A Power-Optimal Repeater Insertion Methodology for Global Interconnects in Nanometer Designs

Kaustav Banerjee, *Member, IEEE*, and Amit Mehrotra, *Member, IEEE*

***Abstract*—This paper addresses the problem of power dissipation during the buffer insertion phase of interconnect performance optimization. It is shown that the interconnect delay is actually very shallow with respect to both the repeater size and separation close to the minimum point. A methodology is developed to calculate the repeater size and interconnect length which minimizes the total interconnect power dissipation for any given delay penalty. This methodology is used to calculate the power-optimal buffering schemes for various ITRS technology nodes for 5% delay penalty. Furthermore, this methodology is also used to quantify the relative importance of the various components of the power dissipation for power-optimal solutions for various technology nodes.**

***Index Terms*—Buffer insertion, delay optimization, leakage power, low-power design, power modeling and optimization, RC interconnects, repeaters, short-circuit power, very large-scale integration (VLSI).**

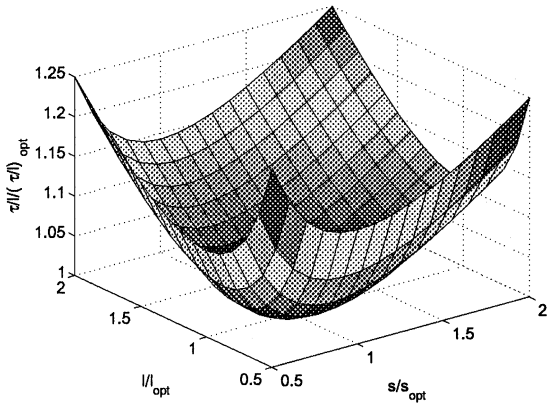


Fig. 1. Normalized delay per unit length as a function of buffer size and interconnect length for 180-nm top layer metal.

I. Introduction

As very large-scale integration (VLSI) circuits continue to be scaled aggressively past the 180-nm technology node, performance of these ICs is being increasingly dominated by the global interconnects [1], [2]. With technology scaling, more and more functionality is being integrated on-chip which results in an increase in the die size in spite of the reduction in minimum feature size [1].[1] As a result, the number of long global lines and the length of these global lines increases with technology scaling. Since the delay of a long unbuffered line is quadratic in its length, long interconnects are divided into a number of segments with repeaters or buffers. The delay of an optimally buffered line is linear in its length [3]. However, for large high-performance designs, the number of such repeaters can be prohibitively high [4] ($> 10^6$ for sub-100-nm designs) and can take up significant fraction of active silicon and routing area [2]. Additionally, as the total chip capacitance (dominated by interconnect network capacitance), operating frequency, and leakage current increases with scaling, total chip power dissipation is increasing rapidly [1], [5]. A significant fraction of the total chip power dissipation arises due to the loading caused by long global- and semi-global-tier interconnect networks, especially in high-performance designs. For example, it has been reported that around 40%-70% of the total power consumption could be due to the clock distribution network [6], [7].

In general, the repeaters are optimally sized and separated to minimize the interconnect delay. However, since these optimally sized repeaters are quite large ($\sim$ 450 times the minimum sized inverter available in the relevant technology for global-tier lines [8]) and also dissipate a significant amount of power, the total power dissipation by such repeaters in large high-performance designs can be prohibitively high. However, as shown in Fig. 1, the interconnect delay is actually very shallow with respect to both the repeater size and separation close to the minimum point. Since, all global interconnects are not on the critical path, a small delay penalty can be tolerated on these noncritical interconnects and there exists a potential for large power savings by using smaller repeaters and larger inter-repeater interconnect lengths.

Some previous work can be found in the literature, which attempt to address the issue of optimizing the repeater design for reduced delay and power [9], [10]. However, these analyses either ignore the leakage power [9], or ignore both the leakage and the short-circuit components of power dissipation [10]. For sub-180-nm VLSI technologies, the leakage power is increasing rapidly [11], and the short-circuit power has also been shown to be a significant fraction (up to 20%) of the total power dissipation for low-power and high-speed CMOS VLSI designs [12]. Hence, ignoring them in the power modeling and optimization process can lead to significant errors and can seriously compromise the validity of the optimized parameters. Furthermore, these analyses do not provide any closed-form expressions for

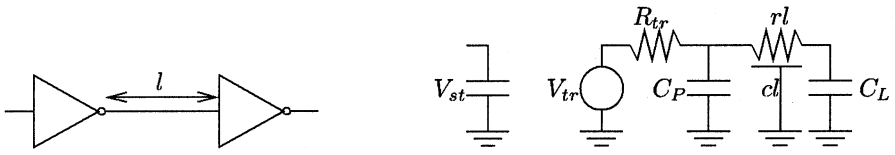
Manuscript received January 30, 2002; revised May 16, 2002. The review of this paper was arranged by Editor T. Skotnicki.

K. Banerjee is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: kaustav@ece.ucsb.edu).

A. Mehrotra is with the Computer and Systems Research Lab, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: amehrotr@uiuc.edu).

