Power Optimal MTCMOS Repeater Insertion for Global Buses

Hanif Fatemi
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-3724
fatemi@usc.edu

Behnam Amelifard
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-9481
amelifar@usc.edu

Massoud Pedram
Department of EE-Systems
University of Southern California
Los Angeles, CA
(213) 740-4458
pedram@ceng.usc.edu

ABSTRACT

This paper addresses the problem of power-optimal repeater insertion for global buses in the presence of crosstalk noise. MTCMOS technique by inserting high-V_{th} sleep transistors to reduce the leakage power consumption in the idle mode is used. We simultaneously calculate the repeater sizes, repeater distances, and the size of the sleep transistors to minimize the power dissipation. The effect of crosstalk coupling capacitance on propagation delay and (switching and short circuit) power dissipation is considered. Experimental results show that depending on the activity factor of the circuit, the proposed technique can significantly reduce the power consumption of the global bus interconnects.

Categories and Subject Descriptors

B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aides

General Terms

Algorithms, Design, Performance

Keywords

Low-power design, buffer insertion, MTCMOS circuits

1. INTRODUCTION

As the CMOS technology continues to scale down toward Ultra Deep Sub-Micron (UDSM) technologies, more functionality is being integrated on a single die. This drastic integration results in increase in the size of the die, and consequently in the number of long global interconnects and in the length of them. The interconnect delay becomes the dominant factor to determine the overall performance of the integrated circuits. Since the delay of an interconnect is quadratic in its length, repeater insertion has been widely used to reduce the delay. As shown in [1] the repeaters can be optimally sized and separated to minimize the interconnect delay. The size of an optimal repeater is typically much larger than a minimum-sized repeater. Since millions of repeaters will be inserted to drive global interconnects, significant power will be consumed by these repeaters, particularly if delay-optimal repeaters are used [3]. Several works used the extra tolerable delay for power saving in interconnects. Authors in [2] and [3] provided analytical methods to compute unit length power optimal repeater sizes and distances. The power analysis should consider switching, leakage and short circuit accurately. As the technology scales down wires are laid out closer to each other which in turn increases the capacitive coupling noise on the interconnection lines. This will affect both delay and power consumption in interconnects. In addition to switching power on the coupling capacitances, the authors of [4] showed that the short circuit power consumption is increased significantly in the presence of crosstalk noise. Therefore, one should also consider this effect in the design of power optimal repeaters. Moreover, the technology scaling has resulted in large increase in leakage current. Leakage power has grown exponentially to become a significant fraction of the total chip power consumption [5]. Authors in [6] studied the applicability of MTCMOS to repeater design for leakage power saving, however they did not provide a mathematical solution for the simultaneous optimal sizing of the sleep transistors and repeaters and the insertion length. In addition the effect of crosstalk on delay and power has not been taken into account for the optimal design.

This paper studies the opportunity of minimizing the average power consumption during both active and standby mode of the bus lines by simultaneously computing repeater sizes, repeater insertion lengths, and the size of the sleep transistors subject to a delay constraint in the presence of crosstalk noise. We consider the worst case crosstalk for the delay constraint. However the assumption of worst case crosstalk is not realistic for power optimization. More precisely, the objective is to minimize the average power (in contrast to minimizing the maximum power). Therefore, we show how to estimate the average power as a function of probability of different types of transitions on the coupled lines. We will also discuss the delivery circuitry of the sleep signals to the sleep transistors.

The remainder of this paper is organized as follows. Section 2 presents the delay and power models in the presence of crosstalk. Power optimization of bus lines by utilizing sleep transistors is presented in section 3. Experimental results and conclusions are presented in section 4 and 5 respectively.

2. PRELIMINARIES

This section describes our delay and power model. We also explain the delay-optimal buffer size and insertion length in the presence of crosstalk noise.

This research was sponsored by a grant from Semiconductor Research Corporation (SRC Task ID 1423.001)

ISLPED'07, August 27–29, 2007, Portland, Oregon.

2.1 Delay Model

Consider a uniform interconnection line of resistance r per unit length and capacitance c per unit length, and total length of L. Suppose the line is divided into L/l segments and identical repeaters of unit driving resistance r_s, unit input capacitance c_g, unit output capacitance c_p, and size s are inserted at each segment (c.f. Figure 1 for a pictorial). Figure 2 shows one stage of the repeater chain with the interconnect model in between. The delay and the transition time of a segment comprising of a repeater driving an interconnect segment of length l terminated with a repeater of the same size and driven by a step input are $\ln 2 \cdot \tau$ and $\ln 9 \cdot \tau / 0.8$, respectively. Note that $\tau = r_s\left(c_g + c_p\right) + r_s cl/s + rlsc_g + \frac{1}{2}rcl^2$. With a finite input slew rate, the contribution of the input transition time t_r to the repeater delay can be represented by γt_r [10] where, for a rising input, γ is calculated as: $\gamma = \frac{1}{2} - \left(1 - V_{tn}/V_{dd}\right)/\left(1+\alpha_n\right)$ where V_{tn} is the threshold of the NMOS and α_n is the NMOS alpha-power parameter. Similarly, for a falling transition, γ is calculated from the PMOS parameters. An average value for γ is used. Therefore the delay of one repeater stage is given by $\left(\ln 2 + \gamma \cdot \ln 9/0.8\right)\tau$.

