

A Global Interconnect Optimization Scheme for Nanometer Scale VLSI With Implications for Latency, Bandwidth, and Power Dissipation

Man Lung Mui, Kaustav Banerjee, *Senior Member, IEEE*, and Amit Mehrotra, *Member, IEEE*

Abstract—This paper addresses the critical problem of global wire optimization for nanometer scale very large scale integration technologies, and elucidates the impact of such optimization on power dissipation, bandwidth, and performance. Specifically, this paper introduces a novel methodology for optimizing global interconnect width, which maximizes a novel *figure of merit* (FOM) that is a user-defined function of bandwidth per unit width of chip edge and latency. This methodology is used to develop analytical expressions for optimum interconnect widths for typical FOMs for two extreme scenarios regarding line spacing: 1) spacing kept constant at its minimum value and 2) spacing kept the same as line width. These expressions have been used to compute the optimal global interconnect width and quantify the effect of increasing the line width on various performance metrics such as delay per unit length, total repeater area and power dissipation, and bandwidth for various International Technology Roadmap for Semiconductors technology nodes.

Index Terms—Bandwidth, critical inductance, delay per unit length, global interconnect optimization, interconnect power dissipation optimization, International Technology Roadmap for Semiconductors (ITRS), optimal buffering, technology scaling.

I. INTRODUCTION

WITH aggressive scaling of CMOS technology, gate delay, and local wire delay decreases rapidly [1]. However, the delay of global interconnects increases with technology scaling [1]–[4] because the global interconnect lengths tend to increase with scaling. Repeater insertion is generally used to reduce the delay of long global interconnects [5]. However, with aggressive scaling of global interconnect dimensions to meet the increased connectivity demands in a high performance system-on-a-chip (SoC), the interconnect delay per unit length of optimally buffered minimum sized global wires is also increasing with technology scaling [6]. Therefore, global interconnects tend to limit the performance of high-performance SoCs.

In order to achieve improvement in performance, designers tend to use wires which are wider than minimum-sized global interconnects prescribed by the technology. Increasing the width of the interconnect proportionally reduces its resistance per unit length and also increases the line capacitance per unit length. However, for global interconnects in nanometer technologies, where the aspect ratio of wires is approximately 2–2.5, the increase in width results in a reduction in the resistance–capacitance (RC) time constant of the line and therefore improves delay per unit length [7]. However, these “fat” wires take up a lot of routing resources and using fat wires can adversely affect the wireability of the chip. For further improvement in performance, the spacing of global interconnects can also be increased which, to some extent, offsets the increase in line capacitance due to increasing line width. However, this increase in spacing will further degrade the wireability of the chip. Furthermore, the delay per unit length for wide wires may degrade due to inductance effects as well. Therefore, in determining the wire widths at the global tier, the number of interconnects per unit chip edge should also be taken into account along with the delay per unit length. For instance, the ratio of the number of interconnects per unit chip edge and the delay per unit length, which represents the rate of data transfer per unit chip edge, can be a useful metric to optimize.

This paper introduces a new methodology for determining the optimum width of global interconnects for a given technology, which maximizes a user-defined figure of merit (FOM), which is a known function of delay per unit length and the rate of data transfer per unit chip edge. As a first step, we develop semi-analytical expressions for line capacitance per unit length as a function of line width and spacing. Using these models, in Section II we obtain the functional dependence of delay per unit length of an optimally buffered interconnect on line width. This, in turn, results in the functional dependence of the given FOM on the line width which is analytically optimized to yield the optimum interconnect width. We carry out this optimization for various FOMs and various International Technology Roadmap for Semiconductors (ITRS) technology nodes for two extreme scenarios: 1) Interconnect spacing is kept at its minimum, and 2) interconnect spacing is kept equal to the interconnect width. The optimization results indicate that the rate of data transfer per unit chip edge is very close to optimum when the line width is minimum as prescribed by the ITRS. However, in order to optimize a different FOM, line width needs to be increased. We also quantify the improvement in delay per unit length, total repeater

Manuscript received February 24, 2003; revised September 23, 2003. This work was supported in part by the University of Illinois and by the University of California–MICRO program. The review of this paper was arranged by Editor R. Singh.

M. L. Mui and A. Mehrotra are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: manmui@uiuc.edu; amehrotr@uiuc.edu).

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

Digital Object Identifier 10.1109/TED.2003.820651

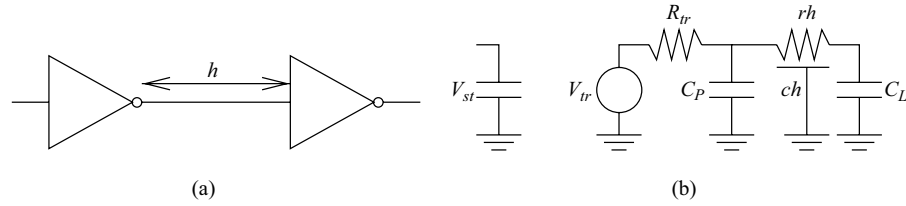


Fig. 1. Interconnect of length h between two identical inverters. (a) Schematic representation. (b) Equivalent RC circuit.

area and power dissipation and the degradation in the per unit width bit transfer rate for this optimum width compared to minimum width lines. We show that these improvements are fairly insensitive to technology scaling.

