Optimal Resizing of Bus Wires in Layout Migration

Shay Michaely[1], Shmuel Wimer[2], and Avinoam Kolodny[1]

1. Electrical Engineering Department, Technion, Haifa, Israel Email: kolodny@ee.technion.ac.il
2. Intel Corp., Mobile Processor Group, Haifa, Israel, Email: shmuel.wimer@intel.com

Abstract – The effect of wire delay on circuit timing typically increases when an existing layout is migrated to a new generation of processing technology, because wire resistance and cross capacitances become more important with scaling. In this paper, timing optimization of signal busses is performed by resizing and spacing individual bus wires, while the total area of the whole bus structure is regarded as a fixed constraint. Properties of optimal bus layouts are proven, and an iterative algorithm to find the optimal wire widths and spaces is presented. Examples of solutions are shown. Guidelines for design are derived from these results.

I. Introduction

Interconnect delays have become dominant in CMOS VLSI digital systems as a result of technology scaling[1][2]. In recent generations, wire resistance and cross-capacitance between adjacent wires have become increasingly important in their effect on signal delay. For a given metal layer, wire resistance and cross-capacitance depend on wire width and inter-wire spacing, respectively. In process migration of existing mask layouts, allocation of wire widths and spaces for bus structures under a total area constraint is an important problem. The nature of the problem allows tradeoff between the resistance of a wire and the coupling capacitances to adjacent wires. Wire resistance affects only the delay of the signal carried by the wire, while coupling capacitances affect the delays of both the wire and its neighbors. For multiple nets, the optimal solution involves simultaneous tradeoffs among all wires sharing a given common area.

The wire sizing problem has been addressed in [3] for a single wire and for a single-net tree structure. The problem of sizing and spacing multiple nets with consideration of coupling capacitance in global interconnect has been addressed in [4], considering general tree structures for nets with fixed terminals, without a total area constraint. The authors modeled coupling between nets by converting cross-capacitance into an effective fringe capacitance, which resulted in a decoupled delay model for each net. The routing tree for each net was sized independently using an algorithm based on dynamic programming [5].

Coupling capacitance has been considered more explicitly in the context of physical design algorithms for minimizing crosstalk noise [6][7][8] or dynamic power [9]. The authors of [10] derived layout rules and presented a simultaneous multiple-net spacing algorithm for area minimization in general layouts under a noise-constraint. The strategy of allocating width and spacing to maximize performance in bus structures was proposed by [6] without formal analysis and solution.

This paper addresses the problem of simultaneously assigning widths and spaces to n parallel wires, representing a bus or several interleaved busses, as illustrated in Figure 1. This geometry is commonly used in practice, and its simplicity enables straightforward mathematical analysis. With given drivers, load capacitances and timing requirements for the individual signals, wire widths and spaces are allocated to maximize circuit speed. The total sum of widths and spaces is a given constraint. Signal delays are expressed by an Elmore model using first-order approximations for capacitances. Cross-coupling capacitances between wires are not multiplied by a "Miller factor"[11]. Hence, nominal delays are considered without crosstalk-induced delay uncertainty. The paper focuses on delay of the critical signal as a goal function in optimization. The goal function is shown to be convex. Properties of the optimum solution are proven, leading to an iterative algorithm to find the optimal wire sizes and spaces.

The rest of this paper is organized as follows: Section II presents the delay model and problem definition, section III analyzes properties of the problem, which provide a basis for constructing an iterative optimization algorithm in section IV. Examples of applying the algorithms are shown in section V, and section VI concludes the paper.

II. Delay model and problem definition

Consider a bus of n signal nets $\sigma_1,...,\sigma_n$, residing between two side walls (wires at fixed locations, connected to Vcc or Vss) as shown in Figure 1. S_{i-1} and S_i, respectively, denote spacings to the right and left neighbors of wire W_i. The length of all the wires is L.

