

# Analysis and Optimization for Global Interconnects for Gigascale Integration (GSI)

A thesis  
Presented to  
The Academic Faculty

by

Azad Naeemi

In Partial Fulfillment  
of the Requirements for the degree of  
Doctor of Philosophy in Electrical and Computer Engineering




School of Electrical and Computer Engineering  
Georgia Institute of Technology  
July 2003

Copyright © 2003 by Azad Naeemi

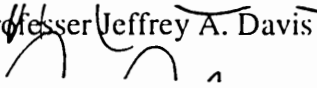
# Analysis and Optimization for Global Interconnects for Gigascale Integration (GSI)

Approved by:




  
Professor James D. Meindl, Advisor

Professor D. Scott Wills 

Professor Jeffrey A. Davis 

  
Professor David Citrin

  
Professor Paul A. Kohl

Date Approved 09/08/2003

To my parents,  
for their wisdom, persistence, and dedication.

## Acknowledgments

I would like to sincerely thank my advisor, Professor James Meindl, for guiding me through this thesis. It has been a great honor and privilege to work under his supervision and benefit from his keen insights. His deep and enlightening questions have always led me to astonishing ideas.

I am also grateful to Dr. Jeff Davis for his thoughtful guidance and interesting discussions, and also for his setting the example as a former student of the gigascale integration (GSI) group. I appreciate the invaluable guidance and attention that I received from Professors Thomas Gaylord and Scot Wills. I am also in great debt to all my ECE professors, especially Professors Glenn Smith and the late John Uyemura, for amazingly teaching me difficult courses in simple language.

My sincere thanks also go to all GSI group members, especially our program manager, Ms. Jennifer Tatham, without whose efforts I could not even start my graduate studies. She has taken care of all GSI students and their families, and because of her efforts, the GSI group has attained a unique pleasant atmosphere.

Most importantly, I would like to thank my wife Nassim whose unconditional love, passion and trust have been the inspiration for this work. Her beautiful smile brought joy and motivation to my life. Thank you for everything.

# Table of Contents

Chapter 1: Introduction and Background.....	1
1.2 Optimal Global Interconnects for GSI.....	4
1.3 Compact Physical Models.....	5
1.3.1 Distributed RLC Lines above an Ideal Ground Plane .....	5
1.3.2 Co-Planar RLC Lines above Orthogonal Lines .....	5
1.3.3 Multi-Level Crosstalk Noise.....	6
1.4 Chip-Package Co-Design Methodologies.....	7
1.4.4 Signal Interconnection .....	7
1.4.5 Power and Ground Interconnection .....	8
1.5 Optical Interconnection versus Electrical Interconnection .....	8
Chapter 2: Optimal Global Interconnects for GSI.....	9
2.2 Optimal On-Chip Wire Width .....	12
2.2.1 Impact of Wire Width on Latency .....	12
2.2.2 Impact of wire width on bandwidth.....	15
2.2.3 Optimal Wire Width .....	17
2.3 Optimal On-Chip Wire Width .....	22
2.3.1 Power Consumption.....	22
2.3.2 Repeater Area.....	26
2.3.3 Via Blockage.....	28

2.4	Conclusions.....	30
Chapter 3: N-Coupled RLC Lines above an Ideal Ground Plane.....		32
3.2	Single, Two and Three Coupled RLC Lines.....	33
3.3	Five or More Coupled RLC Lines .....	35
3.4	Verification and Results.....	38
3.5	Conclusions.....	43
Chapter 4: Modeling of Co-Planar RLC Lines.....		45
4.1	Introduction.....	45
4.2	Periodic Structures .....	47
4.3	Modeling On-Chip Co-Planar Interconnects .....	49
4.4	Two Signal Lines between Power and Ground Lines.....	52
4.5	More Than Two Signal Lines between Power and Ground Lines.....	59
4.6	Conclusions.....	60
Chapter 5: Optimization of Co-Planar RLC Lines .....		62
5.1	Introduction.....	62
5.2	Optimal Signal-Ground Spacing to Signal-Signal Spacing Ratio .....	63
5.3	Optimal Wire Width .....	68
5.3.1	Optimal Wire Width for a Single Signal Line .....	69
5.3.2	Optimal Wire Width for Two Signal Lines .....	72

5.4	Optimal Metal thickness and Spacing.....	79
5.5	Conclusions.....	85
Chapter 6: Multilevel Interconnect Crosstalk Modeling .....		87
6.1	Introduction.....	87
6.2	Methodology .....	88
6.3	Identical Victim and Aggressor Lines .....	89
6.4	Non-Identical Victim and Aggressor Lines .....	95
6.5	Noise Voltage-Time Integral .....	100
6.6	Near and Far Aggressors.....	102
6.6.1	Near and Intra-Level Far Aggressors.....	102
6.6.2	Near and Inter-Level Far Aggressors.....	107
6.7	Integrated Crosstalk Model.....	108
6.8	Impact of Wire Width Optimization .....	112
6.9	Conclusions.....	114
Chapter 7: Chip-Package Co-Design.....		115
7.1	Introduction.....	115
7.2	Signal Interconnection through PWB .....	116
7.2.1	Maximum On-Chip Interconnect Length to Achieve the Highest Performance .	117
7.2.2	Cost Estimation.....	119
7.2.2.1.	Extracting the Wiring Distribution.....	119

7.2.2.2.	Estimating the Required Area and I/O Pads.....	123
7.2.2.3.	Estimating Total PWB Area.....	124
7.3	Power Distribution.....	133
7.4	Conclusions.....	135
Chapter 8: Optical Versus Electrical Interconnection.....		136
8.1	Introduction.....	136
8.2	Electrical Interconnects.....	138
8.3	Optical Interconnects .....	142
8.4	The Partition Length between Electrical and Optical Interconnects .....	144
8.5	Conclusions.....	147
Chapter 9: Future Work and Conclusion .....		149
9.1	Advancing Compact Physical Models .....	148
9.2	Optimizing Other Interconnecting Techniques.....	149
9.3	Developing Stochastic Models for Crosstalk Noise .....	150
9.4	Extending Chip-Package Design Methodologies .....	151
9.5	Analysis and Design of Clock Distribution Networks.....	151
9.6	Conclusion of Dissertation.....	152
Appendix A: Derivation of Optimal Wire Width .....		155
Appendix B: Physical Explanation for Crosstalk of Identical Lines.....		157



Appendix C: Peak Crosstalk Voltage in Terms of $W/W_{opt}$ .....	159
Appendix D: Derivation of Bit-Rate Limit of RLC Lines .....	163
References.....	164
Vita.....	170

## List of Tables

Table 2.1: Key assumptions and design goals of different methodologies for optimizing interconnecting devices. ....	11
Table 2.2. Impact of using optimal wire width on a 24 mm-long interconnect in a projected ASIC chip at the 45-nm node of technology. The chip area is 572mm <sup>2</sup> . The global repeater area and number of vias correspond to two global metal levels with a wiring efficiency of 0.5. ....	30
Table 7.1: Key parameters for the projected microprocessor in the year 2011. ....	120
Table 7.2: Costs and advantages of optimal partition between interconnects and exterconnects. ....	132
Table 7.3: The key parameters and the required number of power/ground pins for various technology generations. Aspect ratios 1 and 2 are considered for the top global interconnects ( $A_r=1, 2$ ). ....	134

# List of Figures

Figure 2.1:	The cross-section of a wire and its neighbors.....	13
Figure 2.2:	Time delay of a 24 mm long interconnect with optimal repeaters versus wire width. The cross-section of the wire is shown in Figure 2.1. ....	13
Figure 2.3:	A layer of global interconnects. (a) All interconnects have the same length. (b) Length of interconnects are different. ....	16
Figure 2.4:	Data flux density versus wire width for a 24 mm long interconnect. Interconnect delay is also plotted. Data-flux density is constant in the RC regime and in the RLC regime, it drops as wire width decreases. ....	17
Figure 2.5:	Data flux density-reciprocal latency product versus the wire width for a 24 mm long interconnect with optimal repeaters. ....	18
Figure 2.6:	Time delay versus the interconnect width for different interconnect lengths. Optimal wire width is independent of length. ....	19
Figure 2.7:	Optimal wire width for different generations [1] and aspect ratios. Using optimal wire width makes the skin effect negligible. The skin depth is calculated for the third harmonic of the global clock frequency.....	20
Figure 2.8:	The required energy to transfer one bit of data along a 24 mm long interconnect versus the wire width. At the optimal wire width, $\Phi_D/E_b$ is maximized. ....	25
Figure 2.9:	The required silicon area for a 24 mm long interconnect and the total silicon area for global repeaters versus the wire width. The Chip area is assumed to be 572 mm <sup>2</sup> , and two levels of metal with a wiring efficiency of 0.5 are considered for global interconnects. ....	27
Figure 2.10:	Number of vias required for the repeaters associated with a pair of global interconnect levels versus the wire width. ....	29
Figure 3.1:	Crosstalk Voltage at the end of the middle quiet victim line when far and near aggressors switch in-phase. ....	39
Figure 3.2:	Crosstalk Voltage at the end of the quiet line when the far and near aggressors switch anti-phase.....	40
Figure 3.3:	Effect of increasing Number of Aggressors from one to four when all aggressors switch in-Phase .....	41

Figure 3.4:	Normalized noise voltage versus cm/cg ratio. Neglecting far lines causes a negligible error in the worst-case crosstalk especially for the range that crosstalk is less than $0.2V_{dd}$ .....	43
Figure 4.1:	Two realizations of periodic structures.....	48
Figure 4.2:	A double sided shielded signal line. (a) there are no orthogonal lines and therefore the three lines form an ideal transmission line (b) there are orthogonal lines which make the signal and ground lines form a periodic structure. ....	50
Figure 4.3:	Voltage at the end of signal line for ideal and non-ideal return path cases. For the ideal case a signal line is above a ground line and in the non-ideal case a signal line is shielded between power/ground lines above orthogonal lines.....	51
Figure 4.4:	Normalized voltage at the end of open-ended signal lines when one of them is excited with a step input and the other one is quiet. An out-of-phase noise appears at the end of the victim line due to different propagation speeds for common and differential modes. ....	55
Figure 4.5:	A segment of the equivalent circuit used for HSPICE simulations. 1000 segments are used in all simulations.....	56
Figure 4.6:	Peak in and out-of-phase noise voltages versus interconnect resistance per unit length. ....	57
Figure 4.7:	Noise voltage at the end of a middle victim line when there are 2, 3 or 5 signal lines between power/ground lines. Unlike the ideal return path case, far lines have a large impact on the worst case crosstalk.....	60
Figure 5.1:	A cross-sectional view of two signal lines between two power/ground lines.....	63
Figure 5.2:	The differential mode characteristic impedance versus the spacing.....	65
Figure 5.3:	Worst case time delay versus the ratio of signal-ground to signal-signal spaces.....	66
Figure 5.4:	Normalized delay variation versus the spacing ratio $S_g/S_m$ for the aspect ratio of two.....	67
Figure 5.5:	Maximum peak crosstalk versus the spacing ratio, $S_g/S_m$ for aspect ratios of 1 and 2. The maximum peak crosstalk is pessimistic because it is for a case that $R_l=0$ , $C_L=0$ , and the line resistance is such that crosstalk is maximized.....	68

Figure 5.6: RC and RLC model latency and data flux density versus wire width for an interconnect 24mm long implemented at the 45 nm technology node. It has been assumed that optimal repeaters are used, and $W_G=2W$ , $T=S=W$ .	70
Figure 5.7: Data flux density-reciprocal latency product versus wire width for an interconnect 24 mm long implemented at the 45 nm technology node. In the shallow RLC region, the data flux density-reciprocal product attains its maximum.	72
Figure 5.8: Data flux density-reciprocal latency product for in-phase and anti-phase switching cases. The optimal wire width for common and differential modes are different. The worst case $\Phi_D/\tau$ is maximized when the difference between in-phase and anti-phase switching latencies is minimum.	73
Figure 5.9: Normalized delay variation versus $c_{orth}/c_g$ ratio when the optimal wire width is used. For a wide range of $c_{orth}/c_g$ ratios, common and differential mode delays differ less than 10%.	75
Figure 5.10: Normalized delay variation versus wire width for an optimally buffered interconnect implemented at the 45 nm technology node. At the optimal design point, delay variation is less than 3%.	76
Figure 5.11: HSPICE simulations showing normalized crosstalk at the end of a quiet line when its adjacent line switches from low to high. Time zero corresponds to one time-of-flight delay. Peak and duration of out-of-phase noise match very well with what (5.25) and (5.31) predict with less than 3% error.	77
Figure 5.12: Interconnect latency versus total spacing for four the aspect ratios of 0.5, 1, 2, and 3. The total spacing ( $S_T = S_g + 2S_m$ ) is normalized to the wire width, and it is assumed that the optimal spacing ratio ( $S_m/S_g = 0.45$ ) and the optimal wire width are used.	81
Figure 5.13: Data flux density versus total spacing for the aspect ratios of 0.5, 1, 2, and 3. Interconnects are 10mm long and are implemented at the 45 nm technology node.	82
Figure 5.14: Energy per bit versus total spacing between interconnects for the aspect ratios of 0.5, 1, 2, and 3. Interconnects are assumed to be 10 mm long.	83
Figure 5.15: Data flux density-reciprocal energy per bit product versus total spacing between interconnects for aspect ratios of 0.5, 1, 2, and 3. Interconnects are 10mm long and are implemented at the 45nm technology node.	84

Figure 5.16: The energy efficient spacing versus aspect ratio. The energy efficient spacing, which maximizes $\Phi_D/E_b$ , increases as the aspect ratio increases.....	84
Figure 6.1: A cross-sectional view of 4 top metal levels. Top two levels are relatively fat and due to inductive effects each signal line has a nearby power/ground line as a return path. The spaces between signal and ground lines are 0.45 times smaller than signal to signal spaces to reduce crosstalk and minimize worst-case delay .....	88
Figure 6.2: A quiet victim line is attacked by some far aggressors. It has been assumed that there is no nearby aggressor and far lines are inductively coupled to the victim line. The resistance and capacitance matrices can be substituted by scalar resistance and capacitance values.....	90
Figure 6.3: The induced noise on the quiet line when all signal lines are shielded by two power/ground lines. HSPICE simulations are compared with compact models for different load capacitances. The input of the aggressors is a positive step voltage, $V_{dd}u_o(t)$ . .....	94
Figure 6.4: The peak and duration of far inductive noise versus the ground line width for the structure shown in Figure 6.2. Increasing ground line width reduces all mutual inductances approximately by the same ratio. The peak noise of an open-ended line is independent of ground line width. ....	95
Figure 6.5: The induced noise at the end of a quiet line for three different load capacitances. HSPICE simulations are compared with compact models. $S_g$ and $S_m$ are optimized so that the worst-case delay is minimized [6].....	97
Figure 6.6: Voltage at the end of a quiet victim line when it is affected by either three or five far aggressors that are two metal levels below the victim line. Line resistance for the two top levels is 45 $\Omega/\text{cm}$ and for the lower metal levels is 90 $\Omega/\text{cm}$ . ....	99
Figure 6.7: Noise voltage-time integral versus interconnect length. $V_{dd}$ is assumed to be 1V. The value of this integral is independent of the load capacitance. ....	101
Figure 6.8: Noise voltage at the end of a quiet victim line when intra-level far lines switch upward and a near aggressor stays quiet.....	103
Figure 6.9: Noise voltage at the end of a quiet victim line when intra-level far lines are quiet and a near aggressor switches upward.....	105

Figure 6.10: Noise voltage at the end of a quiet victim line when all intra-level near and far aggressors switch simultaneously.....	106
Figure 6.11: Noise voltage at the end of quiet victim line when near and intra-level far aggressors switch simultaneously.....	108
Figure 6.12: Total noise caused by near, far intra- and inter-level aggressors. Far Lines switch in the opposite direction and intra-level far aggressors switch 20 ps after near and inter-level far aggressors. This shows the worst case scenario for crosstalk.....	109
Figure 6.13: Worst case noise voltage and interconnect latency versus resistance per unit length of interconnects in the top metal level for the structure shown in Fig. 13. It has been assumed that interconnects in the two lower metal levels have a resistance per unit length two times larger than that of the top two metal levels. ....	110
Figure 6.14: Worst case noise voltage and interconnect latency versus resistance per unit length of interconnects in the top metal level when optimal repeaters are used. All parameters are the same as those in Fig. 14. ....	111
Figure 7.1: Net length distribution for the projected microprocessor in the year 2011.....	120
Figure 7.2: A typical net with fan out 4.....	120
Figure 7.3: The effective length depends on the location of the driver.....	122
Figure 7.4: Effective global net-length distribution for the projected SoC in the year 2011.....	122
Figure 7.5: The required number of pads versus the maximum on-chip interconnect length.....	124
Figure 7.6: By transferring just the longest parts of the nets to PWB the required number of pads can be reduced.....	124
Figure 7.7: PWB layers are used for routing the standard I/Os to other chips, power and ground distribution, and exterconnects. ....	125
Figure 7.8: A channel with two lanes.....	126
Figure 7.9: A quadrant of the standard I/Os routed to the other chips through PWB. They can be routed either through two edges (left) or four edges (right). The dashed lines show the wires in the next layer. ....	127
Figure 7.10: The required number of PWB layers to route standard I/Os versus the number of exterconnect pads. ....	129

Figure 7.11: The via blockage. ....	130
Figure 7.12: Total number of PWB layers versus the maximum on-chip interconnect length. Four additional layers are required to achieve the highest global clock frequency. ....	131
Figure 8.1: Maximum data flux density and optimal wire width for 20, 30 and 40 <i>cm</i> long interconnects implemented in different technology generations. Both maximum data flux density and optimal wire width increase in future generations because of faster transistors. ....	140
Figure 8.2: Maximum data flux density and optimal wire width versus interconnect length for the 45 <i>nm</i> technology node. The dashed line shows the minimum line width available on-board as projected by the ITRS. ....	141
Figure 8.3: Maximum data flux density for electrical and optical interconnects implemented in 130 and 45 <i>nm</i> technology nodes. Performance of optical waveguides is length independent. ....	143
Figure 8.4: Unlike electrical interconnects, optical waveguides can not be routed arbitrarily because sharp bends and inter-level communication should be minimized due to power budget constraints. ....	144
Figure 8.5: Partition length between optical and electrical interconnects for different generations of technology. Partition length decreases in future generations because of faster transistors and finer PWB line width. ....	145
Figure 8.6: The partition length between optical and electrical interconnects versus minimum available board-level line width. Having a $W_{min}$ of about 10 $\mu m$ makes the partition length shorter than almost all board-level interconnects; hence, justifies replacing all board-level electrical interconnects with optical waveguides. ....	146



# Summary

The main objective of this thesis is to develop new interconnect-centric methodologies to optimize global interconnects in a GSI chip. A length-independent optimal wire width is rigorously found that simultaneously maximizes data flux density and minimizes latency. Data flux density is the product of interconnect bandwidth and reciprocal pitch and represents the number of bits per second that interconnects can transfer per unit width. Other cross-sectional dimensions are also optimized to minimize crosstalk, energy-per-bit and the dynamic delay variation caused by different switching patterns. The optimization process is based on novel compact physical models that are derived in this thesis for latency and crosstalk of co-planar distributed RLC lines. Rigorous physical models are derived for multi-level crosstalk noise that take into account virtually all near as well as inter- and intra-level far aggressors. These models prove that crosstalk remains small and constant in all generations of technology if optimal wire dimensions are utilized. Chip-package co-design methodologies are also developed that show the optimal partition between chip-level and package-level signal and power distribution. Finally, the lengths beyond which optical waveguides can outperform electrical wires in terms of data flux densities are identified for various technology generations.

