Gate Sizing and V_t Assignment for Active-Mode Leakage Power Reduction *

Feng Gao and John P. Hayes
Advanced Computer Architecture Lab.
University of Michigan, Ann Arbor, MI, 48109, USA
{fgao,jhayes}@eecs.umich.edu

ABSTRACT

Leakage current is a key factor in IC power consumption even in the active operating mode. We investigate the simultaneous optimization of gate size and threshold voltage to reduce leakage power. We assume a standard-cell-based design flow where the available cell sizes and threshold voltages (V_t's) are given, and model the optimization as a mixed-integer linear programming (MLP) problem. In addition to the exact model, two faster approximate MLP models are proposed, along with CAD tools that generate the models automatically. We present experimental results which show that optimal designs derived from the exact MLP model can achieve the same performance as all-low-V_t unit-size designs, but with only one third the leakage power. The approximate MLP models can be solved about 25 times faster than the optimal model with negligible errors. All the proposed models can be extended to take dynamic power and multiple supply voltages into consideration.

1. INTRODUCTION

The continuous shrinkage of IC feature size is enabling the integration of more and more devices on a single chip. The chip supply voltage is also being scaled down to reduce dynamic power consumption, which necessitates a decrease in the threshold voltage V_t to ensure that performance targets can be reached. Smaller V_t can cause an exponential increase in leakage power consumption, which may soon become the dominant factor in overall power consumption [10]. Dual threshold voltage assignment and gate sizing are two important gate-level techniques to reduce leakage power in both the standby [7], [11] and active [9], [12] operating modes.

Various sensitivity-based heuristic methods to minimize leakage or total power consumption have been proposed for adjusting design parameters such as cell size, threshold voltage and supply voltage [9] [11] [13]. The heuristic approaches, although showing good power reduction in some benchmark circuits, easily fall into local minima and cannot take full advantage of their design parameters.

Nguyen et al. [8] consider the minmization of total power consumption, given two V_t's. Their method first selects the low V_t for all gates and runs a heuristic algorithm to size the gates. Then it uses a linear programming (LP) model to distribute slacks among the gates. Gate sizing and V_t selection are performed to account for the slack distribution. This process is iterated until no further power reduction is possible. The heuristic nature of the sizing process makes the overall design approach of [8] non-optimal. In [12], Srivastava considers tuning V_t's only. Threshold voltage assignment under delay constraints is modeled as an LP problem. It can hence assign V_t's optimally in terms of leakage power consumption, assuming that each gate can have a different threshold voltage. However, the available V_t's are usually determined by the process technology, and it is costly to use many V_t levels.

In our work, we focus on sub-threshold leakage power optimization and take both threshold voltage and gate size into consideration during the optimization process. Furthermore, we accommodate a typical standard-cell-based design flow, where the available V_t's and gate sizes are pre-determined by a cell library. In the sequel, we assume two V_t's are available, namely, V_t^L and V_t^H. Unlike [8] and [12], we directly model the leakage optimization problem under given delay constraints and construct an MLP model [4], which considers both gate sizing and V_t selection. If gate sizes are adjustable, the delay of a gate *G* depends on *G*'s size and the sizes of gates driven by *G*. This complicates the modeling problem which, nevertheless, we can solve very efficiently. We present experiments with this model which show that the optimal designs have performance comparable to all- V_t^L designs[1], but with about one-third the latter's leakage currents. In fact, the leakage power consumption is just a few percent higher than the lower bounds on leakage power, which are determined by the all- V_t^H designs. Furthermore, the optimal designs achieve about 14% higher performance with just half the

* This research was supported by the National Science Foundation under Grant No. CCR-0073406.

[1] In the sequel, the cells referred to as all-low-V_t and all-high-V_t designs are assumed to be of unit size.

leakage current of the all-V_t^L ones. In addition, approximate MLP models are also proposed, which show significant speedup (25x) over the original MLP model, with minor increase in leakage current or delay. Our models can also be easily extended to consider total power consumption and multiple supply voltages [13] [1].

The remainder of this paper is as follows. We define our system model in Section 2. The exact and approximate MLP models for simultaneous selection of threshold voltages and gate sizes are described in Section 3. We present our experimental results in Section 4, and Section 5 concludes this paper.

2. SYSTEM MODEL

We start by describing the assumptions underlying our system model. We assume that there is a basic unit cell for each cell type in the given cell library, and use $\overline{G}$ to denote the unit cell with the same type as G. If $U(G)$ represents the physical size of gate G, then we define the *size* $S(G)$ of G to be $U(G)/U(\overline{G})$.

