

Power Minimization by Simultaneous Dual- V_{th} Assignment and Gate-sizing

Liqiong Wei, Kaushik Roy and Cheng-Kok Koh
School of ECE, Purdue University, West Lafayette, IN 47907

Abstract— Gate-sizing is an effective technique to optimize CMOS circuits for dynamic power dissipation and performance while dual- V_{th} (threshold voltage) CMOS is ideal for leakage power reduction in low voltage circuits. This paper focuses on simultaneous dual- V_{th} assignment and gate-sizing to minimize the total power dissipation while maintaining high performance. An accurate power dissipation model that includes short-circuit, switching, and leakage power is derived and used in our optimization. Results show that more than 20% and 40% power reductions are achievable for circuits at high and low switching activities, respectively, compared to single low- V_{th} CMOS circuits while maintaining performance.

I. INTRODUCTION

Low voltage CMOS designs are becoming very attractive for portable computing and wireless communication systems [3], [4]. With the lowering of supply voltage, the transistor threshold voltage (V_{th}) has to be scaled down to meet the performance requirements. Unfortunately, such scaling increases the sub-threshold leakage current through a transistor, thereby increasing leakage power. Dual- V_{th} design technique can be used to reduce leakage power in low voltage and high performance circuits [9]. A low threshold voltage is used for transistors in the critical path(s) to maintain high performance, while transistors in non-critical paths are selectively assigned a high threshold voltage to suppress the leakage.

Gate sizing is another technique to improve the performance of the circuit. Earlier approaches focused on the minimization of the area under delay constraints [1][13][14][15]. Since the power consumption has become one of the major concerns, approaches to gate (transistor) sizing for low power were considered in [7] and [2]. However, they focused only on dynamic power dissipation and used a zero-load short circuit power model [6] that overestimated the short circuit power significantly. Recently, a simultaneous threshold voltage selection and circuit sizing technique was presented to minimize the standby leakage power [10].

Since gate-sizing changes the dynamic power by varying the load capacitance and dual- V_{th} CMOS effectively reduces the leakage power, the total power can be minimized by simultaneous dual- V_{th} assignment and gate-sizing. In order to estimate the power accurately, an analytical short-circuit current model that considers load capacitance is derived. A sensitivity-based TILOS (Timed LOGic Synthesis [1])-like algorithm is presented for simultaneous gate-sizing and dual- V_{th} assignment.

This research was supported in part by SRC (98-HJ-638) and by Intel Corp.

II. OVERVIEW OF POWER ESTIMATION

For a CMOS circuit, the total power dissipation includes both dynamic and static components. Dynamic power consists of the switching power due to charging and discharging of load capacitance and the short circuit power due to non-zero rise and fall time of input waveforms. Static power of circuit is mainly determined by the sub-threshold leakage current through each transistor. The dynamic switching power, short circuit power, static leakage power and the total average power of a circuit can be expressed as follows:

$$P_{dyn} = \sum_i P_{dyn,i} = \sum_i \left(\frac{1}{2} \alpha_i C_{L,i} V_{dd}^2 f \right) \quad (1)$$

$$P_{sc} = \sum_i P_{sc,i} = \sum_i (\alpha_i V_{dd} I_{sc,i}) \quad (2)$$

$$P_{leak} = \sum_i P_{leak,i} = \sum_i (I_{leak,i} V_{dd}) \quad (3)$$

$$P_T = P_{dyn} + P_{sc} + P_{leak} \quad (4)$$

where the summations in eqn.(1)-(3) are taken over all the gates in the circuit. C_L is the load capacitance, f is the frequency and V_{dd} is the supply voltage. α is the average switching activity of each gate (average number of signal switching per unit time), which can be determined by a Monte Carlo based statistical simulation [11]. I_{sc} and I_{leak} are the average short circuit and leakage current of each gate, respectively. A leakage current model, which considers the the body effect, drain induced barrier lowering (DIBL) and transistor stacking effect, is used in our analysis [12]. However, for the short circuit current, the commonly used model [6] ignores the load capacitance and hence, significantly overestimates the short circuit current. In the following section, we will derive a short circuit current model that considers the effect of load capacitance.