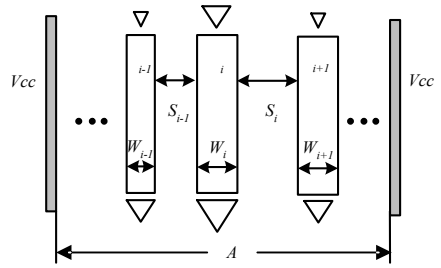

Figure 1. Structure of the bus wires

The delay of signal σ_i can be calculated from the π-model equivalent circuit shown in Figure 2, where R_{d_i} is the effective output resistance of the driver, R_{w_i} is the wire resistance, C_{w_i} is the wire area and fringe capacitance, $C_{c_{i-1}}$ and C_{c_i} are coupling capacitances to the right and left neighboring signals, and C_{l_i} is the capacitive load presented by the receiver's input. Using technology parameters these can be expressed as $R_{w_i} = R_s W_i / L$, $C_{w_i} = C_a L W_i + C_f L$ and $C_{c_i} = k_c L / S_i$, where C_a is area capacitance coefficient, C_f is fringe capacitance coefficient, k_c is a line-to-line coupling coefficient, and R_s is the metal sheet resistance.

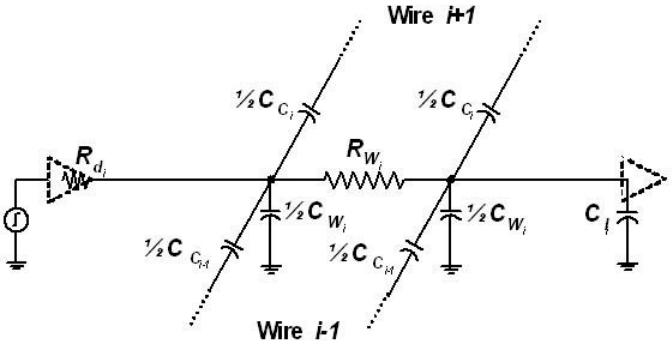

Figure 2. Equivalent circuit for calculating the i^{th} signal delay

Under Elmore delay model, the delay Δ_i of signal σ_i from driver's input to receiver's input is given as follows:

$$\Delta_i = \overbrace{R_{d_i} C_a L}^{a} W_i + \overbrace{\left(D_i + R_{d_i} C_f L + R_{d_i} C_{l_i} + \frac{1}{2} C_a R_s L^2 \right)}^{b} + \overbrace{\left(\frac{1}{2} R_s C_f L^2 + R_s L C_{l_i} \right)}^{c} \frac{1}{W_i} + \overbrace{R_{d_i} k_c L}^{d} \left(\frac{1}{S_{i-1}} + \frac{1}{S_i} \right) + \overbrace{\frac{1}{2} R_s k_c L^2}^{e} \frac{1}{W_i} \left(\frac{1}{S_{i-1}} + \frac{1}{S_i} \right) \quad (1)$$

The coefficients of wire width and spaces in (1) will be marked as *a,b,c,d,e*. The delay expression can be rearranged as

$$\Delta_i \left(\bar{W}, \bar{S} \right) = a_i W_i + b_i + \frac{c_i}{W_i} + \left(d_i + \frac{e}{W_i} \right) \times \left(\frac{1}{S_{i-1}} + \frac{1}{S_i} \right) \quad (2)$$

Note that in (2) the coefficient e is not indexed since it encapsulates only technology parameters, which are common to all nets. The other coefficients are indexed since they include parameters related to the signal's driver and receiver. The sum of wire widths and spaces between the fixed left and right side walls is given in the following constraint, which represents the available area for laying out the signal bus.

$$g\left(\bar{W}, \bar{S} \right) = \sum_{j=1}^{n} W_i + \sum_{j=0}^{n} S_i = A \quad (3)$$

Another set of constraints on wire sizing is geometric design rules, which are imposed by the manufacturing technology. In modern processes of 90 nanometers and below, the width and the space of wires are bounded in some range as follows:

$$S' \le S_i \le S'', 0 \le i \le n \quad (4)$$

$$W' \le W_i \le W'', 1 \le i \le n \quad (5)$$

We are looking for width and space allocation yielding "best timing". The definition of "best timing" depends on the design scenario. In the following we'll define two commonly used timing objectives.