Let $R(G)$, $I_l(G)$, $C_g(G)$, $C_p(G)$ and $D(G)$ denote the resistance, leakage current, gate capacitance, source/drain capacitance, and gate delay of G, respectively. Neglecting high-order effects, we assume the following.

$$R(G) = R(\overline{G})/S(G)$$
$$I_l(G) = I_l(\overline{G})*S(G)$$
$$C_g(G) = C_g(\overline{G})*S(G)$$
$$C_p(G) = C_p(\overline{G})*S(G)$$

We further assume an RC delay model where the gate delay $D(G)$ is linear in the load capacitance C_L [14]. Therefore, viewing a logic gate in terms of resistors and capacitors at the transistor level, as illustrated by Figure 1, we have

$$\begin{aligned} D(G) &= R(G)(C_p(G) + C_L) \\ &= R(\overline{G})/S(G)\ (C_p(\overline{G})*S(G) + C_L) \\ &= R(\overline{G})*C_p(\overline{G}) + R(\overline{G})*C_L/S(G) \\ &= D_p(\overline{G}) + R(\overline{G})/S(G)*C_L \end{aligned} \quad (1)$$

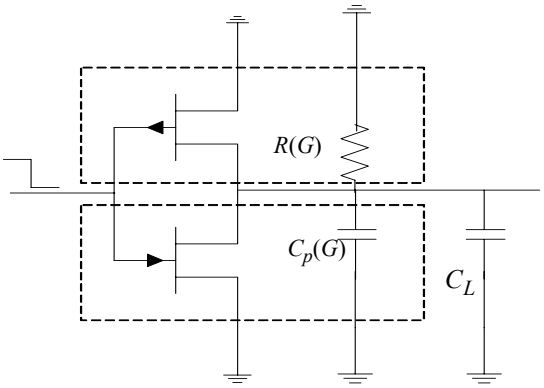

Figure 1. RC gate model used for delay calculation

where $D_p(\overline{G})$ is independent of cell size, and is called the *parasitic delay* of G. The second term of Equation (1) is linear in the load capacitance C_L, and is referred to as *load delay*. To simplify the MLP model, we represent C_L as a multiple of $C_0 = C_g$(INV1), the gate capacitance of a unit inverter. Consequently, the delay $D(G)$ can be expressed as

$$\begin{aligned} D(G) &= D_p(\overline{G}) + R(\overline{G})/S(G)*C_L \\ &= D_p(\overline{G}) + (R(\overline{G}) * \mathrm{C}_0) * (C_L/\mathrm{C}_0)/S(G) \\ &= D_p(\overline{G}) + D_l(\overline{G}) * (C_L/\mathrm{C}_0)/S(G) \end{aligned} \quad (2)$$

where $D_l(\overline{G}) = R(\overline{G})*\mathrm{C}_0$, is the load delay when a unit cell of the same type as G drives a unit inverter.

Although they are independent of G's size, $D_p(\overline{G})$ and $D_l(\overline{G})$, change with G's threshold voltage. If two different threshold voltages are available, we can use a binary variable $V_t(G)$ to represent G's $\mathrm{V_t}$ selection, with $V_t(G) = 1$ for V_t^L and $V_t(G) = 0$ for V_t^H. Consequently, we use $D_p(\overline{G}, V_t(G))$ and $D_l(\overline{G}, V_t(G))$ to explicitly denote dependency on G's threshold voltage. Let $D_p^0(G) = D_p(\overline{G}, 0)$ and $D_p^{10}(G) = D_p(\overline{G}, 1) - D_p(\overline{G}, 0)$. Then, we can rewrite $D_p(\overline{G}, V_t(G))$ as

$$\begin{aligned} & D_p(\overline{G}, 0)\ \overline{V_t(G)} + (D_p(\overline{G},1)\ V_t(G) \\ &= D_p(\overline{G}, 0) + (D_p(\overline{G}, 1) - D_p(\overline{G}, 0))\ V_t(G) \\ &= D_p^0(G) + D_p^{10}(G)\ V_t(G) \end{aligned} \quad (3)$$

Similarly, let $D_l^0(G) = D_l(\overline{G},0)$ and $D_l^{10}(G) = D_l(\overline{G},1) - D_l(\overline{G}, 0)$. We then have

$$D_l(\overline{G}, V_t(G)) = D_l^0(G) + D_l^{10}(G)\ V_t(G) \quad (4)$$

Adding $V_t(G)$ as a new parameter for $D(G)$ yields

$$\begin{aligned} D(G,V_t(G)) = {} & D_p^0(G) + D_p^{10}(G)\ V_t(G) \\ & + D_l^0(G)\ (C_L/\mathrm{C}_0)/S(G) \\ & + D_l^{10}(G)\ V_t(G)\ (C_L/\mathrm{C}_0)/S(G) \end{aligned} \quad (5)$$