II. METHODOLOGY

Consider a uniform interconnect of resistance r per unit length and capacitance c per unit length buffered by identical repeaters as shown in Fig. 1. 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 k , the total output resistance $R_{tr} = (r_s/k)$, the total output parasitic capacitance $C_p = c_p k$ and the total input capacitance is $C_L = c_0 k$. If the line segment is of length h and the repeater size is k , then the time-constant of that segment is [8]

$$\tau = r_s(c_0 + c_p) + \frac{r_s}{k}ch + rhkc_0 + \frac{1}{2}rch^2$$

and the latency or the delay of that section is $\tau \log 2$. Now consider a long interconnect of a given length L which is uniformly buffered with inter-buffer interconnect length h . Therefore the total number of segments is L/h . The total delay through that line is given by

$$\text{delay} = \frac{L}{h} \times \tau \log 2 \propto \frac{\tau}{h}$$

where τ/h is the delay per unit length which is given by

$$\frac{\tau}{h} = \frac{1}{h}r_s(c_0 + c_p) + \frac{r_s}{k}c + rk c_0 + \frac{1}{2}rch.$$

Note that optimizing the delay of the interconnect of a fixed length is equivalent to optimizing τ/h . This delay per unit length is optimal when

$$h_{\text{opt}} = \sqrt{\frac{2r_s(c_0 + c_p)}{rc}} \quad k_{\text{opt}} = \sqrt{\frac{r_sc}{rc_0}}$$

and is given by

$$\left(\frac{\tau}{h}\right)_{\text{opt}} = 2\sqrt{r_sc_0rc} \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0}\right)}\right).$$

Note that this optimal delay per unit length is a function of interconnect parameters r and c which in turn are a function of interconnect width W and spacing S . In the present study we are not explicitly considering cross-talk. The effect of cross-talk would be to change the value of c depending on whether the neighboring interconnects are quiet or are making a transition. For

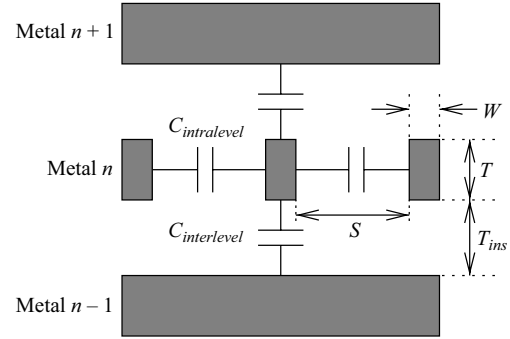


Fig. 2. VLSI interconnect cross section (not to scale).

global interconnects, this assumption is somewhat justified because long global interconnects would be properly shielded to yield predictable delays and therefore c can be assumed to be a function of interconnect geometry only.

Earlier studies for quantifying the optimal buffering schemes for optimal delay per unit length [5] always considered minimum sized global wires. However, for further improvement in performance, the designers have a option of increasing the wire width and/or spacing. This increase in W for a given S will result in a decrease in line resistance per unit length r and an increase in line capacitance per unit length c . However, the decrease in r is much more than the increase in c and therefore the optimal delay per unit length will decrease. However, increased pitch ($P = W + S$) will imply a decrease in wireability of the chip. Some previous work such as [9] can be found in the literature on wire sizing for delay and power optimization. However, these authors considered a discrete set of wire widths and also neglected both the leakage and the short-circuit power in their power estimations. Wire width optimization has also been considered by [10], and [11]. However, as pointed out in the next paragraph, the model for interconnect technology in [10] is not realistic and the metric for optimization in both these approaches is also not flexible and is not applicable for a wide variety of design characteristics and hence their optimization results may not be meaningful. Furthermore [11] do not provide any model for the interconnect delay and power dissipation and therefore their formulation is not very transparent.

We will consider two scenarios. In the first case, line width can be changed but the line spacing S is kept constant at S_{min} . In this case, increasing the wire width will not strongly degrade the wireability of the chip. In the second case, the line spacing will be kept the same as line width for all W s. The second case is a less popular option for designers but will act as a limiting case. We assume that the line thickness T and the interlayer dielectric thickness T_{ins} (Fig. 2) cannot be changed. This is in contrast

with [10] where it was assumed that $W = T = T_{ins} = S$ and can be arbitrarily varied, which is not realistic, since, for a given process technology and a given layer, T and T_{ins} typically cannot be changed by the designers while they are free to choose any $W \geq W_{\min}$ and $S \geq S_{\min}$.

It has been shown that with ITRS scaling scheme, the minimum sized global interconnects are becoming increasingly resistive and the inductive effects are decreasing rapidly [12]. It was also shown that inductive effects on delay may become significant only if line widths are greater than $10 W_{\min}$. Therefore, initially we will assume that inductance effects can be ignored for the purpose of delay and power dissipation calculation and verify for the computed optimum line widths whether this is indeed the case or not.

Let ψ denote the bandwidth, i.e., the rate at which bits can be transmitted across a unit length of interconnect in a given chip edge or width. The rate at which bits can be transmitted per unit length by one interconnect is inversely proportional to the delay per unit length, i.e.,

$$\text{rate bit of transmission} \propto \frac{1}{\tau/h}$$

We assume that the lines are always optimally buffered for a given line width. Therefore

$$\text{rate bit of transmission} \propto \frac{1}{\left(\frac{\tau}{h}\right)_{\text{opt}}}$$

The number of such lines present in a given chip edge is

$$\frac{\text{chip edge}}{\text{metal pitch}}$$

and metal pitch = $W + S$. Therefore

$$\psi \propto \frac{1}{(W + S) \frac{\tau}{h}}$$

The aim of a global interconnect design scheme is to have a large ψ while having a small delay per unit length. As an example, an appropriate FOM to maximize can be

$$\frac{\psi}{\left[\left(\frac{\tau}{h}\right)_{\text{opt}}\right]^i}$$

for some $i \in \mathbb{I}^+$. Larger values of i would imply more importance to delay per unit length at the expense of the rate of bit transfer per unit width. In our study we carry out the analysis for $i = 0, 1$, and 2 . In other words

$$\text{FOM} = \frac{1}{(W + S) \left[\left(\frac{\tau}{h}\right)_{\text{opt}}\right]^{i+1}}$$