# Chapter 1

## Introduction and Background

### 1.1 Introduction

Since their invention in 1958, integrated circuits have grown exponentially in terms of number of transistors, speed and functionality. This trend, which is known as Moore's law [1], has survived thanks to the fact that the performance of a transistor improves as its physical dimensions scale down; a smaller transistor runs faster and dissipates less power. On the contrary, interconnect performance degrades as its cross-sectional dimensions scale down; a thinner wire has a larger latency. For instance, as technology advances from 1  $\mu m$  to the 100  $nm$  node, the RC delay of a 1  $mm$  long interconnect devolves from 6 times faster to 20 times slower than the intrinsic delay of a transistor [2]. This turns into a serious problem for inter-macrocell or global interconnects whose lengths do not scale with technology [3, 4]. To avoid prohibitively large interconnect latencies, designers scale down global wire dimensions more slowly than the transistor dimensions [5, 6], and this causes a rapid growth in the gap between transistor density and interconnect density. Thereby, as technology advances, global interconnect resource becomes more and more valuable than the transistor resource. The central hypothesis of this thesis is that the design of GSI chips should shift from transistor-centric to interconnect-centric to facilitate efficient use of valuable on-chip metal levels.

The existing interconnect design methodologies use a pre-defined wiring distribution to determine the required wire dimensions on each metal level which satisfy specified

system performance targets [7, 8, 9, 10]. However, as GSI chips become interconnect-limited, such methodologies cannot optimally exploit the available wiring resource. In contrast, the proposed interconnect-centric approach suggests optimizing wire dimensions without making any assumption about wiring distribution, net-list or placement. Wire dimensions are optimized such that minimum latency and maximum data flux density (bandwidth per unit width of interconnects) are simultaneously achieved. The structure of global interconnects, which are mainly data buses between macro-cells, should then be designed based on the optimal wire dimensions.

As the clock frequency of GSI chips reaches several gigahertz, lengths of many on-chip global interconnects become comparable with the signal wavelength. Thereby, it would be quite unrealistic to ignore the impact of inductance on latency and crosstalk of global interconnects [11]. Because of lack of on-chip ground planes, the classic transmission line theory cannot be directly applied to the modeling of on-chip distributed RLC lines. Numerical solutions and circuit simulations can be used to analyze on-chip RLC lines. However, compact physical models are more insightful, and they can be easily incorporated in the system-level optimization. A major part of this thesis is, therefore, devoted to deriving compact physical models for delay and crosstalk of on-chip distributed RLC lines. These models are then used to optimize the design of global interconnects.

Today's chips and packages are so inter-related that they can no longer be designed independently. A package can partially distribute signal, clock and power across the chip, and by doing so, it can significantly improve the performance of a chip. In this thesis,

optimal partitions between chip-level and package-level signal and power interconnects are identified.

Optical interconnection has been suggested as an alternative for electrical interconnection for many years. The lengths beyond which optical interconnects outperform electrical interconnects in terms of data flux density are identified for various technology generations. The partition lengths are found assuming that optical emitters and detectors will mature enough to become comparable with their electrical counterparts. The partition lengths show that optical interconnects are more promising for chip-to-chip interconnection, and if high-resolution printed boards are available, optical waveguides can outperform almost all typical chip-to-chip wires.

The outline of this thesis is as follows. In Chapter 2, a new interconnect-centric design methodology is proposed that demonstrates how global wire width can be optimized to minimize latency and maximize data flux density simultaneously. The major emphasis of Chapter 2 is on the concept of wire width optimization, and hence in this chapter, interconnects are modeled as simple RLC interconnects above an ideal ground plane. N-coupled RLC lines above an ideal ground plane are modeled in Chapter 3, and co-planar transmission lines above orthogonal lines that are typically used for on-chip global interconnects are modeled in Chapter 4. In Chapter 5, these physical models are used to identify the optimal cross-sectional dimensions of real on-chip global interconnects. In Chapter 6, novel models are presented for crosstalk in a multi-level interconnect network considering all near as well as intra- and inter-level far aggressors. It is also illustrated that by using optimal wire dimensions, crosstalk remains small and constant in all generations of technology. The impact of on-chip global interconnect

optimization on package design is studied in Chapter 7. Chapter 8 discusses optical and electrical interconnection wherein a partition length is identified that shows the lengths beyond which optical interconnects offer a larger data flux density. Finally, conclusions and future work are portrayed in Chapter 9.

## 1.2 Optimal Global Interconnects for GSI

Performance of a high-speed chip is largely affected by both latency and bandwidth of global interconnects, which connect different macrocells. Therefore, one of the important goals is to design high-bandwidth and fast buses that connect a processor and its on-chip cache memory, or link different processors within a multiprocessor chip. The width of global interconnects is optimized to achieve a large “data flux density” and a small latency simultaneously. “Data flux density” is the product of interconnect bandwidth and reciprocal wire pitch, which represents the number of bits per second that can be transferred across a unit length bisectonal line. The optimal wire width, which maximizes the product of data flux density and reciprocal latency, is independent of interconnect length, and can be used for virtually all global interconnects. It is proved that the optimal wire width is the width that results in a delay 33% larger than the time-of-flight. Using the optimal wire width decreases latency, energy dissipation, and repeater area considerably compared to a sub-optimal wire width (e.g. 42% smaller latency, 30% smaller energy-per-bit, and 84% smaller repeater area compared with the  $W_{opt}/2$  case) at the cost of a small decrease in data flux density (e.g. 14% smaller compared with  $W_{opt}/2$  case). A super-optimal wire width, however, causes a slight decrease in latency (e.g. 14% for  $2W_{opt}$ ) at the cost of a large decrease in data flux density (e.g. 35% for  $2W_{opt}$ ).

## 1.3 Compact Physical Models

### 1.3.1 Distributed RLC Lines above an Ideal Ground Plane

To study the impact of far aggressors on board-level interconnects, the set of differential equations of  $n$ -coupled RLC lines above an ideal ground plane is solved rigorously. The models prove that a nearby ground plane is key in controlling far inductive coupling, and the impact of far lines on the worst-case crosstalk is negligible. In other words, in most practical cases, a three RLC line model can accurately predict the worst-case crosstalk. This result is contrary to what is found for on-chip interconnects, and highlights a major distinction between board-level and chip-level interconnects.

### 1.3.2 Co-Planar RLC Lines above Orthogonal Lines

Compact physical models are derived for the delay and crosstalk of on-chip co-planar transmission lines, which are employed in state-of-the-art high-speed microprocessors. These lines are mainly used for long global interconnects that are relatively thick and wide, and have prominent inductive effects. Using the existing models for the periodic structures, it is shown that the orthogonal lines increase the capacitance per unit length of interconnects without changing their inductance per unit length because current cannot return through the orthogonal lines. The wave propagation speed is, therefore, smaller than the speed of light in the interconnect dielectric. Simplified compact expressions are also presented that offer insight and accurate estimation for latency and crosstalk of global interconnects.

These models are then used to optimize the design of co-planar global interconnects. For the case that there are two signal lines between power and ground lines, it is proved

that interconnect latency is minimized if the ratio of signal-ground spacing to signal-signal spacing is equal to 0.45. This optimal spacing ratio is independent of interconnect aspect ratio, and reduces the crosstalk and dynamic delay variation by 38% and 48%, respectively, as compared to the equal spacing case. The optimal wire width that simultaneously maximizes data flux density and minimizes latency is then identified. Using the optimal wire width together with the optimal spacing ratio limits the crosstalk and dynamic delay variation to less than  $0.2V_{dd}$  and 10%, respectively.

### 1.3.3 Multi-Level Crosstalk Noise

For the first time, compact physical models are derived for crosstalk noise of co-planar RLC lines in a GSI chip that simultaneously consider far and near aggressors in both the same metal level and distant metal levels. Since both the amplitude and duration of noise are important, the noise voltage-time integral can be defined as a figure-of-merit for crosstalk, and it is shown that this integral attains its maximum at the length at which interconnect resistance becomes equal to twice characteristic impedance. It is also established that crosstalk can be prohibitively large if interconnects have small resistances. There is, therefore, a trade-off between interconnect latency and crosstalk. Finally, it is proved that using the optimal wire width results in small and constant crosstalk noise in all generations of technology.

## 1.4 Chip-Package Co-Design Methodologies

### 1.4.4 Signal Interconnection

High-quality low-loss transmission lines that are available at the package or board level can be exploited to enhance global signal interconnection. Since on-board interconnects are typically wide and thick, their loss is small, and their latency is ToF limited if they are properly driven. Some of the long global interconnects can be routed through exterconnects (external interconnects), and in this manner, the largest interconnect delay can be reduced that improves the global clock frequency. The main disadvantage of board-level interconnects is their small wiring density that limits the use of exterconnects. An optimal partition is identified between interconnects and exterconnects to achieve the largest possible global clock frequency with minimum number of exterconnects. The partition length is the length of an on-chip interconnect that has a latency equal to the ToF of the longest exterconnect. Knowing that optimal on-chip wire width results in a latency 33% larger than the ToF, the partition length would be equal to 75% of the longest synchronized interconnect.

Using stochastic wire length models, the total length of all nets that should be routed by exterconnects is found. The total additional board layers is then found considering the layers required to route standard signal I/Os and power and ground planes. For a projected chip at the 45 nm technology node, 4 additional board levels and 1800 additional I/Os can improve the global clock to the maximum possible value which is 33% greater than the design with no exterconnects.



### 1.4.5 Power and Ground Interconnection

For a given IR drop budget, there is a trade-off between the number of power and ground pins of a package and the on-chip metal area that should be dedicated to power and ground distribution. Knowing the optimal on-chip wire dimensions and also considering that every global signal line should have a nearby power/ground line as a return path, the required number of power and ground pins is identified for various generations of technology. The results show that for future technology generations, more I/O pins should be dedicated to power and ground due to larger current densities and smaller power supplies.

## 1.5 Optical Interconnection versus Electrical Interconnection

The lengths beyond which board-level optical waveguides are capable of transferring a larger number of bits per second than electrical interconnects are found for various technology generations. Based on the ITRS projections, as technology scales from the 130 *nm* technology node to the 45 *nm* technology node, the partition length falls from 29 *cm* to 8.3 *cm* due to 7 times faster drivers and 25% finer waveguide pitch.

## Chapter 2

### Optimal Global Interconnects for GSI

#### 2.1 Introduction

As CMOS technology advances towards gigascale integration (GSI), the delays of transistors, and local interconnects with negligible resistance scale down. As a result, the local clock frequency is projected to increase significantly [12, 3]. Delay of global interconnects, however, increases with technology scaling because their lengths do not scale down. In fact, since the chip size and the number of transistors are projected to increase, the length and number of global interconnects will increase. Consequently, global interconnects limit the overall performance of a system-on-a-chip [4], and global clock frequency is projected to increase more slowly than local clock frequency. Repeater insertion and reverse scaling are two key solutions to reduce the delay of long interconnects [13]. Using fat wires, however, reduces the wiring density, which may reduce “data flux density”, the product of bandwidth and reciprocal pitch of an interconnect. The data flux density represents the number of bits per second that an interconnect can transfer per unit width. It determines the chip bisectional bandwidth and therefore, the total number of bits per second that global interconnect levels can potentially transfer. It should be noted that bandwidth is as important as latency, for high-performance systems. For instance, embedded memory plays a critical role in the performance of a microprocessor. The ITRS has therefore projected that 90% of the transistors in a GSI chip of the 45 nm technology node will be used as embedded memory

[12]. Two key parameters that define the performance of a cache memory are the number of bits per second that can be transferred between the cache and processor, and its access time [14, 15]. Hence, bandwidth and latency of interconnects that connect the cache and processor are equally important. Likewise, in a multiprocessor chip, different processors should be able to transfer large data packets with a small latency. *The main goal of this chapter is to determine the wire dimensions which offer the best trade-off between latency and data flux density.*

Various approaches to solve the wire sizing and wire layer assignment problems have been described in [7, 8, 9, 10]. All these algorithms assume a pre-defined wiring distribution to determine the required wire dimensions on each metal level which satisfy specified system performance targets. However, as the design of high-performance GSI systems becomes increasingly interconnect-centric [16], new opportunities arise to optimize the global interconnect design without assuming a wiring distribution and before any detailed floor planning or layout is done. A typical GSI chip would have tens of macrocells, with each macrocell consisting of several hundred thousand to a few million gates. This allows greater flexibility in the design of the inter-macrocell global interconnects. For instance, two low bandwidth interconnects can be used instead of one high bandwidth interconnect, or vice versa. Therefore, unlike the local intra-macrocell interconnects, the global interconnect routing need not be restricted to a fixed netlist. *This flexibility permits optimization of the global interconnections for achieving a large data flux density and a small latency.* The key assumptions and design goals of this new methodology are compared to previous works in Table 2.1.

Table 2.1: Key assumptions and design goals of different methodologies for optimizing interconnecting devices.

Methodologies	Assumptions	Goals
Davis [7] & Venkatesan [8]	<ul style="list-style-type: none"> <li>▪ Intra-macrocell interconnect design.</li> <li>▪ A <i>pre-defined</i> homogenous wiring distribution.</li> </ul>	Minimize area, # of layers, clock frequency, or power.
Kahng [9]	<ul style="list-style-type: none"> <li>▪ Intra-macrocell interconnect design</li> <li>▪ A <i>pre-defined</i> wiring distribution.</li> </ul>	Minimize the # of metal levels.
Zarkesh-Ha [10]	<ul style="list-style-type: none"> <li>▪ Inter-macrocell interconnect design.</li> <li>▪ A <i>pre-defined</i> heterogeneous global wiring distribution.</li> </ul>	Meet wiring demand, crosstalk, and bandwidth requirements.
New Approach	<ul style="list-style-type: none"> <li>▪ Inter-macrocell interconnect design.</li> <li>▪ <i>No pre-defined</i> wiring distribution.</li> <li>▪ Flexibility in inter-macrocell interface design.</li> </ul>	Maximize chip bisectonal bandwidth-reciprocal latency product.

The wire width at which the delay of an interconnect is 33% larger than the time-of-flight (ToF) is identified to be the optimal wire width because it maximizes the product of data flux density and reciprocal latency. It is shown that this optimal wire width is determined by the resistivity of the metal and the intrinsic delay of the repeaters, and is independent of the interconnect length, which facilitates its use for all global interconnects. Utilizing the optimal wire width reduces the latency by 42%, the energy dissipation of global interconnects by 30%, and the global repeater area by 84% compared to the case that half the optimal wire width is used. The price is only 14% smaller data flux density. On the other hand, further reduction of latency reduces data flux density, considerably. For instance, increasing the wire width from  $W_{opt}$  to  $2W_{opt}$  decreases latency by only 14%, but data flux density reduces by 35%.

The major emphasis of this chapter is the concept of wire width optimization, and hence interconnects are modeled as simple RLC lines above an ideal ground plane. In Chapter 4, compact physical models are derived for co-planar transmission lines that are

typically utilized for global interconnects. These models are then used in chapter 5 to optimize on-chip global interconnects. Also, in this chapter, all cross-sectional dimensions are considered to be equal, and the aspect ratio and the spacing between interconnects are optimized in Chapter 5.

The impact of wire width on latency and data flux density is shown in Section 2.2, in which the optimal wire width is introduced as the design with the best trade-off between latency and data flux density. The utility of using the optimal wire width in reducing the power consumption, repeater area, and via blockage is demonstrated in Section 2.3. The conclusions are summarized in Section 2.4, and the rigorous derivation of the optimal wire width is provided in the Appendix A.

## 2.2 Optimal On-Chip Wire Width

### 2.2.1 Impact of Wire Width on Latency

The delay of an interconnect with optimal number and size of repeaters using the RC model is [13]

$$\tau = 2.5 \frac{\ell}{W} \sqrt{\xi \rho \epsilon_r \epsilon_0 R_0 C_0}, \quad (2.1)$$

where  $\ell$  and  $W$  are the wire length and width respectively,  $\rho$  is the metal resistivity,  $R_0$  and  $C_0$  are the output resistance and input capacitance of a minimum size repeater,  $\epsilon_r \epsilon_0$  is the dielectric permittivity, and  $\xi$  is a dimensionless constant determined by the wire geometry and it is given by

$$\xi \equiv \frac{r W^2}{\rho} \times \frac{c}{\epsilon_r \epsilon_0}, \quad (2.2)$$

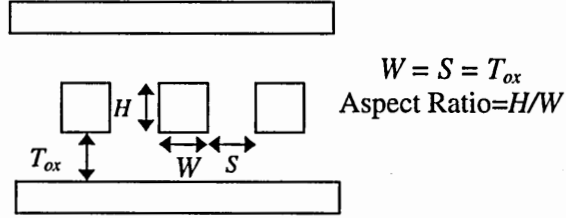


Figure 2.1: The cross-section of a wire and its neighbors.

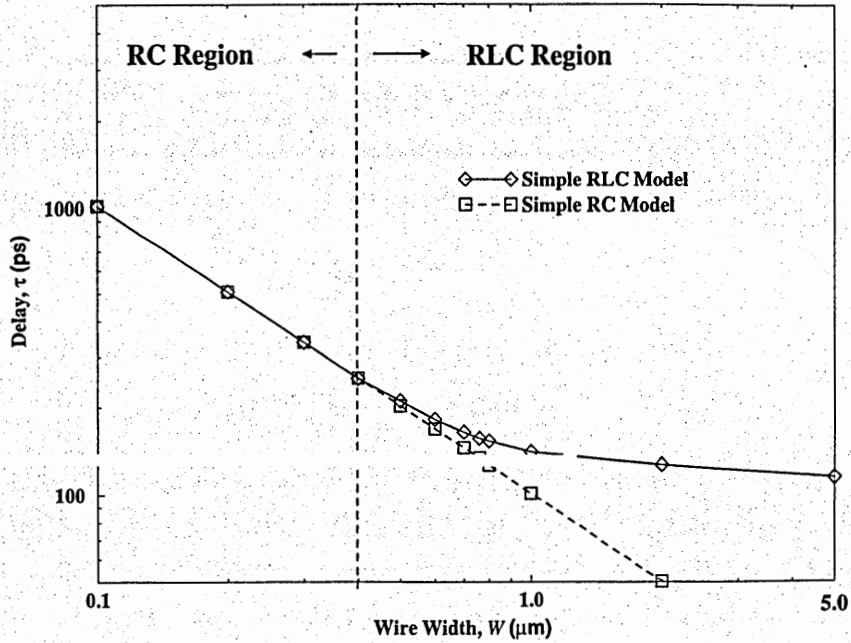


Figure 2.2: Time delay of a 24 mm long interconnect with optimal repeaters versus wire width. The cross-section of the wire is shown in Figure 2.1.

where  $r$  and  $c$  are the resistance and capacitance per unit length. It should be noted that  $\xi$  is determined by the geometry of the wires and is independent of the resistivity of the metal or the dielectric constant. For instance, if the metal resistivity  $\rho$  changes, the resistance per unit length  $r$  changes proportionally, and  $\xi$  remains unchanged. Wire width  $W$ , spacing  $S$ , and dielectric thickness  $T_{ox}$  are assumed to be equal, as shown in Figure

2.1. For the unity aspect ratio ( $H/W = 1$ ),  $\xi$  is found to be 6.05 using RAPHAEL [17] for extracting the capacitance per unit length. The dashed line in Figure 2.2 shows the time delay of an interconnect versus the wire width using the RC model described by (2.1). All technology parameters are projections of the International Technology Roadmap for Semiconductors (ITRS) [12] for the year 2010 (45nm node).

