Wire Width Planning for Interconnect Performance Optimization

Jason Cong, *Fellow, IEEE,* and Zhigang (David) Pan, *Member, IEEE*

***Abstract*—In this paper, we study wire width planning for interconnect performance optimization in an interconnect-centric design flow. We first propose some simplified, yet near-optimal wire sizing schemes, using only one or two discrete wire widths. Our sensitivity study on wire sizing optimization further suggests that there exists a small set of "globally" optimal wire widths for a range of interconnects. We develop general and efficient methods for computing such a "globally" optimal wire width design and show rather surprisingly that using only two "predesigned" widths for each metal layer, we are still able to achieve close to optimal performance compared with that by using many possible widths, not only for one fixed length, but also for all wire lengths assigned at each metal layer. Our wire width planning can consider different design objectives and wire length distributions. Moreover, our method has a predictable small amount of errors compared with optimal solutions. We expect that our simplified wire sizing schemes and wire width planning methodology will be very useful for better design convergence and simpler routing architectures.**

***Index Terms*—Interconnect optimization, wire planning, wire sizing.**

I. Introduction

For deep submicron (DSM) very large scale integration (VLSI) designs, interconnect has become a dominant factor in determining the overall circuit performance, reliability, and cost [1]–[4]. As a result, many interconnect optimization techniques have been proposed in recent years for interconnect performance optimization. Among these techniques, wire sizing optimization is to find proper wire width tapering or sizing function for an interconnect so that a certain objective function, such as the distributed RC delay, is minimized.

The optimal wire sizing (OWS) was first studied in [5] and [6]. Dividing each wire into smaller wire segments and assuming that each wire segment has a uniform wire width (to be selected from a set of discrete wire widths), their work presented an elegant algorithm to obtain optimal wire width for each wire segment, under the weighted delay objective. Later on, continuous wire shaping for a wire was studied, which corresponds to the case of discrete wire sizing formulation in [5] and [6] such that each wire can be chopped into infinitely fine wire segments and arbitrary wire widths can be used. Closed-form wire shaping functions were obtained to minimize the Elmore delay, first without fringing capacitance [7], [8], then with fringing capacitance [9], [10] and were later extended to handle bidirectional wires [11]. There are other variations on wire sizing optimizations, such as [12] for multiple-source nets, [13] and [14] for minimizing the maximum delay objective, and [15] and [16] considering high-order moments. Most of these studies, however, did not consider the coupling capacitance which becomes the dominant capacitance component in DSM designs. In [17]–[19], the coupling capacitance is taken into consideration explicitly by performing interconnect sizing and spacing (ISS) optimization and considerable delay reduction over OWS is obtained. Interested readers can refer to [2] and [3] for a comprehensive survey and tutorial.

Although these wire sizing/spacing optimizations have been shown to be very effective for interconnect delay reduction, there are still a lot of difficulties or limitations for current design flows to take full advantage of them due to the following reasons: i) These wire sizing optimization will lead to the usage of many discrete [5], [6], [12], [13] or even infinite [7]–[11] number of different wire widths. They usually form a wire width tapering that is much wider near the source while much thinner near the sink (e.g., in an exponential shaping function when no fringing capacitance is considered [7], [8]). This will make the overall routing structure irregular and the routing area utilization low. In addition, it needs the support of a full-blown gridless router, which is usually expensive to maintain. ii) To make these interconnect optimization algorithms (which are mainly at the routing level) feasible, proper high level wire planning is needed for the overall design convergence (e.g., to allocate adequate routing resources). However, the usage of many different wire widths (even for the same net) will make the interconnect planning very difficult.

In this paper, we first seek to simplify wire sizing optimizations. We then study wire width planning with performance/area optimizations. The main contributions of this paper include the following.

- We present two simple wire sizing schemes, namely single-width sizing (1-WS) and two-width sizing (2-WS). We show that delay and area of OWS [6] can be reasonably approximated by these two simplified wire sizing schemes. When the coupling capacitance is considered explicitly, 2-WS can provide further delay and area reduction than 1-WS and achieve close-to-optimal solution quality as compared to running an ISS algorithm [19] directly.
- We explore the tradeoff between delay and area, using a set of design metrics in the form of A^iT^j (where A denotes

Manuscript received April 24, 2000; revised January 15, 2001. This paper was recommended by Associate Editor M. Sarrafzadeh. This work was supported in part by Semiconductor Research Corporation under Contract 98-DJ-605 and by a grant from Intel Corporation.

J. Cong is with the Computer Science Department, University of California, Los Angeles, CA 90095 USA (e-mail: cong@cs.ucla.edu).

Z. Pan is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: dpan@watson.ibm.com).

Publisher Item Identifier S 0278-0070(02)01785-2.

area and T denotes delay). In particular, we show that the metric AT^4 is very effective to guide area-efficient performance optimization, with up to 60% area reduction but less than a 10% delay increase compared to a delay-only optimization metric.

- Our delay sensitivity study further suggests that there exists a small set of "globally" optimal wire widths for each layer with a wide range of interconnect lengths so that we can perform early wire width planning. We develop efficient methods for computing such "globally" optimal wire width design and show rather surprisingly that using only two "predesigned" widths for each metal layer, we are still able to achieve close to optimal performance compared with that by using many possible widths, not only for one *fixed* length, but also for *all* wire lengths assigned at each metal layer.
- Furthermore, we provide sample wire-width design recommendations for current and future technologies.

The rest of the paper is organized as follows. Section II states the preliminaries. Section III presents two simplified wire sizing schemes and shows their effectiveness. Section IV studies the interconnect delay/area tradeoff and proposes a new design metric that is performance driven, yet area efficient. Then in Section V, we propose a general and effective wire width planning methodology. We demonstrate that an optimized two-width design for each metal layer shall be enough to achieve near optimality. The conclusions and discussions follow in Section VI. The preliminary results of this work were presented in [20] and a U.S. patent was filed for it [21].

II. Preliminaries

This section presents the preliminaries, including the models and key parameters used in the paper. We model the driver as an effective resistance R_d connected to an ideal voltage source and the sink as a load capacitance C_L. The well-known Elmore delay model [22], [23] is used to compute the device and interconnect delays. Although the Elmore delay model may give too conservative a delay estimation in DSM designs, especially for near-source sinks in a routing tree with many branches due to resistance shielding [24], it is still a good delay measurement for two-pin nets (the majority of all nets in real designs and thus the focus of our wire width planning work) and for general high-level estimation and planning purposes. Note that for high-level estimation and planning, other sources of errors, such as estimation of coupling capacitance due to unknown neighborhood structures, may outweigh the inaccuracy due to the Elmore delay model. Also, our wire width planning methodology can easily adapt to more complex and accurate models. The notations for key interconnect and device parameters are:

$W_{\min}$ minimum wire width, in μm;
$S_{\min}$ sheet resistance, in μm;
r sheet resistance, in $\Omega/\square$;
c_a unit area capacitance, in fF/μm^2;
c_f unit effective-fringing capacitance[1], in fF/μm;
t_g intrinsic device delay in ps;
c_g input capacitance of a minimum device, in fF;
r_g output resistance of a minimum device, in kΩ.

TABLE I
Basic Parameters

Tech. (μm)		0.25	0.18	0.13	0.10	0.07
W_{min}		0.25	0.18	0.13	0.10	0.07
S_{min}		0.34	0.24	0.17	0.14	0.10
t_g		86.6	66.4	54.4	50.1	29.8
c_g		0.282	0.234	0.135	0.072	0.066
r_g		16.2	17.1	22.1	23.4	22.1
Tier-1	r	0.073	0.068	0.081	0.092	0.095
	c_a	0.059	0.060	0.046	0.053	0.056
	c_f	0.082	0.064	0.043	0.045	0.040
Tier-2	r	0.016	0.011	0.018	0.022	0.030
	c_a	0.021	0.0176	0.0128	0.0136	0.0163
	c_f	0.206	0.160	0.103	0.103	0.089
Tier-3	r	0.013	0.0088	0.011	0.011	0.012
	c_a	0.0125	0.0097	0.0067	0.0074	0.0077
	c_f	0.154	0.119	0.104	0.103	0.088
Tier-4	r	-	-	0.0088	0.0088	0.0075
	c_a	-	-	0.0043	0.0043	0.0035
	c_f	-	-	0.0782	0.0782	0.0904

The device and the first metal layer parameters used in this study are extracted based on the *1997 National Technology Roadmap for Semiconductors* (NTRS'97) [25]. As NTRS'97 only provides the first metal layer information, to study the effect of interconnect reverse scaling [26]–[28] at higher metal layers, we extract a set of RC parasitics for higher metal layers, based on the geometry information from UC Berkeley's Strawman technology [29] and from SEMATECH [30]. Similar to [26], [28], [29] we define a routing *tier* to be a pair of adjacent metal layers with the same cross-sectional dimensions. Thus, from bottom to top, Tier-1 refers to metal layers 1 and 2, Tier-2 refers to metal layers 3 and 4,… and Tier-4 refers to metal layers 7 and 8. For capacitance extraction, we use the 2.5-dimensional capacitance extraction methodology reported in [31], which uses a three–dimensional (3-D) field solver to generate accurate capacitance values for interpolation and extrapolation. The values of these basic parameters are shown in Table I. Note that these parameters are used mainly to illustrate our wire width planning and optimization methodology. More complete sets of process parameters, if necessary, can be used in the same manner for wire width planning and optimization.