Line resistance per unit length r is inversely proportional to line width W

$$r = \frac{\rho}{WT}$$

Line capacitance per unit length c is also a function of W , i.e., $c \equiv c(W)$. Using the above, the expression for the optimal delay per unit length in terms of W can be written as

$$\begin{aligned} \left(\frac{\tau}{h}\right)_{\text{opt}} &= 2\sqrt{r_s c_0 r c} \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0}\right)}\right) \\ &= 2\sqrt{r_s c_0 \frac{\rho}{T}} \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0}\right)}\right) \sqrt{\frac{c(W)}{W}} \\ &= k \sqrt{\frac{c(W)}{W}} \end{aligned}$$

where k is a constant for the given metal layer. Therefore

$$\text{FOM} = \frac{W^{\frac{i+1}{2}}}{(W + S) k^{i+1} (c(W))^{\frac{i+1}{2}}}$$

Setting the derivative of this with respect to W to zero, it follows that W_{opt} satisfies the following equation

$$\begin{aligned} 0 &= \frac{i+1}{2} (W_{\text{opt}} + S) \left(c(W_{\text{opt}}) - W_{\text{opt}} \frac{dc(W)}{dW} \Big|_{W_{\text{opt}}} \right) \\ &\quad - c(W_{\text{opt}}) W_{\text{opt}} \left(1 + \frac{dS}{dW} \Big|_{W_{\text{opt}}} \right). \end{aligned} \quad (1)$$

Note that the optimum width is only dependent on line capacitance and line spacing. The optimum delay per unit length for this optimum line width is given by

$$\left(\frac{\tau}{h}\right)_{\text{opt}} = k \sqrt{\frac{c(W_{\text{opt}})}{W_{\text{opt}}}}$$

As expected, as W_{opt} increases, the delay decreases and asymptotes to a constant value for large values of W_{opt} .

The interbuffer interconnect length can be written in terms of interconnect width as

$$h_{\text{opt}} = \sqrt{\frac{2r_s(c_0 + c_p)}{\frac{\rho}{W_{\text{opt}} T} c(W_{\text{opt}})}} \propto \sqrt{\frac{W_{\text{opt}}}{c(W_{\text{opt}})}}$$

As the optimum line width increases, the interbuffer interconnect length increases initially and then asymptotes to a constant. This implies that for a given line length, the number of repeaters reduces. The buffer size is given by

$$k_{\text{opt}} = \sqrt{\frac{r_s c(W_{\text{opt}})}{c_0 \frac{\rho}{W_{\text{opt}} T}}} \propto \sqrt{c(W_{\text{opt}}) W_{\text{opt}}}$$

The repeater area of a *single* interconnect is proportional to k_{opt} and is inversely proportional to h_{opt} , i.e.,

$$\text{repeater area of a single interconnect} \propto \frac{k_{\text{opt}}}{h_{\text{opt}}}$$

The total repeater area for a given metal layer is the product of the number of interconnects and the repeater area of a single

interconnect. The number of interconnects on a metal layer is inversely proportional to the pitch. Therefore

$$A_{\text{total}} \propto \frac{k_{\text{opt}}}{h_{\text{opt}}(W_{\text{opt}} + S)} \propto \frac{c(W_{\text{opt}})}{W_{\text{opt}} + S}$$

Power dissipation per unit length for a single line is given by [13]

$$\frac{P}{l} = \kappa_1 \left(\frac{k_{\text{opt}}}{h_{\text{opt}}} (c_0 + c_p) + c \right) + \kappa_2 \frac{k_{\text{opt}}}{h_{\text{opt}}} + \kappa_3 k_{\text{opt}} \frac{\tau_{\text{opt}}}{h_{\text{opt}}}$$

where

$$\begin{aligned} \kappa_1 &= \alpha V_{DD}^2 f_{clk} \\ \kappa_2 &= \frac{1}{2} V_{DD} (I_{\text{off}_n} + 2I_{\text{off}_p}) W_{n_{\text{min}}} \\ \kappa_3 &= \alpha V_{DD} W_{n_{\text{min}}} I_{\text{short circuit}} f_{clk} \log_e 3. \end{aligned}$$

Here, V_{DD} is the power supply voltage, f_{clk} is the clock frequency, α is the switching factor (or activity factor), which is the fraction of repeaters on a chip that are switched during an average clock cycle, I_{off_n} (I_{off_p}) is the leakage current per unit NMOS (PMOS), $W_{n_{\text{min}}}$ is the width of the NMOS transistor in minimum sized inverter, and $I_{\text{short circuit}}$ is the per unit width short circuit current. We assume $\alpha = 0.15$, $I_{\text{short circuit}} = 65 \mu\text{A}/\mu$ and $V_{t_n} = V_{t_p} = (1/4)V_{DD}$. The total power dissipation per unit length in global interconnects of a given layer is the product of the above quantity with the number of global lines which is inversely proportional to $W_{\text{opt}} + S$. The total power dissipation can be expressed as

$$\begin{aligned} \frac{P_{\text{total}}}{l} &\propto \frac{1}{W_{\text{opt}} + S} \left[\kappa_1 c(W_{\text{opt}}) \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0} \right)} \right) \right. \\ &\quad + \frac{\kappa_2 c(W_{\text{opt}})}{\sqrt{2c_0(c_0 + c_p)}} + 2\kappa_3 r_s \\ &\quad \left. \times \left(1 + \sqrt{\frac{1}{2} \left(1 + \frac{c_p}{c_0} \right)} \right) c(W_{\text{opt}}) \right] \\ &\propto \frac{c(W_{\text{opt}})}{W_{\text{opt}} + S}. \end{aligned}$$

Note that this has the same form and the dependence of total repeater area on line width. Therefore increasing the line width decreases both total repeater area and power dissipation by the same amount.

