Interconnect Estimation and Planning for Deep Submicron Designs

Jason Cong and David Zhigang Pan
Department of Computer Science
University of California, Los Angeles, CA 90095
Email: {cong,pan}@cs.ucla.edu *

Abstract

This paper reports two sets of important results in our exploration of an interconnect-centric design flow for deep submicron (DSM) designs: (i) We obtain efficient yet accurate wiring area estimation models for optimal wire sizing (OWS). We also propose a simple metric to guide area-efficient performance optimization; (ii) Guided by our interconnect estimation models, we study the interconnect architecture planning problem for wire-width designs. We achieve a rather surprising result which suggests that two pre-determined wire widths per metal layer are sufficient to achieve near-optimal performance. This result will greatly simplify the routing architecture and tools for DSM designs. We believe that our interconnect estimation and planning results will have a significant impact on DSM designs.

1 Introduction

For deep submicron (DSM) VLSI designs, interconnect has become a dominant factor in determining the overall circuit performance, reliability, and cost. As a result, in recent years many interconnect optimization techniques, including wire sizing and spacing, buffer insertion and sizing, etc., have been proposed and shown to be very effective (e.g., see [1] for a survey). However, these interconnect optimization algorithms are mainly for physical level, and not efficient to be used in higher level synthesis/planning tools.

Interconnect estimation modeling is to seek fast yet accurate metrics to estimate the optimal performance under various interconnect optimizations. It provides an enabling mechanism to effectively couple the synthesis/planning tools and the interconnect optimization algorithms. [2] developed the first set of interconnect delay estimation models with interconnect optimizations, including optimal wire sizing (OWS), simultaneous driver and wire sizing, and simultaneous buffer insertion/sizing and wire sizing.

However, [2] does not provide wiring area estimation under interconnect optimization. The wiring resources must also be planned at high levels to make sure that the planned interconnect optimization is realizable at the layout level.

In this paper, we study interconnect estimation for both delay and area, with consideration of coupling capacitance. Based on our simple but accurate estimation modeling, we also propose a novel interconnect architecture planning methodology for wire-width design. Our main contributions include the following:

- First, we develop a very simple closed-form area estimation model for OWS [3]. In addition, we find that two simple wire sizing schemes, namely single-width sizing (1-WS) and two-width sizing (2-WS) can be used to approximate OWS reasonably well to certain extend.
- We study the tradeoff between area and delay and propose a metric AT^4 (A–area, T–delay) to guide the area-efficient performance optimization. This metric usually leads to more than 60% area reduction but with only about 10% delay increase compared with pure delay-driven metric.
- Our delay sensitivity study further suggests that there exist some small set of "globally" optimal widths for a wide range of interconnect lengths. We obtain such "globally" optimal wire width design, and show rather surprisingly that using two "pre-designed" widths, we are still able to achieve close to optimal performance compared with those by using many possible widths.

The rest of the paper will be organized as follows. Section 2 states the preliminaries. Section 3 studies interconnect estimation modeling. Section 4 presents results on interconnect architecture planning, specifically the wire-width planning, followed by the conclusion in Section 5. Due to the space limitation, details of this work are left out in a technical report [4].

2 Preliminaries

The key parameters used by our interconnect estimation and planning programs are listed below.

- W_{min}: the minimum wire width
- S_{min}: the minimum wire spacing
- r: the sheet resistance
- c_a: the unit area capacitance
- c_f: the unit effective-fringing capacitance [5]
- t_g: the intrinsic device delay
- c_g: input capacitance of a minimum device
- r_g: output resistance of a minimum device
- R_d: the driver effective resistance
- l: interconnect length
- C_L: loading capacitance.