III. Simplified Wire Sizing Schemes

In this section, we present two simple wire sizing schemes, namely single-width sizing (1-WS) and two-width sizing (2-WS), which will be used later for wire width planning. We show that both 1-WS and 2-WS provide good approximation to OWS that uses many different wire widths, under the assumption of fixed effective-fringing capacitance coefficient [6], [9]. In the scenario of variable effective-fringing capacitance coefficients such as under fixed pitch-spacing between neighboring wires, 2-WS provides more flexibility than 1-WS and still achieves near-optimal performance compared to running an optimal ISS algorithm with many different wire widths [19].

[1] It is the sum of fringing and coupling capacitances [17].

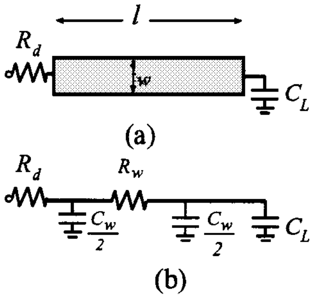


Fig. 1. (a) Single-width sizing to determine the optimal uniform width w. (b) The one-segment π-type RC model for the interconnect.

A. Single-Width Sizing

Given an interconnect of length l with loading capacitance C_L and driver resistance R_d, as shown in Fig. 1(a), the 1-WS problem is to determine the *best* uniform width that minimizes the source-to-sink delay. To compute the distributed Elmore delay, the original wire is often divided into many small wire segments and each wire segment is modeled as a π-type RC circuit. For uniform-width wire with a π-type model, it can be shown that the Elmore delay is the same no matter how the wire is divided into shorter wire segments [12], [32]. Therefore, we can just use the one-segment π-model as in Fig. 1(b), where R_w denotes the total wire resistance and C_w denotes the total wire capacitance. The Elmore delay from the driver to the load in Fig. 1(a) can then be written as follows:

$$T(w,l) = R_d c_f l + R_d C_L + \frac{1}{2} r c_a \cdot l^2 + R_d c_a l \cdot w + \left(\frac{1}{2} r c_f l^2 + r l C_L\right) \cdot \frac{1}{w}. \quad (1)$$

Thus the best wire width to minimize $T(w,l)$ is

$$w^*(l) = \sqrt{\frac{r\left(c_f l + 2C_L\right)}{2R_d c_a}}. \quad (2)$$

From this, we can see that larger c_f and C_L lead to larger wire sizes, while larger R_d (weaker driver) and c_a lead to a smaller wire sizing solution. This simple analytical formula confirms some previous results, including the *wire-sizing/driver-sizing* relation (i.e., larger driver size leads to larger wire sizes), *wire-sizing/capacitive-loading* relation (i.e., larger capacitive loading leads to larger wire sizes) in [33], and the *effective-fringing property* (i.e., larger effective-fringing capacitance leads to larger wire sizes) in [17]. The optimal delay for 1-WS using w^* is

$$T_{1ws}(l) = R_d C_L + R_d c_f l + \sqrt{2R_d c_a r\left(c_f l + 2C_L\right)} \cdot l + \frac{1}{2} r c_a \cdot l^2. \quad (3)$$

The four terms at the r.h.s. of (3) are $O(1)$, $O(l)$, $O(l\sqrt{l})$, and $O(l^2)$ in terms of l, respectively. It can be easily shown that T_{1ws} is a quadratic convex function of the interconnect length l. Therefore, the equally spaced buffer insertion algorithm as in [34] can be used to perform simultaneous buffer insertion and uniform wire sizing.

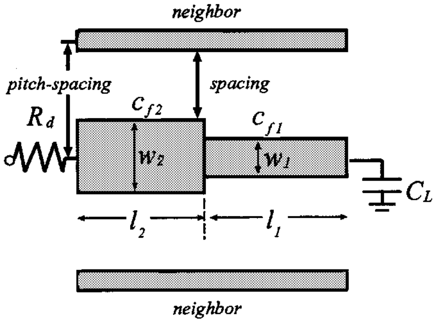


Fig. 2. Two-width sizing to determine the optimal w_1, w_2, l_1, and l_2 with $l_1 + l_2 = l$.

B. Two-Width Sizing

Compared to 1-WS that allows only *one* uniform wire width, the optimal 2-WS provides slightly more flexibility by allowing up to *two* discrete wire widths. As shown in Fig. 2, 2-WS is to determine the optimal two widths w_1 and w_2, together with their lengths l_1 and l_2 (with the constraint of $l_1 + l_2 = l$) for performance optimization.

The Elmore delay under 2-WS can be written as follows:

$$\begin{aligned} T\left(w_1, w_2, l_1, l_2\right) =& R_d \cdot \left(c_{f2} l_2 + c_a w_2 l_2 + c_f 1 l_1 \right. \\ & \left. + c_a w_1 l_1 + C_L\right) \\ & + \frac{1}{2} r c_a \left(l_2^2 + l_1^2\right) + r c_a l_1 l_2 \frac{w_1}{w_2} \\ & + \frac{r c_{f1} l_1 l_2}{w_2} + \frac{r c_{f2} l_2^2}{2w_2} + \frac{r c_{f1} l_1^2}{2w_1} \\ & + r C_L \left(\frac{l_2}{w_2} + \frac{l_1}{w_1}\right). \end{aligned}$$

The above delay formula can be rewritten as a quadratic function of l_2 in the following form, after substituting $l_1 = l - l_2$:

$$T\left(w_1, w_2, l_2\right) = K_2 \cdot l_2^2 + K_1 \cdot l_2 + K_0 \quad (4)$$

where

$$\begin{aligned} K_2 =& r c_a \left(1 - \frac{w_1}{w_2}\right) + \frac{1}{2} r \left(\frac{c_{f1}}{w_1} + \frac{c_{f2}}{w_2} - \frac{2c_{f1}}{w_2}\right) \\ K_1 =& R_d \left(c_{f2} - c_{f1}\right) + R_d c_a \left(w_2 - w_1\right) + r c_a l \left(\frac{w_1}{w_2} - 1\right) \\ & + r l c_{f1} \left(\frac{1}{w_2} - \frac{1}{w_1}\right) + r C_L \left(\frac{1}{w_2} - \frac{1}{w_1}\right) \\ K_0 =& R_d \left(c_{f1} l + c_a w_1 l + C_L\right) \\ & + \frac{1}{2} r \left(c_a + \frac{c_{f1}}{w_1}\right) l^2 + \frac{r C_L l}{w_1}. \end{aligned}$$

Then, the optimal length for l_2, denoted as l_2^*, to minimize $T(w_1, w_2, l_2)$ is either $-K_1/2K_2$ when $K_2 > 0$ and $0 \leq -K_1/2K_2 \leq l$, or the better one of 0 and l that gives smaller delay for all other cases. The optimal delay for given (w_1, w_2) is then

$$T^*\left(w_1, w_2\right) = K_2 \cdot l_2^{*2} + K_1 \cdot l_2^* + K_0 \quad (5)$$