We now consider the following two cases separately.

- 1) Minimum-spaced lines.
- 2) Line spacing is the same as line width.

A. Minimum-Spaced Lines

In an interconnect system, the technology determines the interlayer dielectric thickness, the metal line thickness, minimum metal width and the minimum spacing at a given metal layer. For higher performance or throughput, one can increase the line width in order to decrease the line resistance. However, in order not to severely limit the wireability of the chip, the wires should be minimum spaced. This is specially true for deep submicron technologies where the designs are mostly wire-limited at the global tiers. Since the aspect ratio

of minimum sized global interconnects is two to three, the interlayer dielectric thickness is 2–3 times larger than the minimum inter-wire spacing on a given metal layer and the adjacent metal layers are orthogonal to each other [1] it implies that $C_{\text{intralevel}}$ is typically much larger than $C_{\text{interlevel}}$ (Fig. 2). Therefore, increasing the line width without changing the spacing is not going to significantly increase the interconnect capacitance. For instance, as shown in Fig. 3, for the 130 nm technology, if the global line width is increased from W_{min} to $4W_{\text{min}}$, the interconnect capacitance per unit length increases only by 22%. This is due to the fact that the parallel plate component of $C_{\text{interlevel}}$ to the upper and lower metal layers, which is proportional to line width, is a small fraction of the total line capacitance for a minimum sized wire (Fig. 2).

Line capacitance per unit length c can be written as

$$c(W) = c_a + c_b W$$

where c_a represents the total fringing capacitance and sidewall capacitance which is independent of W and $c_b W$ represents the parallel plate capacitance to the top and bottom layers. Also $S = W_{\text{min}}$. Therefore from (1)

$$W_{\text{opt}} = \frac{\frac{i-1}{2}c_a + \sqrt{\left(\frac{i-1}{2}c_a\right)^2 + 2c_b(i+1)c_a W_{\text{min}}}}{2c_b}.$$

Note that for $i = 0$, the FOM is the rate of bit transfer per unit width itself. Therefore for minimum spaced lines, the rate of bit transfer per unit width itself has a maximum for a particular line width given by the above expression. This is in sharp contrast to the findings in [10], where it was reported that ψ asymptotes to a fixed value as width decreases.

B. Line Spacing Equal to Line Width

This case is similar to the previous one except that the line capacitance is given by

$$c = c'_a + c'_b W + \frac{c_c}{W}$$

where the first term represents the constant fringing capacitance, the second term represents the parallel plate capacitance to top and bottom layers of metal which is proportional to the width and the last term represents the parallel plate capacitance to the neighboring wires which is inversely proportional to the spacing. Also $S = W$. For this case, from (1)

$$W_{\text{opt}} = \frac{\frac{i-1}{2}c'_a + \sqrt{\left(\frac{i-1}{2}c'_a\right)^2 + 4ic_c c'_b}}{2c'_b}.$$

For $i = 0$, the optimum width is zero, which means that the FOM which is also the rate of bit transfer per unit width keeps increasing as W reduces and should be kept minimum sized for maximum ψ .

III. PARAMETER EXTRACTION

We used FASTCAP [14] to extract the capacitance per unit length for global interconnects for ITRS2001 technology nodes up to 45 nm. For both cases, (i.e., when lines are assumed to be minimum spaced and when line spacing is equal to

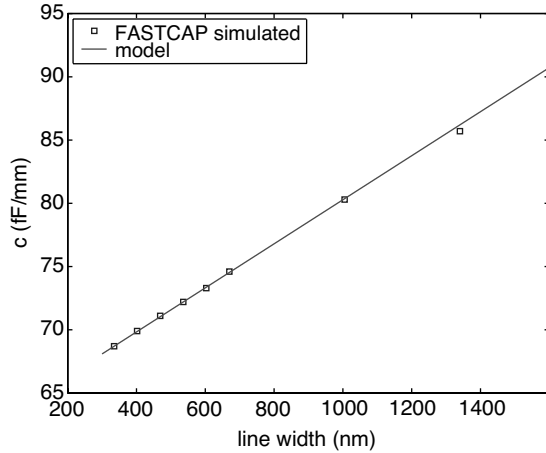


Fig. 3. FASTCAP simulated and fitted data for c as a function of W for 130-nm global line with $S = W_{\min}$.

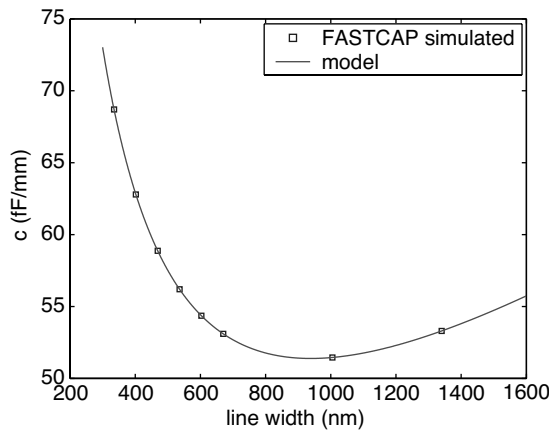


Fig. 4. FASTCAP simulated and fitted data for c as a function of W for 130-nm global line with $S = W$.

line width), capacitance per unit width was extracted using FASTCAP for a dense three layer interconnect mesh and c_a , c_b , c'_a , c'_b and c_c were obtained by curve fitting. Figs. 3 and 4 show the FASTCAP and the curve fitted values for the 130 nm technology node. It can be observed that the line capacitance model fits the FASTCAP simulated data very well. Similar agreement was obtained for other technology nodes.