III. SHORT CIRCUIT CURRENT MODEL

First, let us consider a rising step input to an inverter. The NMOSFET is "on" but PMOSFET is "off". The NMOS drain current equals the current through the load capacitance, which is the switching current responsible for switching power dissipation. However, for a ramp input, both PMOSFET and NMOSFET may be "on" simultaneously during transition. Let us look at Fig. 1 (a), in which the load capacitance (C_L) is the summation of the diffusion capacitance of the driving transistors (C_{diff}) and the output capacitance (C_O). Suppose the current through PMOSFET, NMOSFET and the load capacitance are represented by I_p , I_n and I_c , respectively, then $I_p + I_c = I_n$. I_p , which is non-zero due to the ramp input, is the short circuit current. I_c is the switching current and I_n is the

summation of the short circuit current and the switching current.

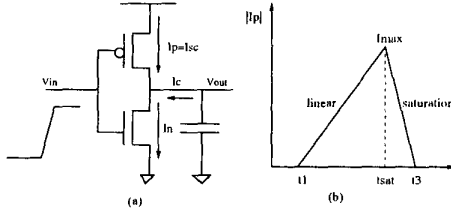


Fig. 1. Short circuit current for the rise input

At the beginning of the input transition, PMOSFET is in the linear region. Increasing the input voltage will increase the source to drain voltage of the PMOS transistor, thereby increasing I_p . Consider $t = t_{sat}$ when $V_o = V_i + |V_{tp}|$. The PMOSFET is in the saturation region. Now as the input voltage increases, I_p decreases due to the reduction of the source to gate voltage. We assume I_p reaches its maximum value (I_{max}) when $t = t_{sat}$. $|I_p|$ is linearized and sketched in Fig. 1 (b), where $t_1 = \frac{V_{in}}{V_{dd}}\tau_i$, $t_3 = (1 - \frac{V_{tp}}{V_{dd}})\tau_i$. τ_i is the input transition time.

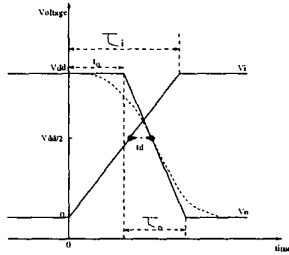


Fig. 2. input and output waveforms

To obtain the analytical model for short circuit current, we approximate the real input/output waveforms by ramp waveforms (see Fig. 2). In Fig. 2, τ_o and t_d are the output transition time and the propagation delay (50% input to 50% output), respectively, which can be evaluated based on the alpha-power law MOSFET model [5]. The input and output voltages can be expressed as follows:

$$V_i = \begin{cases} \frac{V_{dd}}{\tau_i} t & (0 < t \leq \tau_i) \\ V_{dd} & (t > \tau_i) \end{cases} \quad (5)$$

$$V_o = \begin{cases} \frac{V_{dd}}{\tau_o} t & (0 < t \leq t_0) \\ -\frac{V_{dd}}{\tau_o} t + (\frac{\tau_i}{2} + t_d + \frac{\tau_o}{2}) \frac{V_{dd}}{\tau_o} & (t_0 < t \leq \frac{\tau_i + \tau_o}{2} + t_d) \\ 0 & (t > \frac{\tau_i + \tau_o}{2} + t_d) \end{cases} \quad (6)$$

where t_0 is defined to be $(\tau_i - \tau_o)/2 + t_d$. Based on equations (5), (6) and the alpha power law MOSFET model [5], t_{sat} and I_{max} can be derived as follows:

$$t_{sat} = \frac{1}{2}\tau_i + \frac{\tau_i}{\tau_i + \tau_o} t_d - \frac{|V_{tp}|}{V_{dd}} \frac{\tau_i \tau_o}{\tau_i + \tau_o} \quad (7)$$

$$I_{max} = \frac{i_{D0} W}{(V_{dd} - V_{th})^m} (V_{dd} - \frac{V_{dd}}{\tau_i} t_{sat} - |V_{tp}|)^m \quad (8)$$

where m is the saturation index¹. V_{D0} is the drain saturation voltage at $V_{GS} = V_{dd}$ and $I_{D0} = i_{D0} \cdot W$ is the drain

¹In [5] the velocity saturation index is α . In order to distinguish from the switching activity, we use m to represent the saturation index.

current at $V_{GS} = V_{DS} = V_{dd}$.