and the corresponding interconnect area is

$$A^*\left(w_1, w_2\right) = w_2 l_2^* + w_1 \left(l - l_2^*\right). \quad (6)$$

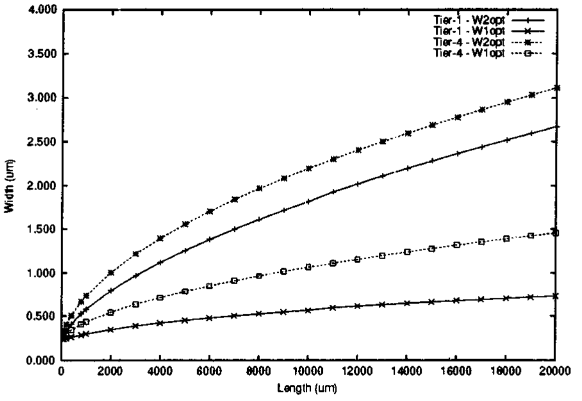


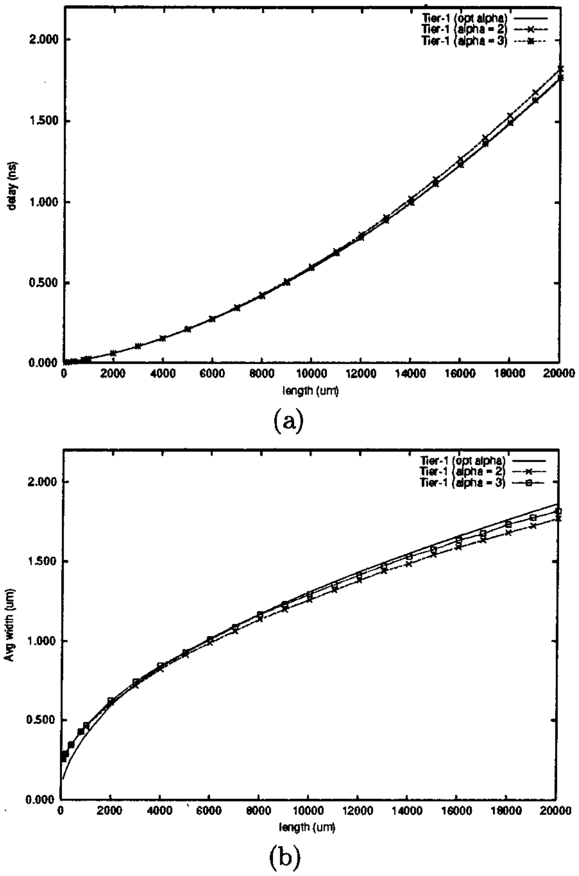
Fig. 3. The two optimal widths w_2^* and w_1^* for Tier-1 and Tier-4 under the 0.10-μm technology. $R_d = r_g/100, C_L = c_g \times 100$.

The 2-WS optimization program will search for the best wire width pair (w_1^*, w_2^*) from given technology and design specification. Let $w_2^* = \alpha w_1^*$, with $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. For optimal 2-WS solution, α is usually within a small range. Fig. 3 shows the optimal widths of w_1^* and w_2^* for Tier-1 and Tier-4 using the 0.10-μm technology for a wide range of interconnect lengths (from 100 μm to 2 cm).[2] It can be seen that for all cases, the ratio of w_2^*/w_1^* is between 1.2 to 3.6. Thus, we can set a conservative search range for $[\alpha_{\min}, \alpha_{\max}]$ to be from 1 to 5 during the 2-WS computation.

In fact, it is very interesting to observe that the 2-WS solution is not sensitive to fairly big variation of α around its optimal value. For example, we can just set α to be a fixed nearby integer, such as 2 or 3 and still achieve comparable performance. Note that for different α, 2-WS optimization will have different w_1^* and w_2^* for delay minimization. Fig. 4 shows the delay and average wire width comparisons for 2-WS using optimal α (see Fig. 3) and two fixed α of 2 and 3. There is very little difference for both delay and area using these two fixed integer ratios versus the optimal α. Therefore, in practice, one can choose to use a fixed integer ratio of two wire widths as this will simplify the overall routing structure and wire planning (see a more detailed discussion in Section V).

To enumerate different α from $\alpha_{\min}$ to $\alpha_{\max}$, an incremental step $\Delta\alpha = 0.5$ is usually adequate (with less than 0.1% delay difference compared to the very fine incremental step of $\Delta\alpha = 0.01$). The optimal wire width enumeration for w_1^* is bounded by the design specification of the minimum wire width $W_{\min}$ and the maximum width $W_{\max}$. In practice, $W_{\max}$ is usually not greater than $10 \cdot W_{\min}$. Again, it is accurate enough for an enumeration step of $\Delta w = W_{\min}/2$ (with less than 0.1% delay difference compared to a very fine increment of $\Delta w = W_{\min}/100$).

To summarize, since for a given (w_1, w_2), the best delay can be computed in closed-form formula from (5) and the number of (w_1, w_2) choices, bounded by $(\alpha_{\max} - \alpha_{\min})/\Delta\alpha \cdot (W_{\max} - W_{\min})/\Delta w$, is constant in practice, the optimal 2-WS can then be computed in constant time as well.



(a)

(b)

Fig. 4. (a) Delay and (b) average wire width comparisons for 2-WS using optimal α, fixed $\alpha = 2$, or $\alpha = 3$ for Tier-1 of the 0.10-μm technology. $R_d = r_g/100, C_L = c_g \times 100$.

C. Comparison of 1-WS and 2-WS With Many-Width Optimal Sizing

In this section, we compare the performances of 1-WS and 2-WS with that of optimal wire sizing with *many* discrete wire widths. There are two common scenarios when performing wire sizing optimization: i) Fixed effective-fringing capacitance coefficient c_f for different wire widths. It essentially assumes some fixed nominal spacing to neighboring nets (i.e., when a net is sized up, its neighboring nets will be pushed away). This simple capacitance model was widely used by early works of wire sizing optimizations [6], [9], [12], [13]. ii) Fixed pitch-spacing, defined to be the distance between the center lines of neighboring wires (see Fig. 2). It essentially assumes that when one net is sized, its neighboring nets are fixed. Then, different wire widths of the net to be sized will lead to different edge-to-edge spacings and thus different coupling and effective-fringing capacitances. This model explicitly considers coupling capacitance, as in [17]–[19].

1) Comparison With OWS Under Fixed c_f: Assuming that each wire has a set of wire width selections, [6] presented an OWS algorithm under the Elmore delay model, by iterative local refinement to compute lower and upper bounds of the optimal wire widths, followed by a dynamic programming algorithm to obtain the final OWS solution. The OWS solution depends on the range and granularity of the *given* wire width choices. Obviously, a larger wire width choice leads to better OWS solution,

[2] Note that from Fig. 3 to Fig. 7, we arbitrarily set the maximum length to be 2 cm, which is roughly the chip dimension in the current and future technologies in NTRS'97. The trend in each figure, however, shall go beyond the 2 cm length.

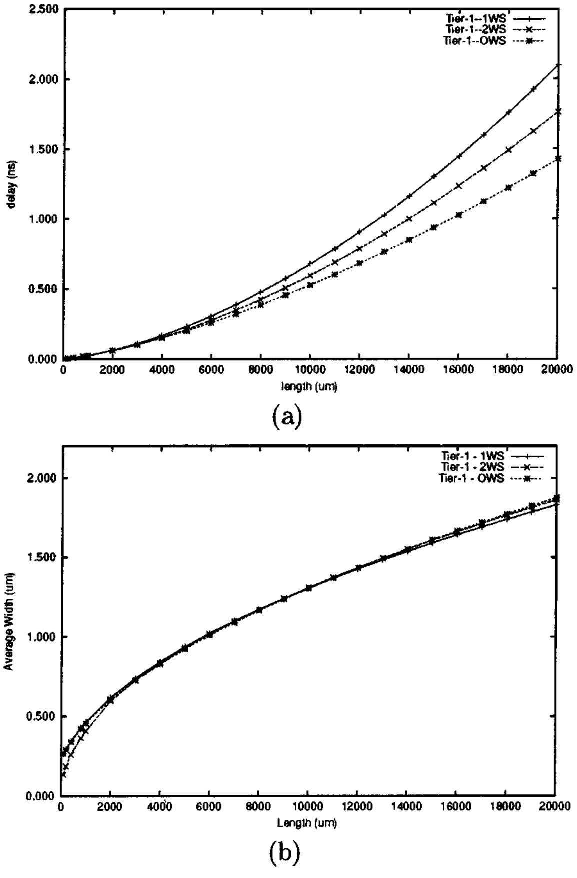


Fig. 5. (a) The delay and (b) average wire width comparisons of 1-WS, 2WS, and OWS for Tier-1 using the 0.10-μm technology. $R_d = r_g/100$, $C_L = c_g \times 100$. To run the OWS algorithm, we set $W_{\max} = 50 \times W_{\min}$ with the width incremental to be $(1/2)W_{\min}$ and the wire is segmented in every 100 μm (same for other figures).

which in the extreme case implies continuous wire shaping (i.e., infinite number of wire widths) as in [7], [8], [9], and [11]. The question is then, how many wire widths are "good" enough?

Our experiments show surprisingly that the optimized delays under 1-WS and 2-WS are close to that from running OWS algorithm [6] using a wide range of parameters from NTRS'97. Figs. 5 and 6 show the optimized delay and average wire width comparison of 1-WS, 2-WS, and OWS for an interconnect of length up to 2cm, for Tier-1 and Tier-4 under the 0.10-μm technology, respectively. For Tier-1 (in Fig. 5), both 1-WS and 2-WS have very comparable delays to OWS up to a wire length of 4 mm. For longer wires in Tier-1, the differences between 1-WS and 2-WS versus OWS become larger (up to 46% for 1-WS and 23% for 2-WS for the 2-cm interconnect). But in practice, we will not have long wires (e.g., > 4 mm) in Tier-1, because for a critical global interconnect, buffers will be inserted and/or upper metal layer will be used to route it. Fig. 6 shows that both 1-WS and 2-WS obtain almost the same delay as OWS for all wire lengths up to 2 cm (the chip dimension) for Tier-4. Figs. 5(b) and 6(b) also show the comparison of average wire widths under 1-WS, 2-WS, and OWS. It is interesting to observe that both 1-WS and 2-WS give very similar average wire widths compared to OWS, even for long wire lengths at Tier-1 where OWS has much better delay than 1-WS and 2-WS.