The parasitic delay $D_p(\overline{G}, V_t(G))$ is a constant for a given cell type and $\mathrm{V_t}$ selection, and so is the load delay $D_l(\overline{G}, V_t(G))$ with $C_L = C_0$. Hence, we can measure them both in advance. We run Hspice with different sizes and load capacitances, and average the measured delay values. Hence we can view $D(G, V_t(G))$ as a linear function of $V_t(G)$, $C_L/S(G)$, and $V_t(G)C_L/S(G)$.

Because the switching time of a gate is just a small portion of a clock cycle, the gates stay in stable states and keep leaking for most of a cycle, even if they switch during the cycle. Hence we assume that leakage currents always exist in an active mode of operation. Let $I_l(G, I)$ be gate G's leakage current under input pattern I, and let $P(G, I)$ be the probability that input pattern I is applied to G; this is usually referred to as the *signal probability*.

The average leakage current $I_l(G)$ of G can be expressed as follows.

$$I_l(G) = \sum_I P(G, I) I_l(G, I) = (\sum_I P(G, I) I_l(\overline{G}, I)) S(G)$$

Taking threshold voltage $V_t(G)$ into consideration, we transform the above equation to

$$\begin{aligned} I_l(G, V_t(G)) &= \sum_I P(G, I) I_l(\overline{G}, I, 0) S(G) \\ &\quad + \{\sum_I P(G, I) I_l(\overline{G}, I, 1) \\ &\quad - \sum_I P(G, I) I_l(\overline{G}, I, 0)\} S(G) V_t(G) \\ &= I_l^0(G) S(G) + I_l^{10}(G) S(G) V_t(G) \end{aligned} \quad (6)$$

where $I_l^0(G) = \sum_I P(G, I) L(\overline{G}, I, 0)$, and

$$I_l^{10}(G) = \sum_I P(G, I) I_l(\overline{G}, I, 1) - \sum_I P(G, I) I_l(\overline{G}, I, 0)$$

$P(G, I)$ can be calculated using BDD-based [5] or simulation-based [2] approaches. We resort to the latter approach because of its simplicity. $I_l(\overline{G}, I, i)$, where $i = 0$ or 1, can also be measured by running Hspice. Therefore, $I_l(G, V_t(G))$ is linear in $S(G)$ and $S(G)\ V_t(G)$.

We now consider linearizing the multiplication and division functions $C = B * A$ and $C = B / A$, where A, B, and C are non-negative variables. This linearization is key to the performance of our approach. Note that popular methods like piece-wise linear approximation cannot be directly used to linearize multiplication or division.

To tackle this problem, we propose a very efficient technique, which is similar to that used in [3] to formulate floorplan minimization as a convex programming problem. Here our goal is to approximate C with linear functions. Consider $C = B / A$. We define two variables a and b such that $A = 2^a$ and $B = 2^b$. Equation $C = B / A$ is thus transformed to $C = 2^{b-a}$. Piece-wise linearization can then be applied to this exponential function. We use the lines determined by points $(k, 2^k)$ and $(k + 1, 2^{k+1})$ for this purpose:

$$C \geq 2^k (b - a) + 2^k(1 - k) \quad (7)$$

where $k = B_1, B_1 + 1, \ldots, B_2$, and $B_1 \leq b - a \leq B_2$. These constraints are not exactly equivalent to $C = B / A$, as Figure 2 illustrates. However, we will show that the approximation is quite accurate, when we develop our MLP models in the next section. The same general approach can be used to linearize multiplication.

3. PROBLEM MODELING

Using the assumptions and the linearization techniques discussed in the preceding section, we now derive an exact MLP model for simultaneous V_t selection and gate sizing. Several faster but approximate MLP models are then described.

