

An Interconnect-Centric Design Flow for Nanometer Technologies

Jason Cong

Department of Computer Science
University of California, Los Angeles, CA 90095
Email: cong@cs.ucla.edu *

Abstract

As the integrated circuits (ICs) are scaled into nanometer dimensions and operate in giga-hertz frequencies, interconnect design and optimization have become critical in determining system performance and reliability. This paper presents the ongoing research effort at UCLA to develop an interconnect-centric design flow, including interconnect planning, interconnect synthesis, and interconnect layout, which allows interconnect design and optimization to be properly considered at every level of the design process. Efficient interconnect performance estimation models and tools at various levels are also being developed to support such an interconnect-centric design flow.

1 Introduction

The conventional IC design flow is device/logic-centric, which places much emphasis on design and optimization of device and logic. The interconnection among various circuit components was done by either layout designers or automatic Place-&-Route tools very late in the overall design process. As the IC technology moves to nanometer device dimensions and gigahertz clock frequencies, interconnects play the dominating role in determining the performance, power, reliability, and cost of the system. Therefore, it is necessary to explore an interconnect-centric design flow which considers interconnect estimation and planning, optimal interconnect synthesis, and efficient interconnect layout implementation at each level of the design process. Early interconnect estimation and planning are critical to assure the proper coupling between synthesis and layout, and enable the design convergence. Optimal interconnect synthesis is key to delivering the best possible interconnect performance and reliability under various design constraints. An efficient and flexible interconnect layout system is the foundation to support various interconnect design constraints and optimization techniques, and to cope with the rapid increase in layout design complexity.

In the past several years, my research group at UCLA has been developing a novel interconnect-centric design flow and methodology, which emphasizes interconnect planning and synthesis throughout the design process. Such a flow

*This research is partially sponsored by Semiconductor Research Corporation under Contract 98-DJ-605, National Science Foundation Young Investigator Award MIP-9357582, and a grant from Intel Corporation.

goes through the following major phases: (1) interconnect planning, which includes interconnect architecture planning, RT-level interconnect planning, and physical-level interconnect planning; (2) interconnect synthesis, which determines the optimal or near-optimal interconnect topology, wire ordering, buffer locations and sizes, wire width and spacing, etc., to meet the performance and signal reliability requirements of one or multiple nets; (3) interconnect layout, which will be achieved by a flexible and highly efficient multi-layer general-area gridless routing system. In addition, efficient interconnect performance estimation models and interconnect verification techniques are needed at every step of the design process. This paper highlights some of the results we have achieved in these areas.

2 Interconnect Synthesis

Interconnect synthesis determines the optimal interconnect structure of each net in terms of interconnect topology, wire width and spacing, buffer locations and sizes, etc., to meet the performance and signal reliability requirements. Our group has started a systematic study of the interconnect synthesis problems since 1990 and developed a number of efficient optimal or near-optimal algorithms for various interconnect synthesis problems, including

- interconnect topology optimization
- wire-sizing optimization
- global interconnect sizing and spacing
- simultaneous driver, buffer, and interconnect sizing
- simultaneous interconnect topology construction with buffer insertion and/or wire-sizing

and other possible combinations of these optimization techniques.

These algorithms have been developed and integrated into an interconnect synthesis package, named TRIO (tree, repeater, and interconnect optimization), which is available from [1]. Two types of device models are considered in the TRIO package – a simple switch-level RC model or a table-based device delay model that models device delay as a function of input waveform slope, device size and output load. Two types of interconnect capacitance models are considered in the TRIO package – a simple model that assumes

constant unit area and unit fringing capacitance or a table-based interconnect capacitance model that considers area, fringe and coupling capacitance as functions of wire width and spacing [2]. Most of the algorithms in TRIO uses the Elmore delay to guide the optimization process. Some of them also use high-order moments based delay models. The optimization engines of the algorithms in TRIO use either bottom-up dynamic programming or efficient iterative local refinement based on the CH-posynomial formulation. Both can guarantee optimal or near-optimal quality with polynomial time complexity in most cases. More detailed discussions of the interconnect synthesis algorithms in TRIO are available from the two survey papers [3, 4].