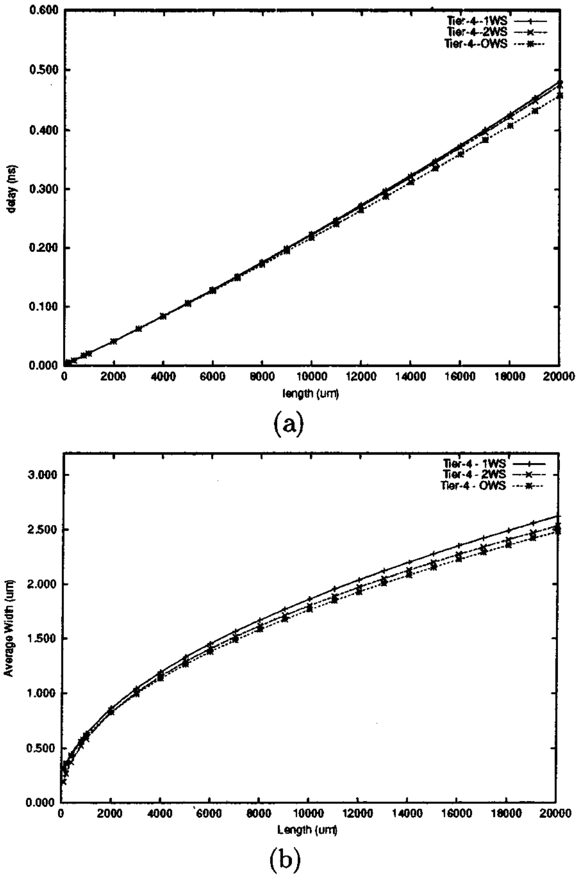


Fig. 6. (a) The delay and (b) average wire width comparisons of 1-WS, 2WS, and OWS for Tier-4 using the 0.10-μm technology.

Note that in theory, T_{1ws} in (3) is still a quadratic function of l, while T_{ows} is a subquadratic function of l [34]. For 1-WS to be a "good" approximation of OWS, the length l shall be smaller than certain threshold length such that the quadratic term becomes less important and dominated by other terms. We observe that as long as the quadratic term in (3), i.e., $(1/2)rc_al^2$, is smaller than the $O(l)$ and $O(l\sqrt{l})$ terms in (3), 1-WS approximates OWS well (usually within 90% accuracy). That is, 1-WS can be used to estimate the delay for OWS provided that $(1/2)rc_al^2 < R_dc_fl$ and $(1/2)rc_al^2 < \sqrt{2R_dc_ar(c_fl + 2C_L)} \cdot l$. It can be shown that if $(1/2)rc_al < R_dc_f$, then

$$\begin{aligned}\sqrt{2R_dc_ar\left(c_fl + 2C_L\right)} &> \sqrt{2R_dc_arc_fl}\\ &> \sqrt{2c_arl\frac{1}{2}rc_al} = rc_al\\ &> \frac{1}{2}rc_al.\end{aligned}$$

Therefore, both inequalities are met if $l < 2R_dc_f/rc_a$. For Tier-1, $2R_dc_f/rc_a =$ 4.3 mm; for Tier-4, $2R_dc_f/rc_a = 96$ cm which is much larger than the chip dimension. This explains why 1-WS and OWS delays are so close for wires shorter than 4 mm in Tier-1 and for wires up to chip dimension in Tier-4. Since 2-WS always achieves better performance than 1-WS, if 1-WS works well (e.g., 90% accuracy compared with OWS), 2-WS shall have a better approximation to OWS.

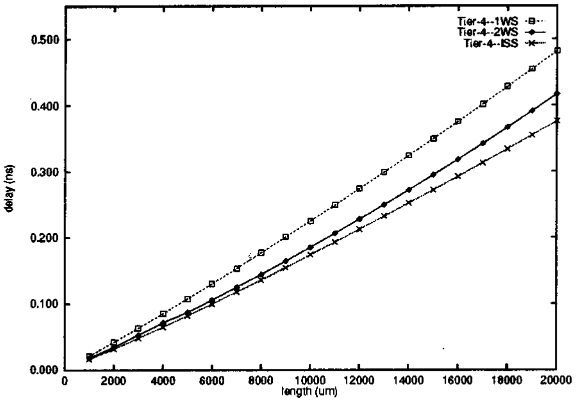


Fig. 7. Comparison of 1-WS, 2-WS, and ISS with variable c_f. $R_d = r_g/100$, $C_L = c_g \times 100$. To run ISS, $W_{\max} = 50 \times W_{\min}$ with the width incremental as $(1/2)W_{\min}$ and ten segments for each wire are used.

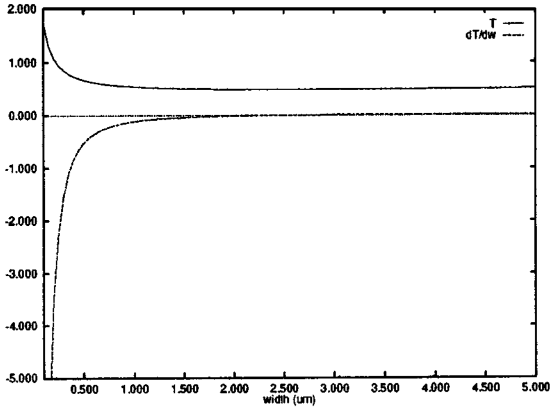


Fig. 8. The delay T and its sensitivity to w, dT/dw, using different uniform wire widths for a 2-cm global interconnect using the 0.10-μm technology. $R_d = r_g/100, C_L = c_g \times 100$.

2) Comparison With ISS Under Variable c_f: So far the validation of 1-WS and 2-WS is under the scenario of fixed effective-fringing capacitance. Another common scenario for wire sizing optimization is to fix the pitch-spacing. Then, different wire widths will lead to different edge-to-edge spacings and thus different coupling and effective-fringing capacitances.

In this scenario, wire tapering will show more advantages since downsizing wire segments near sinks will reduce the coupling capacitances. As a result, the 1-WS solution may not be flexible enough. However, we show that the 2-WS solution still achieves near-optimal performance. Fig. 7 shows the delay comparison of the optimal 1-WS and 2-WS solutions with an ISS solution [19] using many different wire widths under Tier-4 of the 0.10- μm technology. A table-based capacitance model [31] is used to look up the area, fringing, and coupling capacitances for different wire widths. We can see that the delay from 1-WS is about 20% to 30% larger than that from ISS. The 2-WS solution, however, has up to a 15% delay reduction compared to 1-WS and less than a 5% difference compared to ISS using 100 different wire widths.

To briefly summarize, we propose two simplified wire sizing optimization schemes, namely 1-WS and 2-WS. Both 1-WS and 2-WS provide good approximation to OWS [6], [9] with many or even an infinite number of different wire widths, assuming a fixed effective-fringing capacitance coefficient (essentially fixed edge-to-edge spacing). Under a fixed pitch-spacing scenario, 2-WS is superior to 1-WS and still provides good approximation to ISS [19] with many different wire widths. A conservative range for the optimal ratio w_2^*/w_1^* is between 1 and 5. Since the optimal 2-WS solution is not sensitive around the optimal w_2^*/w_1^*, in practice, we can just take the nearby integer to simplify routing structure.

IV. Delay-Area Tradeoff and Area Performance-Driven New Design Metric

The simple closed-form delay formula of 1-WS enables us to study the delay-area tradeoff and the sensitivity of delay versus wire width. From (1), we can compute the differential

$$\frac{dT}{dw} = R_d c_a - \frac{\frac{1}{2} r c_f l^2 + r l C_L}{w^2}.$$

As shown in Fig. 8, delay decreases sharply as width increases from the minimum wire width (i.e., 0.10 μm) since $dT/dw \ll 0$ when $w \approx W_{\min}$, then flattens as dT/dw slowly achieves zero where the delay is the minimum and after that the delay increases slowly as $dT/dw > 0$. The optimal width w^* is about 2.6 μm for a 2 cm global interconnect in Tier-4 under 0.10 μm. It is not difficult to see that in order to achieve the minimum delay, the cost, in terms of wire area, is high. For example, using wire width of 1 μm has only 10% more delay than the optimal OWS, but saves 62% area. Therefore, delay minimization only could lead to significantly larger area!