First objective aims at minimizing the delay of the slowest signal in the bus. It is used in design of a block whose environmental timing constraints are not known yet. There, only the delays of signals are of interest. The corresponding objective function is the following:

$$f_1\left(\bar{W}, \bar{S} \right) = \max_{1 \le i \le n} \left\{ \Delta_i \left(\bar{W}, \bar{S} \right) \right\} \quad (6)$$

A second objective is used when environmental timing requirements are known. When clock frequency is a primary consideration, main focus is meeting timing goal by each signal. Frequency is then dictated by the signal that doesn't meet its timing target to the largest extent. Formally, let T_i denote the delay target for signal σ_i, the goal is to maximize the worst negative slack among all signals. For the sake of dealing with convex functions only, the inner term in (7) is taken with a minus sign, and the goal is to minimize f_2.

$$f_2\left(\bar{W}, \bar{S} \right) = \max_{1 \le i \le n} \left\{ -\min\left\{ 0, T_i - \Delta_i \left(\bar{W}, \bar{S} \right) \right\} \right\} \quad (7)$$

III. Properties of delays and slacks in min max delay problems

Optimization problems whose objective and constraints are convex functions have nice properties, such as a unique, global minimum. There are many techniques to solve such problems, so we may benefit from proving that the problems of interest are convex.

Proposition 1: The problems of minimizing f_1 and f_2, subject to area and design rule constraints are both convex, therefore possessing a unique global minimum.

Proof: The function $\Delta_i \left(\bar{W}, \bar{S} \right)$ in (2) is a sum of terms depending on the variables W_i, S_i and S_{i-1}. In order to prove the convexity it is sufficient to see the convexity for each term, since a sum of convex functions is convex too. Convexity exists if the second order derivatives are non-negative. It is not difficult to see that the terms: $\partial W / \partial^2 W, \partial(1/W) / \partial^2 W, \partial(1/WS) / \partial^2 W, \partial(1/WS) / \partial W \partial S$ and

$\partial(1/WS)/\partial^2 S$ are all non negative, thus implying the convexity of (2).

Objective function $f_1(\overline{W}, \overline{S})$ is convex since maximum of convex functions is convex too. The objective $f_2(\overline{W}, \overline{S})$ is convex too. The inner negative min term is convex as it is a maximum of two convex functions. The outer max is also convex as claimed before.

Finally, the area and design rules constraints given in (3), (4) and (5) are all linear equalities or inequalities. Altogether they define a convex feasible region on which the above objective functions are defined. ●

In the subsequent discussion we'll expose more useful properties of the underlying optimization problem. Such properties will suggest efficient solutions.

Lemma 1: Let us ignore design rules (4) and (5), namely, any width and space of wires are allowed. Then, in the solution of minimizing the maximal delay in (6) subject to the area constraint (3), all the delays are equal.

Proof: Assume in contrary that the above assertion doesn't hold. Namely, in the optimal solution, there exists a wire i whose associated delay is maximal.

There exist therefore signals σ_{i-1}, σ_i and σ_{i+1}, such that their corresponding delays, Δ_{i-1}, Δ_i and Δ_{i+1}, respectively, satisfy $\Delta_{i-1} < \Delta_i$ and $\Delta_{i+1} < \Delta_i$, and $\Delta_j < \Delta_i$, $\forall 1 \le j \le n$.

Therefore, we may narrow wires $i-1$ and $i+1$ slightly, thus increasing their delay, say by a magnitude that doesn't exceed the minimum among $0.5(\Delta_i - \Delta_{i-1})$ and $0.5(\Delta_i - \Delta_{i+1})$. This further reduces Δ_i since the width of wire i didn't change, but its spacing from neighbors was increased. But Δ_i was the maximal delay in the optimal solution. Thus a contradiction follows. ●

In a similar manner, the worst slack minimization problem defined in (7) satisfies the following:

Lemma 2: Let us ignore design rules (4) and (5). Then, in the solution of minimizing f_2 subject to the area constraint (3), all the slacks are equal.