Increasing the wire width decreases the RC model delay by the same ratio, as shown in Figure 2.2. Note that this is true only if all dimensions are simultaneously scaled larger ( $W=S=T_{ox}=H$ ); Otherwise, by only increasing wire width, for example, the delay will not decrease proportionally because a wider line will have a larger capacitance which negates the resistance reduction. Since the metal and the dielectric thicknesses of each interconnect layer have to be constant across the chip, the wire dimensions should be optimized for the whole layer, and it is not feasible to optimize them for each interconnect separately.

Although (2.1) is valid for a small  $W$ , the impact of inductance cannot be neglected for large values of  $W$ . Optimal number and size of repeaters are determined considering the impact of inductance [18]. Assuming that these optimal values are used, the RLC model delay of an interconnect versus the wire width is shown in Figure 2.2. The skin effect is also considered when the wire dimensions are comparable with the skin depth. The skin depth is calculated for the third harmonic of the global clock frequency.

It can be inferred from Figure 2.2 that as long as the RC model is valid, increasing the wire width decreases the delay by the same ratio. However, when the wire width is large enough so that the RC model deviates from the RLC model, increasing the wire width does not decrease the delay proportionally.

### 2.2.2 Impact of wire width on bandwidth

Data flux density (per unit width) is defined as the product of bandwidth and reciprocal pitch of an interconnect, and represents the number of bits per second that can be transferred across a unit length bisectional line. The data flux density can be written as

$$\Phi_D(W) \equiv \frac{Bw}{P} = \frac{1/\tau}{2W}, \quad (2.3)$$

where  $Bw$  is the bandwidth of an interconnect and  $P$  is the wiring pitch. It has been assumed that  $Bw$  is determined by the reciprocal latency and the spacing between wires is equal to the wire width. It will be shown here that maximizing the data flux density, maximizes the “chip bisectional bandwidth,” and therefore, the total number of bits per second that can be transferred through the global interconnects. One level of the global interconnects is shown in Figure 2.3.a, where all interconnects are assumed to have a length equal to the chip edge dimension,  $D_{chip}$ . The bisectional bandwidth of the layer (the number of bits per second that pass through the dashed line of length  $D_{chip}$ ) is

$$Bw_{bisec} = D_{chip} \Phi_D; \quad (2.4)$$

therefore, maximizing  $\Phi_D$  maximizes the bisectional bandwidth. This is also true when interconnects have different lengths because in general, the bisectional bandwidth is (Figure 2.3.b)

$$Bw_{bisec} = \int_0^{D_{chip}} \Phi_D(x) dx. \quad (2.5)$$



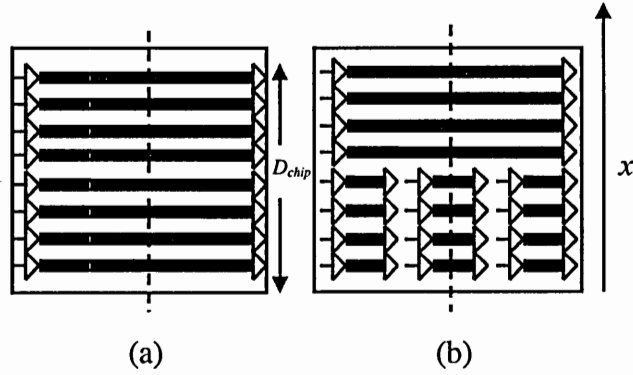


Figure 2.3: A layer of global interconnects. (a) All interconnects have the same length. (b) Length of interconnects are different.

For the case that the global communication is synchronized with a global clock, the delay of the longest interconnect determines the clock cycle and the data flux density is the same for interconnects with different lengths. In the asynchronous case, the data flux density can be different. In all cases, however, maximizing the data flux density maximizes the chip bisectional bandwidth or the total bandwidth regardless of the net length distribution. Note that the dashed line in Figure 2.3 is an imaginary line that can be positioned at any location on the die.

The data flux density,  $\Phi_D$ , is plotted against the wire width in Figure 2.4. It can be seen that as long as the RC model is valid (the difference between the RC and RLC models is small),  $\Phi_D$  is almost constant. This is reasonable since  $W$  is canceled out in (2.3) when  $\tau$  is substituted by (2.1). In the deep RLC region, however,  $\Phi_D$  drops rapidly as  $W$  increases.

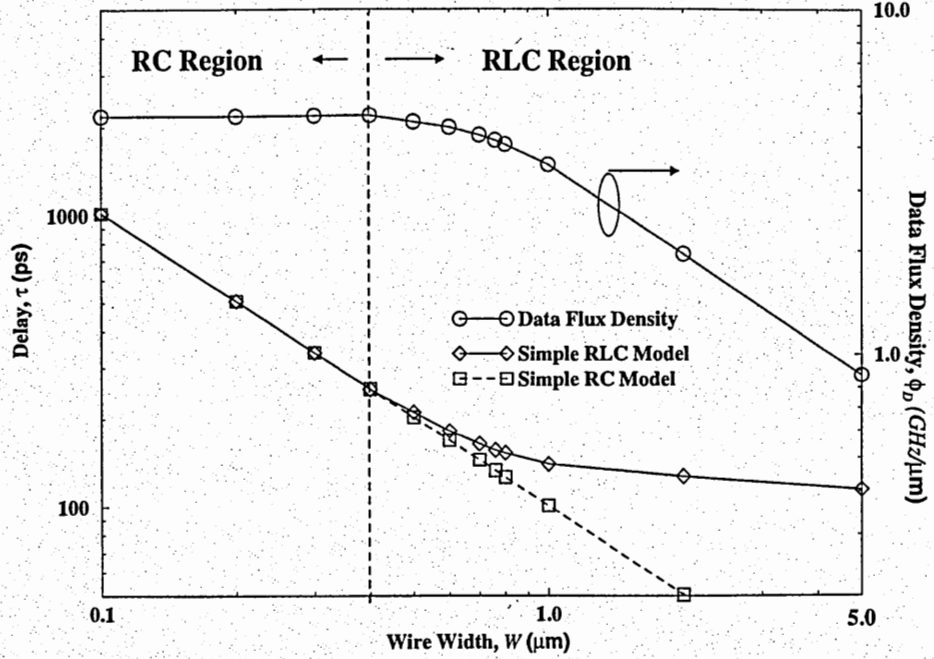


Figure 2.4: Data flux density versus wire width for a 24 mm long interconnect. Interconnect delay is also plotted. Data-flux density is constant in the RC regime and in the RLC regime, it drops as wire width decreases.

### 2.2.3 Optimal Wire Width

The quintessential purpose of an interconnect is communication between distant points with small latency [2]. It is therefore desirable to transfer as many bits as possible per unit time through the interconnects with small latency. In other words, large data flux density and small latency are simultaneously desired. Assuming that latency and data flux density are equally important, the data flux density-reciprocal latency product,  $\Phi_D/\tau$ , can be defined as the figure of merit, and optimal wire width would be the width at which  $\Phi_D/\tau$  is maximized. Using (2.3), this figure-of-merit can be written as

$$\frac{\Phi_D}{\tau} = \frac{1}{2W\tau^2}. \quad (2.6)$$

This new metric is plotted against wire width in Figure 2.5. It is shown that  $\Phi_D/\tau$  attains its maximum in the “shallow” RLC region, where the difference between the RC and RLC model delays is 13%. At this design point, data flux density is only 14% smaller than its maximum value and latency is 33% larger than its minimum value (1.33ToF).

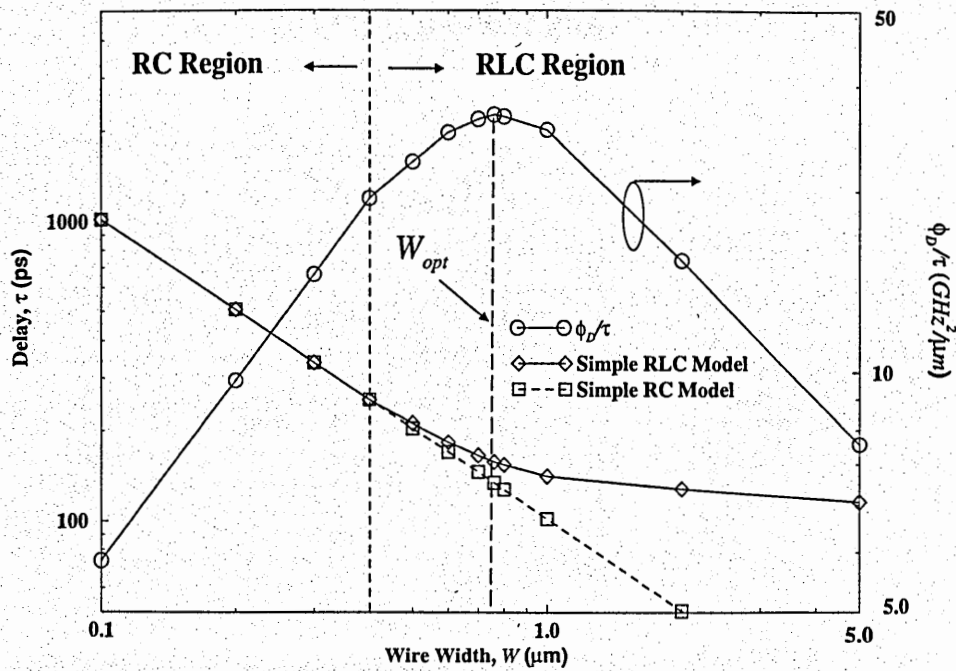


Figure 2.5: Data flux density-reciprocal latency product versus the wire width for a 24 mm long interconnect with optimal repeaters.

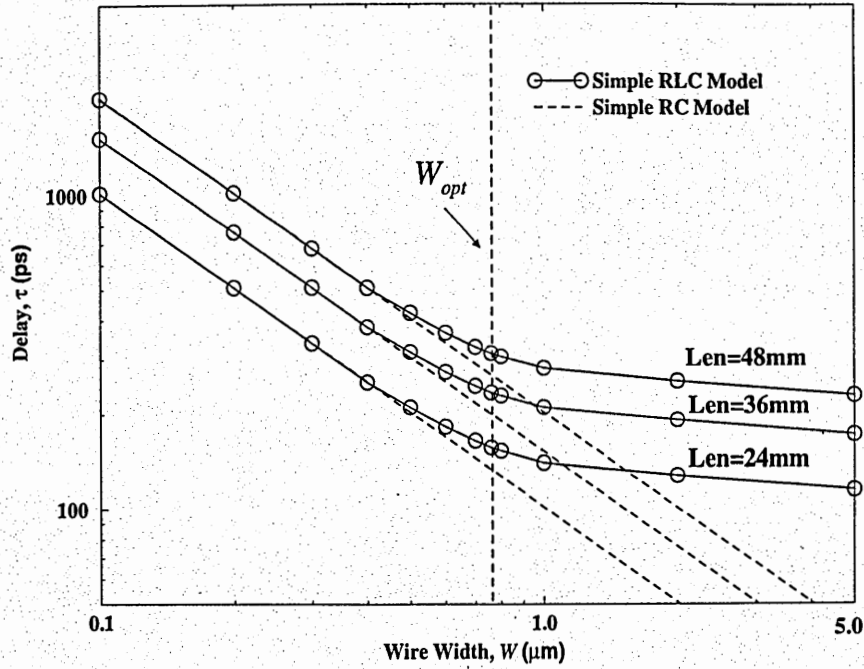


Figure 2.6: Time delay versus the interconnect width for different interconnect lengths. Optimal wire width is independent of length.

The optimal wire width is independent of interconnect length as can be seen from Figure 2.6 in which the RC and RLC model delays of three interconnects with different lengths (48, 36, and 24 mm) are plotted. This can be explained qualitatively by observing that in both RC and RLC regimes, interconnect delay is linearly proportional to interconnect length when optimal repeaters are used. Hence,  $1/\tau^2 W$  for interconnects with different lengths differs by a constant factor ( $\ell_1^2 / \ell_2^2$ ). Its maximum, however, occurs at the same wire width. This important fact facilitates using the optimal wire width for all global interconnects regardless of their lengths.

It is rigorously proved that the optimal wire width happens when the RC model delay becomes equal to 1.18 times the ToF delay (proof is in the Appendix A). This fact was also verified through HSPICE simulations for varying values of interconnect lengths,

aspect ratios and  $R_0$ ,  $C_0$  values. By substituting the above mentioned condition in (2.1), the optimal wire width is found to be

$$W_{opt} = 2.12c_0\sqrt{\xi\rho\epsilon_0R_0C_0}, \quad (2.7)$$

where  $c_0$  is the speed of light in free space. Equation (2.7) shows that the optimal wire width,  $W_{opt}$  depends solely on the resistivity of metal  $\rho$ , the intrinsic delay of a repeater  $R_0C_0$ , and the geometry of the wires ( $\xi$  factor).

The optimal wire widths for different generations of technology and aspect ratios are shown in Figure 2.7 based on the ITRS projections for  $\rho$  and  $R_0C_0$ . The intrinsic delay of a repeater,  $R_0C_0$  is found assuming equal pFET and nFET current drive capabilities.

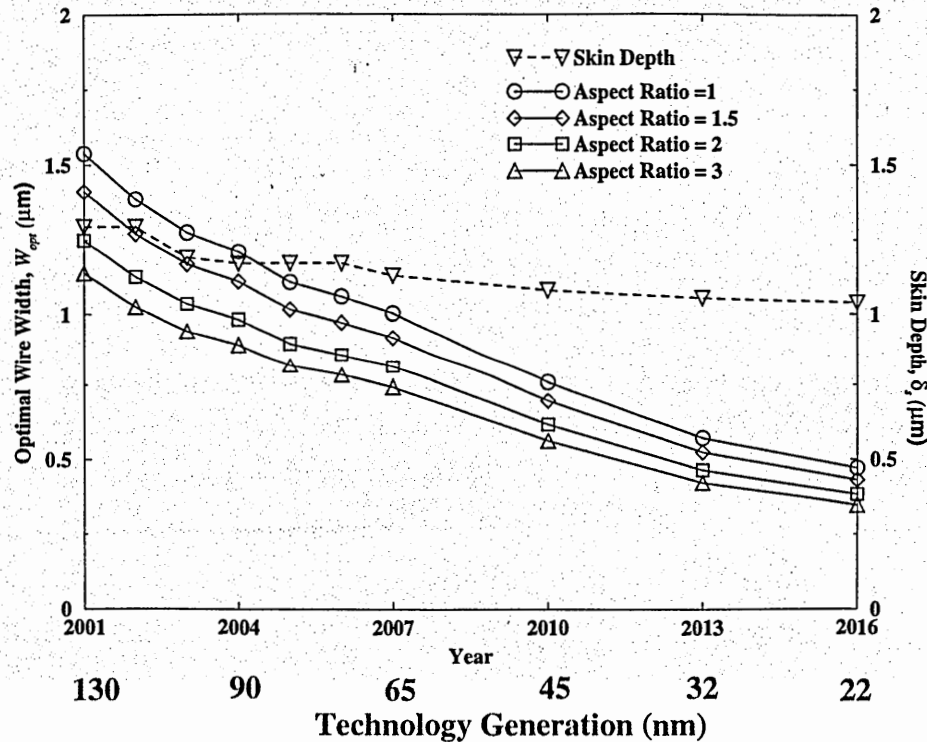


Figure 2.7: Optimal wire width for different generations [12] and aspect ratios. Using optimal wire width makes the skin effect negligible. The skin depth is calculated for the third harmonic of the global clock frequency.

Skin effect can be neglected if the wire dimensions are smaller than three times the skin depth [19]. Therefore, as shown in Figure 2.7 for all technology generations, using the optimal wire width makes the skin effect negligible and utilizes the interconnect metal efficiently.

Knowing the value of this optimal wire width can help to design an efficient interface between different logic and memory cores in a GSI chip. The optimal wire width can be found at the early stage of the design because its value is determined by two basic parameters—resistivity of metal and intrinsic delay of repeaters. In this way, the structure of global interconnects can be designed such that the global interconnect levels are used as efficiently as possible. For instance, designers can use an  $n$ -bit data bus with a wire width of  $W_{opt}$  between the processor and the cache memory rather than a  $(2n)$ -bit data bus with  $W_{opt}/2$  wire thickness, or an  $(n/2)$ -bit data bus with  $2W_{opt}$  wire width. The wiring area required in all three cases is the same. However, the  $2n$ -bit bus has a large latency (72% larger than the optimal case) and the  $(n/2)$ -bit bus has a small bandwidth (35% smaller than the optimal case). The global clock frequency can also be optimized by knowing the delay of the longest synchronized interconnect, which has a width equal to the optimal wire width. Operating at a global clock frequency larger than this value requires a wire width larger than the optimal value, which reduces the bisectional bandwidth; on the other hand, using a smaller global clock frequency increases the latency of all the interconnects, and reduces the system performance.

## 2.3 Optimal On-Chip Wire Width

Repeaters are widely used for global interconnects to increase the wiring density and reduce the delay. There are three important concerns about repeaters: power consumption, the required silicon area, and the via blockage. This section investigates the impact of using the optimal wire width on these three issues.

