optimally allocate FBB values to each cluster of gates for minimal leakage in active mode while satisfying the overall delay constraints on the circuit.

Since the gates within the same cluster are assigned the same FBB value, their threshold voltages are the same. From delay equation 1, we can see that if two gates have the same threshold voltage, then their delays are in a constant ratio. This is an additional constraint thats needs to be added to the LP formulations proposed in section 3.2. Let there be n_k gates in cluster k. The following equations show that the delays of these gates will be in a fixed ratio since there threshold voltages are the same.

$$d_{1_k} = \frac{K_{1_k} C_{L1_k} V_{dd}}{(V_{dd} - V_{th})^\alpha} \quad (17)$$

$$d_{2_k} = \frac{K_{2_k} C_{L2_k} V_{dd}}{(V_{dd} - V_{th})^\alpha} \quad (18)$$

$$..... = ... \quad (19)$$

$$d_{n_k} = \frac{K_{n_k} C_{Ln_k} V_{dd}}{(V_{dd} - V_{th})^\alpha} \quad (20)$$

$$\frac{d_{1_k}}{K_{1_k} C_{L1_k}} = \frac{d_{2_k}}{K_{2_k} C_{L2_k}} = ... = \frac{d_{n_k}}{K_{n_k} C_{Ln_k}} \quad (21)$$

Thus we can add the following additional constraints for each cluster k in the circuit to the LP formulation.

$$\frac{d_{1_k}}{K_{1_k} C_{L1_k}} = \frac{d_{2_k}}{K_{2_k} C_{L2_k}} \quad (22)$$

$$\frac{d_{2_k}}{K_{2_k} C_{L2_k}} = \frac{d_{3_k}}{K_{3_k} C_{L3_k}} \quad (23)$$

$$..... = ... \quad (24)$$

$$\frac{d_{n-1_k}}{K_{n-1_k} C_{Ln-1_k}} = \frac{d_{n_k}}{K_{n_k} C_{Ln_k}} \quad (25)$$

We note that these are *linear* constraints as well. Hence, Cluster based FBB allocation is optimally solvable in polynomial time.

5. PLACEMENT DRIVEN FBB ALLOCATION

In this section, we present a novel placement driven FBB allocation algorithm. The clustering scheme discussed in the previous section does not take into account the area overhead in fabrication associated with clustering gates. Let us suppose that we assume that the N-MOS and P-MOS transistors used in our designs are fabricated using p-wells and n-wells on a silicon substrate respectively. If two gates are in the same cluster, the N-MOS transistors (or the P-MOS) have the same FBB value and can be fabricated in the same p-well (or n-well). If two gates are not in the same cluster, the N-MOS transistors (or the P-MOS) could have different FBB value and cannot be fabricated in the same p-well (or n-well). We would ideally like to cluster together gates that have been placed adjacently, such that the area overhead with creating these new n/p wells is minimum. If we consider clustering gates based on their placement information, we will cluster neighboring gates whose transistors can be fabricated in the same n/p wells. If gates that have been placed far apart are clustered together to get the same FBB value, they may need to be fabricated in different n/p wells because their neighbors might have different FBB values. Hence there is a large area cost associated with such a clustering scheme. Therefore, we should consider the post placement information of the gates and cluster them taking into account their physical proximity. The additional area overhead that is associated in making clusters of gates in each placement row is accounted for using the available area slack in the corresponding row. Hence, we minimize the penalty in the increase in the overall area of the chip.

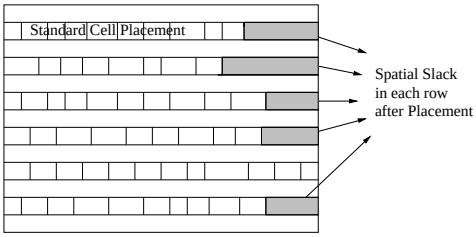

Figure 3: Standard Cell Placement

Let us consider a standard cell based placement scheme that places the circuit into rows as shown in figure 3. All standard cells have the same height but vary in width. We note here that there is some unused area slack in each row which is shown by the shaded areas in figure 3. Hence this available extra area can be used to cluster the standard cells post placement such that they can be allocated one value of FBB. We know that since these cells in each placement row are fabricated as neighbors, if two neighbors need to be given a different FBB value, then their N-MOS (and P-MOS) need to be fabricated in different n-well (p-well). This essentially amounts to an area overhead every time the FBB values are changed within the same placement row. Hence we can formally define this problem as