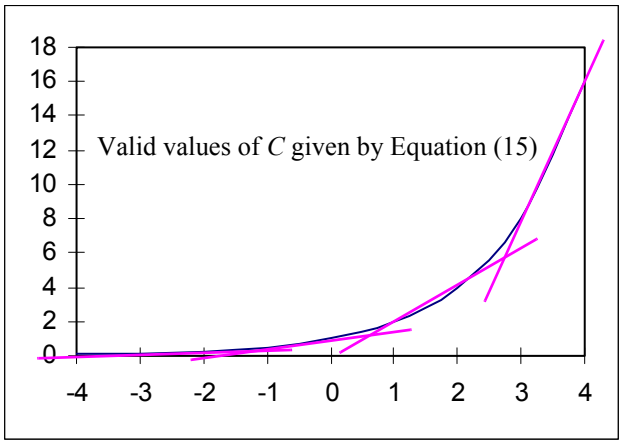

Figure 2. Piecewise linear approximation of an exponential function

We assume that two threshold voltages are available, and that the sizes of the library cells are powers of two, namely, 1, 2, 4, 8,..., $2^{P_{max}}$. Note that this assumption on cell sizes is only used to simplify the discussion. We will generalize the cell sizes later.

Objective Function. We have derived the leakage current for a single gate in Equation (6). Since our goal is to reduce the leakage current of a given circuit, we use the sum of all gate leakage currents as the objective function to be minimized:

$$\sum_G (I_l^0(G) S(G) + I_l^{10}(G) S(G) V_t(G)) \quad (8)$$

Let $SV(G) = S(G)V_t(G)$. As noted earlier, the leakage function is linear in $S(G)$ and $SV(G)$. We will generate a set of linear constraints for $S(G)$ and $SV(G)$, respectively.

Constraints. The constraints of the MLP model fall into two classes: performance and linearization. The performance constraints guarantee that the size and threshold voltage for all gates meet the performance target. We replace any non-linear terms such as $SV(G) = S(G)V_t(G)$ appearing in the objective function and performance constraints with linear inequalities, which constitute the linearization constraints.

First consider the performance constraints. Let real variable $T_a(G)$ be the arrival time of G's output signal. For convenience, we insert a virtual gate in each primary input. To satisfy the overall circuit delay D_{max}, we use constraints $T_a(G_o) \leq D_{max}$, where G_o is any gate driving a primary output, and $T_a(G_I) = 0$ for all primary inputs G_I. We then derive constraints to relate the arrival times of G's input signals to G's output signal. Consider the circuit fragment in Figure 3, where G_1 has two inputs driven by G_2 and G_3, and one output, which drives G_4. The arrival time of G_1's output signal satisfies

$$T_a(G') + D(G_1, V_t(G_1)) \leq T_a(G_1)$$

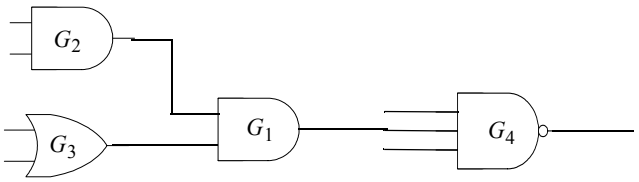

Figure 3. A circuit fragment

where G' is any gate driving G_1's inputs.

We calculate $D(G_1, V_t(G_1))$ as follows. Since G_1 only drives G_4, the load capacitance C_L of G_1 is $C_g(\mathrm{G}_4) = C_g(\overline{G}_4)\ S(G_4)$. The delay $D(G_1, V_t(G_1))$ of G_1 is hence

$$\begin{aligned} D(G_1, V_t(G_1)) = {} & D_p^0(G_1) + D_p^{10}(G_1)\, V_t(G_1) \\ & + D_l^0(G_1)\,(C_g(\overline{G}_4)/\mathrm{C}_0)\ S(G_4)/S(G_1) \\ & + D_l^{10}(G_1)\ (C_g(\overline{G}_4)/\mathrm{C}_0) \\ & * V_t(G_1)\ S(G_4)/S(G_1) \end{aligned} \quad (9)$$

To make Equation (9) linear, we define new variables $SR(G_i, G_j) = S(G_i)/S(G_j)$ and $SRV(G_i, G_j) = SR(G_i, G_j) * V_t(G_i)$. Note that $C_g(\overline{G}_4)/\mathrm{C}_0$ only depends on the type of gate, and so can be calculated beforehand for all gate types in the library. Therefore, we can view Equation (9) as a linear expression with respect to $SR(G_4, G_1)$ and $SRV(G_4, G_1)$.

To confirm the accuracy of the substitutions we made to linearize the objective function and performance constraints, we now analyze the linearization constraints.