Figure 3 shows the delay model for two adjacent bus lines. c_c is the coupling capacitance per unit size. We assume zero skew between the transitions launched into the lines. The worst case delay occurs when transitions on these two lines are in opposite directions.

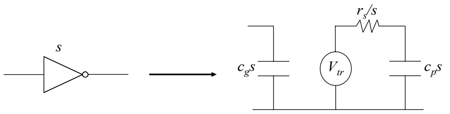

Figure 1. Buffer model

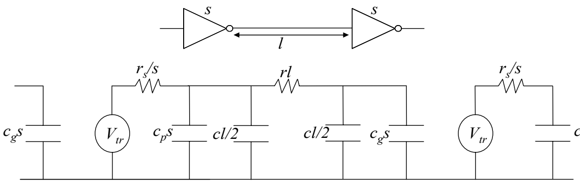

Figure 2. One stage of repeaters with interconnect model

We model the Miller effect in coupling capacitance (to create the worst case delay conditions) by rewriting the formula for the time constant τ as follows:

$$\tau = r_s\left(c_g + c_p\right) + \frac{r_s}{s}\left(c + 2c_c\right)l + rlsc_g + \left(\tfrac{1}{2}c + c_c\right)rl^2 \qquad (1)$$

The total delay of the interconnection line is equal to $\tau.(L/l)$. Therefore, minimizing the total delay is equivalent to minimizing the time constant per unit length i.e., τ/l:

$$\frac{\tau}{l} = \frac{1}{l}r_s\left(c_g + c_p\right) + \frac{r_s}{s}\left(c + 2c_c\right) + rsc_g + \left(\tfrac{1}{2}c + c_c\right)rl \qquad (2)$$

With a derivation similar to that given in [1], the worst case delay per unit length of interconnect line (in the presence of crosstalk) is minimized when:

$$l_{opt} = \sqrt{\frac{2r_s\left(c_g + c_p\right)}{r\left(c + 2c_c\right)}}\ ,\ s_{opt} = \sqrt{\frac{r_s\left(c + 2c_c\right)}{rc_g}} \qquad (3)$$

and,

$$\left(\frac{\tau}{l}\right)_{opt} = 2\sqrt{r_s c_g r\left(c + c_c\right)}\left(1 + \sqrt{\frac{1}{2}\left(1 + \frac{c_p}{c_g}\right)}\right) \qquad (4)$$

It has been shown in [2] and [3] that the optimal delay per unit length (and therefore the optimal total delay) is insensitive to both the size of the repeaters and the distance between repeaters. Hence, significant power and area can be saved by allowing a small delay penalty. Therefore, one can use repeaters with sizes smaller than s_{opt} and segment lengths longer than l_{opt}, and achieve a significant power saving. To accurately address this power optimization problem, we first present the power dissipation model of the global buses and then introduce our power optimal repeater design methodology.

2.2 Power Dissipation Model

The power dissipation of a global bus line has three components: switching power, short circuit power, and leakage power.

2.2.1 Switching Power Dissipation

This is due to charging and discharging of the load capacitance. The switching power for one stage can be calculated as:

$$P = \alpha f V_{dd}^2\left(s\left(c_g + c_p\right) + lc\right) \qquad (5)$$

where α is the switching activity of the inverter, f is the frequency, and V_{dd} is the supply voltage. Note that equation (5) does not consider the switching power consumed on the coupling capacitances. When only one of the lines switches, the coupling capacitance $c_c \cdot l$ charges or discharges with a voltage level change of V_{dd}. Therefore, its coupling energy consumption is $0.5c_c lV_{dd}^2$. When two adjacent bus lines are simultaneously switching in the opposite directions, the coupling capacitance ($c_c \cdot l$) charges or discharges with a voltage level change of $2V_{dd}$. Therefore, the total energy consumption by the drivers of both lines is $0.5c_c \cdot l(2V_{dd})^2$ [16]. Finally when two adjacent bus lines make transitions in the same direction, no coupling energy is consumed. To estimate the average switching power consumption on a single stage of the repeater chain, we make the following assumptions: i) Assume that there is no temporal and spatial correlation between the data which is being transmitted through the two adjacent bus lines. ii) The probability of transmitting a '1' ('0') is equal to 'p'('1–p'). As a result, the probability of the transition between two consecutive data bits on a single bus line can be calculated as k_1=p(1–p). To calculate the average coupling power, we need to calculate the probability of each type of transition on the coupling capacitance between two adjacent lines. Table 1 presents these probabilities for all possible scenarios. Note that $\sum_{i=2}^{5} k_i = 1$.Using the values of k_1 to k_5, we can write the average switching power consumption for one stage of two adjacent bus lines (Figure 3):