To obtain a good metric for area efficient performance optimization, we have performed extensive experiments on different area-delay metrics in the form of $A^i T^j$, including T (delay only), AT (area-delay product), AT^2, AT^3, AT^4, AT^5, and so on. It is obvious that as j gets larger, more weight is given to delay. In particular, our study suggests that AT^4 is a metric that is suited for area-efficient performance optimization, with only about a 10% delay increase from OWS, but significant area reduction. Fig. 9 shows an example. The optimal widths of a 2-cm interconnect for AT, AT^2, AT^3, AT^4, AT^5, and T are 0.10-, 0.30–, 0.60-, 1.0-, 1.15- and 2.6-μm respectively, with a delay of 1.77, 0.84, 0.62, 0.53, 0.52, and 0.48 ns, respectively. The optimal 1-WS solution under the AT^4 metric uses 62% smaller wiring area compared to OWS (20 000 μm^2 versus 52 000 μm^2), with only a 10% increase of delay. Therefore, we will use the performance-driven but area-efficient metric AT^4 in Section V for wire width planning.

V. Interconnect Architecture Planning for Wire Width Design

From our study of 1-WS and 2-WS in the previous sections, a very interesting observation is that the delay is not sensitive to certain degree wire width variations around the optimal solution (see Fig. 8). This not only suggests that we can achieve close to optimal performance with significant area saving (as shown in Section IV), but also suggests that there may exist a small set of "globally" optimal widths for a range of interconnect lengths, so that by just using such a small set of predetermined "fixed" widths, we are still able to get close to optimal performance for all interconnects in given length range! In Fig. 10, we draw the

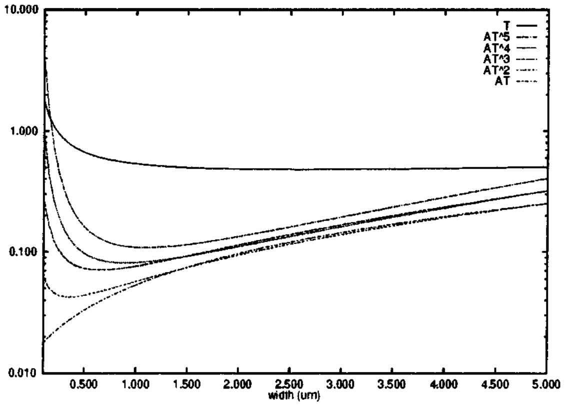


Fig. 9. Different optimization metrics for a 2-cm interconnect in Tier-4 under the 0.10 μm–technology. $R_d = r_g/100, C_L = c_g \times 100$. The y-axis is scaled so that all metrics can be shown in one figure.

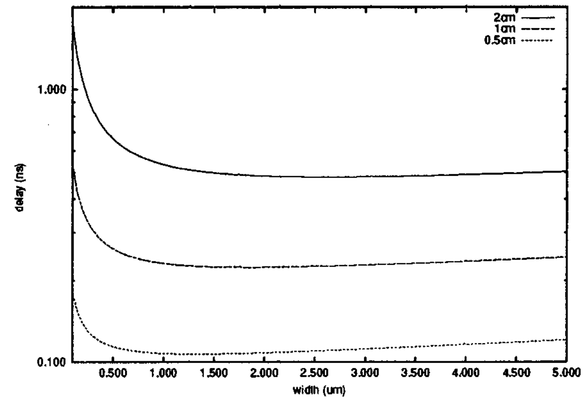


Fig. 10. Delay sensitivity of using different widths for a 0.5-, 1-, and 2-cm interconnect at Tier-4 of the 0.10-μm technology. $R_d = r_g/100, C_L = c_g \times 100$.

delay sensitivity versus wire width for three interconnects of length 0.5, 1, and 2 cm. The optimal widths for them are about 1.0, 1.4, and 2.6 μm. However, any 1-WS with width from 1.0 to 2.0 μm will have less than a 10% delay from that of OWS for all three lengths.

This crucial observation motivates us to study the interconnect architecture planning for optimal wire-width design. In particular, we want to determine a small set of "globally" optimal wire widths (such as only one or two widths) during the design planning phase for a wide range of interconnects (not for just one length!) such that by using these predetermined widths alone, we may still achieve near-optimal performance compared to the full-blown usage of an arbitrary number of wire widths together with complicated wire sizing (and/or spacing) algorithms. This optimal wire-width design, on one hand, still guarantees close to optimal performance; on the other hand, it greatly simplifies the routing architecture and the interaction of layout optimization with other higher level design planning tools and lower level routing tools.

A. Overall Approaches

Given the wire length range for each layer, the wire width planning problem is to find the best wire width design, written in the form of a vector $\vec{W}$, such that the following objective function:

$$\Phi\left(\vec{W}, l_{\min}, l_{\max}\right) = \int_{l_{\min}}^{l_{\max}} \lambda(l) \cdot f\left(\vec{W}, l\right) dl \qquad (7)$$

is minimized, where $\lambda(l)$ is the weighting function for length l and $f(\vec{W}, l)$ is the design objective function to be minimized, such as delay and area. In this paper, the design metric that can explore the delay-area tradeoff, $f(\vec{W}, l) = A^j(\vec{W}, l) \cdot T^k(\vec{W}, l)$ is used, where $A(\vec{W}, l)$ is the area and $T(\vec{W}, l)$ is the optimized delay using only those wire widths from the wire width planning $\vec{W}$. To simplify the routing architecture, we shall use as small a number of wire widths as possible. It is obvious that 1-width design (i.e., $\vec{W}$ has only one component W) and 2-width design (i.e., $\vec{W}$ has two components, W_1 and W_2) are the two simplest ones. So we will start from these two cases and show how the wire width planning works. In fact, as we shall show in Section V-B, the 2-width design is usually good enough to achieve near-optimal performance (only a few percent difference compared to using many widths), thus it is recommended for most designs. In terms of design metrics, when $j = 0$ and $k = 1$, the objective is for performance optimization only. However, as we observe in Section III, delay only minimization tends to use too large a wire width with marginal performance gain, since the delay/width curve becomes very flat while approaching optimal delay. We may use other $A^i T^j$ metrics according to the timing and area constraints. For ease of illustration, we assume $\lambda(l) = 1$.

We use the analytical (if possible) or numerical methods to compute the best 1-width or 2-width design (or a few more widths if necessary). Let us first consider the simplest case, 1-width design using metric T. We need to determine the "globally" best width to minimize

$$\int_{l_{\min}}^{l_{\max}} T(w, l) dl \qquad (8)$$

where

$$T(w, l) = R_d c_f l + R_d C_L + \frac{1}{2} r c_a \cdot l^2 + R_d c_a l \cdot w + \left(\frac{1}{2} r c_f l^2 + r l C_L\right) \cdot \frac{1}{w}$$

is the delay for wire length l using wire width w, the same as (1). So the "globally" optimal width W^* for w is thus

$$W^* = \sqrt{\frac{\int_{l_{\min}}^{l_{\max}} \left(\frac{1}{2} r c_f l + r C_L\right) l dl}{\int_{l_{\min}}^{l_{\max}} R_d c_a l dl}} = \sqrt{\frac{\frac{1}{3} r c_f \left(l_{\max}^3 - l_{\min}^3\right) + r C_L \left(l_{\max}^2 - l_{\min}^2\right)}{R_d c_a \left(l_{\max}^2 - l_{\min}^2\right)}}. \qquad (9)$$

If $l_{\max}^2 \gg l_{\min}^2$, which is the case for our length range for each tier, then W^* can be approximated as

$$W^* \approx \sqrt{\frac{\frac{1}{3} r c_f l_{\max} + r C_L}{R_d c_a}} \qquad (10)$$

which is about $\sqrt{2/3} \cdot w^*(l_{\max})$ from (2) provided that $C_L \ll c_f l_{\max}$.

For the 1-width design under more general design metrics in the form of $A^i T^j$ or 2-width design, a simple analytical formula like (9) or (10) may not be obtained as we have to solve a high-order equation for w. In this case, the numerical method will be used. For the example of the 2-width design, we can obtain the "globally" optimal width pair of (W_1, W_2), denoted as (W_1^*, W_2^*) for all wire lengths from $l_{\min}$ to $l_{\max}$ in a similar manner as computing the optimal 2-WS for a fixed wire length l in Section III-B. Let $W_2^* = \alpha W_1^*$, with $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. As in 2-WS optimization, $\alpha_{\min} = 1$, $\alpha_{\max} = 5$ and $\Delta\alpha = 0.5$ are usually adequate. The width enumeration of W_1^* is bounded by the design specification of minimum width $W_{\min}$ and maximum width $W_{\max}$. Again, the enumeration step of $\Delta w = W_{\min}/2$ is accurate enough (with less than 0.1% delay difference compared to a very fine increment of $\Delta w = W_{\min}/100$). For each (W_1, W_2), we then compute the objective function in (7), using the closed-form formula from (5). Since the total number of wire width choices for (W_1, W_2) is bounded (less than 200 in practice), the optimal (W_1, W_2) can then be computed very efficiently.