Optimal interconnect synthesis can reduce the interconnect delay significantly. Figure 1 shows the impact of optimal interconnect synthesis on a 2cm global interconnect in each technology generation as projected in the 1997 National Technology Roadmap for Semiconductors (NTRS'97) [5]. The three delay curves shown in the figure correspond to a 2cm un-optimized interconnect with driving sizing only to match the load (DS), a 2cm interconnect with optimal buffer insertion and sizing (BIS), and a 2cm interconnect with optimal buffer insertion, sizing and wire-sizing (BISWS), all computed by the TRIO package. As can be seen from the figure, a factor of up to 5X can be achieved with proper interconnect synthesis and optimization. Currently, our research in this area focuses on synthesizing multiple physically related and temporal-related interconnect structures for both delay and noise optimization in nanometer designs.

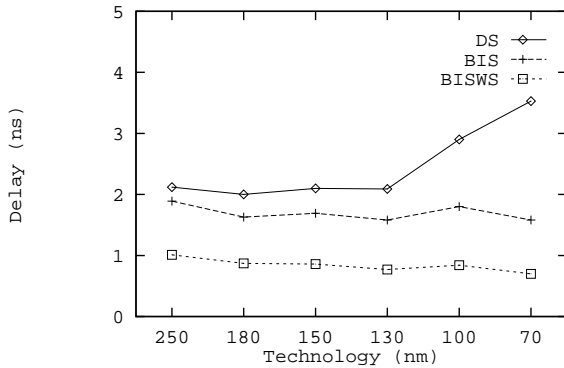


Figure 1: Impact of optimal interconnect synthesis for a 2cm global interconnect in each technology generation in NTRS'97. The maximum buffer during the optimization is limited to $200\times$ minimum feature size and the maximum wire width is set to be $10\times$ minimum wire width.

3 Interconnect Performance Estimation

Although most of the interconnect optimization algorithms discussed in the preceding section have polynomial time complexity and are efficient to use during layout synthesis (TRIO can synthesize/optimize roughly 1 to 100 nets per second),

they are not efficient enough to be used repeatedly during an interconnect planning phase where one may easily explore tens of thousands of floorplan configurations and, for each configuration evaluate the performance of tens of thousands of global and semi-global interconnects. Existing simple interconnect performance models usually consider only the wirelength and ignore various possible interconnect optimization, which leads to very inaccurate results. Therefore, in order to effectively perform high-level interconnect planning, there is a strong need to develop highly efficient interconnect performance estimation models which allow us to quickly evaluate the interconnect performance with consideration of various kinds of optimal interconnect synthesis operations.

In order to satisfy this need, in the past two years we have developed a set of fast and accurate interconnect *delay and area estimation models* (DAEM) [6, 7] with consideration of a number of optimization techniques, including optimal wire sizing (OWS), simultaneous driver and wire sizing (SDWS), and buffer insertion, sizing and wire sizing (BISWS). Our DAEM's provide the following capabilities: (i) they are very efficient (constant run time in practice), (ii) they provide high-level abstraction (for example, they do not require information about the wire segmentation schemes and wire/driver/buffer size granularity, etc., as required by the interconnect synthesis tools), and (iii) they can easily be embedded into synthesis and planning tools. In addition, our DAEM's provide explicit relations between the interconnect performance and layout design parameters under various kinds of optimization, which helps to make design decisions at high levels. These models have been tested on a wide range of parameters and have about 90% accuracy on average compared with those running complex optimization algorithms in TRIO directly followed by HSPICE simulations.