Hence, the average short circuit current can be expressed as follows:

$$I_{sc} = \frac{1}{2} (V_{dd} - V_{tp} - V_{tn}) \frac{\tau_i}{V_{dd}} I_{max} f \quad (9)$$

Similarly, we can get the average short circuit current of an inverter for a falling input.

HSPICE simulations are used to verify our model. Let us consider an input signal of 2ns rise time. Fig. 3 shows the short circuit current at different output capacitances (C_O). The solid lines are the HSPICE simulation results based on 0.25 μ m MOSIS technology. The overshoot at the beginning is due to the overlapped gate to drain capacitance, which is ignored in our model. The dashed lines represent the results based on our model. The supply voltage and threshold voltage for the PMOS transistor are 3V and -0.59V; respectively. i_{d0} is 0.24mA/ μ m and the saturation index is 1.5. Simulation and modeling results for P_{sc} at a cycle time of 24ns are listed in Table I under P_{sc} column. Clearly, the average short circuit current decreases when load capacitance increases. For the zero-load capacitance model [6], the error can be as large as 47.5% and 109% for 50fF and 200fF load capacitance, respectively. The error of our model is only 19-30%. The error in our model is due to: (1) the overshoot of I_p due to the overlapped gate to drain capacitance, (2) the ramped input/output waveforms approximation, and (3) the use of linearized I_p .

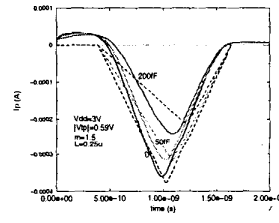


Fig. 3. Short circuit current (I_p)

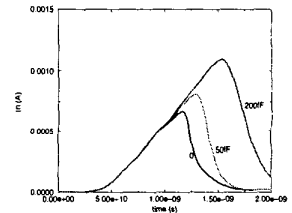


Fig. 4. Total current (I_n)

TABLE I

SIMULATION AND MODELING RESULTS FOR P_{sc} AND $P_{sc}/(P_{dyn} + P_{sc})$

C_O (fF)	t_d (ns)	t_o (ns)	$P_{sc}(\mu W)$		$P_{sc}/(P_{dyn} + P_{sc})$		
			hspice	model	$\tau_i = 2ns$	$1ns$	$0.8ns$
0	0.12	0.45	22.73	29.5	0.48	0.22	0.14
50	0.23	0.57	20	24.8	0.31	0.10	0.053
200	0.46	0.81	14.1	16.8	0.13	0.022	0.003

The simulated I_n curves, which corresponds to the total dynamic power ($P_{dyn} + P_{sc}$), are given in Fig. 4. When C_O is 0, the dynamic switching current is not 0 due to the existence of the diffusion capacitance. Increasing C_O will increase the switching current, thereby increasing the total dynamic power. Table I under $P_{sc}/(P_{dyn} + P_{sc})$ column shows short circuit power as a percentage of the total power at different C_O and input rise times. Clearly, the larger the input rise time, the larger is the percentage of the short

circuit power of the total dynamic power. On the other hand, increasing load capacitance will increase total dynamic power and decrease P_{sc} , thereby decreasing the percentage of short circuit power. Again, we observe that the zero-load short circuit power model [6] significantly overestimates the P_{sc} .

IV. SIMULTANEOUS DUAL- V_{th} ASSIGNMENT AND GATE-SIZING ALGORITHM

In this section, we present a TILOS-like sensitivity-based algorithm for simultaneous dual- V_{th} assignment and gate-sizing to achieve low power dissipation with high performance. TILOS (Timed Logic Synthesis) is a transistor sizing technique for minimization of the area subject to delay constraint [1]. It is an iterative process. First, all the transistors are set to minimal size. Then, static timing analysis is used to find the critical path that determines the maximum speed of the circuit. Critical delay ($T_{critical}$) is the delay along the critical path. In order to reduce the critical delay with minimal increase in the area, TILOS finds the sensitivity of delay with respect to the area for all transistors in the critical path and up-size the transistor with the largest sensitivity. The delay of each transistor is then updated by static timing analysis. The process is iterated until the delay constraint is satisfied.