We use the technique described in Section 2 to linearize $SR(G_i, G_j) = S(G_i)/S(G_j)$. We define variable $p(G)$ for each gate, where $S(G) = 2^{p(G)}$, and $p(G) \in [0, p_{max}]$. Therefore, we obtain

$$SR(G_i, G_j) = S(G_i)/S(G_j) = 2^{p(G_i) \angle p(G_j)} \quad (10)$$

and approximate $SR(G_i, G_j)$ with the inequality

$$SR(G_i, G_j) \geq 2^k\,(p(G_i) - \mathrm{p}(G_j)) + (1-k)2^k,$$

where $k = -p_{max}, -p_{max} + 2, ..., p_{max} + 1$

Note that $SR(G_i, G_j)$ will never be underestimated by the above linearization scheme when $p(G_i) - \mathrm{p}(G_j)$ is an integer; neither will G_i's delay. Consequently, the original performance target is still guaranteed after the linearization.

We also apply the foregoing linearization technique to calculating $S(G) = 2^{p(G)}$. Specifically, we use the inequalities

$$S(G) \geq 2^k\, p(G) + (1-k)2^k, \text{ where } k = 0, 2, 4, ... \quad (11)$$

Similarly, $S(G)$ is at least $2^{p(G)}$ when $p(G)$ is an integer. Furthermore, since we are trying to minimize the objective function (8), $S(G)$ will be assigned the minimum possible value. We can hence conclude that $S(G) = 2^{p(G)}$ in any valid solution. Consequently, we do not sacrifice any optimality by linearizing the objective function and performance constraints, provided the $p(G)$'s are integers.

The other set of equations to be linearized is

$$SV(G_i) = S(G_i) * V_t(G_i)$$
$$SRV(G_i, G_j) = SR(G_i, G_j) * V_t(G_i)$$

where $S(G_i)$ and $SR(G_i, G_j)$ are real variables while $V_t(G_i)$ is binary. Generally, $C = B * A$, where A is a binary variable and M is an upper bound of B, are linearized as follows:

$$0 \leq C \leq B \quad (12)$$
$$C \leq M * A \quad (13)$$
$$C \geq B - M\,(1 - A) \quad (14)$$

With the additional constraints that the $p(G)$'s are integers in the range $[0, p_{max}]$ and the $V_t(G)$'s are binary, we end up with an MLP model for leakage minimization.

Minimize

$$\sum_G (I_l^0(G)\,S(G) + I_l^{10}(G)\,S(G)\,V_t(G))$$

Subject to

{Performance constraints}

$T_a(G_o) \leq D_{max}$, G_o is gate driving a primary output

$T_a(G_I) = 0$, G_I is any virtual gate of a primary input

(We use G_1 as an example for constraints relating gates' inputs and outputs)

$T_a(G_2) + D(G_1, V_t(G_1) \leq T_a(G_1)$

$T_a(G_3) + D(G_1, V_t(G_1) \leq T_a(G_1)$

$$\begin{aligned} D(G_1, V_t(G_1)) = {} & D_p^0(G_1) + D_p^{10}(G_1)\, V_t(G_1) \\ & + D_l^0(G_1)\,(C_g(\overline{G}_4)/\mathrm{C}_0)\ S(G_4, G_1) \\ & + D_l^{10}(G_1)\,(C_g(\overline{G}_4)/\mathrm{C}_0) * SRV(G_4, G_1) \end{aligned}$$