### 2.3.1 Power Consumption

The average energy consumed in each interconnect switching event can be approximated by

$$E_b = \frac{1}{2}(C_{int} + C_{rep})V_{dd}^2. \quad (2.8)$$

where  $C_{int}$  is the capacitance of the interconnect and  $C_{rep}$  is the total capacitance of the repeaters in that interconnect. The capacitance of each repeater is linearly proportional to its  $W/L$  ratio, therefore, assuming that  $W/L$  ratio of repeaters is  $h$  times larger than that of a minimum size repeater, the total capacitance of all repeaters inserted along an interconnect is

$$C_{rep} = hkC_0, \quad (2.9)$$

where  $k$  is the number of repeaters inserted along the interconnect. In the RC regime, the optimal size of repeaters is [13]

$$h_{opt} = \sqrt{\frac{R_0 C_{int}}{C_0 R_{int}}}, \quad (2.10)$$

and the optimal number of repeaters is

$$k_{opt} = \sqrt{\frac{0.4R_{int}C_{int}}{0.7C_0R_0}}. \quad (2.11)$$

Hence, when optimal repeaters are used the total capacitance of the repeaters can be written as

$$C_{rep} = 0.75C_{int}. \quad (2.12)$$

The optimum RLC repeater models should be used when [18]

$$1.33\sqrt{\frac{R_0C_0R_{int}}{C_{int}}} < Z_0, \quad (2.13)$$

and the optimal size and number of repeaters in the RLC regime are given by [18]

$$h_{opt} = 1.15\frac{R_0}{Z_0}, \quad (2.14)$$

and

$$k_{opt} = 0.95\frac{R_{int}}{Z_0}. \quad (2.15)$$

$Z_0$ , the characteristic impedance of the line, is defined as

$$Z_0 = \sqrt{\frac{l}{c}}, \quad (2.16)$$

where  $l$  is the inductance per unit length. Assuming that the propagation speed is equal to the speed of light in the dielectric, the following relation can be derived [20]:

$$\frac{1}{\sqrt{lc}} = \frac{c_0}{\sqrt{\epsilon_r}}, \quad (2.17)$$

where  $c_0$  is the speed of light. Using (2.16) and (2.17),  $Z_0$  can be written as

$$Z_0 = \frac{\sqrt{\epsilon_r}}{c_0} \times \frac{1}{c}. \quad (2.18)$$



In this manner, the total capacitance of the repeaters used for an interconnect operating in the RLC regime is

$$C_{rep} = \frac{1.09 R_0 C_0 c_0^2 (rc)(\ell c)}{\epsilon_r}. \quad (2.19)$$

By using (2.2), the  $rc$  product can be written in terms of the geometry coefficient  $\xi$  as

$$rc = \xi \epsilon_r \epsilon_0 \frac{\rho}{W^2}. \quad (2.20)$$

Substituting (2.20) in (2.19) gives

$$C_{rep} = \frac{1.09 R_0 C_0 c_0^2 \xi \rho \epsilon_0}{W^2} C_{int}. \quad (2.21)$$

Using (2.7), (2.21) can be simplified to

$$C_{rep} = 0.234 \frac{W_{opt}^2}{W^2} C_{int}. \quad (2.22)$$

The capacitance of an interconnect is independent of the wire width because it has been assumed that all cross-sectional dimensions are equal or proportional to the width (constant aspect ratio). Hence, as (2.22) shows, in the RLC regime, making wires fatter reduces the capacitance of the interconnect repeaters, thereby, reduces the power dissipation of the repeaters. By comparing (2.12) and (2.22), it can be found that at the optimal wire width, the power consumed by the repeaters is 67% smaller than the RC case, and the energy per bit, which is given by (2.8), is reduced by 30%. This is shown in Figure 2.8, which plots the consumed energy per bit in a 24 mm long interconnect. Figure 2.8 also shows that the optimal wire width offers a good trade-off between bandwidth and power dissipation by plotting data flux density-reciprocal energy per bit product  $\Phi_D/E_b$ .

Compared to the optimal wire width, at  $W_{opt}/2$  and  $2W_{opt}$ ,  $\Phi_D/E_b$  is smaller by 20% and 30%, respectively.

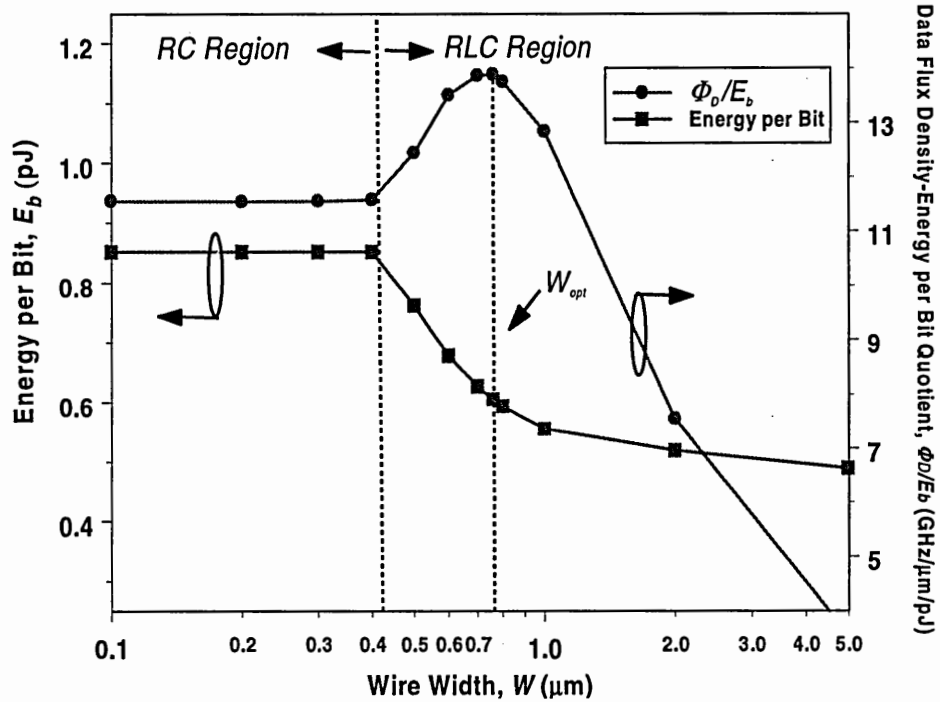


Figure 2.8: The required energy to transfer one bit of data along a 24 mm long interconnect versus the wire width. At the optimal wire width,  $\Phi_D/E_b$  is maximized.

It should be noted that although at the optimal wire width, the RC and RLC model delays are close, the optimal number of repeaters suggested by the RLC model is considerably smaller than that suggested by the RC model. For instance, for a 24 mm long interconnect in the 45 nm generation, the optimal number of repeaters predicted by the RC model is 35 whereas the RLC model predicts 19 repeaters. The power dissipation at the optimal wire width is therefore smaller than that of the RC regime.

### 2.3.2 Repeater Area

Optimal repeater insertion in the RC regime requires a significant number of large repeaters, which consume a large silicon area and increase the via blockage. Some authors have, therefore, suggested tolerating a larger delay by *adequate* or *sub-optimal* repeater insertion instead of optimal repeater insertion [3, 8, 21, 22]. It will be shown here that the optimal wire width is an alternative solution to reduce the repeater demand with no performance loss.

The area occupied by each repeater with the standard cell model is [23]

$$A_{inv} = K_I \left[ 1 + \frac{(1+\beta)(h-1)}{\sqrt{k_I G_{ar}}} \right] F^2, \quad (2.23)$$

where  $k_I$  and  $G_{ar}$  are the minimum sized inverter footprint area and aspect ratio, respectively.  $F$  is the minimum feature size, and  $\beta$  is the ratio of pFET and nFET sizes. In a custom cell,  $K_I$  is 102 and  $G_{ar}$  is 17/6 and a typical value for  $\beta$  is 2 so that the source and sink currents are approximately equal. Using (2.23) and the equations for optimal number and size of repeaters in the RC and RLC regions the required repeater area for a 24 mm long interconnect is plotted versus the wire width in Figure 2.9. In the RC regime, the repeater area is independent of the wire width and in the RLC regime it decreases with the wire width. At the optimal wire width, the required area is 3 times smaller than that of the RC regime. Further reduction of the repeater area, however, requires a considerable sacrifice in wiring density. The total repeater area that one pair of global wiring layers needs can be found assuming that the wiring efficiency is 50%. For a 570 mm<sup>2</sup> chip, which is projected for the 45 nm node [1], 3.78 mm<sup>2</sup> of silicon area is required for the repeaters if the optimal wire width (0.76  $\mu$ m) is used. In contrast, if a wire width

of  $W_{opt}/2$  ( $0.38 \mu\text{m}$ ) is used, then a silicon area of  $22.9 \text{ mm}^2$  is required for the repeaters of the global interconnect levels. Such a large increase in the repeater area is due to 2x increase in the wiring density and 3x increase in the repeater area per unit length of an interconnect. It is worthwhile to note that, in both cases, the global interconnect levels transfer almost the same number of bits per second.

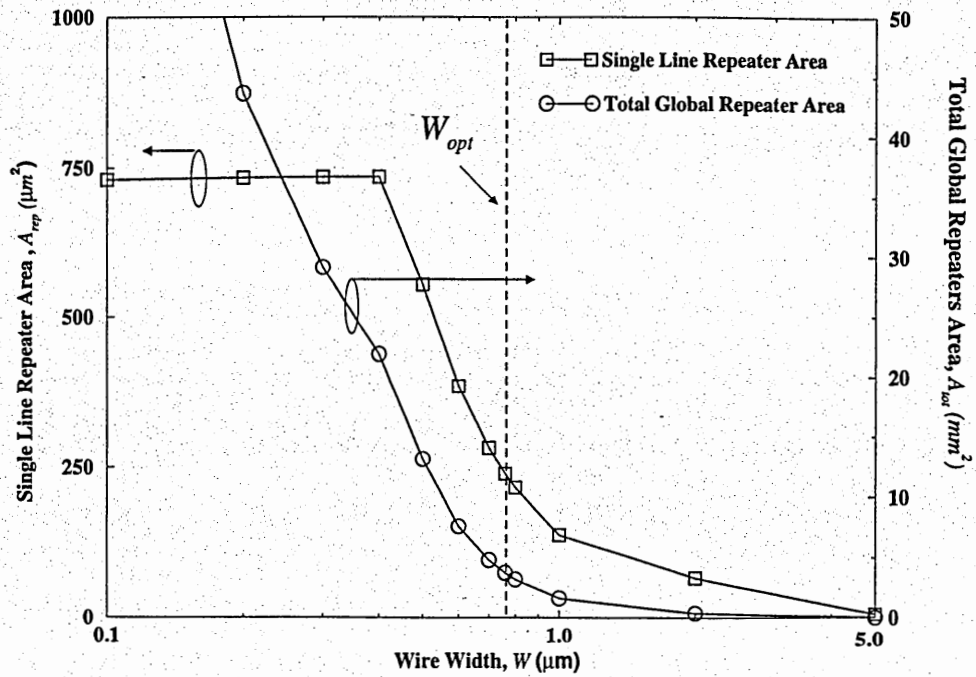


Figure 2.9: The required silicon area for a 24 mm long interconnect and the total silicon area for global repeaters versus the wire width. The Chip area is assumed to be  $572 \text{ mm}^2$ , and two levels of metal with a wiring efficiency of 0.5 are considered for global interconnects.

### 2.3.3 Via Blockage

Each repeater needs two vias to be connected to the wires in the global levels, which cause via blockage for all other metal levels since global wires are on top metal levels.

For a given metal level, the via blockage factor is defined as [24]

$$B_v = A_v / A_{chip}, \quad (2.24)$$

where  $A_v$  is the unused wiring area due to vias and  $A_{chip}$  is the chip area. The via blockage factor  $B_v$  can be estimated by [24]

$$B_v = \sqrt{N_v(2W + s\lambda)^2 / A_{chip}}, \quad (2.25)$$

where  $N_v$  is the number of vias passing through the metal level,  $s$  is the via covering factor, and  $\lambda$  is the layout rule unit. As (2.25) shows the wasted area due to via blockage is proportional to the square root of the number of vias. Figure 2.10 shows the number of vias required for a pair of global interconnect levels versus the wire width for the same projected chip implemented in 45 nm technology. It can be seen that the number of vias at the optimal wire width is 86% smaller than that of  $W_{opt}/2$  wire width.

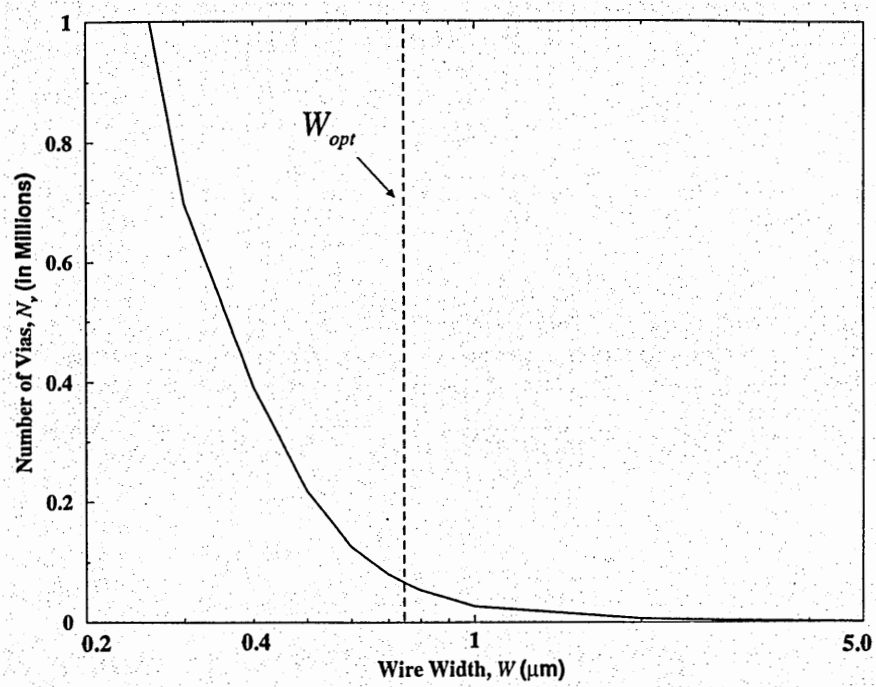


Figure 2.10: Number of vias required for the repeaters associated with a pair of global interconnect levels versus the wire width.

In summary, the optimal wire width not only maximizes the data flux density-reciprocal latency product, but also reduces the power consumption, the repeater area, and the via blockage considerably. The important parameters corresponding to  $W_{opt}/2$ ,  $W_{opt}$ , and  $2W_{opt}$  are summarized in Table 2.2.

Table 2.2. Impact of using optimal wire width on a 24 mm-long interconnect in a projected ASIC chip at the 45-nm node of technology. The chip area is 572mm<sup>2</sup>. The global repeater area and number of vias correspond to two global metal levels with a wiring efficiency of 0.5.

Parameters	$W_{opt}/2$ (0.38 $\mu\text{m}$ )	$W_{opt}$ (0.76 $\mu\text{m}$ )	$2W_{opt}$ (1.52 $\mu\text{m}$ )
Latency, $\tau$ (ps)	266	156	133
Data flux density, $\Phi_D$ (GHz/ $\mu\text{m}$ )	4.9	4.2	2.45
Data flux density- reciprocal latency, $\Phi_D/\tau$ ((GHz) <sup>2</sup> / $\mu\text{m}$ )	18.56	26.8	18.44
Energy per bit, $E_b$ (pJ)	0.85	0.60	0.535
Global repeater area, $A_{rep}$ (mm <sup>2</sup> )	23	3.75	1.0
Number of vias due to global repeaters, $N_V$	$412 \times 10^3$	$62 \times 10^3$	$16 \times 10^3$

## 2.4 Conclusions

An optimal on-chip wire width is identified for GSI global interconnects which offers the best trade-off between data flux density and latency. At this optimal point, the data flux density-reciprocal latency product is maximized so that the global interconnects can transfer as many bits as possible with low latency. The value of the optimal width is independent of the interconnect length and is determined by the resistivity of the metal, the intrinsic delay of the repeaters, and the wire geometry. Therefore, a unique optimal wire width can be used for virtually all global interconnects regardless of their lengths. Using the optimal wire width results in a 42% smaller latency, 30% smaller energy-per-bit, and 84% smaller repeater area at a cost of only a 14% decrease in data flux density, compared to using half the optimal wire width (sub-optimal design). The via blockage at

the optimal design point is also less severe than the half optimal wire width case because of 86% fewer global repeaters. On the other hand, using twice the optimal wire width (super-optimal design) results in only a 14% decrease in latency at the cost of a 35% decrease in data flux density, compared to the optimal wire width design.



## Chapter 3

### N-Coupled RLC Lines above an Ideal Ground Plane

#### 3.1 Introduction

Signal integrity is one of the major issues that designers have to deal with. Although SPICE simulations or numerical calculations can be used to calculate the noise voltage, having compact physical models is more insightful and enables system-level optimizations more easily. Sakurai has modeled crosstalk in distributed RC interconnects [25]. As signal rise time scales down with technology, inductance can no longer be neglected. Davis et al. have rigorously solved the transient voltage of a distributed RLC line [26]. For two and three coupled lines above a ground plane, Davis et al. have used linear transformations to decouple the equations, and then used the single line solution to find crosstalk caused by near aggressors [27].

In this chapter, the methods used in [26, 27] are extended to solve n-coupled distributed RLC lines above a ground plane. It is rigorously proved that the impact of far lines on worst-case crosstalk is negligible if there is a nearby ground plane. This shows the importance of having a nearby ground plane because as it will be shown in Chapters 4 and 6, far inductive noise can be quite large for on-chip interconnects that have no nearby ground plane.