We modify the TILOS algorithm to minimize the total power and critical delay by simultaneous gate-sizing and dual- V_{th} assignment. Let us first consider some definitions for ease of understanding the algorithm. A gate can be represented by an equivalent inverter. W_{EFFP} and W_{EFFN} are the equivalent width of the PMOSFET and NMOSFET of the inverter. In order to compensate for the different mobilities of electrons and holes, we assume that $W_{EFFP} = 3W_{EFFN}$. In our analysis, discrete gate sizes are considered. For simplicity, we also assume a uniform threshold voltage for all the transistors within a gate.

There may be more than one critical path. The critical path set includes all the gates in the critical path(s). Static timing analysis can be used to determine the critical-path set.

Sizing or changing the threshold of a gate (G) affects not only its delay and power, but also the delay and power of its neighbors. For example, sizing a gate (G) changes the load capacitance of its fanin gates (GI) and the input transition time of its fanout gates (GO). Due to the change of the load capacitance of GI , the output transition time of GI will be different, which in turn change the delay and power of all the fanout gates of GI (GIO). If we change the threshold voltage of a gate (G), the delay and power of GO will change. To summarize, sizing or changing the threshold voltage of G may influence the delay and power of G , GI , GO and GIO . Therefore, we define G , GI , GO and GIO as the neighbor set of the gate G .

Suppose a gate x is in the critical path. Its fanin gate and fanout gate in the critical path set are i and o , respectively. We assume the gate size and threshold voltage of x are w and v , respectively. Sensitivity of delay to power of the

gate x with respect to w and v can be expressed as

$$S_w = \left. \frac{\Delta D}{\Delta P} \right|_{w=w+\Delta w} \quad (10)$$

$$S_v = \left. \frac{\Delta D}{\Delta P} \right|_{v=v-\Delta v} \quad (11)$$

$$D = t_d(x) + t_d(i) + t_d(o) \quad (12)$$

$$P = \sum_j (P_{dyn_j} + P_{sc_j} + P_{leak_j}) \quad (13)$$

where the summation is taken over the gates in the neighbor set of x .

A TILOS-like sensitivity-based algorithm is used to minimize the total power for high performance dual- V_{th} CMOS circuits. We start from the minimal size and single high- V_{th} circuit, which corresponds to minimal dynamic switching power and minimal static leakage power. To optimize the circuit, we seek in the critical path set a gate that can reduce the critical delay the best but with a minimal increase in the total power. S_w and S_v of all the gates in the critical path set are calculated and the most sensitive gate and the corresponding parameter $u \in \{w, v\}$ are determined. We modify the most sensitive gate by either increasing the size of the gate or reducing the threshold voltage depending on u . Then, we update the circuit using static timing analysis. We iterate the process until changing gate size and threshold voltage can not further improve the critical delay. The algorithm is guaranteed to converge due to the following reasons: Increasing the size of a gate can enhance the drive capability of the gate, but will increase the load capacitance of the given gate (increased diffusion capacitance) and its fanin gates (larger transistor gate capacitance). Therefore, beyond a certain point, increasing the size may not improve the performance. Lowering threshold voltage can improve the performance. However, if all the gates in the critical path set are already assigned to low V_{th} , we can not further reduce the delay by changing the threshold voltage. The pseudo code for the algorithm is listed below:

Simultaneous dual- V_{th} Assignment and Gate-sizing ()
 Initialize the gates with minimal size and high- V_{th}
 Static timing analysis
 Find the critical path set and calculate $T = T_{critical}$
 Calculate ΔD , ΔP and S_w, S_v of the critical path set
 While(at least one of the ΔD in critical path set is < 0)
 Find the most sensitive gate and modify it
 Static timing analysis
 Find the critical path set and $T = T_{critical}$
 Update the circuit and calculate total power of the circuit

V. IMPLEMENTATION AND RESULTS

Simultaneous dual- V_{th} assignment and gate-sizing algorithm is implemented in C under the Berkeley SIS environment. In order to simplify the analysis, technology-mapping is used to map the ISCAS benchmark circuits to a library that contains NAND, NOR and Inverter gates. MOSIS 0.25 μm technology is assumed in our analysis. The oxide thickness is 5.8nm. The low and high threshold voltages are 0.2V and 0.3V, respectively. The supply voltage is 1V. The minimal channel width and ΔW are assumed to be 0.5 μm .