Proof: By definition of (7) all terms are either positive or zero. If all terms are zero we are done. If all terms are positive, same arguments as in lemma 1 show that all slacks must be the same, as otherwise we could decrease the maximal slack by narrowing slightly its neighbor wires. Let us show that the case where some terms are zero while some are positive is impossible.

Let $D_{\min}$ be the smallest positive term in (7). We assume that there exists at least one signal σ_i whose associated term in (7) is zero. We may narrow its wire slightly say by δW, without changing its left and right spacing, such that the value of its term in (7) doesn't exceed $0.5 \times D_{\min}$. We may now increase the spacing between all the other wires by $\delta W/(n-1)$. This will reduce slightly all the terms, in particular the one that was the maximal in the optimal solution. This is a contradiction to the optimality of the solution, which concludes the proof. ●

Note that the above lemmas impose necessary but not sufficient conditions on optimal solutions. It is not true to state that a solution whose delays (or slacks) are all equal is optimal.

The convex objective function of minimizing the maximal delay and slack, possess a unique, global minimum and are continuous and piecewise differentiable functions. If they were differentiable everywhere we could state that a sufficient condition for minimum is that all first order partial derivatives are zero. The following lemma is analogous for the piecewise differentiable case.

Lemma 3: Let all the delays (slack, respectively) in (6) ((7), respectively) be equal. Then, (6) ((7), respectively) is minimal if and only if no single wire width can be decreased such that the delay (slack, respectively) of the associated signal is decreased.

Proof: Recall that the objective functions (6) and (7) are convex, thus possessing a global minimum (no local minima). Therefore, among all the solutions having equal delays (slacks) for all signals, the characteristic of the optimal one is that no further reduction of individual signal's delay is possible without increasing the delays of others. Considering delay reduction of an individual signal, only its wire narrowing is in order, as this will not increase the delay of its neighbors. Widening necessarily decreases the spacing to at least one of its neighbor signals, thus increasing neighbor's delay. ●

IV. Iterative algorithm

Lemmas 1 through 3, and the convexity properties discussed earlier suggest an iterative algorithm to obtain a minimum of maximal delay (It can be easily adapted to maximize the most critical slack). The algorithm works in two main phases which repeat themselves until convergence.

First phase equates the delay of all signals by iterations. It picks the signal whose delay is currently maximal. It then reduces the delay by equating it with its two neighbors, a technique used in the proof of lemma 1. This is repeated until all delays are equal.

Second phase checks for existence of the sufficient condition posted in lemma 3. If the condition is satisfied, the algorithm terminates at optimum. Otherwise, a delay reduction of some signal is possible by narrowing it without increasing the delay of the others as stated in the lemma. This reduction is then performed by the algorithm, and phase one is re-invoked.

The algorithm for maximal delay minimization is outlined below. Some heuristics aiming at speeding up convergence are included.

MinMaxDelay ( )
set initial solution;
do {
  while (*not all signal delays are equal* ***) {*** *// first phase*

1. Pick signal with maximal delay, in case of tie pick the one whose average delay with neighbors is the smallest;

2. Average delay of the selected signal with its two neighbors; }

if (** no single wire delay can be reduced by narrowing**)
terminate; // optimum reached

***else** // second phase*

1. find the signal whose delay can be minimized by narrowing its wire

2. narrow the wire found in 1. };

Convergence of the above algorithm can be proven as follows: The inner loop of ***while*** (first phase) iterates over signals and reduces the maximal delay. Therefore, the maximal delay, which is positive, is monotonically decreasing. Hence it must reach a limit. In the outer ***do*** loop the delay (equal for all signals) is also monotonically decreasing, thus it must reach a limit as well.

V. Optimization Examples

In this section we present 2 examples of applying the algorithm on bus layouts in 90nm CMOS technology.

Example 1: 10 bus wires are driven by identical drivers. All receivers are of the same size. The total width of the structure is 6.72 μm, and the length is 500 μm. Initially, all wire widths and spaces are 0.32 μm Figure 3(a) shows the cross section of the initial bus setting, in which the worst delay is 117.55 psec. 0Figure 3(b) shows the resultant structure after applying the algorithm.