Most values of these parameters used in our study are based on the 1997 National Technology Roadmap for Semiconductors (NTRS'97) [6]. We consider the effect of interconnect reverse scaling at higher metal layers. Similar to [7], we define a *tier* to be a pair of adjacent metal layers with the same cross-sectional dimensions. So from bottom to top, Tier1 refers to metal layers 1 and 2, Tier2 refers to metal layers 3 and 4, ..., and Tier4 refers to metal layers 7 and 8. Since NTRS'97 only provides the geometry information for Tier1, for higher metal layers, we adopt the geometry information from UC Berkeley's Strawman technology [8] and from SEMATECH [9]. For capacitance extraction, we use the 2.5D capacitance extraction method in [10]. All values of these parameters can be found in [4].

Since interconnect estimation and planning are intended for early design planning stages, we will use simple but reasonably accurate interconnect, device and delay calculation models. Similar to [2], we model a device as a switch-level RC circuit [1], and use the well-known Elmore delay model [11] for delay computation.

*This research is partially sponsored by Semiconductor Research Corporation under Contract 98-DJ-605.

3 Interconnect Delay and Area Estimation

In this section, we study interconnect estimation modeling. We first present the area estimation model for optimal wire sizing (OWS) [3] algorithm. Then we show that two simplified wire sizing schemes, namely single-width sizing (1-WS) and two-width sizing (2-WS) are good approximation of OWS in a wide range. We further explore the delay/area trade-off and propose a new metric for area-efficient performance optimization.

3.1 Area Estimation for Optimal Wire-Sizing (OWS)

Our extensive study of OWS (details in [4]) shows that the average wire width using OWS algorithm [3] for an interconnect of length l can be expressed in the following simple closed-form formula

$$w_{avg}(l) = \sqrt{\frac{r(c_f l + 2C_L)}{2R_d c_a}} \quad (1)$$

From this, we can see that larger c_f and C_L lead to larger wire sizing; while larger R_d (weaker driver) and c_a lead to smaller wire sizing. It confirms the WS/DS, WS/CL relationships in [12] and the effective-fringing property in [5].

Our model gives very accurate wiring area estimation for OWS (usually within 5% error). As an example, Figure 1 shows the comparison of average wire width from our model with that from running OWS algorithm in the UCLA TRIO package [1]. For Tier1, the average wire width from our estimation model is almost identical to that from TRIO. For Tier4, our model gives just slightly larger (about 5%) estimation than TRIO.

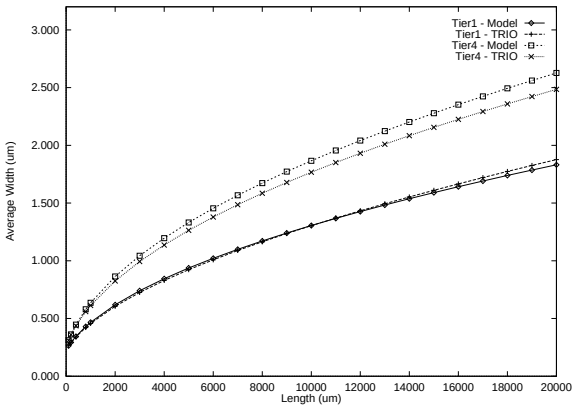

Figure 1: Comparison of our area (width) estimation model and TRIO for Tier1 and Tier4 under 0.10 μm technology with $R_d = r_g/100$, $C_L = c_g \times 100$.

3.2 Two Simple Wire-Sizing Schemes

To simplify the potential routing problem caused by many possible wire widths of using OWS, we study two simple wire-sizing schemes, namely the optimal single-width sizing (1-WS) and optimal two-width sizing (2-WS). As implied by their names, 1-WS computes the best wire sizing solution with one uniform width, and 2-WS computes the best wire sizing solution with two possible widths (together with the length under each width). Derivation of 1-WS and 2-WS solutions is fairly straightforward (see details in [4]). Our study shows, rather surprisingly, that 1-WS and 2-WS are good approximation to OWS, in both delay and area, for a wide range of interconnect lengths (e.g., for interconnects shorter than the critical length in [2]).