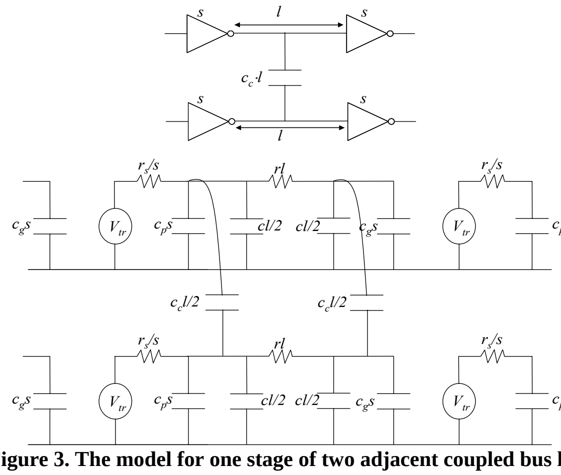

Figure 3. The model for one stage of two adjacent coupled bus lines

$$P_{sw} = 2 \times 0.5 k_1 f V_{dd}^2 \left(s\left(c_g + c_p\right) + lc \right) + 0.5 k_2 f \left(2V_{dd}\right)^2 lc_c + 0.5 k_3 f {V_{dd}}^2 lc_c \tag{6}$$

Without loss of generality and for the sake of the presentation, we will limit ourselves to only two adjacent lines. The analysis for three (and more) bus lines is similar.

In general, if the input pattern and the spatial-temporal correlation between the data bits of a single line or two adjacent lines are available, a number of probabilistic techniques such as [13]-[15] can be used to estimate k_1 to k_5. Furthermore, several encoding techniques have been proposed for minimizing coupling effect for static on chip bus structures [7]-[9]. Some approaches were also introduced to find a permutation for the bus lines for minimizing the crosstalk effects [11],[12]. The impact of these optimization techniques can be captured by appropriately revising the equations for k_1 to k_5. The rest of the analysis remains the same.

Table 1: Probability of different switching scenarios on the coupling capacitances

Transition Type	Occurrence Probability
Opposite direction	$P(\uparrow\downarrow) = k_2 = 2p^2(1-p)^2$
One switches and other is quiet	$P(\uparrow -) = k_3 = 4p(1-p)\left(p^2 + (1-p)^2\right)$
Both quiet	$P(--) = k_4 = p^4 + (1-p)^4 + 2p^2(1-p)^2$
Same direction	$P(\uparrow\uparrow) = k_5 = 2p^2(1-p)^2$

The coupling power is also dependent on the relative switching time of the line drivers [16]. For global buses, we can safely assume zero skew between the drivers' switching times. However, one can consider the relative delay between the transitions of the two lines and use a similar approach as [16] to compute the effect of relative delay on coupling power.

2.2.2 Short-Circuit (SC) Power Dissipation

The SC power is consumed by the current flow between the power rails through a direct current path which is temporarily established during an output transition [19]. Several techniques have been proposed to estimate the SC power dissipation [19]-[22]. The SC power is a function of the input transition time, the output load capacitance, and the size of the transistor. Most of the previous works on power optimal repeater design either ignore the SC power consumption or use an inaccurate approximation of the SC power consumption. We use the closed form formula presented in [20] which captures the dependence of the SC power consumption on the circuit parameters.

The SC power consumption is increased significantly in the presence of crosstalk noise [4]. Therefore similar to switching power, we formulate the average short circuit power consumption based on the transition type probability on adjacent bus lines (Table 1). As shown in [20], the SC energy consumption of an inverter during a full signal switch (such as a falling transition followed by a rising) can be approximated as

$$E_{SC} = \frac{4s^2 {I_{d0}}^2 {t_r}^2 V_{dd}}{V_{dsat} G C_{out} + 2s \cdot H I_{d0} t_r} \tag{7}$$

where H and G are technology dependent parameters [20] and I_{d0} is the average saturated drain current of the NMOS and PMOS transistors of the minimum sized inverter. Due to the shielding effect of the interconnect resistance, the repeater sees a capacitance less than C_{total}, where C_{total} is the summation of repeater parasitic capacitances, interconnect capacitance and the coupling capacitances (considering the miller effect based on the transition type), e.g.,

$$C_{total}(\uparrow\downarrow) = \left(c_p + c_g\right)s + \left(c + 2c_c\right)l \tag{8}$$

Using the effective capacitance approach, the capacitance seen by the repeater for opposite direction transitions is written as:

$$C_{out}(\uparrow\downarrow) = C_{eff}(\uparrow\downarrow) = \left(c_p s + {cl}/{2} + c_c l\right) + \delta \cdot \left(c_g s + {cl}/{2} + c_c l\right) \tag{9}$$

where δ<1 and depends on l and s. The ratio of C_{eff} to C_{total} is also a function of l and s. Similar to [3], we calculate ω, the average ratio of C_{eff} to C_{total} for different types of transitions. This average ratio is used for short circuit evaluation. In addition, due to the impact of crosstalk on transition time, different values for t_r are used (by considering different τ values due to different coupling capacitances). Therefore, the average short circuit power consumption of the repeater (for one falling or rising transition) can be estimated as:

$$P_{SC} = k_2 \cdot \frac{2fs^2 {I_{d0}}^2 \cdot {t_{r_{(\uparrow\downarrow)}}}^2 \cdot V_{dd}}{V_{dsat} G \cdot \omega_{(\uparrow\downarrow)} C_{total_{(\uparrow\downarrow)}} + 2s \cdot H I_{d0} t_{r_{(\uparrow\downarrow)}}} + {k_3}/{2} \cdot \frac{2fs^2 {I_{d0}}^2 \cdot {t_{r_{(\uparrow -)}}}^2 \cdot V_{dd}}{V_{dsat} G \cdot \omega_{(\uparrow -)} C_{total_{(\uparrow -)}} + 2s \cdot H I_{d0} t_{r_{(\uparrow -)}}} + k_5 \cdot \frac{2fs^2 {I_{d0}}^2 \cdot {t_{r_{(\uparrow\uparrow)}}}^2 \cdot V_{dd}}{V_{dsat} G \cdot \omega_{(\uparrow\uparrow)} C_{total_{(\uparrow\uparrow)}} + 2s \cdot H I_{d0} t_{r_{(\uparrow\uparrow)}}} \tag{10}$$

2.2.3 Leakage Power Dissipation

The third source of the power dissipation is the leakage current. In the present CMOS technologies, the major components of the leakage current are sub-threshold and gate-tunneling currents [17]. The sub-threshold leakage is the drain-source current of a transistor operating in the weak inversion region which can be expressed as follows [17],

$$I_{sub} = A_{sub} \mu_0 C_{ox} \left(\frac{W}{L_{eff}} \right) e^{q\left(V_{GS} - V_{t0} - \gamma' V_{SB} + \eta V_{DS}\right)/n'kT} \left(1 - e^{-qV_{DS}/kT}\right) \tag{11}$$

where A_{sub}= $(kT/q)^2 e^{1.8}$, μ_0 is the zero bias mobility, C_{ox} is the gate oxide capacitance per unit area, W and L_{eff} denote the width and effective length of the transistor, k is the Boltzmann constant, T is the absolute temperature, and q is the electrical charge of an electron. In addition, V_{t0} is the zero biased threshold voltage, γ' is the linearized body-effect coefficient, η denotes the drain-induced barrier lowering (DIBL) coefficient, and n' is the sub-threshold swing coefficient of the transistor.

Since V_{DS} of the OFF transistor is V_{dd} which is more than a few kT/q~$26mV$, the sub-threshold leakage power of an NMOS transistor can be written as,

$$P_{sub,nmos} = A'_{sub} W_{nmos} \mu_{nmos} e^{-\lambda V_{tn}} \tag{12}$$

Where A'_{sub}=$A_{sub}V_{dd}$×C_{ox}/L_{eff}×exp($\lambda\eta V_{dd}$) and λ=$q/n'kT$ are technology constants. A similar formula can be derived for a PMOS transistor. Therefore, sub-threshold leakage power dissipation of a repeater can be written as,

$$P_{sub} = p \cdot P_{sub,pmos} + (1-p) \cdot P_{sub,nmos} \tag{13}$$

where p is the probability that the input of the inverter is at logic 1. If the ratio of the width of the PMOS transistor to that of the NMOS transistor is β, equation (13) can be re-written as:

$$P_{sub} = \frac{A'_{sub} \cdot s W_{\min}}{1+\beta} \left(p\beta \mu_{pmos} e^{-\lambda V_{tp}} + (1-p) \mu_{nmos} e^{-\lambda V_{tn}} \right) = K_{sub} \cdot s \tag{14}$$

where $W_{\min}$ is the minimum size of the inverter. Gate tunneling is the other major source of the leakage power. The major source of gate tunneling leakage in CMOS circuits is the gate-to-channel tunneling current of the ON NMOS transistors, which can be modeled as [17]-[18],

$$I_{tunnel} = A_{tunnel} W_{nmos} L_{eff} \left(\frac{V_{ox}}{t_{ox}} \right)^2 e^{-B\frac{t_{ox}}{V_{ox}}} \tag{15}$$

where A_{tunnel} and B are technology constants, and t_{ox} is the oxide thickness. V_{ox} is the potential drop across the oxide. When the transistor is ON, $V_{ox}=V_{gs}-\psi_s$, where ψ_s is the surface potential of the transistor [24]. Ignoring the gate-tunneling leakage of the PMOS, the gate tunneling leakage power dissipation of an inverter, P_{tunnel}, can be calculated by,

$$P_{tunnel} = \frac{A'_{tunnel}}{1+\beta} p \cdot s \cdot W_{\min} = K_{tunnel} \cdot s \tag{16}$$

where $A'_{tunnel} = A_{tunnel} L_{eff} V_{dd} \left(V_{dd} - \psi_s \right)^2 / t_{ox}{}^2 \exp\left(-Bt_{ox} / \left(V_{dd} - \psi_s \right) \right)$ is a coefficient independent of the size and threshold voltage of the inverter [17],[18].