In this chapter, the models for the transient voltage of single, two and three distributed RLC lines are discussed in Section 3.2. A general methodology to solve n-

coupled RLC lines with ideal return paths is presented in Section 3.3. The models for five coupled lines are verified against HSPICE simulations in Section 3.4, and it is shown that far aggressors have a negligible impact on the worst-case crosstalk. Finally the key results are summarized in 3.5

### 3.2 Single, Two and Three Coupled RLC Lines

The differential equation for a single RLC conductor is

$$\frac{\partial^2}{\partial x^2} V(x, t) = rc \frac{\partial}{\partial t} V(x, t) + lc \frac{\partial^2}{\partial t^2} V(x, t), \quad (3.1)$$

where  $l$  is inductance per unit length,  $c$  is capacitance per unit length and  $r$  is resistance per unit length. Davis et al. have rigorously solved the problem when interconnects are open-ended and are stimulated by a step input [26]. The output waveform can be written in terms of a generating function which has the following form

$$V_{gen}(x, t, m) = V_{dd} \left[ \frac{Z_0}{Z_0 + R_r} \left( \frac{t - x\sqrt{lc}}{t + x\sqrt{lc}} \right)^{\frac{m}{2}} e^{\frac{Rr}{2L} I_0 \left[ \frac{r}{2l} \sqrt{t^2 - (x\sqrt{lc})^2} \right]} + \frac{1}{2} \sum_{k=1}^{\infty} \left( \frac{t - x\sqrt{lc}}{t + x\sqrt{lc}} \right)^{\frac{k}{2}} e^{\frac{Rr}{2L} I_0 \left[ \frac{r}{2l} \sqrt{t^2 - (x\sqrt{lc})^2} \right]} (4 - \Gamma^{k-1} (\Gamma + 1)^2) \right] u(t - x\sqrt{lc}), \quad (3.2)$$

where  $Z_0$  is the characteristic impedance of the line,  $I_k$  is the modified Bessel function of the order of  $k$ , and  $\Gamma$  is the reflection coefficient at the driver side  $(\frac{R_r - Z_0}{R_r + Z_0})$ . The

transient voltage at the position  $x$  along a semi-infinite line is

$$V_{inf}(x, t) = V_{gen}(x, t, m = 0), \quad (3.3)$$

and the voltage at the end of a finite line can be calculated taking into account all waves reflected from the source and load boundaries as

$$V_{fin}(l, t) = 2V_{gen}(x = l, t, m = 0) + 2e^{\frac{R}{2L}t} \sum_{n=1}^q \sum_{i=0}^n \sum_{j=0}^{\infty} \frac{n(n-1+j)!}{i!j!(n-1)!} (-1)^j \Gamma^{n-i+j} V_{gen}(x = (2n+1)l, t, m = i+j). \quad (3.4)$$

Davis et al. have used the single line solution to calculate the voltage waveform of two and three coupled RLC lines. For instance, for the three coupled RLC lines, the set of differential equations can be written in matrix form as

$$\frac{\partial^2}{\partial x^2} \begin{bmatrix} V_1(x, t) \\ V_2(x, t) \\ V_3(x, t) \end{bmatrix} = r \begin{bmatrix} c_g + c_m & -c_m & 0 \\ -c_m & c_g + 2c_m & -c_m \\ 0 & -c_m & c_g + c_m \end{bmatrix} \frac{\partial}{\partial t} \begin{bmatrix} V_1(x, t) \\ V_2(x, t) \\ V_3(x, t) \end{bmatrix} + \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \begin{bmatrix} V_1(x, t) \\ V_2(x, t) \\ V_3(x, t) \end{bmatrix}. \quad (3.5)$$

Assuming that outer lines switch in-phase (means  $V_1$  and  $V_3$  are equal) and adding 2 times of the first row to the second row, the following equation is obtained

$$\frac{\partial^2}{\partial x^2} (2V_1 + V_2) = rc_g \frac{\partial}{\partial t} (2V_1 + V_2) + \frac{1}{v^2} \frac{\partial^2}{\partial t^2} V (2V_1 + V_2), \quad (3.6)$$

which represents one of the propagation modes. If the second row is subtracted from the first row, another mode can be described as

$$\frac{\partial^2}{\partial x^2} (V_1 - V_2) = r(c_g + 3c_m) \frac{\partial}{\partial t} (V_1 - V_2) + \frac{1}{v^2} \frac{\partial^2}{\partial t^2} V (V_1 - V_2). \quad (3.7)$$

These two equations are exactly in the form of (3.1) with different coefficients. Hence they can be solved by using the single line solution as

$$(2V_1 + V_2)(x, t) = V_{fin}(x, t, c = c_g, l = \frac{1}{c_g v^2}), \quad (3.8)$$

and

$$(V_1 - V_2)(x, t) = V_{fin}(x, t, c = c_g + 3c_m, l = \frac{1}{(c_g + 3c_m)v^2}). \quad (3.9)$$

Having the solutions for these two modes, voltage of each line can be found by

$$V_1(x, t) = \frac{1}{3}[(2V_1 + V_2)(x, t) + (V_1 - V_2)(x, t)], \quad (3.10)$$

and

$$V_2(x, t) = \frac{1}{3}[(2V_1 + V_2)(x, t) - 2(V_1 - V_2)(x, t)]. \quad (3.11)$$

### 3.3 Five or More Coupled RLC Lines

While for two and three coupled lines it is possible to identify the propagation modes by inspecting the set of differential equations, for a larger number of interconnects it can be quite difficult to do so. The three coupled RLC solution, is useful to find the crosstalk induced by near aggressors (one aggressor on each side). To find the impact of far aggressors, however, more interconnects should be taken into account. In this section, a general method is presented to decouple the set of differential equations for n-coupled RLC lines, and in this way, the five-coupled RLC line case is solved. The solution for five conductors shows that the impact of far aggressors on the worst case crosstalk is negligible if there is a nearby ground plane.

The set of differential equations for n-coupled RLC lines can be written as

$$\frac{\partial^2}{\partial x^2}[V(x, t)] = r[C]\frac{\partial}{\partial t}[V(x, t)] + \frac{1}{v^2}\frac{\partial^2}{\partial t^2}[V(x, t)]. \quad (3.12)$$

The set of differential equations described by (3.12) can be decoupled by finding the eigenvectors of the capacitance matrix [28]. If  $M$  is an  $n \times n$  matrix and there is an  $n \times 1$

matrix  $X$  such that  $MX=\lambda X$  where  $\lambda$  is a scalar,  $X$  is called an eigenvector of  $M$ , and  $\lambda$  is the eigenvalue corresponding to  $X$ .

Assuming that  $[V_\lambda]$  is an eigenvector of the capacitance matrix  $[C]$ , with an eigenvalue of  $\lambda$ , by multiplying both sides of (3.12) by  $[V_\lambda]^T$ , (3.12) can be written as

$$\frac{\partial^2}{\partial x^2} [V_\lambda]^T [V] = r [V_\lambda]^T [C] \frac{\partial}{\partial t} [V] + \frac{1}{v^2} \frac{\partial^2}{\partial t^2} [V_\lambda]^T [V]. \quad (3.13)$$

Since  $[C]$  is a symmetrical matrix, it is equal to its transpose,  $[C] = [C]^T$ , and therefore,

$$[V_\lambda]^T [C] = ([C]^T [V_\lambda])^T = ([C][V_\lambda])^T = \lambda [V_\lambda]^T, \quad (3.14)$$

and, (3.13) can be simplified to

$$\frac{\partial^2}{\partial x^2} [V_\lambda]^T [V] = r\lambda \frac{\partial}{\partial t} [V_\lambda]^T [V] + \frac{1}{v^2} \frac{\partial^2}{\partial t^2} [V_\lambda]^T [V] \quad (3.15)$$

where  $[V_\lambda]^T [V_\lambda]$  is a  $1 \times 1$  matrix which means that (3.15) is exactly like the differential equation of a single RLC line with different coefficients. In other words, (3.15) describes one of the propagation modes of signal in  $n$ -coupled distributed RLC lines. Hence, solving the problem of  $n$ -coupled RLC line can be done by finding the eigenvectors and eigenvalues of the capacitance matrix.

For the five coupled RLC line case, the capacitance can be written as

$$[C] = \begin{bmatrix} c_g + c_m & -c_m & 0 & 0 & 0 \\ -c_m & c_g + 2c_m & -c_m & 0 & 0 \\ 0 & -c_m & c_g + 2c_m & -c_m & 0 \\ 0 & 0 & -c_m & c_g + 2c_m & -c_m \\ 0 & 0 & 0 & -c_m & c_g + c_m \end{bmatrix}. \quad (3.16)$$

Since capacitance is a local effect, the mutual capacitances between far lines are relatively small, and are therefore ignored in (3.16).

Eigenvectors and eigenvalues of capacitance matrix should be calculated to decouple the set of differential equations. Assuming that the far aggressors switch in-phase, because of the symmetry that exists, they will have equal voltages ( $V_{far}(x,t)$ ). Similarly, the near aggressors that switch in-phase have equal voltages ( $V_{near}(x,t)$ ). Based on these two assumptions, there would be only three propagation modes and they correspond to the following eigenvectors:

$$V_{\lambda 1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \quad V_{\lambda 2} = \begin{bmatrix} \frac{-2}{3+\sqrt{5}} \\ 1 \\ \frac{-4}{1+\sqrt{5}} \\ 1 \\ \frac{-2}{3+\sqrt{5}} \end{bmatrix}; \quad V_{\lambda 3} = \begin{bmatrix} \frac{-2}{3-\sqrt{5}} \\ 1 \\ \frac{-4}{1-\sqrt{5}} \\ 1 \\ \frac{-2}{3-\sqrt{5}} \end{bmatrix}, \quad (3.17)$$

where the corresponding eigenvalues are

$$\lambda_1 = c_g; \quad \lambda_2 = c_g + \frac{5+\sqrt{5}}{2} c_m; \quad \lambda_3 = c_g + \frac{5-\sqrt{5}}{2} c_m,$$

respectively. The three propagation modes can be written as

$$V_{sum} \equiv [V_{\lambda 1}]^T [V] = 2V_{far} + 2V_{near} + V_Q, \quad (3.18)$$

and

$$V_{dif1} \equiv [V_{\lambda 2}]^T [V] = \frac{-4}{3+\sqrt{5}} V_{far} + 2V_{near} + \frac{-2}{3+\sqrt{5}} V_Q, \quad (3.19)$$

and

$$V_{dif2} \equiv [V_{\lambda 3}]^T [V] = \frac{-4}{3-\sqrt{5}} V_{far} + 2V_{near} + \frac{-2}{3-\sqrt{5}} V_Q, \quad (3.20)$$

where  $V_Q$  is the voltage of the middle victim line. Voltage of each mode can be calculated by the single RLC line solution:

$$V_{sum}(x, t) = V_{fin}(x, t, V_{in} = V_{in-sum}, c = c_g, l = \frac{1}{c_g v^2}), \quad (3.21)$$

and

$$V_{dif1}(x, t) = V_{fin}(x, t, V_{in} = V_{in-dif1}, c = c_g + \frac{5+\sqrt{5}}{2}c_m, l = \frac{1}{\left[c_g + \frac{5+\sqrt{5}}{2}c_m\right]v^2}), \quad (3.22)$$

and

$$V_{dif2}(x, t) = V_{fin}(x, t, V_{in} = V_{in-dif2}, c = c_g + \frac{5-\sqrt{5}}{2}c_m, l = \frac{1}{\left[c_g + \frac{5-\sqrt{5}}{2}c_m\right]v^2}). \quad (3.23)$$

For each mode, the input voltage should be calculated by using the definition of each mode given by equations (3.18)-(3.20). After calculating the voltages of all three modes, voltage of each line can be calculated by the following equations:

$$V_Q = 0.2V_{sum} - 0.1V_{dif1} + 0.1V_{dif2} \quad (3.24)$$

$$V_{near} = 0.2V_{sum} + 0.2618V_{dif1} + 0.0382V_{dif2}, \quad (3.25)$$

$$V_{far} = 0.2V_{sum} - 0.3236V_{dif1} + 0.1236V_{dif2}. \quad (3.26)$$

### 3.4 Verification and Results

To verify the compact expressions, the results of the compact expressions are compared with the HSPICE simulations for various cases. The noise voltage at the end of the victim line is plotted Figure 3.1 when all near and far aggressors switch simultaneously from low to high. HSPICE simulations verify the compact expressions and the slight ringing

in HSPICE simulations is due to finite number of segments that are used for simulating distributed RLC lines.

$$r = 38\Omega / cm, c_g = 0.177 pf / cm, c_m = 0.11 pf / cm, len = 3.6cm, R_{lr} = 133\Omega.$$

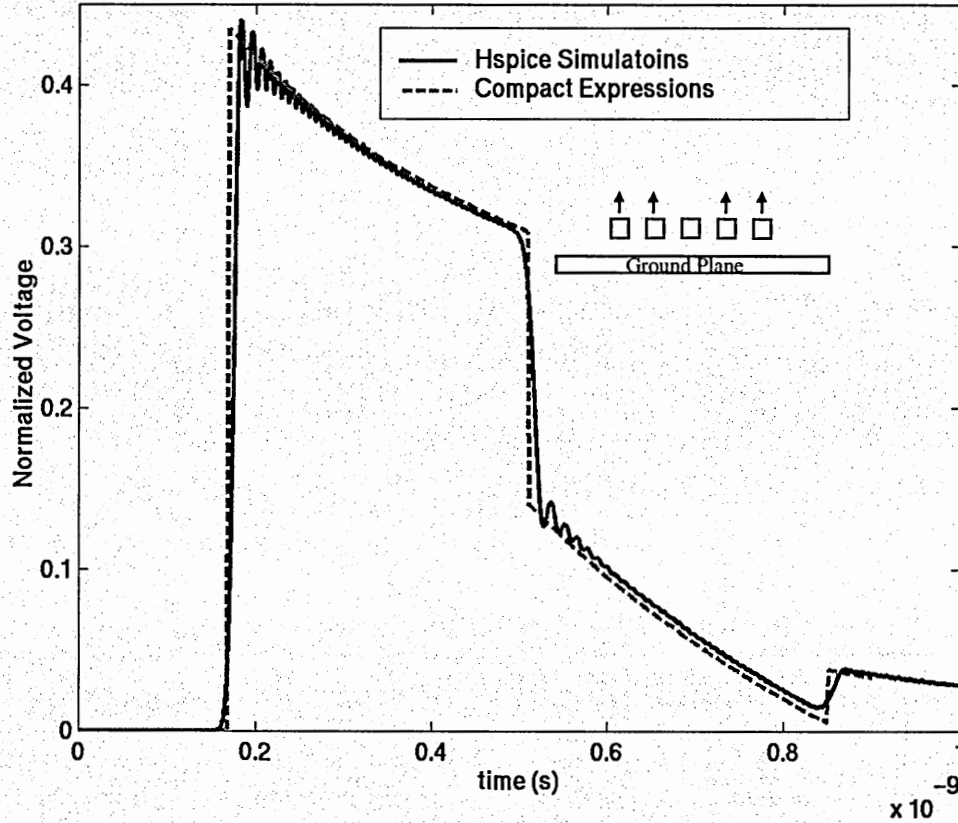


Figure 3.1: Crosstalk Voltage at the end of the middle quiet victim line when far and near aggressors switch in-phase. All cross-sectional dimensions are  $2.4 \mu m$ .

To identify the worst case scenario for crosstalk, noise voltage is plotted in Figure 3.2 for the cases that near and far lines switch in- and anti-phase. It can be acquired from Figure 3.2 that the worst case is when near and far lines switch in-phase. The reason is that when far and near lines switch in the same direction, the voltage swing in the near aggressors would be larger as compared to the case that the near and far lines switch anti-



phase. A larger voltage swing in the near aggressors causes a larger noise voltage in the middle quiet line.

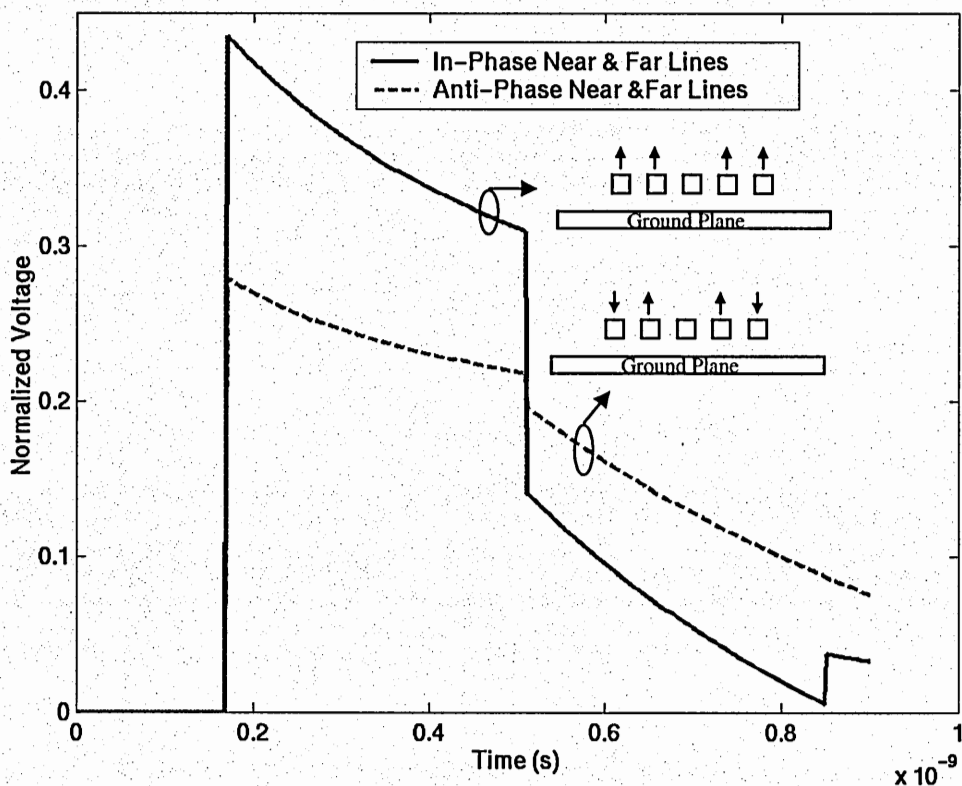


Figure 3.2: Crosstalk Voltage at the end of the quiet line when the far and near aggressors switch anti-phase. All cross-sectional dimensions are  $2.4 \mu\text{m}$ .

To evaluate the impact of far lines on crosstalk, the worst case crosstalk is plotted in Figure 3.3 when a victim line is attacked by 1, 2 and 4 aggressors. Figure 3.3 shows that using the two-line model (only one aggressor) underestimates crosstalk significantly. The three-line model, however, predicts crosstalk with negligible error. In other words, impact of far lines on the worst-case crosstalk is negligible. It is worthwhile to note that

this is true only when there is a nearby ground plane, and as it will be shown in Chapter 4, far inductive noise can be large when ground plane is replaced by orthogonal lines.

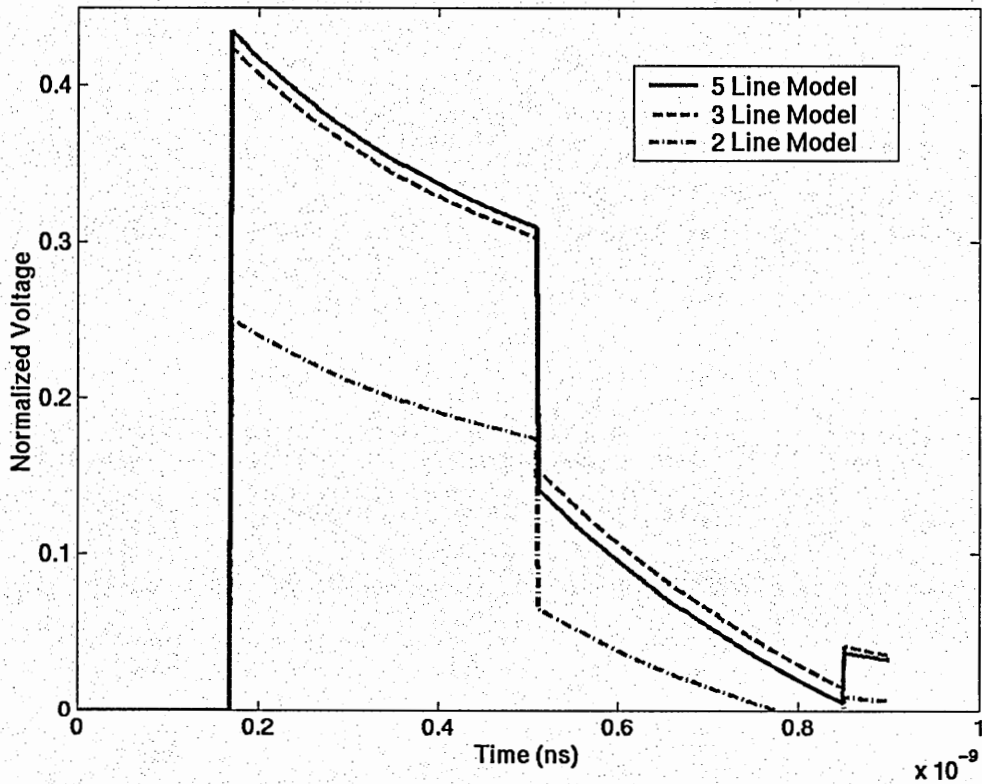


Figure 3.3: Effect of increasing Number of Aggressors from one to four when all aggressors switch in-Phase. All cross-sectional dimensions are  $2.4 \mu\text{m}$ .