Given a standard cell row based placement of the gate level circuit,the arrival time at each primary input and a required time constraint at each of the primary outputs, each row has to be split into clusters of adjacent standard cells such that the area overhead is within the spatial slack available for that row and each cluster is assigned a FBB value such that the leakage of the circuit is minimized and the overall delay constraints are satisfied.

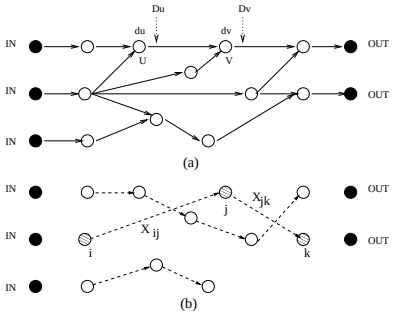

Figure 4: Modified DAG Representation

The problem can be modeled by adding a set of extra edges E' to the original DAG $G(V, E)$ shown in figure 4(a). Each row can be represented by a path of new edges denoting the adjacency in their placement of the standard cells as shown in figure 4(b). An edge $e(i, j)$ is added to E' for every pair of adjacently placed cells i and j. We note here that if we have a placement with N rows of standard cells, we will now have N additional *disjoint* paths in the DAG. For any two adjacent nodes on any path p, as shown by the shaded nodes i and j in figure 4(b), we associate three binary variable X^p_{ij}, Q^p_{ij} and Q^p_{ji} with edge $e(i, j)$. The following equations are added to our earlier LP formulation from section 3.2 to impose the additional placement based constraints for clustering and FBB allocation under a delay constraint.

$$(Q^p_{ij} + Q^p_{ji})M + \frac{d_i}{K_i C_{Li}} - \frac{d_j}{K_j C_{Lj}} \geq 0 \quad \forall\ e(i,j)\ on\ path\ p,\ \forall\ path\ p \tag{26}$$

$$-(Q^p_{ij} + Q^p_{ji})M + \frac{d_i}{K_i C_{Li}} - \frac{d_j}{K_j C_{Lj}} \leq 0 \quad \forall\ e(i,j)\ on\ path\ p,\ \forall\ path\ p \tag{27}$$

$$Q^p_{ij} + Q^p_{ji} = X^p_{ij}\ \forall\ e(i,j)\ on\ path\ p,\ \forall\ path\ p \tag{28}$$

$$1 + \sum_{\forall e(i,j)\ on\ path\ p} X^p_{ij} \leq Max_p\ for\ all\ path\ p \tag{29}$$

$$Q^p_{ij},\ Q^p_{ji},\ X^p_{ij}\ \in\ [0,1]\ \forall\ e(i,j)\ on\ path\ p,\ \forall\ path\ p \tag{30}$$

where M is a very large positive number. We know from section 4 that if two gates are in the same cluster, they have the same FBB value and their delays are in a fixed ratio. We have associated a variable X^p_{ij} with every pair of adjacently placed nodes i, j in row p of the placement. If $X^p_{ij} = 0$, then nodes i and j are in the same cluster and have the same FBB value. From a physical point of view, this would imply that they are fabricated in the same well. If $X^p_{ij} = 1$, then nodes i and j are in different clusters and do not have the same FBB value. These two gates would therefore be fabricated in different wells even though they have been placed adjacent to each other. This results in an extra area overhead. The spatial slack in each row obtained after placement is used to determine the maximum number of allowed clusters in each placement row as denoted by Max_p for each row p in equation 29. Equations 28 and 30 ensure that either both (Q^p_{ij}, Q^p_{ji}) are zero or only one of them is one. From equations 26 and 27, we can see that if both (Q^p_{ij}, Q^p_{ji}) are zero which means X^p_{ij} is zero and nodes i and j are in the same cluster, we impose the condition on the ratio of the delays of two nodes. On the other hand, if either one of (Q^p_{ij}, Q^p_{ji}) is one, then since M is a large positive number, we do not impose any condition on the ratio of the delays of the two nodes. Therefore these additional constraints alongwith the formulations proposed in section 3.2, we have a placement driven FBB clustering and allocation algorithm under a delay constraint for leakage minimization. We note that the additional constraints added to the LP formulation are also linear and hence optimality is retained.