Our experiments show that the 2-width design is usually good enough. Yet, if needed, one can compute a few more wire widths using the same enumeration method described above. For a k-width design (where $k > 2$ is a small constant), denoted as $(W_1, W_2, \ldots, W_k)$, we assume that they form an arithmetic series, i.e., $(W_1, \alpha W_1, (2\alpha - 1)W_1, \ldots, [(k-1)\alpha - (k-2)]W_1)$, so that we limit our search space again to only two variables, α and W_1. For a given set of $(W_1, W_2, \ldots, W_k)$, we can use the efficient *local refinement* (greedy wire sizing algorithm) [6] to compute the optimal delay and the wiring area for each wire length during numerical integration. The granularities for searching α and W_1 are the same as those for the 2-width design. Note that the time complexity for the wire width planning is actually not a major concern, since it only needs to be run *once* for a given design or a set of designs. The key idea, however, is to identify a small number of "globally" optimal wire widths design, such that by using these predetermined wire widths, the near optimality can still be met, rather than using a large number of wire widths.

The weighting function $\lambda(l)$ provides a lot of flexibility. It can naturally be the wire length distribution function, or it can be a weight that a designer wants to put on different wire lengths (for example, larger weight for global interconnects). Our wire width planning, nonetheless, is flexible for *any weighting function*, with bounded maximum error compared with the "true" optimal solution by using many possible widths. It is justified by the following *maximum error theorem*. In the theorem, $f(\vec{W}, l)$ denotes the optimized design metric using an arbitrary number of wire widths, while $f(\vec{W}^*, l)$ denotes the optimized design metric using only our small set of "planned" wire widths. Usage of the maximum error theorem will be shown in Section V-B.

Theorem 1 (Maximum Error Theorem: If $|(f(\vec{W}, l) - f(\vec{W}^*, l))/f(\vec{W}^*, l)| \leq \delta_{\max}$ for any $l \in (l_{\min}, l_{\max})$, then for any $\lambda(l)$, we have

$$\left| \frac{\Phi\left(\vec{W}, l_{\min}, l_{\max}\right) - \Phi\left(\vec{W}^*, l_{\min}, l_{\max}\right)}{\Phi\left(\vec{W}^*, l_{\min}, l_{\max}\right)} \right| \leq \delta_{\max}. \quad (11)$$

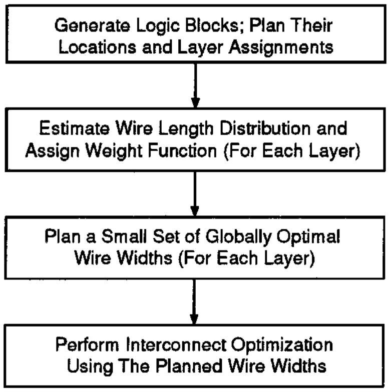

Generate Logic Blocks; Plan Their Locations and Layer Assignments
Estimate Wire Length Distribution and Assign Weight Function (For Each Layer)
Plan a Small Set of Globally Optimal Wire Widths (For Each Layer)
Perform Interconnect Optimization Using The Planned Wire Widths

Fig. 11. An exemplary flow of wire width planning and optimization.

Proof: The left-hand side of (11) can be written as

$$\begin{aligned} l.h.s. &= \left| \frac{\int_{l_{\min}}^{l_{\max}} \lambda(l) \cdot \left[f\left(\vec{W}, l\right) - f\left(\vec{W}^*, l\right) \right] dl}{\int_{l_{\min}}^{l_{\max}} \lambda(l) \cdot f\left(\vec{W}^*, l\right) dl} \right| \\ &\leq \left| \frac{\int_{l_{\min}}^{l_{\max}} \lambda(l) \cdot \delta_{\max} \cdot f\left(\vec{W}^*, l\right) dl}{\int_{l_{\min}}^{l_{\max}} \lambda(l) \cdot f\left(\vec{W}^*, l\right) dl} \right| \\ &= \delta_{\max}. \end{aligned}$$

□

Fig. 11 shows an exemplary flow of using our proposed wire width planning and optimization. At the beginning, logic blocks for a design are generated and their locations are roughly planned. Also, the designer may specify some rules for the wiring layer assignment (e.g., short wires are routed in lower metal layers). Then, based on the geometric locations of the logic blocks, the wire length information in the design can be computed. By assigning each interconnect to a specific metal layer, the wire length distribution of each layer can be obtained. Alternatively, if there is no physical locations, the wire length distribution data may be extracted from previous designs of similar characteristics, or obtained using some statistical models like the one in [27]. Note that while wire length distribution function is a natural candidate for the weighting function $\lambda(l)$ in the objective function (7) for wire width planning, the designer may choose to weight $\lambda(l)$ in some other manners (e.g., assign larger weights for global interconnects). Then, for a given design optimization metric, a small set of "globally" optimal wire widths for one or more specified layers are determined and planned (usually two-width design is adequate for both delay and area optimization). These predetermined wire widths for each layer will be used to plan and allocate proper routing resources, perform interconnect layout optimization, and generate final layouts. The reader may refer to [4] for more detailed discussions of how our wire width planning results can be used in an interconnect-centric design flow.

B. Effectiveness of Wire Width Planning

In this section, the results from using 1-width and 2-width designs are presented to show the effectiveness of wire width planning.

TABLE II
MAX WIRE LENGTH (IN MM) ASSIGNED TO EACH TIER

Tech. (μm)	0.25	0.18	0.13	0.10	0.07
Tier-1	2.50	1.80	1.30	1.00	0.70
Tier-2	6.50	5.85	3.27	2.84	2.30
Tier-3	17.3	19.0	8.23	8.04	7.57
Tier-4	-	-	20.7	22.8	24.9

TABLE III
ONE-WIDTH AND TWO-WIDTH PLANNING UNDER THE T METRIC

Tier		1	2	3	4
1-width	W^* (μm)	0.11	0.55	1.40	3.82
	T_{avg} (ps)	69.2	134.8	160.5	166.8
	ΔT_{max}	3.6%	2.6%	3.7%	6.7 %
2-width	W_1^* (μm)	0.10	0.33	0.84	2.12
	W_2^* (μm)	0.15	0.66	1.68	4.66
	T_{avg} (ps)	69.2	134.0	159.2	163.9
	W_{avg} (μm)	0.11	0.53	1.34	3.68
	ΔT_{max}	2.4 %	1.8%	2.6%	4.4 %
OWS	T_{avg} (ps)	69.2	132.2	156.2	158.9

The experimental setting is as follows. For wire length distribution and layer assignment at each layer (tier), we assume that the maximum wire length ($l_{\max}$) in Tier-1 is 10 000 × feature size and $l_{\max}$ in the top tier is L_{edge}, i.e., the chip dimension [30]. The $l_{\max}$ in the intermediate tiers is then determined by a geometric sequence such that for any tier i, $l_{\max}(i+1)/l_{\max}(i) = l_{\max}(i)/l_{\max}(i-1)$. For example, in 0.10-$\mu$m technology, $l_{\max}(1) = 1000$ μm, $l_{\max}(4) = 22\,800$ μm. Since $22.8^{1/4} = 2.84$, we have $l_{\max}(2) = 2840$ μm and $l_{\max}(3) = 8040$ μm. The minimum wire length for tier i is the maximum length for tier $i-1$, i.e., $l_{\min}(i) = l_{\max}(i-1)$. Table II shows the maximum wire length in each tier for NTRS'97 technologies. We assume a uniform weighting function $\lambda(l) = 1$. We also take a representative driver for each metal tier for our wire width planning. The drivers for Tier-1 through Tier-4 are 10 ×, 40 ×, 100 ×, and 250 × of the minimum gate in the given technology, respectively.

To numerically compute the integral of the objective function (7), we use wire length incremental step to be $\delta l = 10$ μm. On a Sun UltraSPARC 10 machine, less than 0.1 second CPU is needed for our wire width planning (either one-width design or two-width design) for any metal layer.

1) Under Fixed c_f: We first show the effectiveness of our wire width planning under fixed effective-fringing capacitance coefficient, which essentially assumes a fixed spacing between a net and its neighboring wires. Table III shows the optimal 1-width design and 2-width design under the delay-only metric T and the comparison between 1-WS and 2-WS (using selected widths) with OWS for different tiers in the 0.10 μm technology. The OWS results are listed at the last row of the table. The 1-width design (W^*) selects minimum width for Tier-1 and sizes up in a factor of 2.5 to 5 for upper tiers, with 3.82 μm in Tier-4. The average delay (T_{avg}) for each tier is computed for all wire length distribution in each tier. It ranges from 69 to 167 ps, less than 5% larger than that obtained by OWS. The maximum delay difference compared to OWS at each tier ($\Delta T_{\max}$) is only up to 6.7% (for Tier-4). According to Theorem 1, it can be used as a maximum error bound under *any weighting function* $\lambda(l)$.