{Linearization constraints for performance constraints}

$SR(G_4,G_1) \geq 2^k\,(p(G_4) - \mathrm{p}(G_1)) + (1-k)2^k$,

for $k = -p_{max}, -p_{max} + 2, ... p_{max} + 1$

$0 \leq SRV(G_4, G_1) \leq 2^{p_{max}}$

$SR(G_4,G_1) - 2^{p_{max}}\,(1 - V_t(G_1)) \leq SRV(G_4, G_1)$

$SRV(G_4, G_1) \leq SR(G_4,G_1)$

{Linearization constraints for objective function}

$SV(G) \geq 2^k\, p(G) + (1-k)2^k$, $k = 0, 2, ..., p_{max}$

$0 \leq SV(G_4, G) \leq 2^{p_{max}}$

$S(G) - 2^{p_{max}}\,(1 - V_t(G)) \leq SV(G)$

$SV(G) \leq S(G)$

Bounds

$0 \leq p(G) \leq p_{max}$, for all gate G

$p(G)$'s are integers and $V_t(G)$'s are binary variables

Figure 4. Summary of the exact MLP model MLP-0

Circuit	Gate count	A (All- V_t^L design)		B (All-V_t^H design)				C (MLP-0-based design)			
		Leakage current (pA) L_A	Delay D_A (ps)	Leakage current (pA) L_B	L_B/L_A	Delay (ps) D_B	D_B/D_A	Leakage current (pA) L_C	L_C/L_A	Delay (ps) D_C	D_C/D_A
C432	121	1257	2857	434	0.345	3287	1.151	665	0.525	2855	1.000
C880	345	3439	2017	1144	0.332	2271	1.126	1310	0.381	1944	0.964
C1908	435	4262	3145	1367	0.320	3587	1.141	1524	0.358	3145	1.000
C2670	746	7218	1946	2398	0.332	2227	1.144	2503	0.346	1942	0.998
C3540	892	8903	3323	2987	0.336	3770	1.135	3084	0.364	3322	1.000
C7552	2066	20638	2789	6279	0.304	3143	1.127	6626	0.321	2798	1.000
Pair	1299	12715	3094	4311	0.339	3475	1.123	4895	0.386	3093	1.000
Apex6	615	5927	1064	1899	0.320	1198	1.126	2070	0.354	1061	0.997
Ave.					**0.329**		**1.134**		**0.379**		**0.995**

Figure 5. Leakage currents and delays for three versions of the benchmark circuits

The general form of this exact MLP model, referred to as *MLP-0*, is summarized in Figure 4.

It turns out that if we are willing to accept slight inaccuracies in the results, we can remove the restriction that the cell sizes are powers of two. To do this, we eliminate the constraints that the $p(G)$'s are integers and leave only the $V_t(G)$'s as integer variables, which are binary in this specific case. Changing the variable types from integer to real significantly speeds up the solution of the MLP formulation. The resulting errors are very minor, as we will see from the experimental results.

Solving the modified MLP model, we obtain the $V_t(G)$'s as binary numbers which are 1 for V_t^L and 0 for V_t^H, and the $p(G)$'s as real numbers. We then determine $S(G)$ using various approximation approaches according to the cell sizes provided by the library. We consider two particular cell sizing scenarios here. In scenario one, we assume that the library supports cell sizes of any real number, and set cell size $S(G) = 2^{P(G)}$. The circuit produced by this scenario provides a lower bound on all possible gate sizes and V_t assignments for the given circuit and delay constraints. We call this MLP model with unrestricted cell sizes the *MLP-1* model. In scenario two, we assume that cell sizes are powers of two, and call the result the *MLP-2* model. The cell size of gate G is chosen as $S(G) = 2^{\lceil P(G) \rceil}$. The MLP-2 model turns out to be a very good approximation to the original MLP-0 model with the assumption that the library cell sizes are powers of two.

4. EXPERIMENTAL RESULTS

We have developed a CAD tool to automatically generate our various MLP models, which are then solved using the commercial LP solver cplex [6]. The program first reads the circuit netlist and calculates signal probabilities by random simulation. Next, it traverses the netlist to generate the MLP models, using the cell library data we have measured beforehand. The automatically generated MLP models are written to a file that can be directly imported into cplex.

We have applied the proposed methods to a set of representative ISCAS and MCNC benchmark circuits. The circuits were synthesized using the Synopsys Design Compiler tool and a TSMC 0.18µm standard cell library. We assume the high V_t's for PMOS and NMOS are 0.44V and 0.50V, respectively. The low V_t's are 0.1V less than their high-V_t counterparts. In addition, we ran 1,000 random patterns to calculate signal probabilities for the given circuits.