Figure 2 shows the delay comparison of 1-WS, 2-WS and OWS for Tier1 and Tier4 under 0.10μm technology. For Tier1, 1-WS and 2-WS have at most 10% more delay than OWS for wire length up to 5mm. For Tier4, 1-WS and 2-WS have almost the same delay as OWS for all wire length up to 2cm. In this figure, we assume constant c_f for different widths as in [3]. However, when we take coupling capacitance into consideration, c_f will be a function of wire width (e.g., in the fixed pitch-spacing scenario of [5]), and 2-WS will have significant advantage over 1-WS, as we shall see in Section 4.3.

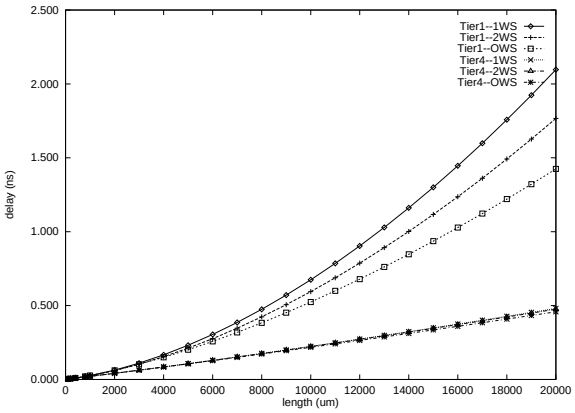

Figure 2: Comparison of 1-WS, 2WS and OWS for Tier1 and Tier4 using the 0.10 μm technology. $R_d = r_g/100$, $C_L = c_g \times 100$.

3.3 Delay/Area Trade-off and Sensitivity Study

So far, our objective is pure delay minimization, which may lead to significant area overhead. To obtain a good metric for area efficient performance optimization, we have performed extensive experiments on different area-delay metrics, including T (delay only), AT (area-delay product), AT^2 (area-delay-square product), AT^3, AT^4, AT^5, etc. Our study concludes that AT^4 is a suitable metric for area-efficient performance optimization, resulting only marginal delay increase, but significant area reduction. Figure 3 shows an example. The optimal widths of a 2cm interconnect for T, AT^5, AT^4, AT^3, AT^2, AT are 2.6, 1.15, 1.0, 0.6, 0.3, and 0.1μm, with delays of 0.48, 0.52, 0.53, 0.62, 0.84, and 1.77ns, respectively. The optimal width under the AT^4 metric uses 62% smaller wiring area than that under the T metric (20,000μm^2 vs. 52,000μm^2), but with only 10% increase of delay. Therefore, we propose AT^4 as a performance-driven yet area-efficient metric for interconnect optimization. It will be used in Section 4 for interconnect architecture planning.

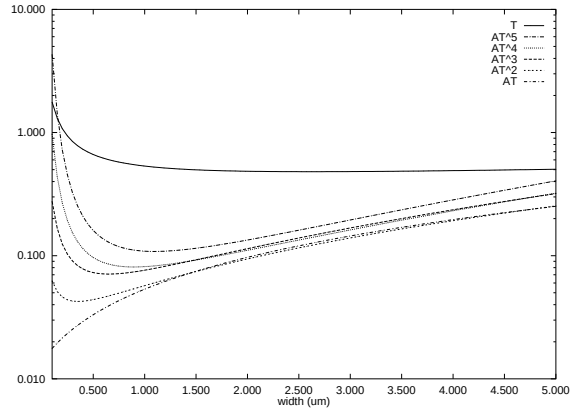

Figure 3: Different optimization metrics for a 2cm interconnect in Tier4 under the 0.10 μm technology. $R_d = r_g/100$, $C_L = c_g \times 100$. The y-axis is scaled to compare all metrics in one figure.