Device parameters were extracted using SPICE simulation similar to [12]. A five stage ring oscillator with a given length of global interconnect of width W_{\min} in between each stage was simulated. The interconnect length h and inverter size k were varied to obtain the minimum stage delay per unit length. r_s , c_0 and c_p were calculated from these values of k_{opt} , h_{opt} and $(\tau/h)_{\text{opt}}$.

IV. RESULTS

The methodology outlined above was used to optimize global interconnect width for maximum FOM for ITRS 2001 technology nodes up to 45 nm. Device models were found to be extremely unreliable at 32 nm and 22 nm nodes and therefore were not included in this study. NMOS and PMOS off currents

TABLE I
TECHNOLOGY AND EQUIVALENT CIRCUIT MODEL PARAMETERS FOR TOP LAYER METAL FOR DIFFERENT TECHNOLOGY NODES BASED ON THE ITRS 2001

Tech. node (nm)	130	90	65	45
width (nm)	335	230	145	103
thickness (nm)	670	482	319	236
t_{ins} (μ)	6.3	4.7	3.9	2.9
ϵ_r	3.3	2.8	2.5	2.1
c_a (fF/mm)	207	181	165	143
c_b (fF/ μ^2)	0.057	0.071	0.103	0.116
c'_a (fF/mm)	70.95	58.24	65	52.08
c'_b (fF/ μ^2)	0.053	0.065	0.057	0.072
c_c (fF)	0.046	0.029	0.015	0.0098
r_s (k Ω)	6.23	9.04	9.6	13.2
c_0 (fF)	1.33	1.1	1.03	0.9
c_p (fF)	3.32	2.04	1.22	0.6
V_{DD} (V)	1.1	1	0.7	0.6
I_{off_n} ($\mu\text{A}/\mu$)	2	3.56	20	35.5
I_{off_p} ($\mu\text{A}/\mu$)	1.34	2.38	13.4	23.83
f_{clk} (GHz)	1.68	3.99	6.73	11.51

TABLE II
RATIO OF OPTIMUM INTERCONNECT WIDTH FOR VARIOUS TECHNOLOGY NODES WITH W_{\min}

Tech. node (nm)	$S = W_{\min}$			$S = W$	
	$i = 0$	$i = 1$	$i = 2$	$i = 1$	$i = 2$
130	0.86208	3.28286	7.53422	2.80179	5.09633
90	0.86485	3.32698	7.69269	2.88095	5.16434
65	0.86421	3.31664	7.65540	3.57768	7.37231
45	0.87332	3.47002	8.21825	3.60953	7.17473

were estimated similar to [15]. The relevant technology parameters are shown in Table I. S_{\min} was assumed to be equal to W_{\min} across all technology nodes.

Table II shows the calculated optimum width as a ratio of W_{\min} for various technologies for all cases. Note that for minimum spaced lines the optimum value of W which maximized ψ is approximately 13% less than W_{\min} for all technologies. This is clearly not feasible, however, as shown in Fig. 5, ψ is only 0.3% lower at $W = W_{\min}$ than the optimal value. This was also found to be true across all the technologies considered.

Also note that for all cases, the optimum interconnect width is less than $8.5 W_{\min}$ so inductance effects are not significant. To further verify this, Fig. 6 plots the *critical inductance* (see Appendix) as a function of line inductance for minimum width and $7.5 \times$ minimum width global line for 130 nm technology node. As pointed out in [12], if line inductance is less than l_{crit} then the interconnect system is *overdamped* and inductive effects are negligible. From Fig. 6, we observe that even for $W = 7.5 W_{\min}$, the interconnect is overdamped for most practical range of line inductance values ($0 \leq l \leq 2$ nH/mm). However, this may not be true for $i > 2$.

Further, note that W_{opt}/W_{\min} values are similar across all technology nodes when line spacing is kept minimum for $i = 1$

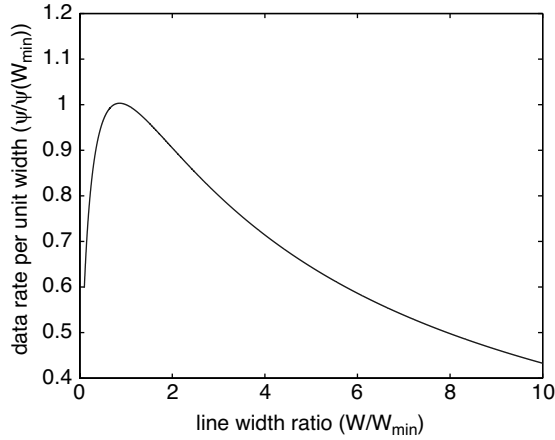


Fig. 5. Data rate per unit width as a function of line width for minimum spaced lines in 180-nm technology node.

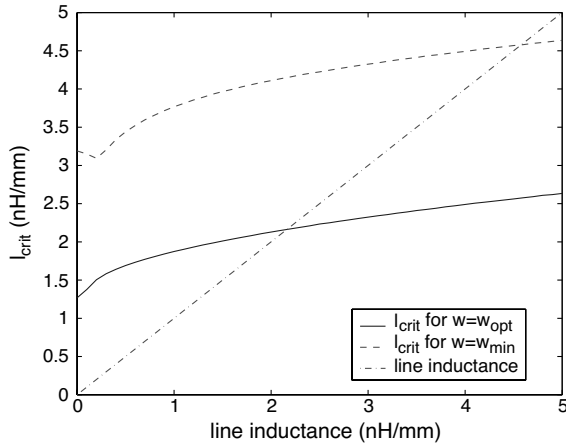


Fig. 6. Critical inductance as a function of line inductance for minimum spaced 130-nm global interconnect of $W = W_{\min}$ and $W = 7.5 W_{\min}$.