The fact that impact of far lines is negligible when there is a nearby ground plane can be proved using low-loss approximation which accurately predicts the peak noise voltage. For a single RLC line the low-loss approximate solution for a step input voltage is

$$V(x,t) = \frac{Z_0}{R_{tr} + Z_0} V_{in} e^{\frac{rx}{2Z_0}} u(t - x/v), \quad (3.27)$$

where  $Z_0$ , the loss-less characteristic impedance, is

$$Z_0 \equiv \sqrt{\frac{l}{c}}. \quad (3.28)$$

Knowing that the wave propagation speed is equal to the speed of light in the dielectric:

$$v \equiv \frac{1}{\sqrt{lc}} = \frac{c_0}{\sqrt{\epsilon_r}}, \quad (3.29)$$

the characteristic impedance can be written as

$$Z_0 = \frac{\sqrt{\epsilon_r}}{c_0 c}. \quad (3.30)$$

Having the propagation modes for three and five coupled RLC lines, the noise voltage for each case can be calculated. For the five conductor case, the peak crosstalk for open-ended lines is

$$V_{peaknoise} = \left( 0.8 \frac{Z_{com}}{Z_{com} + R_{tr}} e^{-\frac{rl}{2Z_{com}}} - 0.4 \frac{Z_{dif1}}{Z_{dif1} + R_{tr}} e^{-\frac{rl}{2Z_{dif1}}} - 0.4 \frac{Z_{dif2}}{Z_{dif2} + R_{tr}} e^{-\frac{rl}{2Z_{dif2}}} \right) V_{dd}, \quad (3.31)$$

where  $Z_{sum} = \frac{\sqrt{\epsilon_r}}{c_g c_0}$ ,  $Z_{dif1} = \frac{\sqrt{\epsilon_r}}{\left( c_g + \frac{5 + \sqrt{5}}{2} c_m \right) c_0}$ , and  $Z_{dif2} = \frac{\sqrt{\epsilon_r}}{\left( c_g + \frac{5 - \sqrt{5}}{2} c_m \right) c_0}$ . A similar

expression can be found for the three line case:

$$V_{peaknoise} = \left( \frac{4}{3} \frac{Z_{sum}}{Z_{sum} + R_{tr}} e^{-\frac{rl}{2Z_{com}}} - \frac{4}{3} \frac{Z_{dif}}{Z_{dif} + R_{tr}} e^{-\frac{rl}{2Z_{dif}}} \right) V_{dd}, \quad (3.32)$$

where  $Z_{dif} = \frac{\sqrt{\epsilon_r}}{(c_g + 3c_m) c_0}$ . To show that the three-line model predicts the worst case

crosstalk with a small error, it should be shown that (3.31) and (3.32) have similar results

for the typical cases. This is done in Figure 3.4 wherein the normalized noise voltage is plotted versus  $c_m/c_g$  ratio. It is evident from Figure 3.4 that using the three-line model has a small error (less than 15%) for most typical cases. It should be noted that crosstalk is normally limited to  $0.2V_{dd}$  and for those cases error of using the three-line model is less than 1%.

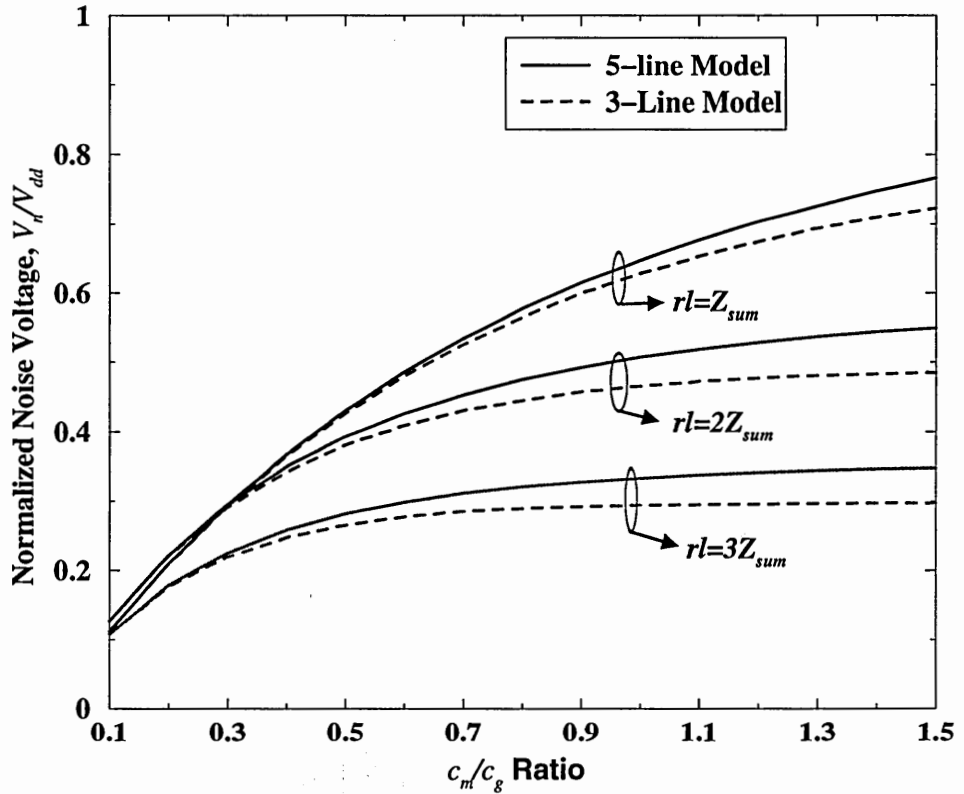


Figure 3.4: Normalized noise voltage versus  $c_m/c_g$  ratio. Neglecting far lines causes a negligible error in the worst-case crosstalk especially for the range that crosstalk is less than  $0.2V_{dd}$ .

### 3.5 Conclusions

A general technique is presented to decouple the set of differential equations for n-coupled RLC lines above a ground plane. In this manner, the transient solution for a

single RLC line can be used to identify the transient voltage of  $n$ -coupled RLC lines. It is shown that the worst case crosstalk occurs when all near and far aggressors switch in-phase. It has been also proved that the impact of far lines on worst-case crosstalk is negligible when there is a nearby ground plane. In other words, a nearby ground plane localizes inductance.

## Chapter 4

### Modeling of Co-Planar RLC Lines

#### 4.1 Introduction

As the integrated circuit technology advances, transistor dimensions scale down, which leads to faster transistors and smaller signal rise-times. In contrast, delay of interconnects increases with down scaling their cross-sectional dimensions. This is a major problem for long global interconnects that connect different macro-cells within a chip because their lengths do not scale with technology. To avoid large global interconnect latencies, low driver resistances and larger cross-sectional dimensions are used [29]. Having a small resistive loss accompanied with small signal rise-times makes it quite unrealistic to ignore the impact of inductance on crosstalk and latency of global interconnects.

Partial electrical equivalent circuit (PEEC) is a general technique to solve RLC interconnects [30, 31]. Since return paths can be unknown a priori, all self and mutual inductance values are calculated assuming that the return paths are at infinity. A circuit simulator or a numeric algorithm is then used to determine the latency and crosstalk of interconnects. PEEC, however, is suitable for CAD tools, and not for system-level design. It is also not numerically stable, and careful sparse approximations are needed to have a robust analysis [31]. Kopocsay et al. have shown that long global interconnects in high-speed chips can be accurately modeled as 2-D transmission lines because dense power and ground grids provide nearby return paths for signal interconnects [32]. They have also presented circuit models that can be used for SPICE or numerical simulations.

For system-level interconnect optimization, however, designers need compact analytical models that accurately predict latency and crosstalk of global interconnects.

In Chapter 3, compact physical models have been derived for  $n$ -coupled distributed RLC interconnects above an ideal ground plane. While on-board or chip-to-chip interconnects are typically sandwiched by power and ground planes, on-chip interconnects typically have no nearby ground planes. Also global interconnects that are on top metal levels are far from the silicon substrate, which to some extent can behave like a ground plane. Hence, models presented in Chapter 3 do not accurately model real GSI global interconnects.

In this Chapter, compact physical models are presented for delay and crosstalk of coplanar on-chip transmission lines which are used in state-of-the-art high performance microprocessors. In these microprocessors, power and ground lines are inserted between every one or two global signal interconnects to provide adequate return paths for signal interconnects and distribute power across the chip [33, 34]. Simplified expressions are also derived which provide designers with physical insight and accurate estimation of noise and latency. In the next chapter, these models are used to optimize the structure of global interconnects.

In this chapter, it has been assumed that the power and ground lines that are inserted between signal lines are wide enough to isolate signal lines from far aggressors, and hence, only near lines are taken into account. The models that are found in this chapter are extended in Chapter 6 to describe the crosstalk noise caused by virtually all near and far aggressors.

In Section 4.2, the periodic structures that are widely used in microwave circuits are introduced. In Section 4.3, it is shown how the existing models for a periodic structure can be used to model an on-chip co-planar transmission line. Two coupled signal lines are then studied in Section 4.4, wherein crosstalk and delay variation are modeled. Impact of having more than two signal lines between power and ground lines on crosstalk is discussed in Section 4.5. Finally, the results are summarized in Section 4.6.

## 4.2 Periodic Structures

The inductance and capacitance of a physically smooth transmission line in a homogenous media have a reciprocity relation independent of the geometry of the line. Hence, the propagation speed is equal to the speed of light in a dielectric [28, 35]:

$$\frac{c_0}{\sqrt{\epsilon_r}} = \frac{1}{\sqrt{lc}}, \quad (4.1)$$

where  $c_0$  is the speed of light in free space,  $\epsilon_r$  is the dielectric constant, and  $l$  and  $c$  are inductance and capacitance per unit length, respectively. Any change in geometry that results in an increase in capacitance decreases inductance, and vice versa. Using (4.1), the characteristic impedance defined as the square root of inductance to capacitance ratio can be written as

$$Z_0 = \frac{\sqrt{\epsilon_r}}{c_0 c}. \quad (4.2)$$

There are, however, structures in which inductance and capacitance can be changed independently. Waveguides and transmission lines loaded at periodic intervals with identical obstacles e.g. a reactive element, are referred as periodic structures. Periodic



structures are widely used in microwave circuits and are well characterized [35]. Figure 4.1 shows two realizations of periodic structures, one is a strip-line and the other one is a coaxial line, both are periodically loaded with lumped capacitances.

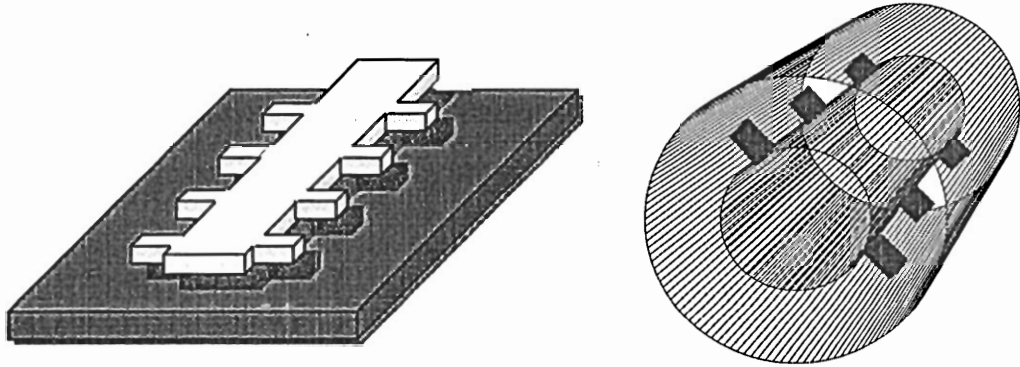


Figure 4.1: Two realizations of periodic structures.

At ultra-high frequencies, when the signal wavelength is comparable with the periodic interval (frequencies above  $15\text{GHz}$  for a period interval of  $0.1\text{mm}$  and above  $1.5\text{GHz}$  for a period interval of  $1.0\text{mm}$ ), a periodic structure behaves like a bandpass filter which passes some frequencies and stops some others. The reason is that the reflected wave at each discontinuity can constructively or destructively interfere with the traveling wave. For lower frequencies, when the signal wavelength is much larger than the periodic interval, a periodic structure behaves like a smooth transmission line. The wave propagation speed, however, is smaller than the speed of light in the interconnect dielectric because discontinuities provide no current path and hence have no impact on inductance. Capacitance per unit length, however, is increased by the discontinuities and as a result, the propagation speed is smaller than that of a regular transmission line [35].

Since the propagation speed for the unloaded line is equal to the speed of light, it can be written

$$\frac{c_0}{\sqrt{\epsilon_r}} = \frac{1}{\sqrt{l(c - C_0/d)}}, \quad (4.3)$$

where  $c$  is the total capacitance per unit length and  $C_0$  is the lumped capacitance that is added at intervals  $d$ . The wave propagation speed in a periodic structure can therefore be written as [35]

$$v = \frac{1}{\sqrt{lc}} = \frac{c_0}{\sqrt{\epsilon_r}} \sqrt{1 - \frac{C_0/d}{c}}, \quad (4.4)$$

and similarly, using (4.3), the characteristic impedance of a periodic structure can be written as

$$Z_0 = \frac{1}{\sqrt{1 - \frac{C_0/d}{c}}} \frac{\sqrt{\epsilon_r}}{c_0 c}. \quad (4.5)$$

Equation (4.4) shows that the propagation speed decreases as the percentage increase in the line capacitance due to the lumped elements becomes larger.

### 4.3 Modeling On-Chip Co-Planar Interconnects

To see how the existing models for periodic structures can be used for on-chip interconnects, a single signal line sandwiched by two power/ground lines is shown in Figure 4.2.a. The three lines form a transmission line and since they are in a homogenous medium, the loop inductance and capacitance per unit length are related as (4.1) shows. Orthogonal lines, however, make the medium inhomogeneous, and the structure shown in Figure 4.2.b becomes a periodic structure. The pitch of the orthogonal lines is typically

less than a few microns and hence a few orders of magnitude smaller than the typical on-chip signal wavelength. For instance, for  $10\text{GHz}$ , signal wavelength is around  $1.5\text{cm}$ . Hence, the on-chip co-planar structure shown in Figure 4.2.b can be modeled as a smooth transmission line with a smaller wave propagation speed. In other words, the orthogonal lines have no impact on inductance because current cannot return through them; they, however, increase the interconnect capacitance per unit length. The loop resistance can also be calculated by adding the equivalent resistance of the power and ground lines to the signal line's resistance [32]. It should be noted that in the gigahertz regime, due to proximity effect, current returns mainly through the adjacent power and ground lines [29, 32].

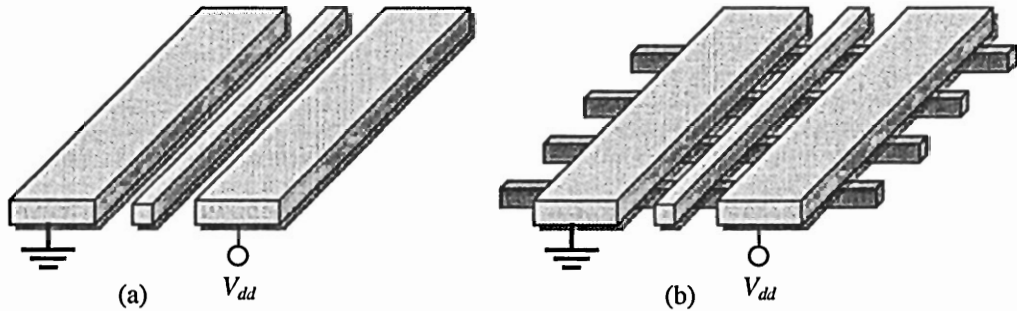


Figure 4.2: A double sided shielded signal line. (a) there are no orthogonal lines and therefore the three lines form an ideal transmission line (b) there are orthogonal lines which make the signal and ground lines form a periodic structure.

Having the interconnect inductance, capacitance and resistance per unit length, the rigorous solutions for a single RLC line can be used to find the interconnect latency. Orthogonal lines increase the latency because they reduce the propagation speed and the characteristic impedance. A smaller characteristic impedance results in a larger attenuation, and a smaller propagation speed results in a larger time-of-flight (ToF). This

is shown in Figure 4.3 in which the output voltages of two interconnects are plotted versus time, one above a ground plane and the other one above orthogonal lines.

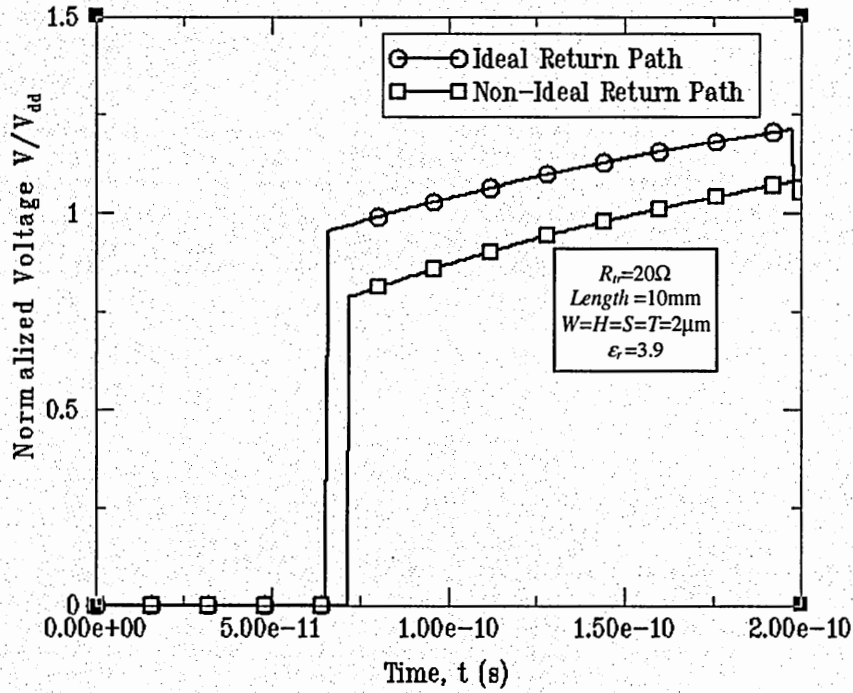


Figure 4.3: Voltage at the end of signal line for ideal and non-ideal return path cases. For the ideal case a signal line is above a ground line and in the non-ideal case a signal line is shielded between power/ground lines above orthogonal lines.

For a co-planar transmission line above orthogonal lines, the propagation speed can be written as

$$v = \frac{c'_0}{\sqrt{\epsilon_r}}, \quad (4.6)$$

and the characteristic impedance is

$$Z_0 = \frac{\sqrt{\epsilon_r}}{cc'_0}, \quad (4.7)$$

where  $c'_0$  is the modified speed of light given by

$$c'_0 = c_0 \sqrt{1 - \frac{c_{orth}}{c}}, \quad (4.8)$$

where  $c_{orth}$  is the capacitance per unit length to orthogonal lines. By comparing (4.6) and (4.7) with their counterparts for interconnects with ideal return paths, (4.4) and (4.5), it can be seen that the only difference is that the speed of light factor is reduced. By using this transformation, all equations that are found for delay of interconnects with ideal return paths can be used for on-chip co-planar transmission lines.