4 Interconnect Architecture Planning

4.1 Motivation

From our study of delay-area sensitivity study in the previous section, a very interesting observation is that the delay is not sensitive to certain degree wire width variations around the optimal solution. This not only suggests that we can achieve close to optimal performance with significant area saving (as we show in Section 3.3), but also suggests that there may exist a small set of "globally" optimal widths for a range of interconnect lengths, so that we may use a small set of pre-determined "fixed" widths to get close to optimal performance for all interconnects in a wide range of wire lengths (not just one length!), as opposed to the wire sizing solution with many different widths obtained from running complicated wire sizing/spacing algorithms [3, 5]. This optimal wire-width design, on one hand, still guarantees near-optimal performance; on the other hand, greatly simplifies the detailed routing problem and the interaction between higher level design planning/optimization tools and lower level routing tools. In particular, if only one or two fixed widths are used for every metal layer, a full-blown gridless router may not be necessary. This may significantly simplify many problems, including RC extraction, detailed routing and layout verification.

4.2 Overall Approach

Our wire-width planning is tier-based, i.e., we will determine the best width designs for each tier. In general, local interconnects are routed in the lower tier (Tier1), while global interconnects are routed in the higher tier (Tier3 or Tier4). The wire length distribution on different tiers usually varies from design to design, and also depends on the layout tools and optimization objectives. In our study, we assume that the maximum wire length (l_{max}) in Tier1 is $10,000\times$ feature size, and l_{max} in the top tier is L_{edge}, i.e, the chip dimension under a given technology [9]. The l_{max} in the intermediate tiers will be determined by a geometric sequence such that for any tier i, $l_{max}(i+1)/l_{max}(i) = l_{max}(i)/l_{max}(i-1)$. Table 1 shows l_{max} of each tier for NTRS'97 technologies. The minimum wire length for tier i is the maximum length for tier $i-1$, i.e., $l_{min}(i) = l_{max}(i-1)$. We also take a representative driver for each metal tier for our wire width planning. The drivers for Tier1 through Tier4 are $10\times$, $40\times$, $100\times$, and $250\times$ of the minimum gate in the given technology, respectively.

Tech (μm)	0.25	0.18	0.13	0.10	0.07
Tier1	2.50	1.80	1.30	1.00	0.70
Tier2	6.50	5.85	3.27	2.84	2.30
Tier3	17.3	19.0	8.23	8.04	7.57
Tier4	-	-	20.7	22.8	24.9

Table 1: Maximum wire length (in mm) assigned to each tier.

Given the wire length range for each tier, the wire-width design problem is to find the best width vector $\vec{W}$ such that the following objective function

$$\Phi(\vec{W}, l_{min}, l_{max}) = \int_{l_{min}}^{l_{max}} \lambda(l) \cdot f(\vec{W}, l) dl \qquad (2)$$

is minimized, where $\lambda(l)$ is the distribution function of l, 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 1-width design, $\vec{W}$ has only one component W. For 2-width design, $\vec{W}$ has two components W_1 and W_2. For $j = 0$ and $k = 1$, the objective is performance optimization only. However, as we observe in Section 3, this tends to use large wire width with marginal performance gain, since the delay/width curve becomes very flat while approaching optimal delay. So we use the AT^4 (i.e., $j = 1$ and $k = 4$) metric for area-efficient performance optimization.

For our current study, we assume $\lambda(l)$ is a uniform distribution function. Other distribution functions such as wire length distribution function in [13], can also be used. Yet, our results in Section 4.3 show that our 2-width design is so robust that it can be applied to *any length distribution function*, with predictable small amount of errors compared with the optimal solution using many possible widths.