TABLE III
RATIO OF OPTIMUM DELAY PER UNIT LENGTH AT $W = W_{\text{opt}}$
WITH OPTIMUM DELAY PER UNIT LENGTH AT $W = W_{\min}$

Tech. node (nm)	$S = W_{\min}$		$S = W$	
	$i = 1$	$i = 2$	$i = 1$	$i = 2$
130	0.60304	0.45428	0.51668	0.40311
90	0.59877	0.44953	0.50405	0.39565
65	0.59976	0.45064	0.43992	0.32586
45	0.58546	0.43478	0.43324	0.32576

and 2, while they increase with technology scaling when line spacing is equal to line width. Also note that we have not included the trivial and infeasible result $W_{\text{opt}} = 0$ for $S = W$ and $i = 0$. Also for $S = W_{\min}$ and $i = 0$, the FOM (ψ) at $W = W_{\min}$ is only 0.3% lower than the optimal value at $W = W_{\text{opt}}$ which is approximately $0.87 W_{\min}$. In the following series of results (shown in Tables III–VI) we always report the ratio of performance metrics at $W = W_{\text{opt}}$ and the corresponding value at $W = W_{\min}$. Therefore we will exclude $i = 0$ case for $S = W_{\min}$ from now on.

Table III shows the optimum delay per unit length at $W = W_{\text{opt}}$ as a fraction of the optimum delay per unit length when

TABLE IV
RATIO OF TOTAL REPEATER AREA AT A GIVEN LEVEL AT $W = W_{\text{opt}}$
WITH TOTAL REPEATER AREA AT A GIVEN LEVEL AT $W = W_{\min}$

Tech. node (nm)	$S = W_{\min}$		$S = W$	
	$i = 1$	$i = 2$	$i = 1$	$i = 2$
130	0.55750	0.36437	0.26696	0.16250
90	0.55134	0.35767	0.25407	0.15654
65	0.55277	0.35922	0.19353	0.10618
45	0.53217	0.33705	0.18770	0.10612

TABLE V
RATIO OF ψ AT $W = W_{\text{opt}}$ WITH ψ AT $W = W_{\min}$

Tech. node (nm)	$S = W_{\min}$		$S = W$	
	$i = 1$	$i = 2$	$i = 1$	$i = 2$
130	0.77437	0.51588	0.69079	0.48676
90	0.77194	0.51182	0.68864	0.48941
65	0.77251	0.51276	0.63537	0.41626
45	0.76423	0.49902	0.63947	0.42785

TABLE VI
RATIO OF OPTIMIZED FIGURE OF MERIT AT $W = W_{\text{opt}}$
WITH FIGURE OF MERIT AT $W = W_{\min}$

Tech. node (nm)	$S = W_{\min}$		$S = W$	
	$i = 1$	$i = 2$	$i = 1$	$i = 2$
130	1.2841	2.4998	1.3370	2.995
90	1.2892	2.5327	1.3662	3.1264
65	1.2880	2.5250	1.4443	3.9203
45	1.3053	2.6399	1.4760	4.0317

$W = W_{\min}$. As expected, increasing the width of the wires reduces the delay per unit length significantly. Note that when the line width is increased from W_{\min} to $\sim 3 W_{\min}$, (corresponding to $i = 1$ for both cases in Table II), the delay improvement is significant. However, as the line width is increased further to $\sim 7\text{--}8 W_{\min}$ (corresponding to $i = 2$ for both cases in Table II), the incremental improvement in delay is not as significant. This is expected since

$$\left(\frac{\tau}{h}\right)_{\text{opt}} \propto \sqrt{\frac{c(W)}{W}}$$

and as W becomes very large, the line capacitance is dominated by the parallel plate component of $C_{\text{interlayer}}$, i.e., $c(W) \propto W$. Also note that the relative improvements in delay are not very sensitive to technology scaling.

Table IV shows the total repeater area for all interconnects at the global tier when $W = W_{\text{opt}}$ as a fraction of the total repeater area for all interconnects at the global tier when $W = W_{\min}$. As pointed out earlier, this fraction is also the ratio of the total power dissipation of all repeaters at the global tier when $W = W_{\text{opt}}$ and the total power dissipation of all repeaters at the global tier when $W = W_{\min}$. It can be observed that as the line width is increased (as i increases as per Table II), the total repeater area (and power dissipation) decreases dramatically, even though the

size and therefore the area (and power dissipation) of a single repeater increases. This is due to the fact that the wider wires result in a large increase in optimal interbuffer interconnect length and also fewer number of interconnects at a given tier. Therefore the total repeater power dissipation reduces dramatically.

Table V shows the rate of bit transfer per unit width ψ at $W = W_{\text{opt}}$ as a fraction of ψ at $W = W_{\text{min}}$. As indicated in Fig. 5 ψ peaks at $W = 0.87 W_{\text{min}}$ and is only 0.3% lower at $W = W_{\text{opt}}$. Therefore if the primary goal of the design is to maximize ψ , then minimum sized, minimum space wires as prescribed by the ITRS should be used. However, if the delay needs to be improved, then wire width should be increased at the expense of ψ . Note that the ratio of ψ with $W = W_{\text{opt}}$ and $W = W_{\text{min}}$ is fairly insensitive to technology scaling.