Digital Object Identifier 10.1109/TED.2002.804706

[1]Note that even if the die size were to remain constant for future technology nodes, continuous device scaling will make interconnects the main performance bottleneck.



(a) Schematic representation. (b) Equivalent RC circuit.

Fig. 2. Interconnect of length l between two identical inverters.

their proposed optimization techniques and therefore, they are not very suitable for integration in a CAD tool flow.

In this work, we develop a methodology to estimate the repeater size and inter-repeater interconnect length which minimizes the total interconnect power dissipation for a given delay penalty. We use this methodology to find the power-optimal buffering schemes for various ITRS technology nodes for a given delay penalty. Furthermore, we use this methodology to show the relative importance of the various components of the power dissipation for various technology nodes. We show that for a given delay penalty, the relative power saving increases as the technology scales. This is shown to be due to the fact that leakage power dissipation becomes the dominating component of the total power dissipation, and therefore reducing the repeater size and the number of repeaters results in large power savings.

II. Preliminaries

Consider a uniform interconnect of resistance r per unit length and capacitance c per unit length buffered by identical repeaters, as shown in Fig. 2. Assume that for a minimum sized repeater, the input capacitance is c_0, the output parasitic capacitance is c_p, and output resistance is r_s. Therefore, for a repeater of size s, the total output resistance $R_{tr} = r_s/s$, the total output parasitic capacitance $C_p = c_p s$ and the total input capacitance is $C_L = c_0 s$. If the line segment is of length l and the repeater size is s, then the delay of that segment which is defined as the time difference between the input and output waveforms crossing 50% of their full-swing value is given by $\tau \log_e 2$, where the time constant τ is [3]

$$\tau = r_s(c_0 + c_p) + \frac{r_s}{s}cl + rlsc_0 + \frac{1}{2}rcl^2$$

and the delay per unit length is given by

$$\log_e 2\frac{\tau}{l} = \log_e 2\left(\frac{1}{l}r_s(c_0 + c_p) + \frac{r_s}{s}c + rsc_0 + \frac{1}{2}rcl\right).$$

This delay per unit length is optimal when [3]

$$l_{\text{opt}} = \sqrt{\frac{2r_s(c_0 + c_p)}{rc}} \quad s_{\text{opt}} = \sqrt{\frac{r_s c}{rc_0}}$$

and

$$\left(\frac{\tau}{l}\right)_{\text{opt}} = 2\sqrt{r_s c_0 rc}\left(1 + \sqrt{\frac{1}{2}\left(1 + \frac{c_p}{c_0}\right)}\right).$$

Note that minimizing the 50% delay per unit length is equivalent to minimizing τ/l.

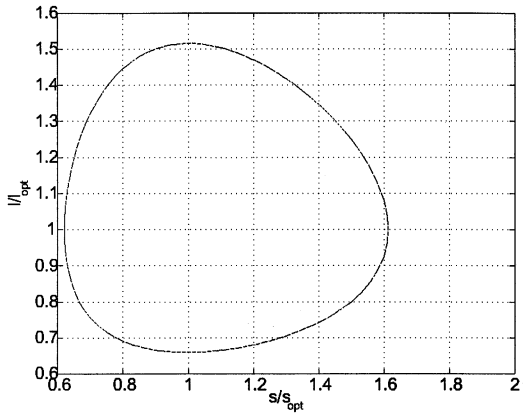


Fig. 3. Set of s/s_{opt} and l/l_{opt} values for which $\tau/l = 1.05\,(\tau/l)_{\text{opt}}$.

It should be pointed out that effect of line inductance on the delay of the interconnect segment has not been included in the above expression. In other words, we considered the interconnect segment as an RC element and not an RLC element. This has been done due to the fact that it has been shown in [13] and [14] that the effect of line inductance reduces with technology scaling for minimum sized global interconnects. It has also been shown in [13] and [14] that global line widths need to be increased by a large factor ($16\times$) before inductive effects become important. Therefore, RC delay is used throughout this paper.

It is widely believed that the total power dissipation due to optimum repeater insertion scheme can be excessive. As shown in Fig. 1, the minima of τ/l is very shallow both with respect to s and l. For this example, if the repeater size is $(1/2)s_{\text{opt}}$ and the interconnect length is $2l_{\text{opt}}$, the delay penalty is only 25%. Therefore, in practice the repeater size is smaller than s_{opt} and the interconnect length is larger than l_{opt} in the hope that power dissipation of such a configuration will be small with minimal impact on delay.

We would therefore like to quantify the reduction in power dissipation when repeater sizes smaller than s_{opt} and interconnect lengths larger than l_{opt} are used for a fixed delay penalty. It is obvious from Fig. 1 that for a given value of $(\tau/l) > (\tau/l)_{\text{opt}}$, there is a family of values of s and l which satisfy this equation which would be the closed curve formed by the intersection of the surface of solutions in Fig. 1 with a plane parallel to the s-l axis. As an illustration, Fig. 3 shows the set of solutions for which $\tau/l = 1.05\,(\tau/l)_{\text{opt}}$, i.e., a delay penalty of 5%. From this family of solutions, we would like to select the one which gives the minimum total power dissipation for the line.