TABLE IV
TWO-WIDTH PLANNING UNDER DIFFERENT METRICS

Metrics	T_{avg}	W_1^*	W_2^*	W_{avg}	ΔT_{avg}	ΔT_{max}
T	164	2.12	4.66	3.68	2.84%	4.36%
AT^6	168	1.38	2.76	2.51	5.07%	9.61%
AT^5	171	1.21	2.42	2.25	6.63%	12.3%
AT^4	176	1.00	2.00	1.90	9.78%	17.3%
AT^3	190	0.77	1.46	1.42	17.5%	29.2%
AT^2	239	0.46	0.78	0.77	46.6%	71.1%

The 2-width design optimally selects two wire widths W_1^* and W_2^*. In general, $W_1^* < W^* < W_2^*$ for each metal layer. The optimal width ratio $\alpha^* = W_2^*/W_1^*$ for Tier-1 to Tier-4 are 1.5, 2, 2, and 2.2, respectively.[3] As expected, the two-width design obtains even better approximation to OWS than the one-width design, with a few percent delay and area reduction. Note that in the table, we show the average wire width for *all* wire lengths at each tier. For individual wires, 1-width design may have to use a much larger average wire width, especially for shorter wires at each tier. For an example of wire length 8.04 mm (shortest wire in Tier-4), 1-width design still has to use 3.82 μm, while a more flexible 2-width design has an average wire width of only 2.99 μm, which is a 22% reduction of wiring area.

As seen in Section IV, a performance-only wire planning metric may lead to excessive wire area. Our wire width planning methodology, however, can easily explore the tradeoff between performance and wiring area (for routability consideration). Table IV shows the results of using several optimization metrics in the form of AT^j and compares the average delay T_{avg}, W_1^*, W_2^* ($= \alpha^* W_1^*$), W_{avg}, ΔT_{avg}, and $\Delta T_{\max}$ (the average and maximum error compared to OWS) for Tier-4 of the 0.10-μm technology. The area-aware metrics of AT^6, AT^5, and AT^4 all have within 7% average delay difference compared to the performance-only metric T, but reduce area (i.e., W_{avg}) by 32%, 39%, and 48%, respectively.

2) Under Variable c_f: When we assume fixed pitch-spacing and consider variable coupling capacitance during wire sizing optimization, the 2-width design shows much more flexibility than the 1-width design. Table V shows the comparison of using the optimal 1-width and 2-width designs under metrics AT^4 versus using many different wire widths (denoted as m-width in Table V), where 100 discrete widths are used by running ISS algorithm [19]. We compare the average delay (T_{avg} in nanoseconds), the maximum delay difference compared to ISS ($\Delta T_{\max}$ in percentage), and the average wire width (W_{avg} in μm) of using 1-width design, 2-width design, and many discrete wire widths (m-width) by running the ISS algorithm. Tier-4 of 0.10-μm technology with different pitch-spacings (pitch-sp) is used for the experiments. For pitch-spacing of 2.0 μm, the 1-width design has average delay about 14% and 20% larger than those from 2-width and m-width. Moreover, it has an average wire width (thus area) about 1.83 × and 1.92 × those from the 2-width and m-width results. The 2-width design, however, has close to optimal delay compared to the solution obtained from many widths (m-width) by running the optimal ISS algorithm (just 3%–6% larger), with only slightly bigger area

[3] Again, we can simply set a fixed ratio of 2, with almost no difference from using α^*.

TABLE V
WIRE WIDTH PLANNING UNDER VARIABLE c_f

Scheme	pitch-sp=2.0 μm			pitch-sp=2.9 μm			pitch-sp=3.8 μm		
	T_{avg}	ΔT_{max}	W_{avg}	T_{avg}	ΔT_{max}	W_{avg}	T_{avg}	ΔT_{max}	W_{avg}
1-width	0.245	28.2%	1.98	0.177	15.7%	1.83	0.143	5.9%	1.63
2-width	0.215	7.0%	1.08	0.167	5.9%	1.23	0.140	3.9%	1.41
m-width	0.204	-	1.03	0.159	-	1.19	0.136	-	1.38

TABLE VI
SAMPLE WIRE WIDTH PLANNING

Tech. (μm)		0.25	0.18	0.13	0.10	0.07
Tier-1	W^*	0.25	0.18	0.13	0.10	0.07
Tier-2	W_1^*	0.25	0.18	0.13	0.10	0.08
	W_2^*	0.50	0.36	0.26	0.20	0.16
Tier-3	W_1^*	0.65	0.47	0.24	0.22	0.23
	W_2^*	1.30	0.94	0.48	0.44	0.46
Tier-4	W_1^*	-	-	0.98	1.00	1.06
	W_2^*	-	-	1.96	2.00	2.12

(less than 5%) than that of the m-width. Note that when the pitch-spacing becomes larger, the difference between 1-width, 2-width, and m-width results will get smaller.

In Table V, we also list the maximum delay difference ($\Delta T_{\max}$) of 1-width and 2-width designs compared to m-width. This is an important metric which provides the maximum error bound under *any weighting function* $\lambda(l)$ in our objective function. Note that although we derive the optimal 1-width or 2-width design using the uniform weighting functions $\lambda(l) = 1$, our maximum delay difference $\Delta T_{\max}$ using 2-width design is only 3.9%–7%. Therefore, from Theorem 1, this 2-width design differs from many-width optimal solution by 3.9–7% for *any weighting function* $\lambda(l)$.

C. Sample Wire Width Planning for Technology Generations in NTRS

We have further performed wire width planning for all major technology generations listed in NTRS'97 from 0.25 to 0.07 μm. Our recommendation is based on the optimal 2-width design with the area-efficient performance optimization metric AT^4. The results are shown in Table VI. It suggests the minimum width for local interconnects in Tier-1. For Tier-2 to Tier-4, there are two different predetermined wire widths with width ratio of 2:1.[4] Therefore, we have a wiring hierarchy on different metal layers such that Tier-2 is about 1–2 times wider than Tier-1, Tier-3 is about 2–3 times wider than Tier-2, and Tier-4 (if available) is about 4–5 times wider than Tier-3. Such a wiring hierarchy can effectively minimize the interconnect delays for all local, semiglobal, and global interconnects while ensuring high routing density and simplified routing solutions.

VI. CONCLUSION AND DISCUSSIONS

In this paper, we present two simplified wire sizing schemes (1-WS and 2-WS) for VLSI interconnect optimization. Our sensitivity study on wire sizing optimization reveals an interesting delay-area tradeoff and suggests that there exists a small set of "globally" optimal wire widths for a range of interconnects. We develop a general and efficient wire width planning methodology to obtain them. We demonstrate that using two predetermined wire widths for each metal layer, one can achieve near-optimal performance compared to that from running complex wire sizing/spacing algorithms with many possible wire widths.

With the usage of these "predetermined" small number of wire widths for each metal layer from our wire width planning methodology, many interconnect-centric problems become much easier, such as interconnect performance estimation, interconnect planning (routing resource allocation at high levels and so on), and performance-driven global and detailed routing. In particular, if only one or two fixed widths are used for every metal layer, a full-blown gridless router may be unnecessary or can be much simplified. Note that a straightforward method to realize a gridless detailed router is to use a grid-based router with very fine grids.[5] The grid size is determined by the largest common divisor of all the wire widths (assuming the grid for wire spacing is the same) and it will be the manufacturing grid in the extreme case. It is obvious that using one or two fixed widths (with integer ratio of 2:1 as shown in this paper), the grid size is just the planned wire width itself for one-width design or the smaller one for two-width design. It is much larger than the manufacturing grid and sometimes even larger than the minimum wire width allowed at each metal layer. Thus, the routing grid is much smaller and problem complexity is much reduced. This, in turn, will significantly simplify several other problems, including RC extraction, detailed routing, and layout verification.

In this paper, fixed-size drivers and loads are used to derive one-width and two-width designs. That is, we assume that all drivers are of the same size in each layer, and so are the loads. Our wire width planning methodology, however, can be extended to handle more general cases with a range of drivers and loads using similar numerical integration. Depending on input parameter ranges, a few more widths may be needed to achieve near-optimal results. We can also extend the method to perform interconnect architectural planning for other parameters, such as wire spacing or metal or dielectric thickness and for other design metrics such as noise and power optimizations.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Brayton and his research group at University of California, Berkeley, for providing the Strawman technology. We also thank the anonymous reviewers for their constructive comments.