Table VI shows the ratio of the optimized FOM when $W = W_{\text{opt}}$ and the FOM when $W = W_{\text{min}}$. If this ratio was very close to 1, it would imply that the above-mentioned optimizations were not significantly improving the user-specified FOM and therefore were not very useful. However, in Table VI we find that these ratios are very different from 1, indicating a nontrivial improvement in the FOM at W_{opt} compared to W_{min} which further emphasizes the utility of these optimizations. Also note that except for the second case with $i = 2$, the improvement in FOM at the optimum width is fairly insensitive to technology scaling.

V. CONCLUSION

In conclusion, we have developed a new methodology for optimizing global interconnect width which maximizes a user-specified FOM, which is a function of the data-rate per unit chip edge and interconnect delay per unit length. Using this methodology we have developed expressions for optimum interconnect widths for typical FOMs for two extreme scenarios regarding line spacing: 1) spacing kept constant at its minimum value and 2) spacing kept the same as line width. We have used these expressions to compute the optimal global interconnect width and quantified the effect of increasing the line width on delay per unit length, total repeater area and power dissipation and bandwidth. As expected, an increase in the line width decreases the optimal delay per unit length (i.e., decreases latency), total buffer area and power dissipation, but severely degrades the rate at which bits can be transmitted per unit chip edge, i.e., bandwidth. We also observed that in most cases, the relative increase in the line width (from W_{min} to W_{opt}), the relative improvement in delay per unit length, total repeater area and power dissipation, and the relative degradation in the datarate per unit chip edge are fairly insensitive to technology scaling.

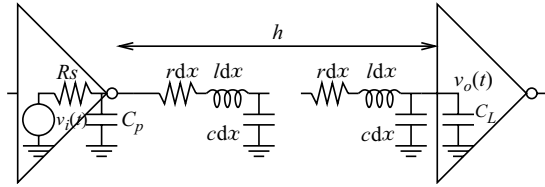


Fig. 7. Equivalent circuit of a driver-interconnect-load segment. The interconnect is uniform with resistance, capacitance and inductance per unit length of r , c and l respectively.

This work will have significant implications for signaling and design optimization for global interconnects in future nanometer-scale technologies.

APPENDIX CRITICAL INDUCTANCE

Consider a uniform line with resistance, capacitance and inductance per unit length of r , c , and l , respectively, driven by a repeater of series resistance R_S and output parasitic capacitance C_P , and driving an identical repeater with load capacitance C_L (Fig. 7). For a given technology, let the output resistance, output parasitic capacitance and input capacitance of a minimum-sized repeater be r_s , c_p and c_0 respectively. Therefore if the repeater size is k times the size of a minimum sized repeater, $R_S = r_s/k$, $C_P = c_p k$ and $C_L = c_0 k$. The transfer function derivation is outlined here from [12] for completeness. The ABCD parameter matrix for a uniform RLC transmission line of length h is given by [12]

$$\begin{bmatrix} \cosh(\theta h) & Z_0 \sinh(\theta h) \\ \frac{1}{Z_0} \sinh(\theta h) & \cosh(\theta h) \end{bmatrix}$$

where

$$Z_0 = \sqrt{\frac{r + sl}{sc}}$$

s is the complex frequency $j\omega$, and

$$\theta = \sqrt{(r + sl)sc}$$

Therefore the ABCD parameter matrix of the configuration in Fig. 7 is given by the equation shown at the bottom of this page, and the input-output transfer function is given by the first equation shown at the top of the next page. The step response of this system is given by $V_o(s) = (1/s)H(s)$ in the Laplace domain. However, computing the response in the time domain is

$$\begin{aligned} & \begin{bmatrix} 1 & R_S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC_P & 1 \end{bmatrix} \begin{bmatrix} \cosh(\theta h) & Z_0 \sinh(\theta h) \\ \frac{1}{Z_0} \sinh(\theta h) & \cosh(\theta h) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ sC_L & 1 \end{bmatrix} \\ & = \begin{bmatrix} (1 + sR_S C_P) [\cosh(\theta h) + sC_L Z_0 \sinh(\theta h)] & (1 + sR_S C_P) Z_0 \sinh(\theta h) + R_S \cosh(\theta h) \\ + \frac{R_S}{Z_0} \sinh(\theta h) + sC_L R_S \cosh(\theta h) & \\ sC_P [\cosh(\theta h) + sC_L Z_0 \sinh(\theta h)] & \\ + \frac{1}{Z_0} \sinh(\theta h) + sC_L \cosh(\theta h) & sC_P Z_0 \sinh(\theta h) + \cosh(\theta h) \end{bmatrix} \end{aligned}$$

$$H(s) = \frac{V_o(s)}{V_i(s)} = \frac{1}{[1 + sR_S(C_P + C_L)] \cosh(\theta h) + \left[\frac{R_S}{Z_0} + sC_L Z_0 + s^2 R_S C_P C_L Z_0 \right] \sinh(\theta h)}$$

$$l_{\text{crit}} = \frac{\frac{b_1^2}{4} - \frac{r^2 c^2 h^4}{4!} - R_S(C_P + C_L) \frac{rch^2}{2!} - (R_S ch + C_L rh) \frac{rch^2}{3!} - R_S C_P C_L rh}{\frac{ch^2}{2!} + C_L h}$$

analytically intractable. The above transfer function is therefore approximated by a second order Padé approximation as

$$H(s) \approx \frac{1}{1 + sb_1 + s^2 b_2} = \frac{1}{\left(1 + \frac{s}{s_1}\right) \left(1 + \frac{s}{s_2}\right)} \quad (2)$$

where

$$\begin{aligned} b_1 &= R_S(C_P + C_L) + \frac{rch^2}{2!} + R_S ch + C_L rh \\ b_2 &= \frac{lch^2}{2!} + \frac{r^2 c^2 h^4}{4!} + R_S(C_P + C_L) \frac{rch^2}{2!} \\ &\quad + (R_S ch + C_L rh) \frac{rch^2}{3!} + C_L lh + R_S C_P C_L rh. \end{aligned}$$