For example, the following two formulae give delay and area estimation of a wire of length l driven by a driver of effective resistance R_d with loading capacitance C_L under optimal wire sizing (OWS):

$$T_{ows}(R_d, l, C_L) = (\alpha_1 l / W^2 (\alpha_2 l) + 2\alpha_1 l / W (\alpha_2 l) + R_d c_f + \sqrt{R_d r c_a c_f l}) \cdot l \quad (1)$$

$$A_{ows}(R_d, l, C_L) = \sqrt{\frac{r(c_f l + 2C_L)}{2R_d c_a}} \cdot l \quad (2)$$

where $\alpha_1 = \frac{1}{4} r c_a$, $\alpha_2 = \frac{1}{2} \sqrt{\frac{r c_a}{R_d C_L}}$, and $W(x)$ is Lambert's W function [8] defined as the value of w that satisfies $w e^w = x$ (r is the sheet resistance, c_a and c_f are the unit area and fringing capacitance coefficients, respectively). Figure 2 shows the comparison of using our delay estimation model and running TRIO package under OWS optimization. Our model has very high accuracy (about 90% accuracy on average), yet is an order of 10,000 times faster than running the best available OWS algorithm directly.

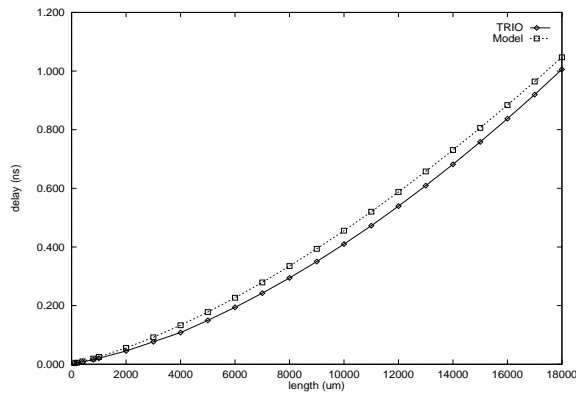


Figure 2: Comparison of our delay estimation model with running TRIO for OWS under the $0.18\mu\text{m}$ technology, with R_d and C_L from a $100\times$ min gate. TRIO uses wire width set $\{W_{min}, 2W_{min}, \dots, 20W_{min}\}$ and $10\mu\text{m}$ -long segments.

4 Interconnect Planning

We further divide the interconnect planning phase into three steps, including interconnect architecture planning, RT-level interconnect planning, and physical-level interconnect planning.

Due to the advance in the VLSI fabrication technology, such as the use of chemical-mechanical polishing (CMP) for global and local planarization of insulator and metal levels, the design rules are no longer completely dictated by the manufacturing capability. The goal of *interconnect architecture planning* is to take advantage of the degree of freedom in the process technology and determine the optimal number of routing layers, the thickness of each interconnect and isolation layer, the metal resistivity and dielectric constant of each layer (assuming different material/process may be used for different layers for performance, yield, cost considerations), the nominal width and spacing in each layer, vertical interconnection schemes (e.g., via dimensions and structures), etc., for overall system-level performance, reliability and power optimization, subject to the manufacturing constraints. Such interconnect architecture planning should consider a given design characterization (specified in terms of target clock rate, interconnect distribution, depths of the logic network, etc.) and will take place prior to the actual design process. Since such optimization requires adjustments in the fabrication process, it is most likely to be suitable and economical for high-volume designs (such as microprocessor designs) or a class of designs with similar design characterizations.

The objective of *interconnect planning at the RT-level* is to interact with a RT-level floorplanner to define the global and local interconnects, provide an estimate of overall interconnect distribution and interconnect performance, and guide the RT-level and logic-level synthesis tools to perform retiming, pipelining, and various logic optimization to meet the performance target. The RT-level design may change as a result of such interconnect planning. For example, a

different micro-architecture might be considered in order to improve the global interconnect distribution. One may want to re-partition the functional hierarchy and/or re-synthesize the control logic for possible retiming and pipelining to cope with long interconnect delays.