2.2.3.1 Average Power Dissipation

Having obtained the equations for different components of the power dissipation in equations (6), (10), (14) and (16), the total average power dissipation for one stage of two adjacent bus lines in the active mode of circuit operation can be written as,

$$P_{active} = P_{sw} + 2P_{sc} + 2P_{sub} + 2P_{tunnel} \tag{17}$$

The factor 2 is due to the presence of two repeaters in one stage of two adjacent lines. Note that we have already considered the two repeaters on adjacent lines in the case of P_{sw} in equation (6). In the standby mode, however, the only sources of the power dissipation are the sub-threshold and gate-tunneling leakage; so,

$$P_{standby} = 2P_{sub} + 2P_{tunnel} \tag{18}$$

The average power consumption can be obtained as a weighted sum of the power consumption in the active and standby modes:

$$P_{total} = \chi P_{active} + (1-\chi) P_{standby} \tag{19}$$

where χ is the *active mode factor* of the circuit, i.e., the percentage of the time the circuit is in the active mode.

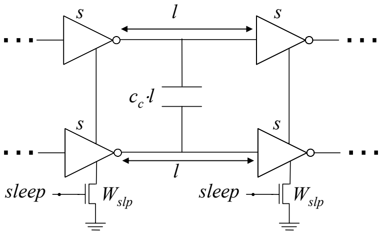

Figure 4. Sharing of sleep transistors among different bus lines

3. POWER OPTIMIZATION FOR MTCMOS DESIGN

3.1 Power and Delay Modeling

MTCMOS technology provides low leakage and high performance operation by utilizing high speed, low V_t transistors for logic cells and low leakage, high V_t devices as sleep transistors [23]. Sleep transistors disconnect logic cells from the supply and/or ground to reduce the leakage in the standby mode. The bus lines spend large percentage of the time in the standby mode. Therefore sleep transistors can be used for total power saving. The drawback is the increase in the delay in the active mode due to the additional resistance of the sleep transistors. Sleep transistors can be shared between the repeaters. Since repeaters are inserted at identical distances, we can share the sleep transistors between repeaters on different data lines. Figure 4 shows the case for only two adjacent bus lines. Similarly we can share the sleep for more than two bus lines. In the presence of sleep transistor both leakage components are substantially smaller in the standby mode. In the standby mode the virtual ground node (i.e., the drain terminal of sleep transistor) charges to a voltage near V_{dd} [23]; hence, the potential drop across the oxide of the ON NMOS transistors becomes very small and, from equation (15), the gate-tunneling leakage of the inverter becomes negligible. The sub-threshold leakage current and power dissipation can be calculated from equation (11) as,

$$\begin{aligned} P_{standby,MTCMOS} &= V_{dd} \cdot I_{sub,standby} \\ &= V_{dd} \cdot A_{sub} \mu_0 C_{ox} V_{dd} \frac{W_{slp}}{L_{eff}} e^{\lambda\left(-V_{t,high} + \eta V_{dd}\right)} \\ &= K_{standby,MTCMOS} \cdot s_{slp} \end{aligned} \tag{20}$$

where W_s and $V_{t,high}$ denote size and threshold voltage of the sleep transistor, $K_{standby,MTCMOS}$ is the sub-threshold current for the minimum size sleep transistor and s_{slp} is the size of the sleep transistor normalized to that of the minimum size transistor.

Using the MTCMOS technique, the total power of one stage of two adjacent bus lines can be written as:

$$P_{total,MTCMOS} = \chi P_{active} + (1-\chi) P_{standby,MTCMOS} \tag{21}$$

In order to consider the effect of the MTCMOS on the worst case delay constraint, we need to consider two cases:

I) Adjacent bus lines are switching in the opposite direction; therefore, the sleep transistor is contributing to a single falling transition. Using equation (1), the time constant for one stage can be written as:

$$\begin{aligned} d_1 = {} & r_s\left(c_g + c_p\right) + \frac{r_s}{s}\left(c + 2c_c\right)l + rlsc_g + \left(\tfrac{1}{2}c + c_c\right)rl^2 \\ & + \frac{r_{slp}}{W_{slp}}\left[s \cdot \left(c_g + c_p\right) + \left(c + 2c_c\right)l \right] \end{aligned} \tag{22}$$

II) Adjacent lines are switching in the same direction; when there are two simultaneous falling transitions, twice as much current has to be sunk through the sleep transistor. Therefore, the resistance of the sleep transistor should be doubled for the delay estimation. More precisely,

$$\begin{aligned} d_2 = {} & r_s\left(c_g + c_p\right) + \frac{r_s}{s}cl + rlsc_g + \tfrac{1}{2}crl^2 \\ & + \frac{2r_{slp}}{W_{slp}}\left[s \cdot \left(c_g + c_p\right) + cl \right] \end{aligned} \tag{23}$$

Note that the sleep transistors result in the delay increase only in the case of falling transitions at the output node of the repeaters. Therefore we introduce a new time constant as $d_1{}'=(\tau_1+d_1)/2$ and $d_2{}'=(\tau_2+d_2)/2$ where τ_1 (as in equation (1)) and τ_2 are the time constants for opposite and same direction transitions without any sleep transistors, respectively. The worst case delay per stage is equal to $\max\{d_1{}', d_2{}'\}$.