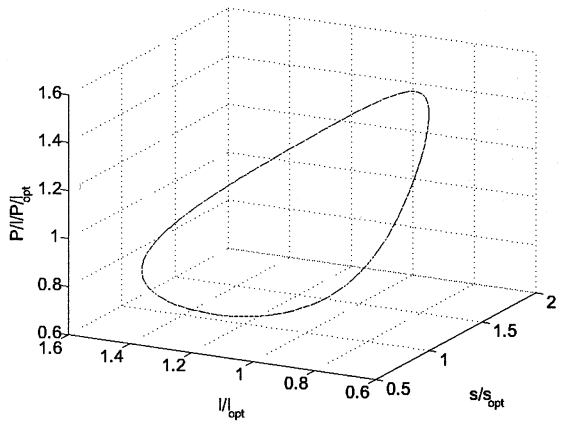


Fig. 4. Normalized power dissipation per unit length for a 5% delay penalty as a function of $s/s_{\rm opt}$ and $l/l_{\rm opt}$.

For a long interconnect of length L which is buffered several times the total power dissipation is

$$P_{\rm line} = nP_{\rm repeater} = \frac{L}{l}P_{\rm repeater}$$

where $n = L/l$ is the number of repeaters for that line. For a fixed L, we therefore seek to minimize $P_{\rm repeater}/l$ in order to minimize the total power dissipation.

Fig. 4 shows the power dissipation per unit interconnect length for the curve shown in Fig. 3. The power dissipation is calculated using (3) derived in the next section. It is obvious from this figure that a optimum value of repeater size s and inter-repeater interconnect length l exists for which the delay penalty criteria is met and power dissipation is minimum.

III. Methodology

The power dissipation of a repeater shown in Fig. 2(a) is given by [15]

$$P_{\rm repeater} = P_{\rm switching} + P_{\rm short-circuit} + P_{\rm leakage}.$$

The various components of the total power are expressed as follows.

A. Switching Power

The switching power of a repeater is given by

$$P_{\rm switching} = \alpha(s(c_p + c_0) + lc)V_{DD}^2 f_{\rm clk}$$

where

- V_{DD} power supply voltage;
- $f_{\rm clk}$ clock frequency;
- α switching factor (or activity factor), which is the fraction of repeaters on a chip that are switched during an average clock cycle.

α can be taken as 0.15 [15]. Note that as the repeater size is reduced and the inter-buffer interconnect length is increased, for a given line length the intrinsic repeater power dissipation reduces whereas the switching power due to total line capacitance remains unchanged.

B. Leakage Power

The average leakage power of a repeater in a long buffered interconnect is given by

$$\begin{aligned} P_{\rm leakage} &= V_{DD} I_{\rm leakage} \\ &= V_{DD}\frac{1}{2}(I_{{\rm off}_n} W_n + I_{{\rm off}_p} W_p) \\ &= V_{DD}\frac{1}{2}\left(I_{{\rm off}_n} W_{n_{\rm min}} + I_{{\rm off}_p} W_{p_{\rm min}}\right) s \end{aligned}$$

where

- $I_{\rm leakage}$ leakage current flowing through the repeater;
- $I_{{\rm off}_n}$ ($I_{{\rm off}_p}$) leakage current per unit NMOS (PMOS) transistor width;
- W_n (W_p) width of the NMOS (PMOS) transistor;
- $W_{n_{\rm min}}$ ($W_{p_{\rm min}}$) width of the NMOS (PMOS) transistor in minimum sized inverter.

The factor 1/2 is included because, in a long buffered interconnect, on an average, half the inverter will have input of one, i.e., the NMOS transistor will be ON and the leakage current will be determined by the PMOS transistors, while the other half of the inverters will have input of zero, i.e., the PMOS transistor will be ON and the leakage current will be determined by the NMOS transistor. Usually $I_{{\rm off}_n} \approx I_{{\rm off}_p}$ and the width of the PMOS transistor is two to three times larger than the NMOS device in an inverter. In this study, we will assume that $I_{{\rm off}_n} = I_{{\rm off}_p}$ and $W_{p_{\rm min}} = 2W_{n_{\rm min}}$ throughout. This implies that

$$P_{\rm leakage} = \frac{3}{2}V_{DD} I_{{\rm off}_n} W_{n_{\rm min}} s.$$

For long-channel devices, this used to be negligible but for nanometer technologies, this can be significant.

The *subthreshold swing* S, which is defined as the change in V_{GS} for the drain current to change by ten times, is given by [16]

$$S = \left(\frac{d\log(I_D)}{dV_{GS}}\right)^{-1} = \frac{kT}{q}\log_e(10)(1+\gamma) \qquad (1)$$

where

- k Boltzmann's constant;
- T temperature;
- q electron charge.