Fig. 5 shows the power dissipation vs. critical delay for the optimization of the single low- V_{th} circuit (C880) by gate-sizing. The switching activity of the primary input (PI) is assumed to be 0.03. The circuit is initialized with

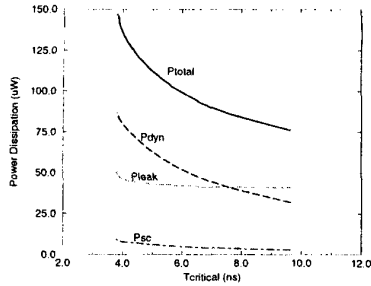


Fig. 5. Power vs. $T_{critical}$ for gate-sizing of single low- V_{th} circuit

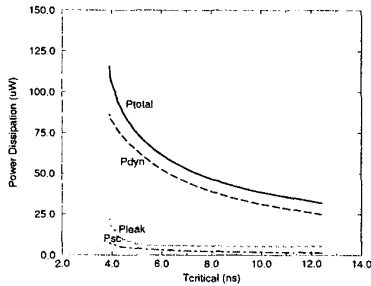


Fig. 6. Power vs. $T_{critical}$ for simultaneous dual- V_{th} assignment and gate-sizing

minimal size and low- V_{th} . The critical delay for the initialized circuit is 9.6ns. Then we keep increasing the size of the gates to reduce the critical delay, while the total power increases. After the gate-sizing, the critical delay is 3.8ns and the total power is $147\mu W$. The power and delay of C880 (by simultaneous dual- V_{th} assignment and gate-sizing) are illustrated in Fig. 6. At the start point of the optimization, all the gates have the minimal size and high- V_{th} and the critical delay is 12.4ns. After the simultaneous dual- V_{th} assignment and gate-sizing, the circuit has almost the same critical delay as the corresponding single low- V_{th} circuit, but the total power is $115.8\mu W$, which is 21% smaller than the corresponding single low- V_{th} circuit.

Table II gives the minimal critical delay and its corresponding total power for the ISCAS benchmark circuits that are obtained by sizing (S) or simultaneous dual- V_{th} assignment and gate-sizing (S+D). It can be seen that the critical delay of the circuit by simultaneous dual- V_{th} assignment and gate-sizing is close to that of the corresponding single low- V_{th} circuit by gate-sizing. The difference is due to the discrete gate sizes we use in our analysis. For circuits C5315 and C1908, the critical delays are even smaller. The total power of most of the circuits can be reduced. For circuits C5315 and C7552, the total power can be reduced by more than 20% and 40% at 0.1 and 0.03 switching activities, respectively. Consider the average case. the total power can be reduced by 14% and 24% at 0.1 and 0.03 switching activities, respectively.

TABLE II

CRITICAL DELAY AND P_T (S: SIZING OF SINGLE LOW- V_{th} CIRCUIT, S+D: SIMULTANEOUS GATE-SIZING AND DUAL- V_{th} ASSIGNMENT)