3.2 Sleep Signal Delivery Circuitry

An important issue in the design of MTCMOS circuits is how to deliver the sleep signal to all MTCMOS transistors in the design. The sleep signal should be fast enough to minimize the transition time of the system from the standby mode to active mode [23]. If the sleep signal driver circuit is improperly designed, it will result in unnecessary switching and leakage power consumption. To minimize the delay of the system for transition from the standby mode to active mode and also to reduce the power consumption of the sleep signal delivery circuit, we use asymmetric inverters in this network as depicted in Figure 5. In this figure, weak

Table 2: Technology Parameters Used in the Simulation Setup

Technology Parameter	V_{dd}	$V_{t,low}$	$V_{t,high}$	β	$K_{sub,NMOS}$ ($\mu W/\mu m$)	$K_{sub,PMOS}$ ($\mu W/\mu m$)	K_{tunnel} ($\mu W/\mu m$)	K_{MTCMOS} ($\mu W/\mu m$)	c_c (fF/mm)	c (fF/mm)	r (Ωmm)
Value	1.1V	0.25V	0.35V	2.2	881	301	273	58	53.68	19.41	1099.99

transistors are minimum-sized and have high threshold voltages. The rationale is that only the rise delay of the sleep signal plays a role in determining the wake-up delay of the circuit. The fall delay of the sleep signal, on the other hand, determines the active to standby mode delay which is not a critical factor. The sleep signal delivery circuit shown in Figure 5 not only minimizes the sleep signal propagation delay, but also linearly reduces the switching power dissipation of the sleep signal delivery circuit due to selective use of minimum-size transistors. At the same time, it exponentially reduces the leakage power of the sleep signal delivery circuit during the active mode of circuit operation by using high threshold voltage transistors in each inverter (which are OFF in the active mode).

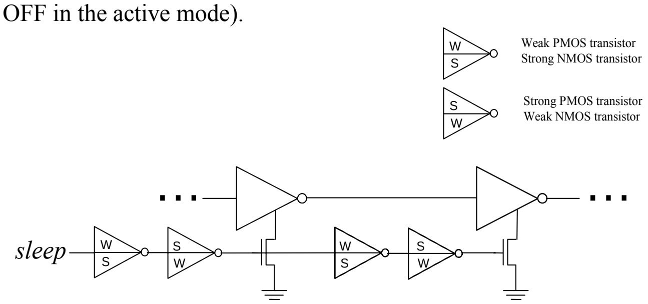

Figure 5. Using asymmetric inverters in the sleep signal delivery circuitry

3.3 Problem Formulation

Equation (4) gives the optimal worst-case delay per unit length for non-MTCMOS bus lines. Next we consider the same problem for the MTCMOS bus lines. Suppose a target end-to-end delay per unit length of interconnection line is given, which is expressed as Δ% more than $(\tau/l)_{opt}$. Given this target delay, we need to calculate the values of *l, s,* and W_{slp}, which minimize the total power dissipation. The total power for an interconnect of length L is equal to $P_{total\text{-}MTCMOS}\cdot(L/l)$ where $P_{total\text{-}MTCMOS}$ was given in equation (21). Therefore, a constrained minimization problem for $P_{total\text{-}MTCMOS}/l$ should be solved:

$$\begin{cases} \min \quad P\left(l,s,W_{slp}\right) \\ s.t \quad (1)\ Q\left(l,s,W_{slp}\right) \le T_{req} \\ \qquad (2)\ R\left(l,s,W_{slp}\right) \le T_{req} \end{cases} \tag{24}$$

$$\text{where } P \equiv \frac{P_{total-MTCMOS}}{l};\quad Q \equiv \frac{d'_1}{l};\quad \mathrm{R} \equiv \frac{d'_2}{l}$$

$$\text{and } T_{req} \equiv (1+\Delta)\left(\frac{\tau}{l}\right)_{opt}$$

The optimization problem can be solved by using the Lagrangian relaxation technique. In this technique, the constraints are relaxed and summed up in the objective function after multiplying them by non-negative coefficients, called the Lagrange multipliers:

$$F = P + \lambda_1 \cdot \left(Q - T_{req}\right) + \lambda_2 \cdot \left(R - T_{req}\right) \tag{25}$$

From the Lagrange method, the solution of the optimization problem (24) should satisfy the following set of conditions (known as the Kuhn-Tucker optimality conditions):

$$\begin{cases} \dfrac{\partial F}{\partial s} = 0;\quad \dfrac{\partial F}{\partial l} = 0;\quad \dfrac{\partial F}{\partial W_{slp}} = 0; \\ \lambda_1 \cdot \left(Q - T_{req}\right) = 0;\quad \lambda_2 \cdot \left(R - T_{req}\right) = 0; \end{cases} \tag{26}$$

These equations are solved numerically and the triplet (l ,s ,W_{slp}) which results in minimum $P_{total\text{-}MTCMOS}/l$ is selected.