Another level of interconnect planning takes place during physical floorplan designs, called *physical-level interconnect planning*. It interacts closely with the interconnect synthesis tools (discussed in Section 2) and plans for the best interconnect topology, wire ordering and width, wire spacing, layer assignment, etc., for all global and semi-global interconnects to meet the required performance. For example, it is estimated that there will be a large number of buffers to be inserted for high-performance designs in future technology generations (close to 800,000 in 50nm technology [9]). If these buffers are distributed over the entire chip in an unstructured way, it will definitely complicate the layout design and verification. One of the research problems that we are currently investigating is to automatically generate buffer blocks during physical-level floorplan to achieve performance, area, and routability optimization. The interconnect performance estimation models presented in the preceding section are very important to guide every step of the interconnect planning process.

As an example, in the remainder of this section we shall present our recent results into wire width planning, as part of our investigation on interconnect architecture planning. As stated in Section 2, wire-sizing is an effective technique for reducing interconnect delays. However, having many different wire widths will considerably complicate the layout design, especially the routing process. As a result, it is interesting to investigate the possibility of using a small set of predetermined “fixed” widths in each layer to get close to optimal performance for all interconnects in a wide range of wirelengths in that layer (not just one length!).

Given the wirelength distribution in each layer, the *wire-width planning problem* is to find the best width vector \vec{W} for that layer such that the following objective function

$$\Phi(\vec{W}, l_{min}, l_{max}) = \int_{l_{min}}^{l_{max}} \lambda(l) \cdot f(\vec{W}, l) dl \quad (3)$$

is minimized, where $\lambda(l)$ is the distribution function of wirelength l , l_{min} and l_{max} are the minimum and maximum wirelengths for this metal layer, and $f(\vec{W}, l)$ is the objective function to be minimized by the design. In this study we choose $f(l) = A^j(\vec{W}, l) \cdot T^k(\vec{W}, l)$, where $A(\vec{W}, l)$ and $T(\vec{W}, l)$ denote the area and delay using \vec{W} . For our one-width design, \vec{W} has only one component W . For two-width design, \vec{W} has two components W_1 and W_2 . For $j = 0$ and $k = 1$, the objective is performance optimization only, which tends to use large wire width with marginal performance gain (since the delay/width curve becomes very flat while approaching optimal delay). Our study indicates that the AT^4 metric (i.e., $j = 1$ and $k = 4$) leads to area-efficient performance optimization in general.

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

Table 1: Comparison of using one-width design, two-width design and many-width design (up to $50 \times$ min width) using GISS for wire-sizing and spacing. Layers 7 and 8 of $0.10 \mu m$ technology are used, with wirelength ranging from 8.04 to $22.8 mm$. Driver size is assumed to $250 \times$ min.

We achieved a rather surprising result which suggests that two predetermined wire widths per metal layer are sufficient to achieve near-optimal performance. For example, for layers 7 and 8 in the $0.10 \mu m$ technology, given the interconnect length distribution model (from $7.57 mm$ to $24.9 mm$) described in [7], our wire-width planning tool suggests that the best one-width is $1.98 \mu m$, and that the best two-width design consists of wires of width $1.0 \mu m$ and $2.0 \mu m$. Table 1 shows the comparison of using the one-width, two-width, and many-width designs by running GISS (global interconnect sizing and spacing) algorithm [10]. Three different pitch-spacings (pitch-sp) between adjacent wires in layers 7 and 8 of the $0.10 \mu m$ technology are used. For each pitch-sp, we compare the average delay, the maximum delay difference (in percentage) from GISS (ΔT_{max}) for all lengths, and the average width. For pitch-spacing of $2.0 \mu m$, one-width design has an average delay about 14% and 20% larger than those from the two-width design and many-width design, respectively. Moreover, it has an average wire width (thus area) about $1.83 \times$ and $1.92 \times$ of those from two-width design and many-width design, respectively. The two-width design, however, has close to optimal delay compared to that of many-width design obtained by the GISS algorithm (just 3-5% larger) and uses only a slightly bigger area (less than 5%) than that of GISS. When the pitch-spacing becomes larger, the difference between one-width design, two-width designs and many-width gets smaller. In the table, we also list the maximum delay difference between the two-width and many-width designs. It is an important metric as it can bound our estimation error under *any length distribution function* $\lambda(l)$ in (3). The reader may refer to [7] for more details.