#### 4.4 Two Signal Lines between Power and Ground Lines

Models presented in Section 4.3 can be extended to model two signal lines between power and ground lines. If there are no orthogonal lines, interconnects are in a homogenous medium and the inductance and capacitance matrices are related as [28]

$$\begin{bmatrix} l_s & l_m \\ l_m & l_s \end{bmatrix} \begin{bmatrix} c_g + c_m & -c_m \\ -c_m & c_g + c_m \end{bmatrix} = (\epsilon_r / c_0^2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (4.9)$$

where  $l_s$  and  $l_m$  are self and mutual inductances per unit length, respectively,  $c_g$  is the capacitance between a signal line and its nearby co-planar ground line per unit length, and  $c_m$  is the mutual capacitance between two signal lines per unit length. Equation (4.9) results in

$$\frac{1}{\sqrt{c_g(l_s + l_m)}} = \frac{1}{\sqrt{(c_g + 2c_m)(l_s - l_m)}} = \frac{c_0}{\sqrt{\epsilon_r}}, \quad (4.10)$$

which shows that the common mode (in-phase switching) and the differential mode (anti-phase switching) propagate with the speed of light in a dielectric. For on-chip

interconnects, however, there are orthogonal lines below and above signal lines that increase the total capacitance without affecting the inductance values. Any switching pattern can be written in terms of common and differential modes. Hence, having the solution for each mode, any switching pattern can be modeled. For the common mode, which is the in-phase switching case, the equivalent inductance and capacitance per unit length are

$$C_{com} = C_g + C_{orth}, \quad (4.11)$$

and

$$L_{com} = L_s + L_m, \quad (4.12)$$

respectively. The mutual capacitance between two signal lines does not contribute to the common mode capacitance because the two lines have equal voltages. For the differential mode, which is the anti-phase switching case, the capacitance and inductance per unit length are

$$C_{dif} = C_g + 2C_m + C_{orth}, \quad (4.13)$$

and

$$L_{dif} = L_s - L_m, \quad (4.14)$$

respectively. Due to Miller's effect, twice the mutual capacitance appears in the differential mode capacitance. Knowing the equivalent capacitance and inductance values, the single RLC line solution [26] can be used to find the solution for each mode.

The propagation speed for each mode is

$$v = \frac{1}{\sqrt{L_{eq} C_{eq}}}, \quad (4.15)$$

where  $l_{eq}$  and  $c_{eq}$  are the equivalent inductance and capacitance per unit length for each mode. Using (4.10)-(4.15), the modified speed of light for the common and differential modes can be written as

$$c'_{0-com} = c_0 \sqrt{1 - \frac{c_{orth}}{c_{com}}}, \quad (4.16)$$

and

$$c'_{0-dif} = c_0 \sqrt{1 - \frac{c_{orth}}{c_{dif}}}, \quad (4.17)$$

respectively. Equations (4.16) and (4.17) show that the differential mode travels faster than the common mode because the common mode has a smaller capacitance, and, therefore, the percentage increase in the common mode capacitance due to orthogonal lines is larger than that of the differential mode. This is important because as it will be shown, it causes an out-of-phase noise at the end of a quiet victim line.

When one of the signal lines switches and the other line stays quiet, both common and differential modes are generated and the voltages of the quiet and active lines are

$$V_Q(x, t) = \frac{1}{2} V_{com}(x, t) - \frac{1}{2} V_{dif}(x, t), \quad (4.18)$$

and

$$V_A(x, t) = \frac{1}{2} V_{com}(x, t) + \frac{1}{2} V_{dif}(x, t), \quad (4.19)$$

respectively. These equations are verified against HSPICE simulations as shown in Figure 4.4, where voltages at the end of active and quiet lines are plotted versus time. For HSPICE simulations, PEEC methodology is used and 1000 segments shown in Figure 4.5

are put in series to model the distributed RLC lines. All self and mutual inductances are extracted by RAPHAEL [17].

As Figure 4.4 shows, an out-of-phase noise appears at the end of the victim line. The reason is that the differential mode travels faster and reaches the end of the line sooner than the common mode. Hence, as (4.18) shows, a negative pulse appears at the end of the victim line and its duration is

$$t_{out-phase} = \frac{\ell}{v_{com}} - \frac{\ell}{v_{dif}}. \quad (4.20)$$

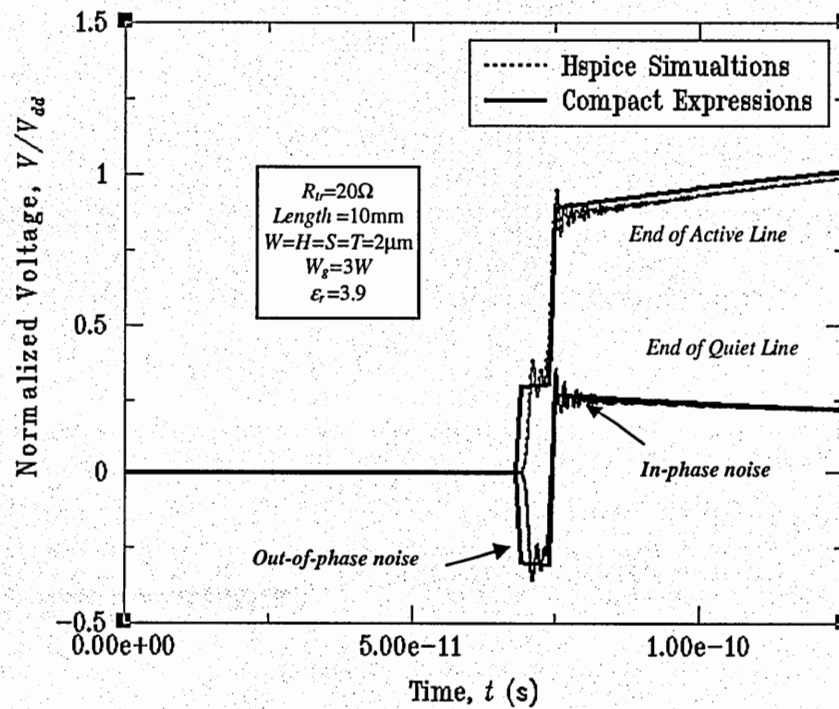


Figure 4.4: Normalized voltage at the end of open-ended signal lines when one of them is excited with a step input and the other one is quiet. An out-of-phase noise appears at the end of the victim line due to different propagation speeds for common and differential modes.



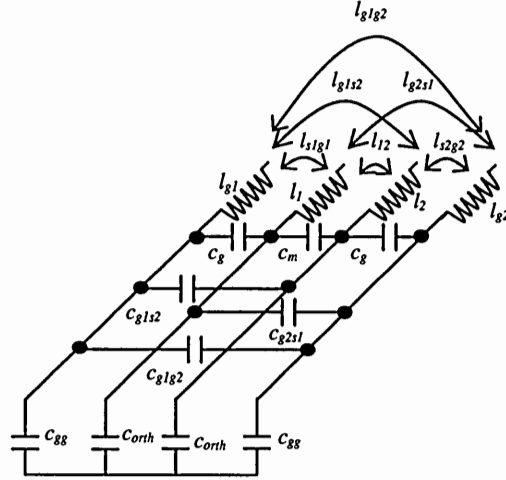


Figure 4.5: A segment of the equivalent circuit used for HSPICE simulations. 1000 segments are used in all simulations.

The rigorous solution for a single RLC line is complicated, and it is not suitable for optimizing the structure of global interconnects. Low-loss approximation, however, is relatively simple and has a small error for the peak voltages even for lossy interconnects [26]. Using low-loss approximation, the noise voltage defined by (4.18) can be written as

$$\begin{cases} V_Q = 0 & t < \ell / v_{dif} \\ V_Q = -\frac{Z_{dif} V_{dd}}{Z_{dif} + R_{tr}} e^{-\frac{rt}{2Z_{dif}}} & \ell / v_{dif} < t < \ell / v_{com} , \\ V_Q = \frac{Z_{com} V_{dd}}{Z_{com} + R_{tr}} e^{-\frac{rt}{2Z_{com}}} - \frac{Z_{dif} V_{dd}}{Z_{dif} + R_{tr}} e^{-\frac{rt}{2Z_{dif}}} & \ell / v_{com} < t \end{cases} \quad (4.21)$$

where  $R_{tr}$  is the driver resistance, and  $Z_{com}$  and  $Z_{dif}$  are the characteristic impedances of the common and differential modes, respectively. It has been assumed that interconnects are open at their end; therefore, noise voltage doubles at the end of the victim line. The

in- and out-of-phase noise voltages are plotted versus resistance per unit length in Figure 4.6. It can be acquired from Figure 4.6 that for small line resistances, the out-of-phase noise is dominant and for large line resistances, the in-phase noise is larger.

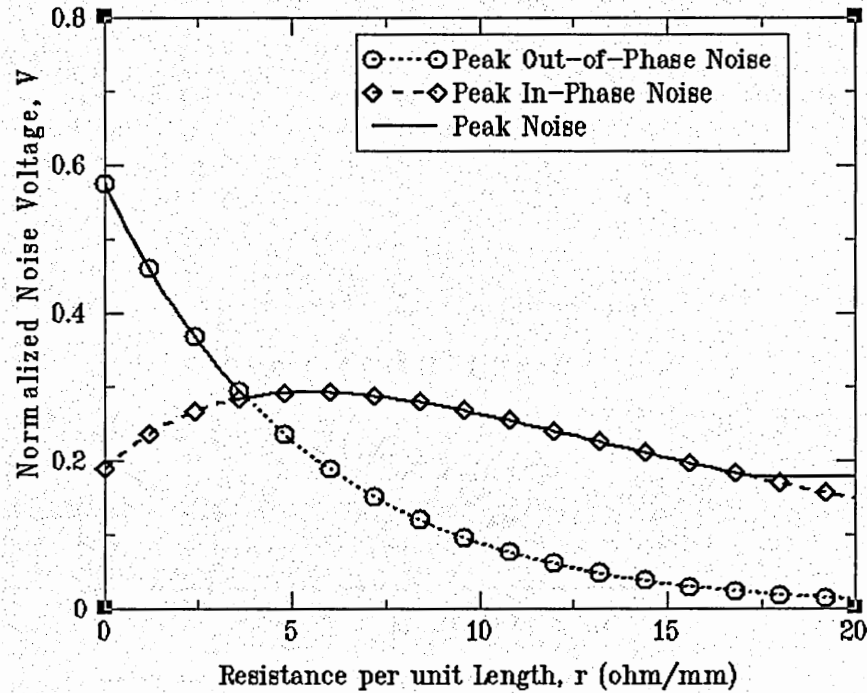


Figure 4.6: Peak in and out-of-phase noise voltages versus interconnect resistance per unit length.

It should be noted that the out-of-phase noise can be either positive or negative depending on the switching direction of the aggressor line. A large out-of-phase noise can, therefore, cause false switching or reduce the reliability of the chip by damaging the gate oxide of the receiver. Hence, a large out-of-phase noise should be avoided, too.

By using (4.21), the minimum line resistance beyond which in-phase noise becomes dominant can be identified as:

$$r_{\min} = \frac{2}{\ell} \times \frac{Z_{com} Z_{dif}}{Z_{com} - Z_{dif}} \ln \left[ 2 \frac{1 + R_{tr}/Z_{com}}{1 + R_{tr}/Z_{dif}} \right], \quad (4.22)$$

and by substituting (4.7), (4.16), and (4.17) in (4.22), it can be written as

$$r_{\min} = \frac{2}{\ell} \times \frac{\sqrt{\epsilon_r}}{rc_0 c_m} \ln \left[ 2 \frac{1 + R_{tr}/Z_{com}}{1 + R_{tr}/Z_{dif}} \right]. \quad (4.23)$$

For the small driver resistance case ( $R_{tr} \approx 0$ ), (4.23) can be approximated by

$$r_{\min} = \frac{2}{\ell} \times \frac{\sqrt{\epsilon_r}}{rc_0 c_m} \ln 2. \quad (4.24)$$

The maximum peak in-phase crosstalk can be found by taking the derivative of (4.21) with respect to  $r$ :

$$V_{peak, \max} = \left[ \frac{Z_{dif} + R_{tr}}{Z_{com} + R_{tr}} \right]^{\frac{-Z_{dif}}{(Z_{dif} - Z_{com})}} \frac{Z_{com} - Z_{dif}}{Z_{com} + R_{tr}}, \quad (4.25)$$

which occurs at

$$r_{\max} = \frac{2}{\ell} \frac{Z_{dif} Z_{com}}{(Z_{com} - Z_{dif})} \ln \left[ \frac{Z_{dif} + R_{tr}}{Z_{com} + R_{tr}} \right]. \quad (4.26)$$

The worst case crosstalk happens when the driver resistance,  $R_{tr}$ , is equal to zero. For the worst case, (4.25) can be rewritten as

$$V_{peak, \max} = \eta^{\frac{1}{1-\eta}} \left( 1 - \frac{1}{\eta} \right), \quad (4.27)$$

where  $\eta$  is the ratio of the common and differential mode characteristic impedances.

Knowing (4.7), (4.16), and (4.17),  $\eta$  can be written as

$$\eta = \sqrt{\left(1 + \frac{2c_m}{c_g}\right) \left(1 + \frac{2c_m}{c_g + c_{orth}}\right)}. \quad (4.28)$$

The worst case crosstalk, which is given by (4.27), can be approximated by

$$V_{peak,max} = \frac{\pi}{4} \frac{c_m}{\sqrt{c_g(c_g + c_{orth}) + c_m}}, \quad (4.29)$$

with less than 3% error for all typical values of  $c_g$ ,  $c_{orth}$  and  $c_m$ . Comparing (4.29) with its counterpart for interconnects above an ideal ground plane [27],

$$V_{peak,max} = \frac{\pi}{4} \frac{c_m}{c_g + c_{orth} + c_m}, \quad (4.30)$$

shows that modeling orthogonal lines with a ground plane underestimates the crosstalk.

For instance, if  $c_g = c_{orth} = c_m$ , (4.30) underestimate the maximum peak crosstalk by 20%.

## 4.5 More than Two Signal Lines between Power and Ground Lines

Having more than two signal lines between power and ground lines makes the modal analysis very complicated because unlike the interconnects over an ideal ground plane, inductance and capacitance matrices have different eigenvectors [28]. High-performance state-of-the-art processors, however, typically do not use such structures for global interconnects with prominent inductive effects because the middle line has no nearby return path and has a large mutual inductance to other signal lines and crosstalk can be prohibitively large if inductive effects are dominant [33, 34]. Figure 9 shows the crosstalk voltage at the end of a quiet victim line when there are two, three, and five signal lines between power and ground lines. It shows that far aggressors have a large impact on

crosstalk. This is in contrast with crosstalk of interconnects over a ground plane where only near aggressors have a considerable impact (Chapter 3).

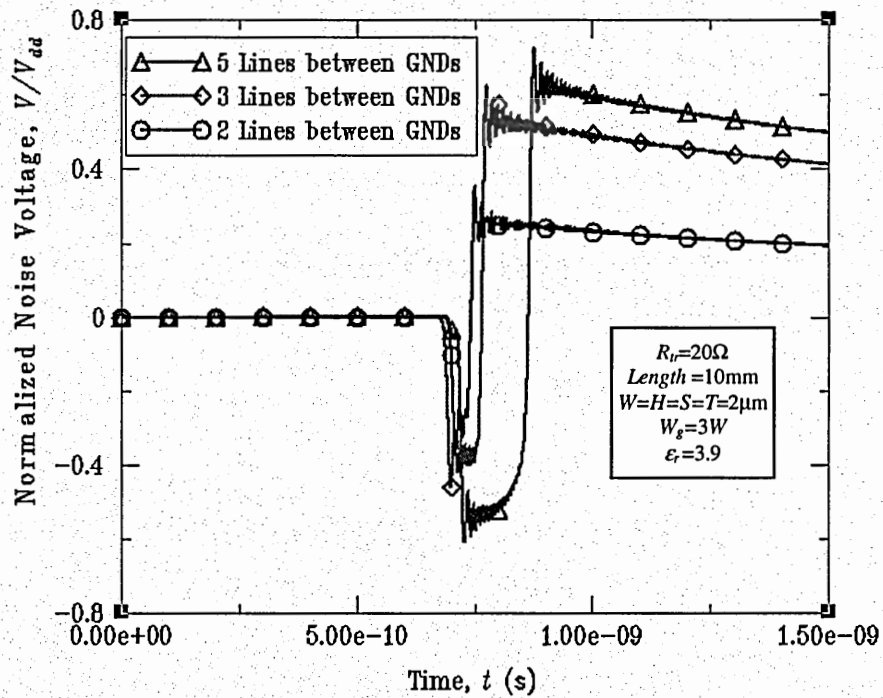


Figure 4.7: Noise voltage at the end of a middle victim line when there are 2, 3 or 5 signal lines between power/ground lines. Unlike the ideal return path case, far lines have a large impact on the worst case crosstalk.

## 4.6 Conclusions

The existing physical models for periodic structures, which are widely used in microwave circuits, are used to derive compact expressions for delay and crosstalk of on-chip global interconnects. Because of lack of on-chip ground planes, power and ground lines are usually inserted between signal lines to provide adequate return paths for signal interconnects and also distribute power across the chip. It is shown that the wave propagation speed is smaller than the speed of light in the interconnect dielectric because

of lower or upper orthogonal lines which increase the capacitance per unit length of interconnects without affecting the inductance per unit length of interconnects. Due to smaller propagation speeds for common than differential modes, an out-of-phase noise appears at the end of a quiet victim line. Compact expressions that are provided for the delay and crosstalk of global interconnects are suitable for system level interconnect optimization. The results of this chapter are used in the next chapter to optimize the design of global interconnect.

## Chapter 5

### Optimization of Co-Planar RLC Lines

#### 5.1 Introduction

In Chapter 2, a new interconnect-centric methodology has been proposed to optimize the width of global interconnects by maximizing data flux density and minimizing latency simultaneously. The optimal wire width is in the shallow RLC region, where interconnect latency is 33% larger than the time-of-flight (ToF). The analysis presented in Chapter 2, however, assumes that there is a nearby ground plane to provide an ideal return path for interconnects. It also assumes that all cross-sectional dimensions are equal.

In this Chapter, compact physical models for delay and crosstalk of on-chip coplanar transmission lines are used to optimize the cross-sectional dimensions of global interconnects. The ratio of signal-to-signal spacing to signal-to-ground spacing is optimized to minimize latency and reduce crosstalk and delay variation. An optimal wire width is identified that simultaneously maximizes the data flux density and minimizes latency. It is finally shown that to have a good trade-off between data flux density and energy dissipation, larger spaces should be used between interconnects if a larger aspect ratio is used.

In Section 5.2, the ratio of signal-to-signal spacing to signal-to-ground spacing is optimized. In Section 5.3 the optimal wire width that maximizes the data flux density-reciprocal latency product is identified. It is also shown that this optimization along with the spacing ratio optimization makes the crosstalk and the dynamic delay variation small

and constant for all generations of technology. The optimal relationship between the metal thickness and the spacing between interconnects is identified in Section 5.4 that simultaneously maximizes data flux density and minimizes energy per bit. Finally, the results are summarized in Section 5.5.