4. EXPERIMENTAL RESULTS

To study the efficacy of the proposed technique, we conducted a comprehensive set of experiments. To extract the parameters which are used in the optimization problems, we performed transistor level simulation of devices in HSPICE [25] on a 45nm predictive technology model. All simulations were carried out at the frequency of 1GHz and die temperature of 100°C. The extracted technology parameters are reported in Table 2. MOSEK optimization toolbox [26] was used to solve the mathematical problem. Two coupled bus lines as described in the paper are used for our experiments. After optimizing the bus lines, the corresponding values of the design were extracted to SPICE netlist and detailed HSPICE simulations were performed to measure the worst-case delay and the average power consumption of the buffer chain. We first calculated the average power consumption when the worst case delay is optimized. These values are reported in Table 3 as P_D. The measurements were done for different active mode factors, χ. The power-optimal solutions with 10% delay penalty and for different χ, without using MTCMOS sleep transistors and with only two degrees of freedom, *s* and *l*, are reported as P_P in the table. Finally, the power optimal solutions with MTCMOS sleep transistors are reported as P_M in the table. When the percentage of the time that the circuit is in the active mode (i.e., χ) is small, the dominant component of the power consumption is the standby leakage. Therefore, MTCMOS technique results in significant power savings compared to P_D and P_P. As χ increases, the power saving diminishes. Since the active mode factor of global buses is usually very small, one can see that the power saving achieved by applying our technique is high. Note that the sleep signal delivery was achieved by the circuit shown in Figure 5 and its power dissipation overhead was considered in the total power consumption results.

In the second set of our experiments, where results are presented in Table 4, we compared the efficacy of the proposed technique for different values of delay penalty. More precisely, here the value of χ assumed to be 10% and the delay penalty Δ was varied from 5% to 40%. For each case, P_P and P_M were measured by HSPICE simulation. As we increase the delay penalty, the power reduction in both P_P and P_M increases. This power saving saturates as we increase Δ. Table 5 reports the optimal parameter values for the power-optimized design using the MTCMOS technique. The design parameters are normalized with respect to the delay-optimized repeater size (s_{opt}) and insertion length (l_{opt}). It is observed that by increasing Δ, both repeater and sleep sizes are decreasing. However, decrease in the sizes diminishes as the delay budget increases.

Finally, we compared our results with a two-step approach to design MTCMOS repeaters. In this two-step approach, first the power-optimal solution with no sleep transistor is found; then the size of the sleep transistors is calculated based on the power-optimal *l* and *s* values of the first step. We assume equal Δ% in each step of this approach. Therefore for a fair comparison we have to compare the two-step approach results with our solution with $(2\Delta+\Delta^2)\% \approx 2\Delta\%$ delay penalty. Table 6 compares the average power consumption achieved by our technique with that of two-step approach, denoted as P_T. It is seen that on average, our approach gives about 9.5% improvements in average power consumption over the two-step solution.

5. CONCLUSION

This paper addressed the problem of power-optimal repeater insertion for global buses in the presence of crosstalk noise. We used MTCMOS technique by inserting high-V_{th} sleep transistors to reduce the leakage power consumption in the idle mode. By accurately modeling different components of the power consumption and the delay, a mathematical problem was formulated for minimizing the average power under a timing constraint. Detailed HSPICE simulation showed that by considering the effect of crosstalk on both delay and power consumption, and by using MTCMOS technique, the average power consumption of the bus lines can be reduced by more than 50% with a small delay penalty of 5%.

Table 3: Power consumption results for different designs activity mode factor χ. Frequency=1GHz.

χ	P_D (μW)	P_P (μW)	P_M (μW)	P_M reduction over P_D (%)	P_M reduction over P_P (%)
1%	59.1	24.2	9.9	83.3	59.3
2%	66.1	28.0	11.6	82.4	58.6
5%	87.3	39.4	22.4	74.4	43.2
10%	122.6	58.4	46.3	62.2	20.7
20%	193.1	96.3	89.3	53.8	7.3
30%	263.7	134.2	132.9	49.6	1.0

Table 4: Power consumption results for different delay penalties. Frequency=1GHz, L=10mm, χ=10%

Δ	P_P (μW)	P_M (μW)	P_M reduction over P_D (%)	P_M reduction over P_P (%)
5%	73.1	56.1	54.2	23.2
10%	58.4	46.3	62.2	20.7
15%	51.2	41.1	66.5	19.7
20%	49.1	36.7	70.0	25.3
25%	43.0	36.1	70.5	15.9
30%	38.0	32.7	73.4	14.0
35%	37.7	29.3	76.1	22.3
40%	33.2	29.0	76.4	12.7

Table 5: Design parameters for the optimized MTCMOS design. Frequency=1GHz, L=10mm, χ=10%

Δ	s/s_{opt}	l/l_{opt}	W_{slp}/s_{opt}
5%	0.79	1.21	3.89
10%	0.70	1.43	2.90
15%	0.63	1.57	2.47
20%	0.57	1.71	2.20
25%	0.53	1.82	2.01
30%	0.51	1.93	1.88
35%	0.48	2.07	1.77
40%	0.45	2.14	1.68

Table 6: Comparing the proposed technique with a two-step approach to design MTCMOS repeaters

Delay Penalty	P_T (μW)	P_M (μW)	P_M reduction over P_T (%)
5%	56.7	56.1	0.9
10%	49.6	46.3	6.8
15%	44.6	41.1	8.0
20%	40.2	36.7	8.7
25%	39.7	36.1	9.0
30%	35.8	32.7	8.7
35%	35.3	29.3	17.1
40%	34.8	29.0	16.8

6. REFERENCES

[1] H. B. Bakoglu and J. D. Meindl, “Optimal interconnection circuits for VLSI,” *IEEE Trans. on Electron Devices*, vol. ED-32, no. 5, pp. 903–909, May 1985.