The 50% delay τ is given by

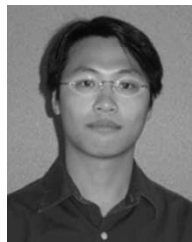
$$0.5 - \frac{s_2}{s_2 - s_1} \exp(s_1 \tau) + \frac{s_1}{s_2 - s_1} \exp(s_2 \tau) = 0.$$

This transfer function can be used to calculate the 50% delay [16]. Long VLSI interconnects are typically broken up into buffered segment of equal lengths and driven by identical repeaters. For minimum total delay in these long interconnects, the delay per unit length in the optimally buffered segment should be minimized. The driver size k and interconnect length h can be numerically optimized to give minimum delay per unit length [16], [17].

The second order transfer function given by (2) and discussed in [16], [17] can be *critically damped*, *overdamped*, and *underdamped* when $b_1^2 - 4b_2$ is equal to, greater than, or less than zero respectively. The response of an overdamped system is very similar to an RC line whereas for an underdamped system, the behavior is significantly different from an RC line, i.e., inductive effects are significant. Since b_1 and b_2 are functions of h and k and b_2 is a function of l , it has been shown [16] that for optimum values of h and k where interconnect delay is minimum for a given line inductance, a value l_{crit} can be obtained for which the system will be critically damped [16]. If line inductance is less than l_{crit} , the system will be overdamped where as if line inductance is greater than l_{crit} , the system will be underdamped, as specified in the second equation shown at the top of the page.

REFERENCES

- [1] International Technology Roadmap for Semiconductors (ITRS), 2001.
- [2] W. J. Dally, "Interconnect-limited VLSI architecture," in *Proc. IEEE Int. Conf. Interconnect Technology*, 1999, pp. 15–17.
- [3] M. T. Bohr, "Interconnect scaling—the real limiter to high performance ULSI," in *IEDM Tech. Dig.*, 1995, pp. 241–244.
- [4] J. D. Meindl, "Beyond moore's law: The interconnect era," *Comput. Sci. Eng.*, pp. 20–24, 2003.
- [5] R. H. J. M. Otten and R. K. Brayton, "Planning for performance," in *Proc. Design Automation Conf.*, 1998, pp. 122–127.
- [6] K. Banerjee and A. Mehrotra, "Inductance aware interconnect scaling," in *Proc. Int. Symp. Quality Electronic Design*, 2002, pp. 43–47.
- [7] C.-K. Cheng, J. Lillis, S. Lin, and N. Chang, *Interconnect Analysis and Synthesis*. New York: Wiley, 1999.
- [8] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.
- [9] J. Cong and C.-K. Koh, "Simultaneous driver and wire sizing for performance and power optimization," *IEEE Trans. VLSI Syst.*, vol. 2, pp. 408–425, Apr. 1994.
- [10] A. Naeemi and J. D. Meindl, "Optimal global interconnecting devices for GSI," in *IEDM Tech. Dig.*, 2002, pp. 319–322.
- [11] T. Lin and L. T. Pileggi, "Throughput-driven IC communication fabric synthesis," in *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, 2002, pp. 274–279.
- [12] K. Banerjee and A. Mehrotra, "Analysis of on-chip inductance effects for distributed *RLC* interconnects," *IEEE Trans. Computer-Aided Design*, vol. 21, pp. 904–915, Aug. 2002.
- [13] —, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, pp. 2001–2007, Nov. 2002.
- [14] 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.
- [15] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. 1999 Int. Symp. Low Power Electronics and Design*, 1999, pp. 163–168.
- [16] K. Banerjee and A. Mehrotra, "Analysis of on-chip inductance effects using a novel performance optimization methodology for distributed *RLC* interconnects," in *Proc. Design Automation Conf.*, 2001, pp. 798–803.
- [17] —, "Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling," in *Proc. IEEE Symp. VLSI Circuits*, vol. 2001, pp. 195–198.



Man Lung Mui was born in Hong Kong, China. He received the B.S. degree in electrical engineering from the University of Illinois, Urbana-Champaign, in May 2002 and is currently pursuing the M.S. degree in electrical engineering with an emphasis in integrated circuit design.

In 2002, he joined the Illinois Center for the Integrated Micro-Systems group, Coordinated Science Laboratory, University of Illinois, as a Research Assistant. His research is focusing on interconnect performance and modeling for VLSI circuit designs.



Kaustav Banerjee (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley, CA, in 1999.

He is an Assistant Professor in the Electrical and Computer Engineering Department, University of California, Santa Barbara (UCSB), CA. He has held various academic and industrial research/visiting positions at Stanford University, Stanford, CA, Swiss Federal Institute of Technology (EPFL), Fujitsu, Intel and Texas Instruments. His present

research interests focus on a wide variety of nanometer scale issues in high-performance VLSI and mixed-signal designs, as well as on circuits and systems issues in emerging nanoelectronics. He is also interested in some exploratory interconnect and circuit architectures including 3-D ICs. At UCSB, he mentors six doctoral and two masters students. He also coadvise graduate students at Stanford University, University of Illinois at Urbana-Champaign, and EPFL-Switzerland. He has codirected two doctoral dissertations at Stanford University, and the University of Southern California. He has published over 85 scientific papers in international journals and 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 Chair of ISQED '04. He has also served 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 ACM Design Automation Conference, and is listed in *Who's Who in America*.



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, University of California at Berkeley, in 1996 and 1999, respectively.

In 1999, he joined the University of Illinois, 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

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 modelling 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.