$\gamma > 0$ can be treated as a process-dependent fitting parameter. The subthreshold current at a given technology node can be computed as

$$I_{\rm off} = I_{{\rm off}_{180}} 10^{(V_{t_{180}} - V_t)/S} \qquad (2)$$

where $I_{{\rm off}_{180}}$ and $V_{t_{180}}$ are the leakage current and threshold voltage, respectively, at the 180-nm technology node; and V_t is the threshold voltage at the given technology node. This indicates that, for a given temperature, as the threshold voltage decreases at $V_{DS} = 0$ V, the subthreshold current increases exponentially. Assuming a die temperature of 100 °C, the subthreshold swing is taken to be 100 mV/decade [11]. The subthreshold leakage current per unit width ($I_{\rm off}$) of NMOS and PMOS transistors for all technologies is given in Table I. Note

TABLE I
TECHNOLOGY AND EQUIVALENT CIRCUIT MODEL PARAMETERS FOR TOP LAYER METAL FOR DIFFERENT TECHNOLOGY NODES BASED ON THE ITRS. c WAS OBTAINED USING FASTCAP [17]

Tech. node (nm)	180	130	100	70	50
width (nm)	525	382.5	280	195	137.5
height (nm)	1155	1033	756	546	399
t_{ins} (nm)	7699	6664	6022	5571	4116
ϵ_r	3.75	3.1	1.9	1.5	1.25
r (kΩ/m)	36.3	60.1	103.9	206.6	401.3
c (pF/m)	269	240	154	125	106
l_{opt} (mm)	3.33	2.5	2.22	1.32	1.06
s_{opt}	174	151	110	82	53
$\left(\frac{\tau}{l}\right)_{opt}$ (ps/mm)	49.5	58.8	56.3	67.4	67.0
r_s (kΩ)	8	9.5	10	15.8	12.5
c_0 (fF)	1.9	1.7	1.5	1.3	1.2
c_p (fF)	4.8	3.5	2.5	1.5	0.75
V_{DD} (V)	1.8	1.5	1.2	0.9	0.6
V_t (V)	0.45	0.375	0.3	0.225	0.15
I_{off_n} (μA/μ)	0.2	1.13	6.33	35.6	200
f_{clk} (GHz)	1.2	1.6	2.0	2.5	3.0

that as the repeater size is reduced and the inter-buffer interconnect length is increased, the leakage power per repeater decreases, as well as the total number of repeaters inserted along the line decreases. Therefore, this results in large savings in leakage power dissipation.

C. Short-Circuit Power

This power dissipation is incurred when the NMOS and PMOS transistors in an inverter are simultaneously ON. Consider the inverter shown in Fig. 5(a). The input and output voltage waveforms are shown in Fig. 5(b). Let t_r denote the time for the input voltage to rise from V_{t_n} to $V_{DD} - V_{t_p}$. Note that, in general, the short-circuit current not only depends on the shape of the input waveform, but also depends on the output waveform, which, in turn, depends on the parasitic output and interconnect capacitance and output resistance. Approximating the short-circuit current waveform by a triangular wave [16], the energy dissipated due to the short-circuit current pulse during a low-to-high transition is

$$E_{l\to h} = \frac{1}{2} t_r I_{\text{peak}} V_{DD}.$$

Assuming symmetric high-to-low and low-to-high transitions both at the input and output of the inverters, the total short-circuit power is given by

$$\begin{aligned} P_{\text{short-circuit}} &= \alpha t_r V_{DD} I_{\text{peak}} f_{\text{clk}} \\ &= \alpha t_r V_{DD} W_{n_{\min}} s I_{\text{short-circuit}} f_{\text{clk}} \end{aligned}$$

where α is the same switching factor as in the switching power expression. It has been empirically observed from SPICE simulations that $I_{\text{peak}} = W_{n_{\min}} s I_{\text{short-circuit}}$ where $I_{\text{short-circuit}}$

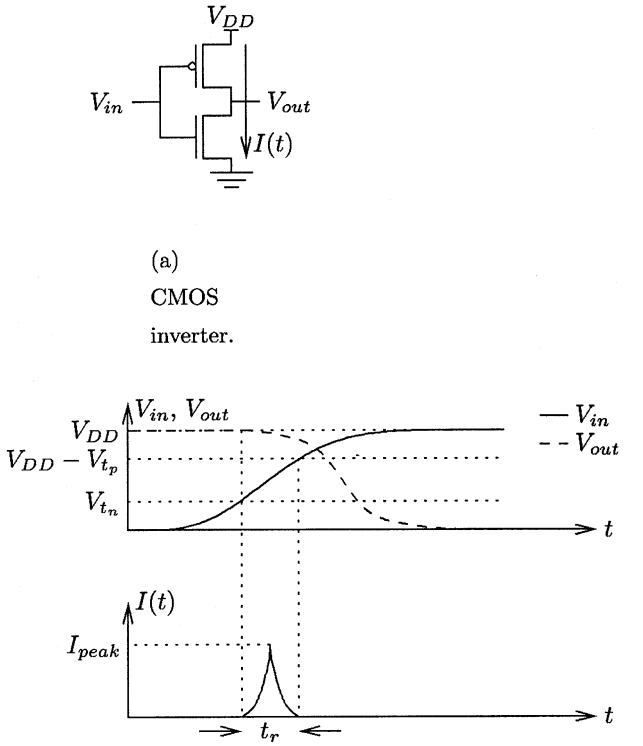


Fig. 5. Voltage and current waveforms of a CMOS inverter.