5 Interconnect Layout

Aggressive interconnect synthesis and optimization often result in complex interconnect structures with many buffers and with variable widths within the same net, or even within the same segment. Different spacing rules are also needed for crosstalk control and minimization. These require a very flexible and efficient multi-layer gridless router for implementation in the layout phase. Currently, we are in the process of developing a scalable and efficient multi-layer gridless routing system based on the ideas of global track ordering and spacing, implicit routing graph representation, and efficient techniques from computational geometry to support gridless routing for delay and noise control and minimiza-

tion. Due to page limitations, we are not able to discuss these results in this paper.

6 Conclusion

In this paper, we discussed our research efforts which focus on developing an interconnect-centric design flow that goes through three major design phases of interconnect planning, interconnect synthesis, and interconnect layout. Efficient interconnect performance estimation models have also been proposed to support careful interconnect estimation and planning in the early design process. We hope that such an interconnect-centric design flow will effectively bridge the gap between high-level design abstraction and physical-level implementation, and reduce or eliminate the uncertainty due to interconnects on system performance and reliability.

Acknowledgments

The author would like to thank the students and researchers (current and former) who are/were part of the UCLA "interconnect team". These include Chin-Chih Chang, Lei He, Kei-Yong Khoo, Cheng-Kok Koh, Hardy Leung, Patrick Madden, Takumi Okamoto, and David Z. Pan, all of whom contributed various interesting and important results on interconnect modeling, design, and optimization that formed the basis of the interconnect-centric design flow and methodology discussed in this paper. Additionally, the author is grateful to David Pan for his assistance in preparing this paper.

References

- [1] J. Cong, L. He, C.-K. Koh, D. Z. Pan, and X. Yuan, "TRIO: Tree, repeater and interconnect optimization package." <http://cadlab.cs.ucla.edu/~trio>.
- [2] 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.*, pp. 40.1.1-40.1.6, June, 1997.
- [3] J. Cong, L. He, C.-K. Koh, and P. H. Madden, "Performance optimization of VLSI interconnect layout," *Integration, the VLSI Journal*, vol. 21, pp. 1-94, 1996.
- [4] J. Cong, L. He, K.-Y. Khoo, C.-K. Koh, and D. Z. Pan, "Interconnect design for deep submicron ICs," in *Proc. Int. Conf. on Computer Aided Design*, pp. 478-485, 1997.
- [5] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, 1997.
- [6] J. Cong and D. Z. Pan, "Interconnect delay estimation models for synthesis and design planning," in *Proc. Asia and South Pacific Design Automation Conf.*, pp. 97-100, Jan., 1999.
- [7] J. Cong and D. Z. Pan, "Interconnect estimation and planning for deep submicron designs," in *Proc. Design Automation Conf.*, June, 1999.
- [8] C.-P. Chen and D. F. Wong, "Optimal wire sizing function with fringing capacitance consideration," in *Proc. Design Automation Conf.*, pp. 604-607, 1997.
- [9] J. Cong, "Challenges and opportunities for design innovations in nanometer technologies," Dec. 1997. <http://www.src.org/research/frontier.dgw>.
- [10] J. Cong, L. He, C.-K. Koh, and D. Z. Pan, "Global interconnect sizing and spacing with consideration of coupling capacitance," in *Proc. Int. Conf. on Computer Aided Design*, pp. 628-633, 1997.