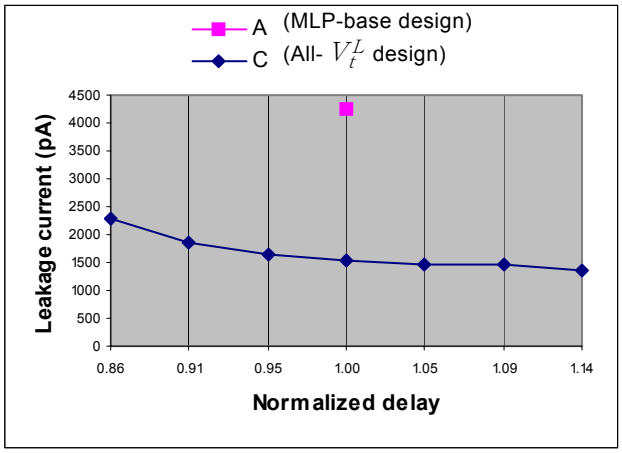

Figure 6. Leakage-performance curve for C1908

Circuit	C (MLP-0-based design)			D (MLP-1-based design)				E (MLP-2-based design)					
	Leakage current (pA) L_C	Delay (ps) D_C	Runtime (s) T_C	Leakage current (pA) L_D	L_D/L_C	Delay (ps) D_D	D_D/D_C	Leakage current (pA) L_E	L_E/L_C	Delay (ps) D_E	D_E/D_C	Runtime (s) T_E	T_C/T_E
C432	665	2855	390.89	659	0.991	2855	1.000	704	1.059	2835	0.993	19.51	20.0
C880	1310	1944	144.19	1307	0.998	1943	0.999	1314	1.003	1948	1.002	7.87	18.3
C1908	1524	3145	48.41	1523	0.999	3145	1.000	1524	1.000	3145	1.000	14.39	3.4
C2670	2503	1942	186.57	2488	0.994	1938	0.998	2503	1.000	1942	1.000	12.11	15.4
C3540	3084	3322	9463.50	3071	0.996	3319	0.999	3084	1.000	3335	1.004	94.20	100.5
C7552	6626	2798	9359.20	6626	1.000	2793	0.998	6628	1.001	2794	0.999	862.00	10.9
Pair	4895	3093	992.96	4886	0.998	3093	1.000	4902	1.001	3094	1.000	38.91	25.5
Apex6	2070	1061	155.22	2070	1.000	1061	1.000	2079	1.004	1061	1.000	11.34	13.7
Ave.					**0.997**		**0.999**		**1.008**		**1.000**		**26.0**

Figure 7. Comparison of designs based on the exact MLP model MLP-0 and two approximate MLP models: MLP-1 (unrestricted cell sizes) and MLP-2 (power-of-two cell sizes)

Since different cell libraries, threshold voltages and baseline circuits are used by different authors, it is hard to make direct comparisons with previous experimental results. We therefore generate a small set of design variants for comparison purposes. Here, we use three variants of each benchmark circuit, denoted A, B, and C. Circuit A is the all- V_t^L unit-size version. Circuit B is the all- V_t^H unit-size design. Circuit C is the case where the solution of the MLP-0 model determines the sizes and threshold voltages of the gates, whose overall delays are set to the same values as in form A.

The results of the experiments are presented in Figure 5. For each circuit, we measure the leakage currents and delays occurring in all three circuit variants. The leakage current and delay ratios with respect to those of case A are also given.

The performance of the circuits generated by the MLP-0-based approach is quite similar to that of the A designs. However, their leakage currents are just one third of their A counterparts. The only exception is C432, whose leakage current is more than half of that of the corresponding A design. The reason is that this circuit is small, and we would need to enlarge the circuit or assign low V_t's to more of its gates to obtain the same speedup. Note that the B designs give lower bounds on the leakage currents of the given circuits. Therefore, the leakage currents of the MLP-0-based designs are very close to the minimum values.

We also applied the exact MLP-0 model to analyze leakage power vs. performance tradeoffs using C1908 as an example. We first calculated the delays of C1908 in forms A and B, denoted D_A and D_B, respectively. We then applied our MLP-0 model by setting the circuit delay to $D_B - k(D_B - D_A)/3$, where $k \in [1, 6]$. We solved these MLP models using cplex, and calculated the leakage current for each design variant with the calculated gate size and V_t. The resulting leakage vs. performance curve is given in Figure 6, where delays are normalized to D_A. Note that the circuit with the longest delay is of form B. The data show that the optimal design can achieve the same performance as form A with far less leakage current. In fact, we can reduce the delay by another 14% with just half the leakage current of form A.

Finally, we constructed circuits using the approximate MLP models discussed in Section 3. For brevity, we refer to a circuit obtained using the MLP-1 model as variant D, while a circuit obtained using the MLP-2 model is denoted by E. The leakage currents and delays for the examined circuits in variants C, D, and E are presented in Figure 7. The ratios of leakage currents and delays of the D and E designs to those of the C designs are also listed. In addition, the runtimes to solve the exact and approximate

	A and B	C	D	E
C432	550.6	636.8	615.61	653.90
C880	1442.23	1524.17	1533.47	1543.61
C1908	1618.62	1678.46	1677.03	1678.46
C2670	2820.47	2840.55	2839.74	2847.75
C3540	3812.85	3907.20	3891.48	3906.84
C7552	7537.2	7601.28	7581.18	7600.56
Pair	5405.76	5443.60	5426.47	5438.76
Apex6	2394.71	2464.50	2452.77	2474.22

Figure 8. Area Comparison

MLP models and the corresponding speedups are given. We also calculated the areas of all design variants, including the unit-size designs A and B, optimized designs C, D, and E. The area is given by the total transistor width (in microns). The results are shown in Figure 8. Except for circuit C432, the overheads of all circuits in forms C, D, and E are within a few percent of one another. As argued previously, C432 is a small design, and we have to size up more of its gates to obtain the same speedup.