is approximately 65 μA/μm across all technologies. Assuming that the input waveform is a single time-constant exponential and $V_{t_n} = V_{t_p} = (1/4)V_{DD}$

$$\begin{aligned} t_r &= \tau \log_e \left(\frac{V_{DD} - V_{t_p}}{V_{t_n}} \right) = \tau \log_e 3 \\ &= \left[r_s(c_0 + c_p) + \frac{r_s}{s} cl + rlsc_0 + \frac{1}{2} rcl^2 \right] \log_e 3. \end{aligned}$$

Note that as the repeater size is reduced and the inter-buffer interconnect length is increased, the rise time t_r increases and therefore, the short-circuit power dissipation for one repeater may increase.

Therefore, the total power can be written as

$$P_{\text{repeater}} = k_1(s(c_p + c_0) + lc) + k_2 s + k_3 s\tau \qquad (3)$$

where

$$\begin{aligned} k_1 &= \alpha V_{DD}^2 f_{\text{clk}} \\ k_2 &= \frac{3}{2} V_{DD} I_{\text{off}_n} W_{n_{\min}} \\ k_3 &= \alpha V_{DD} W_{n_{\min}} I_{\text{short-circuit}} f_{\text{clk}} \log_e 3. \end{aligned}$$

If the fractional delay penalty to be tolerated is f, then

$$\begin{aligned} \frac{\tau}{l} &= (1+f) \left(\frac{\tau}{l} \right)_{\text{opt}} \\ &= \frac{1}{l} r_s(c_0 + c_p) + \frac{r_s}{s} c + rsc_0 + \frac{1}{2} rcl \end{aligned} \qquad (4)$$

TABLE II
POWER PER UNIT LENGTH OPTIMIZATION RESULTS FOR 5% DELAY PENALTY FOR VARIOUS ITRS TECHNOLOGY NODES

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left[\frac{P}{l}\right]_{opt}$
180	0.6676	1.2518	0.9082	0.7255
130	0.6766	1.2660	0.9144	0.7223
100	0.6929	1.2988	0.8839	0.6805
70	0.7150	1.3415	0.8048	0.5999
50	0.7394	1.3893	0.7589	0.5462

or

$$\tau = (1+f)\left(\frac{\tau}{l}\right)_{\text{opt}} l.$$

Therefore

$$\begin{aligned} P_{\text{repeater}} =& k_1(s(c_p+c_0)+lc)+k_2 s + k_3(1+f)\left(\frac{\tau}{l}\right)_{\text{opt}} sl \\ =& k_1(s(c_p+c_0)+lc)+k_2 s + k_3' sl \end{aligned} \tag{5}$$

where

$$k_3' = k_3(1+f)\left(\frac{\tau}{l}\right)_{\text{opt}}$$

and

$$\frac{P_{\text{repeater}}}{l} = k_1\left(\frac{s}{l}(c_p+c_0)+c\right)+k_2\frac{s}{l}+k_3' s. \tag{6}$$

Setting the derivative of this with respect to s to zero we have

$$\begin{aligned} \frac{d\frac{P_{\text{repeater}}}{l}}{ds} =& \frac{k_1(c_p+c_0)}{l}+\frac{k_2}{l}+k_3' \\ & -\left[\frac{k_1 s(c_p+c_0)}{l^2}+\frac{k_2 s}{l^2}\right]\frac{dl}{ds}=0. \end{aligned}$$

dl/ds can be calculated by differentiating (4). Therefore, we have the following three nonlinear equations to solve:

$$\begin{aligned} \frac{k_1(c_p+c_0)}{l}+\frac{k_2}{l}+k_3' - \left[\frac{k_1 s(c_p+c_0)}{l^2}+\frac{k_2 s}{l^2}\right]\frac{dl}{ds} =& 0 \\ \frac{1}{l}r_s(c_0+c_p)+\frac{r_s}{s}c+rsc_0+\frac{1}{2}rcl-(1+f)\left(\frac{\tau}{l}\right)_{\text{opt}} =& 0 \\ \left[\frac{1}{2}rc-\frac{r_s(c_0+c_p)}{l^2}\right]\frac{dl}{ds}+rc_0-\frac{r_s c}{s^2} =& 0 \end{aligned} \tag{7}$$

with three unknown l, s, and dl/ds, out of which we only are interested in l and s. This can be solved numerically using Newton-Raphson. As indicated in Tables I and II, the inverter sizes in the buffered interconnects are very large. A typical minimum-sized VLSI gate will not be able to directly drive this inverter while still meeting the delay constraint. Therefore, intermediate inverters need to be introduced between the minimum sized gate and the interconnect buffer [16]. The ratio of the sizes of successive inverters is typically four in order to minimize the propagation delay [16]. In our analysis, we ignore the power dissipation of these intermediate inverters because this will be a negligible fraction of the total power dissipation for long interconnects.