There, the wires got narrower while the spaces increased, and the corresponding delay of all signals was reduced to 101.64 psec.

Example 2 illustrates the case of non-homogeneous bus, where the fifth wire from the left is driven by a weaker driver. Initial wire widths and spaces are illustrated in Figure 4, and the signal delays are depicted on each wire in the figure. Notice that the delay of the fifth wire is much worse than the others. After applying the algorithm, all the delays were equalized to the value depicted on the fifth wire.

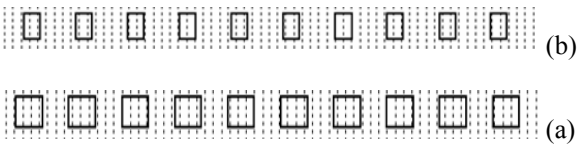

Figure 3. Uniform bus wire sizing:

(a)Initial Setting (b) Optimal Setting

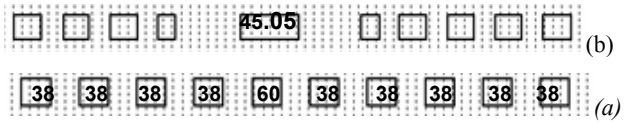

Figure 4. Non-uniform bus wire sizing:

(a) Initial Setting (b) Optimal Setting

VI. Conclusion

This paper studied the optimal allocation of widths and spaces to wires of n-signal bus in the context of layout migration, where area is a major design constraint. A unified combinatorial approach to deal with various timing optimization objectives was presented. Properties of the optimal solution were presented. In particular, all signal delays (or slacks) must necessarily be equal in the optimal solution. A sufficient condition for optimality states that no single wire width can be further reduced in order to obtain delay or slack improvement. A practical algorithm based on the necessary and sufficient conditions has been presented, along with examples.

VII. References

[1] R. Ho, K. Mai and M. Horowitz, "The future of wires", *Proceedings of the IEEE*, Vol. 89, no. 4, Apr. 2001.

[2] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron", *in Proc. ICCAD*, pp. 203-211, 1998.

[3] J. Cong and K. S. Leung, "Optimal Wiresizing Under Elmore Delay Model". *IEEE Trans. on Computer-Aided Design*, vol. 14, no. 3, pp. 321-336, Mar. 1995.

[4] J. Cong, L. He, C. K. Koh and Z. Pan, "Interconnect Sizing and Spacing with Consideration of Coupling Capacitance", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 20, no. 9, pp.1164-1169, September 2001.

[5] J. Lillis, C. K. Cheng, and T. T. Y. Lin, "Simultaneous routing and buffer insertion for high performance interconnect," in *Proc. 6th Great Lakes Symp. VLSI*, Mar. 1996, pp. 148–153.

[6] A. B. Kahng, S. Muddu, E. Sarto and R. Sharma, "Interconnect Tuning Strategies for High-Performance ICs", *Proc. Design, Automation and Testing in Europe*, Paris, February 1998.

[7] M. Becer, D. Blaauw, V. Zolotov, R. Panda, I. Hajj, "Analysis of Noise Avoidance Techniques in Deep-Submicron Interconnects Using a Complete Crosstalk Noise Model", *IEEE/ACM Design Automation and Test in Europe Conference (DATE)*, March 2002, pp. 456-463.

[8] D. A. Kirkpatrick and A. L. Sangiovanni-Vincentelli. "Techniques for crosstalk avoidance in the physical design of high performance digital systems", *In Proc. IEEE/ACM Int. Conf. Computer Aided Design*, pp. 616--619, Nov. 1994.

[9] R. Arunachalam, E. Acar, and S. R. Nassif, "Optimal shielding/spacing metrics for low power design", *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, pp. 167-172, 2003.

[10] J. Cong, D. Z. Pan and P. V. Srinivas, "Improved Crosstalk Modeling for Noise Constrained interconnect Optimization," *ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems,* pp. 14-20, Austin, December 2000.

[11] P. Chen, D. A. Kirkpatrick and K. Keutzer, "Miller Factor for Gate-Level Coupling Delay Calculation", *in Proc. ICCAD*, pp.68-74, 2000.