[4] From Fig. 4, interconnect performance remains almost the same for a fixed width ratio of 2:1, versus the optimal ratios (ranging from 1.5 to 3) for different wire lengths. Using a fixed integer ratio, however, can significantly simplify the routing architecture. Note that one may choose to use another integer ratio 3:1 and still have near-optimal performance.

[5] The reader may refer to [35] for more detailed discussion on gridless routing.

REFERENCES

[1] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Readington, MA: Addison-Wesley, 1990.

[2] J. Cong, L. He, C.-K. Koh, and P. H. Madden, "Performance optimization of VLSI interconnect layout," *Integration VLSI J.*, vol. 21, pp. 1–94, 1996.

[3] J. Cong, L. He, K.-Y. Khoo, C.-K. Koh, and D. Z. Pan, "Interconnect design for deep submicron IC's," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1997, pp. 478–485.

[4] J. Cong, "An interconnect-centric design flow for nanometer technologies," *Proc. IEEE*, vol. 89, pp. 505–528, Apr. 2001.

[5] J. Cong, K. S. Leung, and D. Zhou, "Performance-driven interconnect design based on distributed RC delay model," in *Proc. Design Automation Conf.*, June 1993, pp. 606–611.

[6] J. Cong and K. S. Leung, "Optimal wiresizing under the distributed elmore delay model," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1993, pp. 634–639.

[7] J. P. Fishburn and C. A. Schevon, "Shaping a distributed-RC line to minimize elmore delay," *IEEE Trans. Circuits Syst. I: Fund. Theory Applicat.*, vol. 42, no. 12, pp. 1020–1022, Dec. 1995.

[8] C. P. Chen, Y. P. Chen, and D. F. Wong, "Optimal wire-sizing formula under the elmore delay model," in *Proc. Design Automation Conf.*, June 1996, pp. 487–490.

[9] C.-P. Chen and D. F. Wong, "Optimal wire sizing function with fringing capacitance consideration," in *Proc. Design Automation Conf.*, June 1997, pp. 604–607.

[10] J. P. Fishburn, "Shaping a VLSI wire to minimize elmore delay," in *Proc. European Design and Test Conf.*, Mar. 1997.

[11] Y. Gao and D. F. Wong, "Optimal shape function for a bi-directional wire under elmore delay model," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1997, pp. 622–627.

[12] J. Cong and L. He, "Optimal wiresizing for interconnects with multiple sources," *ACM Trans. Design Automation Electron. Syst.*, vol. 1, no. 4, pp. 478–511, Oct. 1996.

[13] S. S. Sapatnekar, "RC interconnect optimization under the Elmore delay model," in *Proc. Design Automation Conf.*, June 1994, pp. 387–391.

[14] C. P. Chen, Y. W. Chang, and D. F. Wong, "Fast performance-driven optimization for buffered clock trees based on lagrangian relaxation," in *Proc. Design Automation Conf.*, June 1996, pp. 405–408.

[15] N. Menezes, S. Pullela, F. Dartu, and L. T. Pillage, "RC interconnect synthesis—A moment fitting approach," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1994, pp. 418–425.

[16] L. Pileggi, "Coping with RC(L) interconnect design headaches," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1995, pp. 246–253.

[17] J. Cong, L. He, C.-K. Koh, and Z. Pan, "Global interconnect sizing and spacing with consideration of coupling capacitance," in *Proc. Int. Conf. Computer Aided Design*, Nov. 1997, pp. 628–633.

[18] J. Cong and L. He, "Theory and algorithm of local-refinement based optimization with application to device and interconnect sizing," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 18, pp. 406–420, Apr. 1999.

[19] J. Cong, L. He, C.-K. Koh, and D. Z. Pan, "Interconnect sizing and spacing with consideration of coupling capacitance," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 20, pp. 1164–1169, Sept. 2001.

[20] J. Cong and D. Z. Pan, "Interconnect estimation and planning for deep submicron designs," in *Proc. Design Automation Conf.*, June 1999, pp. 507–510.

[21] J. Cong and Z. Pan, "Wire width planning and performance optimization for VLSI interconnects," U.S. Patent pending.

[22] W. C. Elmore, "The transient response of damped linear networks with particular regard to wide-band amplifiers," *J. Applied Phys.*, vol. 19, no. 1, pp. 55–63, Jan. 1948.

[23] J. Rubinstein, P. Penfield Jr., and M. A. Horowitz, "Signal delay in RC tree networks," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. CAD-2, pp. 202–211, July 1983.

[24] L. Pileggi, "Timing metrics for physical design of deep submicron technologies," in *Proc. Int. Symp. Physical Design*, Apr. 1998, pp. 28–33.

[25] "National technology roadmap for semiconductors," Semiconductor Industry Association, 1997.

[26] G. A. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, pp. 20–36, Jan. 1995.

[27] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire length distribution for gigascale integration (gsi)," in *Proc. IEEE Custom Integrated Circuits Conf.*, May 1996, pp. 140–145.

[28] J. A. Davis and J. D. Meindl, "Is interconnect the weak link?," *IEEE Circuits Devices Mag.*, vol. 14, no. 2, pp. 30–36, 1998.

[29] R. H. J. M. Otten and R. K. Brayton, "Planning for performance," in *Proc. Design Automation Conf.*, June 1998, pp. 122–127.

[30] P. D. Fisher and R. Nesbitt, "The test of time clock-cycle estimation and test challenges for future microprocessors," *IEEE Circuits Devices Mag.*, vol. 14, pp. 37–44, Mar. 1998.

[31] J. Cong, L. He, A. B. Kahng, D. Noice, N. Shirali, and S. H.-C. Yen, "Analysis and justification of a simple, practical 2 1/2-d capacitance extraction methodology," in *Proc. ACM/IEEE Design Automation Conf.*, June 1997, pp. 40.1.1–40.1.6.

[32] R. Kay and L. T. Pileggi, "EWA: Efficient wiring-sizing algorithm for signal nets and clock nets," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 17, no. 1, pp. 40–49, Jan. 1998.

[33] C.-K. Koh, "VLSI interconnect layout optimization," Ph.D., Univ. California, Los Angeles, 1998.

[34] J. Cong and Z. (David) Pan, "Interconnect performance estimation models for design planning," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 20, pp. 739–752, June 2001.

[35] J. Cong, J. Fang, and K. Y. Khoo, "Dune—A multilayer gridless routing system," *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.*, vol. 20, pp. 633–647, May 2001.

Jason Cong (S'88–M'90–SM'96–F'00) received the B.S. degree in computer science from Peking University, in 1985, and the M.S. and Ph. D. degrees in computer science from the University of Illinois, Urbana-Champaign, in 1987 and 1990, respectively.

Currently, he is a Professor and Co-Director of the VLSI CAD Laboratory in the Computer Science Department of University of California, Los Angeles. His research interests include layout synthesis and logic synthesis for high-performance low-power VLSI circuits, design and optimization of high-speed VLSI interconnects, FPGA synthesis, and reconfigurable computing. He has published over 150 research papers in those areas.

Dr. Cong received the Best Graduate Award from the Peking University in 1985 and the Ross J. Martin Award for Excellence in Research from the University of Illinois, Urbana-Champaign, in 1989. He received the NSF Research Initiation Award and NSF Young Investigator Award in 1991 and 1993, respectively. He received the Northrop Outstanding Junior Faculty Research Award from UCLA in 1993 and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN Best Paper Award in 1995. He received the ACM Recognition of Service Award in 1997, the ACM SIGDA Meritorious Service Award in 1998, an SRC Inventor Recognition Award in 2000 and the SRC Technical Excellence Award in 2001. He has been an appointed Guest Professor of Peking University since 2000. He served as the General Chair of the 1993 ACM/SIGDA Physical Design Workshop, the Program Chair and General Chair of the 1997 and 1998 International Symposium on FPGA's, respectively, and on program committees of many VLSI CAD conferences, including DAC, ICCAD, and ISCAS. He is an Associate Editor of *ACM Transactions on Design Automation of Electronic Systems* and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS.

Zhigang (David) Pan (S'97–M'00) received the B.S. degree in geophysics from Peking University, in 1992, the M.S. degree in atmospheric sciences, the M.S. degree in computer science, and the Ph.D. degree in computer science, all from University of California at Los Angeles (UCLA) in 1994, 1998, and 2000, respectively.

He was with Magma Design Automation, Inc., during the summer of 1999 and with the IBM T.J. Watson Research Center during the summer of 2000. He is currently a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include VLSI interconnect modeling, synthesis, planning, and their interaction with physical design and logic synthesis, as well as low power designs.

Dr. Pan received the Best Paper in Session Award from SRC Techcon 1998, IBM Research Fellowship in 1999, Dimitris Chorafas Foundation Award in 2000, SRC Inventor Recognition Award in 2000, and Outstanding Ph.D. Award from the UCLA Computer Science Department in 2001.