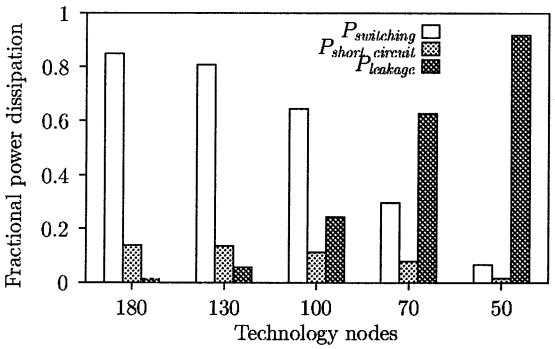


Fig. 6. Relative contributions of the three components of overall power dissipation for 5% delay penalty for various technology nodes.

IV. RESULTS

The methodology outlined in the last section was used to optimize power for global tier interconnects for ITRS technology nodes for a 5% delay penalty as an illustrative example. The ITRS technology parameters are shown in Table I. r_s, c_0, c_p, and $I_{\text{short-circuit}}$ were obtained by SPICE simulations. I_{off} at 100 °C was taken to be 0.2 μA/μm for the 180-nm technology node [11] and, as indicated in Section III, was estimated for other technology nodes using a subthreshold swing of 100 mV/decade at that temperature [11].

The power optimization results are shown in Table II. s/s_{opt} is the new repeater size as a ratio of the delay optimal repeater size, l/l_{opt} is the new interconnect length between successive repeaters as a ratio of the delay optimal interconnect length, P/P_{opt} is the power dissipation of a *single* repeater as a ratio of the power dissipation of the delay optimal repeater, and $P/l/\left(P/l\right)_{\text{opt}}$ is the power dissipation per unit length as a ratio of the power dissipation per unit length of the delay optimal case. From the table, it is obvious that for optimal power dissipation at a given delay penalty, the repeater size needs to be reduced and the interconnect length between successive repeaters needs to be increased. The total power savings increase as the technology scales. This is due to that fact that leakage current I_{off} increases substantially with scaling and therefore reducing the repeater size results in large savings in total power dissipation.

This fact is further illustrated in Fig. 6 which plots the relative contributions of $P_{\text{switching}}$, $P_{\text{short-circuit}}$, and P_{leakage} as the technologies scale. It can be observed that leakage power starts dominating as the technology scales. Also note that the short-circuit power is also nontrivial across all technology nodes. Therefore, short-circuit power needs to be considered in any power optimization.

With this basic framework, various power optimization alternatives can be compared. For instance, a naïve approach would be to minimize the power dissipation of individual repeaters instead of minimizing the repeater power per unit length. For this case, (5) needs to be used instead of (6) in the set of the nonlinear equation (7). The results of this optimization are shown in Table III. Comparing these results with Table II, we observe that if power dissipation of one inverter is minimized, the *power-optimal* inter-repeater interconnect length l is *smaller* than the delay optimal length l_{opt}. Therefore, even though the power dissipation of one repeater is smaller than that in Table II (column

TABLE III
POWER MINIMIZATION OF INDIVIDUAL REPEATERS: RESULTS FOR 5% DELAY PENALTY FOR VARIOUS ITRS TECHNOLOGY NODES

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left[\frac{P}{l}\right]_{opt}$
180	0.6820	0.7776	0.6984	0.8982
130	0.6857	0.7764	0.7000	0.9016
100	0.6682	0.8112	0.6967	0.8589
70	0.6474	0.8992	0.6740	0.7496
50	0.6241	0.9140	0.6007	0.6721

TABLE IV
RESULTS FOR MINIMIZATION ONLY OF THE SWITCHING POWER PER UNIT LENGTH FOR 5% DELAY PENALTY FOR VARIOUS ITRS TECHNOLOGY NODES

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left[\frac{P}{l}\right]_{opt}$
180	0.6967	1.3160	0.9580	0.7280
130	0.7053	1.3287	0.9628	0.7246
100	0.7134	1.3415	0.9145	0.6817
70	0.7257	1.3625	0.8179	0.6003
50	0.7412	1.3926	0.7607	0.5462

4), since the number of repeaters for a given line length is larger for this case, the total power dissipation (or equivalently power dissipation per unit length) (column 5) is higher than that in Table II.

Similarly, the effect of ignoring short-circuit power and leakage power on the optimization can be quantified. For this purpose, it is instructive to review the form of (6) which is repeated here for convenience

$$\frac{P_{\text{repeater}}}{l} = \underbrace{k_1\left(\frac{s}{l}(c_p+c_0)+c\right)}_{\text{switching}} + \underbrace{k_2\frac{s}{l}}_{\text{leakage}} + \underbrace{k_3' s}_{\text{short-circuit}} .$$

Note that both the switching and leakage power terms are of the form

$$f_1 + f_2\frac{s}{l}$$

where f_1 and f_2 are constants. Therefore, if short-circuit power term is negligible compared to the other two terms or is ignored, optimizing driver size and inter-buffer interconnect length for power per unit length is equivalent to optimizing for switching or leakage power per unit length alone.