The overall approach of the wire-width planning is straightforward. Basically, we want to find the best 1-width or 2-width pair to minimize the objective function in (2). We take 1-width planning with metric T as an example to illustrate how the wire-width planning works. For 1-width planning, we need to determine the best width W^* to minimize $\int_{l_{min}}^{l_{max}} T(w,l)dl$ 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 with width w. Then the "globally" optimal width W^* is

$$\begin{aligned} W^* &= \sqrt{\frac{\int_{l_{min}}^{l_{max}} r(\frac{1}{2} r c_f l + r C_L) l dl}{\int_{l_{min}}^{l_{max}} R_d c_a l dl}} \\ &= \sqrt{\frac{\frac{1}{3} r c_f (l_{max}^3 - l_{min}^3) + r C_L (l_{max}^2 - l_{min}^2)}{R_d c_a (l_{max}^2 - l_{min}^2)}} \end{aligned} \qquad (3)$$

For the 1-width design under metric AT^4, a simple analytical formula like (3) cannot be obtained as we need to solve an 8-th order equation for w, which does not have analytical solutions. But since the complexity of our delay and area modeling is very low, we can easily enumerate all available wire widths (provided by a given technology) to find the the best width design.

Similarly for the two-width design, we can obtain the "globally" optimal width pair W_1^* and W_2^* in an exhaustive search manner. Without loss of generality, we assume that $W_2^* = \alpha W_1^*$. Our study shows that α is usually between 2 to 3. Given each α, we can easily search the best W_1^*. In practice, we just need to search two α's[1], $\alpha = 2$ and $\alpha = 3$, which can be done very efficiently. We shall point out that the complexity of the wire-width planning step is not a major concern, since we just need to run it *once* for all future designs under a given technology.

4.3 Case Study for 0.10μm Technology

In this subsection, we present our result of using 1-width and 2-width designs under AT^4 metric. It suggest that the 2-width design under AT^4 metric has both area efficiency and also near-optimal performance.

Table 2 shows the comparison of using our 1-width, 2-width designs from running GISS algorithm [5] with many wire width choices. Three different pitch-spacings (pitch-sp) between adjacent wires in Tier4 of $0.10\mu m$ 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 μm, 1-width design has average delay about 14% and 20% larger than those from 2-width design and GISS. Moreover, it has an average wire width (thus area) about $1.83\times$ and $1.92\times$ of those from 2-WS and GISS. The 2-width design, however, has close to optimal delay compared to the solution obtained from running GISS algorithm (just 3-5% larger) and uses only slightly bigger area (less than 5%) than that of GISS. When the pitch-spacing becomes larger, the difference between 1-width design, 2-width design and GISS will get smaller. In the table, we also list the maximum delay difference from GISS. It is an important metric which can bound our estimation error under *any length distribution function* $\lambda(l)$ in (2) based on the following theorem.

[1] In fact, we try many different α's (not just integers) in our experiments and it turns out that $\alpha = 2$ or 3 is good enough.

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
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
GISS [5]	0.204	-	1.03	0.159	-	1.19	0.136	-	1.38

Table 2: Comparison of using 1-width design, 2-width design and running GISS algorithm with many wire width choices (up to 50× min width). Tier4 of 0.10μm technology is used, with wirelength range from 8.04 to 22.8mm. Driver size is 250 × min.

Theorem 1 *If* $\left|\frac{f(\vec{W},l)-f(\vec{W}^*,l)}{f(\vec{W}^*,l)}\right| \le \delta_{max}$ *for any* $l \in (l_{min}, l_{max})$, *then for any distribution function* $\lambda(l)$, *we have*

$$\left|\frac{\Phi(\vec{W}, l_{min}, l_{max}) - \Phi(\vec{W}^*, l_{min}, l_{max})}{\Phi(\vec{W}^*, l_{min}, l_{max})}\right| \le \delta_{max}. \qquad (4)$$

Since for the 2-width design derived from uniform distribution $\lambda(l) \equiv 1$, the maximum delay difference ΔT_{max} is only 3.9–7%, according to Theorem 1, this 2-width design will differ from the optimal-width design (using possibly many widths) by at most 3.9–7% for *any distribution function* $\lambda(l)$.