[2] K. Banerjee and A. Mehrotra, “A power-optimal repeater insertion methodology for global interconnects in nanometer designs,” *IEEE Trans. on Electron Devices*, vol. 49, pp. 2001–2007, Nov. 2002.

[3] G. Chen and E. Friedman, “Low power repeaters driving RC interconnects with delay and bandwidth constraints,” in *Proc.* of *ASIC/SOC*, pp. 335–339, 2004.

[4] H. Fatemi, S. Nazarian, and M. Pedram, "A current-based method for short circuit power calculation under noisy input waveforms," in *Proc. of ASP-DAC*, pp. 774-779, 2007.

[5] Semiconductor Industry Association, International Technology Roadmap for Semiconductors, 2003 edition, http://public.itrs.net/

[6] R. Rao, K. Agarwal, D. Sylvester, *et al.* “Approaches to run-time and standby mode leakage reduction in global buses,” in *Proc. of ISLPED*, pp. 188-193, 2004.

[7] S. Ramprasad, N. R. Shanbhag, and I. N. Hajj, "A Coding framework for low-power address and data busses." *IEEE Trans. on VLSI*, vol. 7, no. 2, pp. 212-221, June 1998.

[8] L. Benini, G. D. Micheli, E. Macii, *et al*, "Power optimization of core-based systems by address bus encoding," *IEEE Trans. on VLSI*, vol. 6, no. 4, pp. 551-562. Dec. 1998.

[9] Y. Shin, S. Chae, and K. Choi, "Partial bus-invert coding for power optimization of application-specific systems,” *IEEE Trans. on VLSI*, vol. 9, no. 2, pp. 377-383, Apr. 2001.

[10] T. Sakurai and A. R. Newton, "A simple MOSFET model for circuit analysis," *IEEE Trans. on Electron Devices*, vol. 38, no. 4, pp. 887-894, Apr. 1991.

[11] H. Zhou and D. Wong. "Global routing with crosstalk constraints," in *Proc. of DAC*, pp. 374-377, 1998.

[12] I. Jiang, Y. Chang, and I. Jou, "Crosstalk-driven interconnect optimization by simultaneous gate and wire sizing," *IEEE Trans. on CAD*, vol. 19, no. 9, pp. 999-1010. Sep. 2000.

[13] R. Marculescu, D. Marculescu, and M. Pedram, "Switching activity estimation considering spatiotemporal correlations," in *Proc. of ICCAD*, pp. 294-299, 1994.

[14] R. Marculescu, D. Marculescu, and M. Pedram, "Probabilistic modeling of dependencies during switching activity analysis," *IEEE Trans. on CAD*, vol. 17, no. 2, pp. 73-83, Feb. 1998.

[15] M. Xakellis and F. Najm, “Statistical estimation of the switching activity in digital circuits,” in *Proc. of DAC,* pp. 728-733, 1994.

[16] D. Sinha, D. Khalil, Y. Ismail, and H. Zhou, “A timing dependent power estimation framework considering coupling”, in *Proc. ICCAD*, pp. 401-407, 2006.

[17] V. De, A. Keshavarzi, S. Narendra, *et al.* "Techniques for leakage power reduction," in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, *et al.* Eds. Piscataway, NJ: IEEE, 2001.

[18] D. Lee, D. Blaauw, and D. Sylvester, "Gate oxide leakage current analysis and reduction for VLSI circuits," *IEEE Trans. on VLSI*, vol. 12, no. 2, pp. 155-166, Feb. 2004.

[19] H. Veendrick, “Short circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits,” *IEEE Jour. Solid- State Circuits*, vol. SC-19, pp. 468-473, 1984.

[20] K. Nose and T. Sakurai, "Analysis and future trend of short circuit power," *IEEE Trans. on CAD*, vol. 19, no. 9, pp. 1023-1030, Sept. 2000.

[21] M. Pedram, “Power minimization in IC design: principles and applications,” *ACM Trans. on Design Automation of Electronic Systems*, vol. 1, no. 1, pp. 3-56, 1996,.

[22] E. Acar, R. Arunachalam, and S. R. Nassif, “Predicting short circuit power from timing models,” in *Proc. of ASP-DAC*, pp. 277-282, 2003.

[23] A. Abdollahi, F. Fallah, and M. Pedram, "An effective power mode transition technique in MTCMOS circuits," in *Proc. of DAC*, pp. 37-42, 2005.

[24] Y.Taur and T.H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998.

[25] *HSPICE: The Gold Standard for Accurate Circuit Simulation,* www.synopsys.com/products/mixedsignal/hspice/hspice.html

[26] MOSEK Optimization Software, http://www.mosek.com