Table IV shows the optimization considering only the switching component of the power dissipation. However, the power dissipation is calculated considering all three components: switching, leakage and short-circuit, using s and l values from the (incorrect) power optimization. Similarly, Table V shows the optimization considering only the switching and leakage component of the power dissipation. Notice that as explained above, all the entries in these two tables are identical. This also highlights the importance of considering the short-circuit power in the optimization process. Table VI shows the optimization considering only the switching and short-circuit component of the power dissipation. Comparing these results with Table II, it can be observed that ignoring leakage power results in large errors in power optimization at future

TABLE V
RESULTS FOR MINIMIZATION OF ONLY SWITCHING AND LEAKAGE POWER PER UNIT LENGTH FOR 5% DELAY PENALTY FOR VARIOUS ITRS TECHNOLOGY NODES

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left[\frac{P}{l}\right]_{opt}$
180	0.6967	1.3160	0.9580	0.7280
130	0.7053	1.3287	0.9628	0.7246
100	0.7134	1.3415	0.9145	0.6817
70	0.7257	1.3625	0.8179	0.6003
50	0.7412	1.3926	0.7607	0.5462

TABLE VI
RESULTS FOR MINIMIZATION OF ONLY SWITCHING AND SHORT-CIRCUIT POWER PER UNIT LENGTH FOR 5% DELAY PENALTY FOR VARIOUS ITRS TECHNOLOGY NODES

tech. node (nm)	s/s_{opt}	l/l_{opt}	P/P_{opt}	$\frac{P}{l}/\left[\frac{P}{l}\right]_{opt}$
180	0.6668	1.2499	0.9068	0.7255
130	0.6733	1.2573	0.9082	0.7223
100	0.6795	1.2656	0.8621	0.6812
70	0.6797	1.2525	0.7585	0.6056
50	0.6952	1.2836	0.7121	0.5548

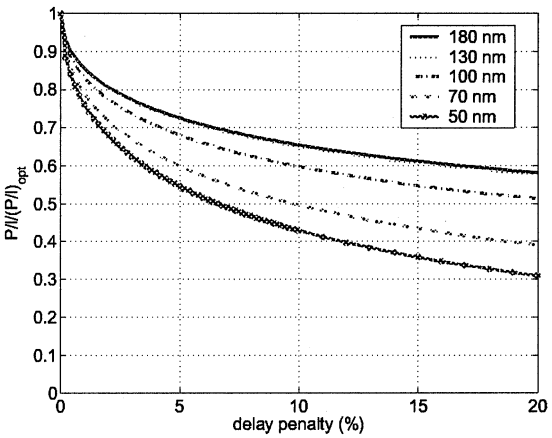


Fig. 7. Power per unit length as a function of delay penalty for various technology nodes.

technology nodes. Similarly, ignoring short-circuit power also results in errors when short-circuit power is nonnegligible, specially for 180-nm to 100-nm technology nodes. For 70-nm and 50-nm technology nodes, however, the optimum power per unit length with and without considering short-circuit power is almost the same for 5% delay penalty. From Fig. 6, it can be observed that short-circuit power is negligible for these technology nodes at 5% delay penalty. However, if the allowed delay penalty is increased, the rise time will increase which increases the short-circuit power.

Fig. 7 shows the power per unit length as a function of delay penalties for various technology nodes. As expected, $P/l/(P/l)_{\text{opt}}$ reduces as the delay penalty increases. Note that the incremental reduction in $P/l/(P/l)_{\text{opt}}$ is high for small values of delay penalty and starts decreasing as the delay penalty increases. Also note that the curves for 180-nm and 130-nm technology nodes are very similar. However, for a

given delay penalty, $P/l/(P/l)_{\text{opt}}$ reduces as the technology is scaled beyond 130 nm. This is entirely due the leakage power. From Fig. 6, it can be observed that for both 180-nm and 130-nm technology nodes, leakage power is a negligible portion of the overall power dissipation whereas for other technology nodes, it becomes progressively significant and is the dominant fraction of total power dissipation for the 70-nm and 50-nm technology node.

V. Conclusions

In conclusion, we have developed a methodology for choosing the repeater size and inter-repeater interconnect length for a given global line which satisfies a given delay penalty criteria and minimizes the total power dissipation. Using this methodology, we have computed the power-optimal buffering schemes for various technology nodes for a 5% delay penalty. Furthermore, we have shown that short-circuit and leakage power are important components of the total power dissipation and ignoring them in power optimization can lead to errors. Short-circuit power becomes important as the allowed delay penalty increases since rise time of the signal increases. Similarly, leakage power increases exponentially with device scaling and is the dominant component of power dissipation for 50-nm technology node. We have also shown that for 180-nm and 130-nm technology nodes where leakage power is not significant, the relative power saving is almost the same for a given delay penalty. However, beyond 130-nm node, leakage power becomes significant and therefore the relative power savings increase with technology scaling for a given delay penalty.

Acknowledgment

The authors would like to thank an anonymous reviewer for meticulously reviewing the manuscript.

References

[1] *International Technology Roadmap for Semiconductors (ITRS)*, Semiconductor Industry Association, San Jose, CA, 1999.