As discussed in the previous section, design variant D provides a lower bound for the circuits with all possible gate sizes and V_t assignments under the given delay constraints. On the other hand, design variant C is the optimal solution if the library cell sizes are powers of two. Comparison between these two design types will hence disclose the effectiveness of supporting a library with relatively few cell sizes such as powers of two. In fact, compared with the D designs, the errors in both leakage current and delay of the C designs are less than 1%.We conclude that libraries with power-of-two cell sizes usually suffice for good performance-leakage tradeoffs. Moreover, comparison of the leakage currents and delays of the E and C designs shows that the E design produced by the MLP-2 model is a very good approximation to the C case, since the differences in both leakage current and delay are within 1%, except for the smallest circuit C432. In addition, solving the approximate MLP models is about 25 times faster than solving the original, exact MLP model. We therefore conclude that the MLP-2 model is an efficient and accurate replacement for the exact MPL-0 model, and is able to handle much larger circuits.

5. CONCLUSIONS

We have presented a mixed linear programming method to simultaneously choose threshold voltages and gate sizes optimally in order to minimize leakage power. The exact MLP model is made possible by a novel way of linearizing gate delay functions. Approximate MLP models are also proposed, which are much easier to solve. Two particular approximate MLP models with different gate sizing scenarios have been investigated. The MLP-1 model places no restrictions on library cell sizes and provides a lower bound for all possible gate sizes and V_t assignments under the given delay constraints. On the other hand, the simplified MLP-2 model approximates the exact MLP-0 model very well.

Our experimental results show that optimal MLP-0-based designs have the same performance as the all-V_t^L designs, with around one third the latter's leakage power. Furthermore, the leakage of the MLP-0 designs is only a few percent higher than that of the all-V_t^H designs, which have the smallest leakage of all possible designs. The performance of the optimal designs can be pushed even higher without incurring significant leakage current penalty. As illustrated by our experiments, the optimal design of C1908 has 14% higher performance than the all-V_t^L one, with only half the leakage current. The bounds provided by the MLP-1 model indicate the sufficiency of employing the MLP approach with power-of-two cell sizes only. In addition, use of the MLP-2 model leads to significant speedup (25x), with negligible loss of optimality.

Both the exact and approximate MLP models can be easily extended to consider both total power consumption and multiple supply voltages. Furthermore, the approximate MLP models scale well and can be used with much larger circuits.

REFERENCES

[1] S. Augsburger et al., "Reducing Power with Dual Supply, Dual Threshold and Transistor Sizing", *Proc. ICCD*, 2002.

[2] R. Burch et al., "A Monte Carlo Approach for Power Estimation", *IEEE Trans. on VLSI*, Mar. 1993, pp. 63 - 71.

[3] T. Chen and M. K. H. Fan, "On Convex Formulation of the Floorplan Area Minimization Problem", *Proc. ISPD*, 1998, pp. 124-128.

[4] F. Gao and J. P. Hayes, "ILP-based Optimization of Sequential Circuits for Low Power", *Proc. ISLPED*, 2003, pp. 140 - 145.

[5] A. Ghosh et al., "Estimation of Average Switching Activity in Combinational and Sequential Circuits", *Proc. DAC*, 1992, pp. 253-259.

[6] ILOG cplex webpage. http://www.ilog.com/products/cplex/.

[7] M. Ketkar and S.S. Sapatnekar, "Standby Power Optimization via Transistor Sizing and Dual Threshold Voltage Assignment", *Proc. ICCAD*, 2002, pp. 375 - 378.

[8] D. Nguyen et al., "Minimization of Dynamic and Static Power Through Joint Assignment of Threshold Voltages and Sizing Optimization", *Proc. ISLPED*, 2003, pp. 158 - 163.

[9] P. Pant, K. Roy and A. Chattejee, "Dual-threshold Voltage Assignment with Transistor Sizing for Low Power CMOS Circuits", *IEEE Trans. on VLSI*, April 2001, pp. 390 -394.

[10] Semiconductor Industry Association. *International Technology Roadmap for Semiconductors*. http://public.itrs.net/, 2000.

[11] S. Sirichotiyakul et al., "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing", *Proc. DAC*, 1999.

[12] A. Srivastava, "Simultaneous V_t Selection and Assignment for Leakage Optimization", *Proc. ISLPED*, 2003, pp. 172-175.

[13] A. Srivastava et al., "Concurrent Sizing, Vdd and Vth Assignment for Low-Power Design", *Proc. DATE*, 2004.

[14] N. H. E. Weste and K. Eshraghian. *Principles of CMOS VLSI Design: A Systems Perspective*, Reading, MA: Addison-Wesley, 1993.