Circuit Chosen	PI's activity	critical delay (ns)			P_T (μW)		
		S	S+D	%	S	S+D	%
C432	0.1	5.52	5.61	1.6	147	133	-9.5
	0.03	5.52	5.57	0.9	77.9	66.7	-14.4
C499	0.1	3.44	3.49	1.5	532	553	3.9
	0.03	3.43	3.45	3	271	278	2.6
C880	0.1	3.8	3.9	2.6	315	270	-14.3
	0.03	3.81	3.9	2.4	147	115.8	-21.2
C1355	0.1	3.77	3.81	1	496	490	-1.2
	0.03	3.81	3.85	1	260	261	0.4
C1908	0.1	6	5.56	-7.3	371	346	-6.7
	0.03	6	5.57	-7.2	205	168	-18
C2670	0.1	8.83	9	1.9	568	459	-19.2
	0.03	8.83	9.1	3	269	173	-35.7
C3540	0.1	7.1	7.2	1.4	517	391	-23.1
	0.03	7.1	7.26	2.2	295	182	-38.3
C5315	0.1	9	8.25	-8	871	695.5	-20.1
	0.03	9	8.25	-8	463	277	-40.2
C6288	0.1	14.4	14.8	2.8	715	567	-20.7
	0.03	14.4	14.8	2.8	489	352	-28
C7552	0.1	8.56	8.66	1.1	1156	850	-26.5
	0.03	8.6	8.8	2.3	627.5	349.5	-44
average	0.1	-	-	-0.14	-	-	-14
	0.03	-	-	0.24	-	-	-24

VI. SUMMARY

In this paper, we present a simultaneous dual- V_{th} assignment and gate-sizing algorithm to optimize power and performance of CMOS circuits. A short circuit current model that considers the effect of load capacitance is derived and a sensitivity-based TILOS-like algorithm is used to minimize the total power for high performance CMOS.

REFERENCES

- [1] J. Fishburn and A. Dunlop "TILOS: a posynomial programming approach to transistor sizing", *IEEE International Conference on Computer-Aided Design*, 1985, pp. 326-328.
- [2] U. Ko and P. T. Balsara, "Short-Circuit Power Driven Gate Sizing Technique for Reducing Power Dissipation", *IEEE Transaction on VLSI Systems*, Vol. 3, No. 3, 1995, pp. 450-455.
- [3] J.D. Meindl, "Low power Microelectronics: Retrospect and Prospect", *Proceedings of the IEEE*, Vol.83, No.4, 1995, pp. 619-635
- [4] A.P. Chandrakasan, et al., "Low-Power CMOS Digital Design", *IEEE Journal of Solid-State Circuits*, Vol.27, No.4, 1992, pp. 473
- [5] T. Sakurai, et al., "Alpha-Power Law MOSFET Model and its Application to CMOS Inverter Delay and Other Formulas", *IEEE Journal of Solid-State Circuits*, Vol.57, No.2, 1990, pp. 584-593.
- [6] H. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on design of buffer circuits", *IEEE Journal of Solid-State Circuits*, Vol.SC-19, 1984, pp. 468-473.
- [7] M. Borah, et al., "Transistor Sizing for Low Power CMOS Circuits", *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 15, No. 6, 1996, pp. 665-671.
- [8] Jan M. Rabaey, *Digital Integrated Circuits*, New Jersey: Prentice-Hall, 1996.
- [9] L. Wei, et al., "Design and Optimization of Dual Threshold Circuits for Low Voltage Low Power Applications," *IEEE Trans. on VLSI Systems*, Vol. 7, 1999, pp. 16-24.
- [10] S. Sirichotiyakul, et al., "Stand-by Power Minimization through Simultaneous Threshold Voltage Selection and Circuit Sizing", *ACM/IEEE Design Automation Conf.*, 1999, pp. 436-441.
- [11] T.L. Chou and K. Roy, "Estimation of Sequential Circuit Activity Considering Spatial and Temporal Correlations," *IEEE Intl. Conf. on Computer Design*, 1995, pp. 577-583.
- [12] M.C. Johnson, D. Somasekhar, and K. Roy, "Models and Algorithms for Bounds on Leakage in CMOS Circuits", *Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol 18, No 6, 1999, pp. 714-725.
- [13] SS. Sapatnekar, et al., "An exact solution to the transistor sizing problem for CMOS circuits using convex optimization", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol.12, No.11, 1993, pp. 1621-34
- [14] A. R. Conn. et al., "Gradient-Based Optimization of Custom Circuits Using a Static-Timing Formulation", *ACM/IEEE Design Automation Conf.*, 1999
- [15] C.-P. Chen, et al., "Fast and Exact Simultaneous Gate and Wire Sizing by Lagrangian Relaxation", *IEEE International Conference on Computer Aided Design*, 1998, pp. 617-624