Furthermore, if we also have an upper limit $LIMIT$ on the total number of FBB values that can be allocated to the circuit due to routing and fabrication constraints, we can additionally impose this limit as shown in equation 31.

$$N + \sum_{\forall e(i,j) in E'} X_{ij} \leq LIMIT \tag{31}$$

where N is the total number of standard cell rows in the placement information of the circuit.

This completes the description of our fine grained placement-driven simultaneous clustering and FBB allocation scheme. We can also apply other popular heuristics to solve this formulation. An interesting approach to solve this convex optimization problem with a set of linear constraints would be to use the Lagrange Multiplier Method.

6. EXPERIMENTAL RESULTS

The objective of these experiments was to prove two main points:

1. There is inherent slack available in circuits which can be used to slow down non-critical gates and get considerable savings in the leakage power without any delay penalty.
2. Placement Driven Clustering for fine-grained FBB allocation is effective in reducing the leakage power without any significant area penalty.

We have implemented our placement driven fine grained FBB allocation scheme in SIS [8]. We have built an integrated software interface that generates a placement for a benchmark from the SIS library using the CAPO placement tool [1]. We then use this placement information to generate the constraints as explained in the formulations in section 5. The delay constraint on the circuit is the best case delay of the circuit with all gates at low threshold voltages (minimum delay), thereby imposing no additional delay penalty on the timing. This illustrates our proposition of using the inherent slack available in the benchmark to assign FBB values under a delay constraint. Every gate is allocated a threshold voltage in the range 0.3V to 0.5V for the N-MOS and -0.3V to -0.5V for the P-MOS transistors. For simplification, we have used piece-wise linearization of our convex objective function and used CPLEX to implement our LP formulation.

Table 1 shows the results from our experiments over a large range of benchmarks. Column 1 shows the various benchmarks from SIS that have been used for the experiments. Column 2 gives the initial leakage current values for the benchmarks using low threshold gates. This is the best case possible delay for the circuit. However, we will show

Bench -mark	Initial (10^{-12}A)	Optimal (10^{-12}A)	Savings %	0% Area (10^{-12}A)	Savings %	4% Area (10^{-12}A)	Savings %	8% Area (10^{-12}A)	Savings %
C432	13507	4312	68.1	10987	18.7	6970	48.4	6124	54.7
C499	27442	9979	63.6	20839	24.1	16043	41.5	13854	49.5
C880	19472	5936	69.5	18101	7.1	8002	58.9	7232	62.8
C1908	31432	9721	69.1	22831	27.4	8618	72.6	8467	73.0
x1	16871	4703	72.1	7670	54.6	4752	71.8	4708	72.1
x3	45734	12632	72.4	16005	65.0	12845	71.9	12822	72.0
x4	27416	7940	71.0	20198	26.3	9245	66.2	8105	70.4
i5	19566	5481	71.9	10436	46.7	5498	71.9	5492	71.9
i6	50571	13938	72.4	16264	67.8	13938	72.4	13938	72.4
ai8	66516	18580	72.0	27310	58.9	18596	72.0	18580	72.0
Avg.			70.2		39.7		64.7		67.1

Table 1: Results from Experiments

that we can still maintain this delay constraint by fabricating high V_{th} gates (minimum leakage and maximum delay) and allocating appropriate FBB values to speed up timing critical gates to meet the delay constraints. Thus, utilizing the available delay slack in the circuit, we can reduce the leakage power of the gates. Column 3 in table 1 shows the result from our optimal FBB allocation formulation discussed in section 3.2. Here we have assumed that we can independently control the FBB value of each gate and do not take into account the area overhead associated with this scheme. Column 4 shows that there is an average 70.2% reduction in the total leakage current across the benchmarks using the optimal FBB allocation scheme. This proves our first claim that there is available delay slack in circuits which can be used for significant savings in leakage without any extra delay penalty.