4.4 Recommendation for Future Technologies

We have further performed wire-width planning for future technology generations listed in NTRS'97 from 0.25 to 0.07μm. Our recommendation is based on the optimal 2-width design under the area-efficient performance optimization metric AT^4. The results are shown in Table 3. It suggests to use the minimum widths for local interconnects in Tier1. For Tier2 to Tier4, it suggests to use two different pre-determined wire widths with 1:2 ratio. Therefore, we have a wiring hierarchy on different metal layers such that Tier2 is about 1-2 times wider than Tier1, Tier3 is about 2-3 times wider than Tier2, and Tier4 (if available) is about 4-5 times wider than Tier3. Such a simple, pre-determined wiring hierarchy can effectively achieve close to optimal RC delays for all local, semi-global and global interconnects while ensuring high routing density and much simplified routing solutions.

Tech. (μm)		0.25	0.18	0.13	0.10	0.07
Tier1	W_1^*	0.25	0.18	0.13	0.10	0.07
	W_2^*	0.25	0.18	0.13	0.10	0.07
Tier2	W_1^*	0.25	0.18	0.13	0.10	0.08
	W_2^*	0.50	0.36	0.26	0.20	0.16
Tier3	W_1^*	0.65	0.47	0.24	0.22	0.23
	W_2^*	1.30	0.94	0.48	0.44	0.46
Tier4	W_1^*	-	-	0.98	1.00	1.06
	W_2^*	-	-	1.96	2.00	2.12

Table 3: Wire-width design (in μm) for area-efficient performance optimization.

5 Conclusion

In this paper, we have presented two sets of important results on an interconnect-centric design flow. First, we obtain a simple wiring area estimation formula for OWS. We further propose two simplified wire sizing schemes (1-WS and 2-WS) and an area-efficient performance optimization metric AT^4 for interconnect optimization. Based on these simple but accurate estimation models, we then study interconnect architecture planning for wire-width designs. We show that using two pre-determined 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. It shall be noted that our wire-width planning assumes a typical driver size for each metal layer. Our study shows that it is valid to certain degree of driver size variation. For example, the average error with 2× driver size variation will be up to 8% (cf. 5% in Table 2) and the maximum error will be up to 18% (cf. 7% in Table 2). If driver variation becomes even larger, our 2-width design under the AT^4 metric may not be adequate to achieve near optimal performance for all interconnect lengths in each tier. More wire width or modified design metric may then be needed, which is currently under investigation.

We expect that our interconnect estimation models be used in many applications such as interconnect-driven synthesis, floorplanning and placement. Our wire-width planning results may greatly simplify the performance-driven routing architecture for DSM designs.

Acknowledgments

The authors would like to thank Wilsin Gosti from UC Berkeley for providing Strawman technology and Lei He from UCLA for helping to generate capacitance parameters. The authors would also like to thank Kei-Yong Khoo from UCLA for prompting us to look into the possibility of using a small set of wire widths for wire sizing optimization for the ease of detailed routing.

References

[1] 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.

[2] 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.

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

[4] J. Cong and D. Z. Pan, "Interconnect estimation and planning for deep submicron designs," Tech. Rep. 980035, UCLA CS Dept, 1998. http://cadlab.cs.ucla.edu/~pan/publications/.

[5] 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.

[6] Semiconductor Industry Association, *National Technology Roadmap for Semiconductors*, 1997.

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

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

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

[10] 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.

[11] W. C. Elmore, "The transient response of damped linear networks with particular regard to wide-band amplifiers," *Journal of Applied Physics*, vol. 19, pp. 55–63, Jan. 1948.

[12] C.-K. Koh, *VLSI Interconnect Layout Optimization*. PhD thesis, University of California, Los Angeles, 1998.

[13] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI) i. derivation and validation," *IEEE Transactions on Electron Devices*, vol. 45, no. 3, pp. 580–9, 1998.