[2] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, pp. 602–633, May 2001.

[3] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addision-Wesley, 1990.

[4] J. Cong and L. He, "An efficient technique for device and interconnect optimization in deep submicron designs," in *Proc. Int. Symp. Physical Design*, 1998, pp. 45–51.

[5] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunties and new frontiers," *Proc. Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 22–25, 2001.

[6] H. Kawaguchi and T. Sakurai, "A reduced clock swing flip-flop (RCFF) for 63% power reduction," *IEEE J. Solid-State Circuits*, vol. 33, pp. 807–811, 1998.

[7] T. Sakurai, "Design challenges for 0.1 μm and beyond," in *Proc. ASP DAC*, 2000, pp. 553–558.

[8] K. Banerjee, A. Mehrotra, A. Sangiovanni-Vincentelli, and C. Hu, "On thermal effects in deep submicron VLSI interconnects," in *Proc. Design Automation Conf.*, 1999, pp. 885–891.

[9] V. Adler and E. G. Friedman, "Repeater design to reduce delay and power in resistive interconnect," *IEEE Trans. Circuits Syst. I*, vol. 45, pp. 607–616, May 1998.

[10] A. Nalamalpu and W. Burleson, "A practical approach to DSM repeater insertion: Satisfying delay constraints while minimizing area and power," *Proc. 14th Annu. IEEE Int. ASIC/SOC Conf.*, pp. 152–156, 2001.

[11] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. Int. Symp. Low Power Electronics and Design*, 1999, pp. 163–168.

[12] K. Nose and T. Sakurai, "Analysis of future trend of short-circuit power," *IEEE Trans. Computer-Aided Design*, vol. 19, no. Sept., pp. 1023–1030, 2000.

[13] K. Banerjee and A. Mehrotra, "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," *Proc. IEEE Symp. VLSI Circuits*, pp. 195–198, 2001.

[14] ——, "Analysis of on-chip inductance effects for distributed *RLC* interconnects," *IEEE Trans. Computer-Aided Design*, vol. 21, pp. 904–915, Aug. 2002.

[15] A. P. Chandrakasan and R. W. Brodersen, "Sources of power consumption," in *Low Power Digital CMOS Design*. Norwell, MA: Kluwer, 1995.

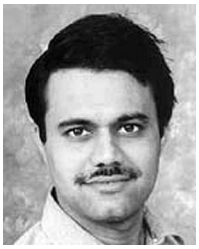
[16] J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[17] K. Nabors and J. K. White, "FASTCAP: A multipole-accelerated 3-D capacitance extraction program," *IEEE Trans. Computer-Aided Design*, vol. 10, pp. 1447–1459, Nov. 1991.

Kaustav Banerjee (S'94–M'99) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley in 1999.

He was with Stanford University, Stanford, CA, from 1999 to 2002 as a Research Associate at the Center for Integrated Systems. In July 2002, he joined the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, as an Assistant Professor. His research interests include nanometer scale circuit effects and their implications for high-performance/low-power VLSI and mixed-signal ICs and their design automation methods. He is also interested in some exploratory interconnect and circuit architectures such as 3-D ICs, integrated optoelectronics, and in nanotechnologies such as single electron transistors. He co-advises several doctoral students at Stanford University, University of Southern California, Los Angeles, and the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. From February 2002 to August 2002 he was a Visiting Professor at the Circuit Research Labs of Intel in Hillsboro, OR. In the past, he has also held summer/visiting positions at Texas Instruments Inc., Dallas, Texas, and EPFL-Switzerland, and has consulted for several EDA companies in the Silicon Valley. He has authored or co-authored over 70 technical papers in archival journals and refereed international conferences and has presented numerous invited talks and tutorials.

Dr. Banerjee served as Technical Program Chair of the 2002 IEEE International Symposium on Quality Electronic Design (ISQED '02), and is the Conference Vice-Chair of ISQED '03. He also serves on the technical program committees of the ACM International Symposium on Physical Design, the EOS/ESD Symposium, and the IEEE International Reliability Physics Symposium. He is the recipient of a Best Paper Award at the 2001 Design Automation Conference.



Amit Mehrotra (S'96–M'99) received the B. Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1994, and the M.S. and Ph.D. degrees from the Department of Electrical Engineering and Computer Science, the University of California at Berkeley in 1996 and 1999, respectively.

In August 1999, he joined the University of Illinois at Urbana-Champaign where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering and a Research Assistant Professor with the Illinois Center for the Integrated Micro-Systems group at the Coordinated Science Laboratory. His research interests include RF, analog and mixed signal circuit design for mobile communication systems, simulation techniques for RF and mixed signal circuits and systems, interconnect performance and modeling issues in VLSI and novel circuits and physical design issues for high-performance VLSI designs, model-order reduction of linear and nonlinear circuits. He has authored and coauthored over 30 technical papers in archival journals and refereed international conferences.

Dr. Mehrotra has served as the Technical Program Committee member of International Symposium on Quality Electronic Design in 2002 and 2003. He received best paper awards at the 1997 International Conference on Computer Design and 2001 Design Automation Conference.