We now consider our placement driven clustering strategy which tried to minimize the area penalty incurred. We first consider the worst case scenario when there is no available area slack in each of the placement rows. This essentially means that all gates in a particular placement row have been clustered together. Columns 5 and 6 show that there is an average 39.7% reduction in leakage for this case. We now consider the scenario with available area slack to compensate for the area penalty from clustering. We have assumed in our experiments, that each new cluster (new n and p well) formed requires an area overhead equal to that of an inverter. Columns $7-10$ show that for 4% and 8% available area slack, there is on an average 64.7% and 67.1% reduction in leakage current respectively. Hence, we have shown the effectiveness of our placement driven fine-grained FBB scheme through these experiments. The available area slack after placement can be effectively used to cluster together gates in each of the placement rows and then we can optimally allocate FBB values to these clusters using our formulation as presented in this work.

7. CONCLUSION AND FUTURE WORK

In this work we have proposed a novel fine-grained FBB allocation algorithm for reducing leakage power of circuits in active mode of operation. We have presented an optimal polynomial time FBB allocation algorithm alongwith a placement-driven FBB allocation algorithm which tries to utilize the available area slack in each of the placement rows for clustering. Our results have shown that the proposed schemes are effective in reducing the active mode leakage power in gate level circuits with no delay penalty. An interesting direction for future work is to evaluate our scheme on other popular heuristics like the Lagrange Multiplier Method and also to study the effect of imposing a delay penalty on the timing constraint.

8. REFERENCES

[1] A. Caldwell et al. "Can Recursive Bisection Alone Produce Routable Placements?". In *Proc. of DAC*, 2000.

[2] A. J. Bhavnagarwala et al. "Dynamic-Threshold CMOS SRAM Cells for Fast, Portable Applications". In *ASIC/SOC Conference*, pages 359–363, 2000.

[3] A. Sedra and K. Smith. "Microelectronic Circuits". Oxford University Press, 1997.

[4] Ali Keshavarzi et al. "Forward body Bias for Microprocessors in 130nm Technology Generation and Beyond". In *VLSI Circuits Symp.*, pages 312–315, 2002.

[5] C. H. Kim et al. "Dynamic Vt SRAM: a Leakage Tolerant Cache Memory for Low Voltage Microprocessors". In *ISLPED*, pages 251–254, 2002.

[6] Chris H. Kim et al. "A Forward Body-Biased Low-Leakage SRAM Cache: Device and Architecture Considerations". In *Proceedings of ISLPED*, August 2003.

[7] D. Hochbaum and J. Shanthikumar. "Convex Separable Optimization is not much harder than Linear Optimization". In *Journal of the ACM*, vol. 37, No. 4, 1974.

[8] E.M. Sentovich, K.J. Singh, L. Lavagno, C. Moon, R. Murgai, A. Saldanha, H. Savoj, P.R. Stephan, R.K. Brayton, A.L. Sangiovanni-Vincentelli. *SIS: A System for Sequential Circuit Synthesis*. Memorandum No. UCB/ERL M92/41, Department of EECS. UC Berkeley, May 1992.

[9] H. Kawaguchi et al. "Dynamic Leakage Cut-Off Scheme for Low-Voltage SRAM's". In *VLSI Circuits Symp.*, pages 140–141, 1998.

[10] R. Gu and M. Elmasry. "Power Dissipation Analysis and Optimization of Deep Submicron CMOS Digital Circuits". In *IEEE JSSC*, May 1996.

[11] 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 2003*, Aug. 2003.

[12] S. Narendra et al. "1.1V 1GHz Communications Router with On-Chip Body Bias in 150nm CMOS". In *ISSCC*, 2002.

[13] T. Sakurai and A.R. Newton. "A Simple MOSFET Model for Circuit Analysis". In *IEEE Transactions on Electron Devices*, April 1991.

[14] Y. Taur et al. "25 nm CMOS Design Considerations". In *IEDM*, pages 789–792, 1998.