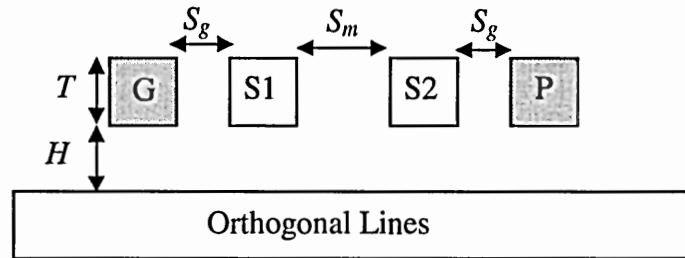


Figure 5.1: A cross-sectional view of two signal lines between two power/ground lines

## 5.2 Optimal Signal-Ground Spacing to Signal-Signal Spacing Ratio

The cross-sectional view of two signal interconnects between power/ground lines is shown in Figure 5.1. This is a typical structure for global interconnects which are in top metal levels. Power and ground lines are inserted between signal lines to provide adequate return paths and distribute power and ground across the chip [33, 34]. Typically, the worst case scenario for delay is when two adjacent signal lines switch anti-phase. The equivalent capacitance for the differential mode (anti-phase switching case) is

$$c_{dif} = c_g + 2c_m + c_{orth}, \quad (5.1)$$



where  $c_g$  is the capacitance between a signal line and its nearby ground line per unit length,  $c_m$  is the capacitance between two signal lines per unit length, and  $c_{orth}$  is the capacitance to lower and upper orthogonal lines per unit length. Due to Miller's effect, the mutual capacitance appears with a factor of two in the differential mode capacitance. Hence, it is beneficial to make the signal lines further apart to reduce the differential mode capacitance. Minimizing the capacitance minimizes the characteristic impedance, which reduces the attenuation and results in minimum delay. Using a simple parallel plate model for capacitance, the optimal ratio between signal-to-signal spacing,  $S_m$ , and signal to power/ground line spacing,  $S_g$  can be identified. In this analysis, the wire width and the distance between the two power/ground lines are assumed to be constant, which means that the wiring density is not affected by this optimization. Using the parallel plate approximation, the differential mode capacitance can be written as

$$c_{dif} = c_{orth} + \epsilon_r \epsilon_0 \frac{T}{S_g} + 2\epsilon_r \epsilon_0 \frac{T}{S_m}, \quad (5.2)$$

and assuming that the power/ground line pitch is constant, the total spacing

$$S_T = 2S_g + S_m, \quad (5.3)$$

should remain constant. By taking the derivative of (5.2), the optimal  $S_g/S_m$  ratio is identified to be 0.5, where  $c_g$  and  $c_m$  become equal. This optimal value is independent of the metal thickness,  $T$ , and the total spacing,  $S_T$ . Although fringing effect is neglected in this analysis, a complete analysis shows that the error is small. Figure 5.2 plots the characteristic impedance for the differential mode versus  $S_g/S_m$  ratio based on the exact capacitance values extracted by RAPHAEL [17]. Figure 5.2 confirms that the optimal ratio is 0.5, the same as what is identified by the parallel plate analysis. It, however, shows that the curves are very flat around the optimal point.

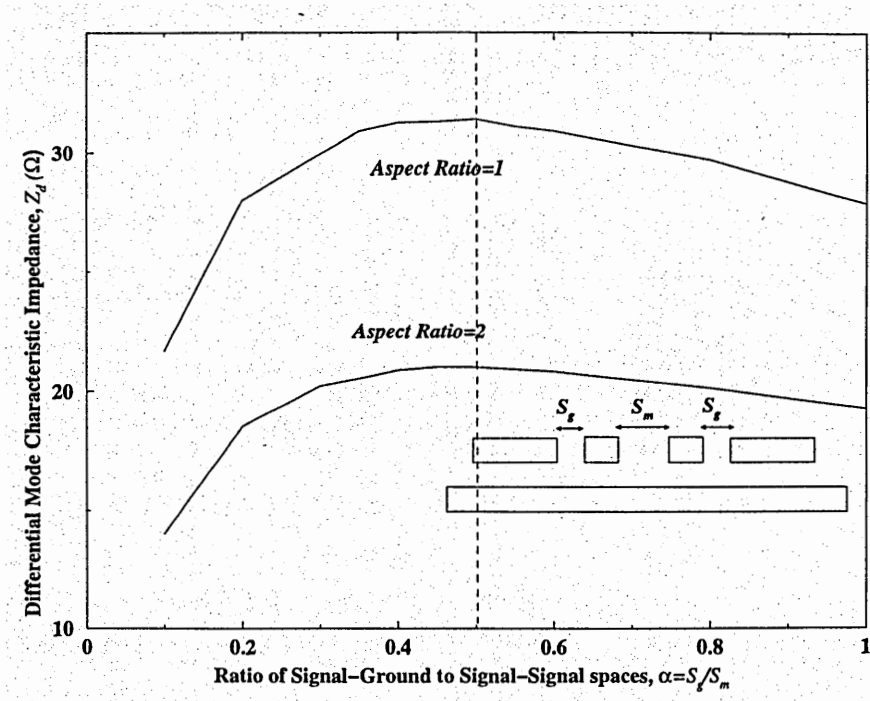


Figure 5.2: The differential mode characteristic impedance versus the spacing.

Minimizing the differential mode capacitance maximizes the characteristic impedance and results in minimum attenuation and delay. Figure 5.3 shows the worst case delay versus the spacing ratio and shows that latency is 12% smaller than the equal spacing case.

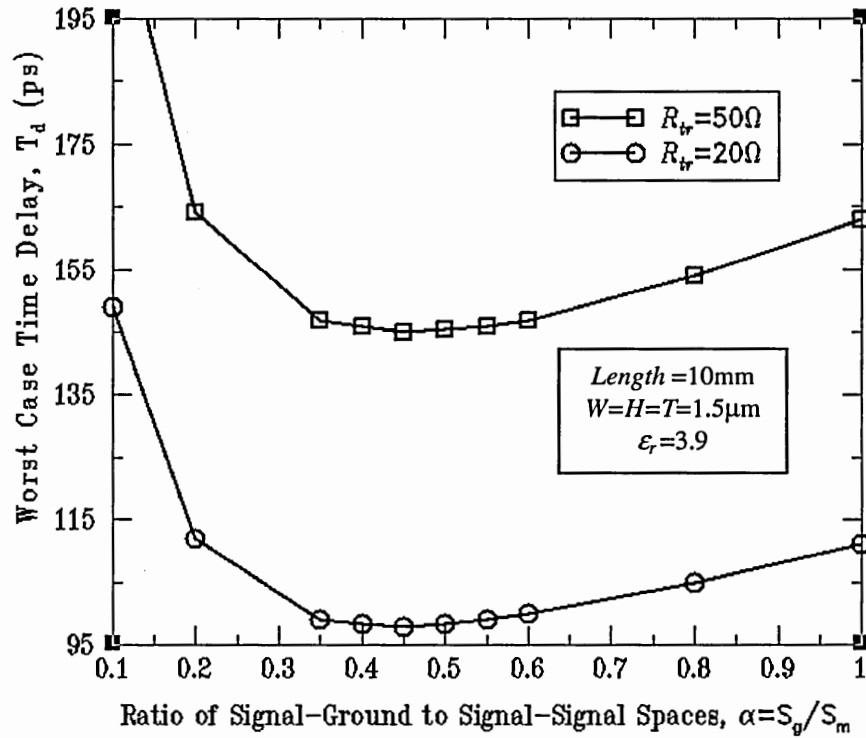


Figure 5.3: Worst case time delay versus the ratio of signal-ground to signal-signal spaces.

This optimization has two important subsidiary advantages, too. It reduces the delay variation and crosstalk considerably. The delay variation caused by different switching patterns and the worst-case peak crosstalk are plotted in Figure 5.4 and Figure 5.5 versus the spacing ratio for different aspect ratios. It can be acquired from Figure 5.4 and Figure 5.5 that at the optimal design point, the delay variation and crosstalk are smaller by 48% and 33%, respectively, in comparison with those of the equal spacing case. It is worthwhile to note that unlike the general intuition that wiring density should be sacrificed to reduce delay, delay variation and crosstalk, wiring density is not affected in this optimization because the total spacing is assumed to be constant.

It can be seen in Figure 5.4 and Figure 5.5 that the delay variation and crosstalk decrease monotonically with the  $S_g/S_m$  ratio decreasing. On the other hand, the differential mode characteristic impedance curve is flat around the optimal design point (Figure 5.2). Hence, by using an  $S_g/S_m$  ratio slightly smaller than the optimal value, the crosstalk and delay variation can be improved considerably with a negligible increase in latency. For instance, by choosing  $S_g/S_m$  ratio of 0.45 instead of 0.5, latency is increased by less than 1% while the delay variation and the maximum peak crosstalk are reduced by 10% and 9%, respectively. From now on, therefore, the  $S_g/S_m$  ratio of 0.45 is used as the optimal  $S_g/S_m$  ratio.

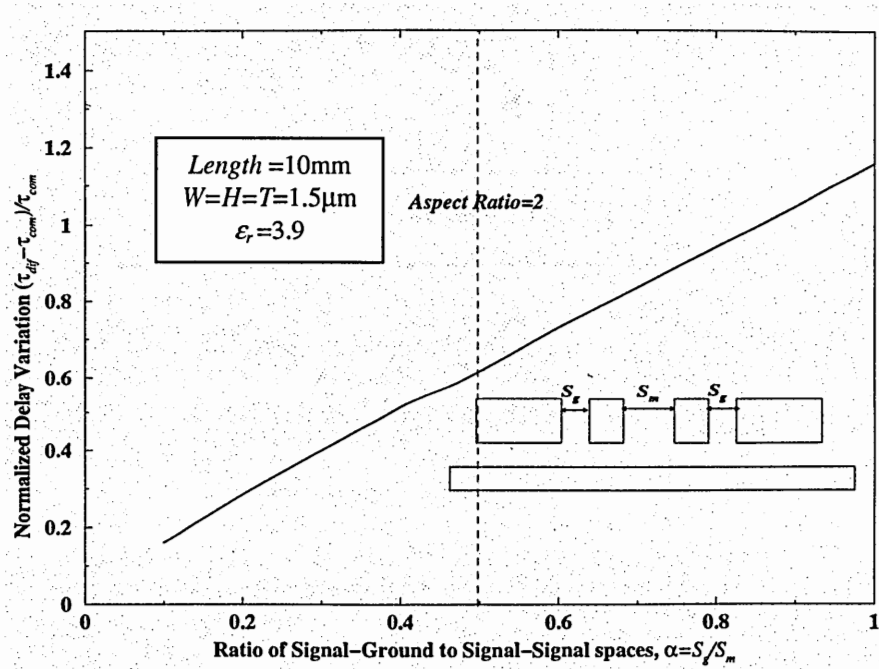


Figure 5.4: Normalized delay variation versus the spacing ratio  $S_g/S_m$  for the aspect ratio of two.

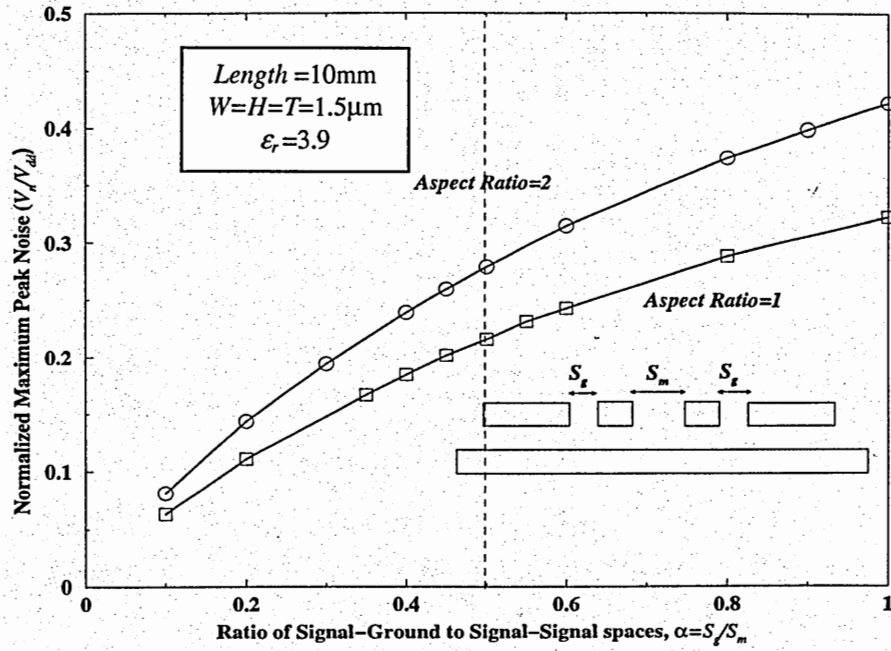


Figure 5.5: Maximum peak crosstalk versus the spacing ratio,  $S_g/S_m$  for aspect ratios of 1 and 2. The maximum peak crosstalk is pessimistic because it is for a case that  $R_{tr}=0$ ,  $C_L=0$ , and the line resistance is such that crosstalk is maximized.

### 5.3 Optimal Wire Width

Performance of a high-speed chip is largely affected by both latency and bandwidth of global interconnects, which connect different macrocells. Therefore, designers always try to have high-bandwidth and fast buses between a processor and its on-chip cache memory, or between different processors within a multiprocessor chip. In Chapter 2, an interconnect-centric methodology is proposed to optimize the width of global interconnects to maximize data flux density and minimize latency simultaneously. It is rigorously shown that the optimal wire width is independent of interconnect length and is in the shallow RLC region, where interconnect latency becomes 33% larger than the

time-of-flight. The analysis presented in Chapter 2, however, is for interconnects above an ideal ground plane where propagation speed is equal to the speed of light in the interconnect dielectric. In this section, it is shown how physical models that are derived in Chapter 3 can be used to find the optimal wire width for global interconnects in a real chip, where there is no nearby ground plane. A single signal line shielded from both sides is first analyzed and the two signal lines between power/ground lines are then studied.

### 5.3.1 Optimal Wire Width for a Single Signal Line

Latency of an optimally-buffered interconnect in the RC region is [13]

$$\tau_{RC} = 2.5 \frac{\ell}{W} \sqrt{\xi \rho \epsilon_r \epsilon_0 R_0 C_0}, \quad (5.4)$$

where  $R_0$  and  $C_0$  are the output resistance and input capacitance of a minimum size repeater, respectively,  $\rho$  is resistivity of metal, and the geometry factor,  $\xi$ , is a dimensionless constant given by

$$\xi \equiv \frac{rW^2}{\rho} \times \frac{c}{\epsilon_r \epsilon_0}, \quad (5.5)$$

where  $r$  and  $c$  are resistance and capacitance per unit length. The value of  $\xi$  is independent of resistivity or dielectric constant. For large wire widths, inductance cannot be neglected, and delay is equal to [18]

$$\tau_{RLC} = (1 + 1.5 \frac{R_0 C_0}{Z_0^2} \frac{r}{c}) ToF. \quad (5.6)$$

The wave propagation speed is smaller than the speed of light in the interconnect dielectric because of lower or upper orthogonal lines which increase the capacitance per unit length of interconnects without affecting their inductance per unit length. Current

cannot return through orthogonal lines and that is why orthogonal lines do not change the inductance per unit length of interconnects. The propagation speed for an on-chip coplanar transmission line can be calculated using a modified speed of light as

$$c'_0 = c_0 \sqrt{1 - \frac{c_{orth}}{c}}, \quad (5.7)$$

where  $c$  is the total capacitance per unit length. Figure 5.6 plots interconnect latency versus wire width assuming that all cross-sectional dimensions scale proportionally. It can be seen that in the RC region, latency decreases linearly with increasing wire width. In the RLC region, however, interconnect becomes time-of-flight limited, and there is a diminishing return in latency reduction.

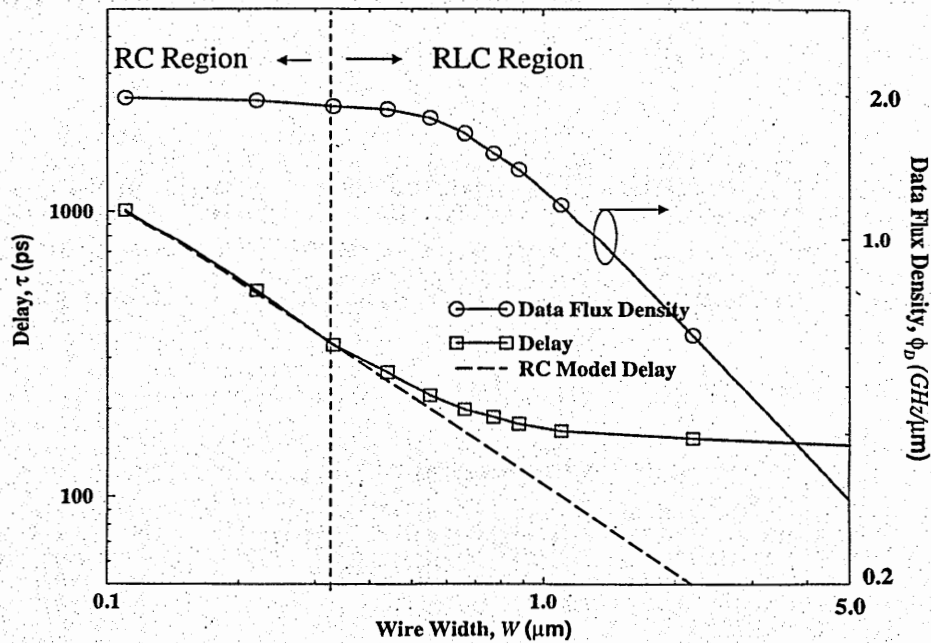


Figure 5.6: RC and RLC model latency and data flux density versus wire width for an interconnect 24mm long implemented at the 45 nm technology node. It has been assumed that optimal repeaters are used, and  $W_G=2W$ ,  $T=S=W$ .

To study the impact of wire width on chip bisectional bandwidth, data flux density is defined as the interconnect bandwidth per unit width of an interconnect:

$$\Phi_D \equiv \frac{Bw}{p} = \frac{1/\tau}{2S + W + W_G}, \quad (5.8)$$

where  $p$  is interconnect pitch,  $Bw$  is interconnect bandwidth,  $W$  is the width of signal lines and  $W_G$  is the width of power/ground lines. Interconnect bandwidth is the number of bits per second that an interconnect can transfer and is assumed to be equal to reciprocal latency. To find the optimal wire width, we assume that all cross-sectional dimensions scale proportionally and later in the next section, the optimal  $T/W$  and  $S/W$  will be found. In this manner, data flux density can be written as

$$\Phi_D = \frac{1/\tau}{\gamma W}, \quad (5.9)$$

where  $\gamma$  is a constant determined by  $S/W$  and  $W/W_g$  ratios. A large data flux density is desired to have a large chip bisectional bandwidth,  $Bw_{bisec}$ , because

$$Bw_{bisec} = \int_0^{D_{chip}} \Phi_D(x) dx. \quad (5.10)$$

Data flux density is plotted versus wire width in Figure 5.6. Data flux density is constant in the RC region, and it drops in the RLC region. Hence, to have a large data flux density and a small latency simultaneously, data flux density-reciprocal latency product should be maximized (Figure 5.7). The optimal wire width is

$$W_{opt} = 2.12c'_0 \sqrt{\xi \rho \epsilon_0 R_0 C_0}, \quad (5.11)$$

which is the same form as the optimal wire width for interconnects above an ideal ground plane. The speed of light, however, is substituted by the modified speed of light,  $c'_0$ . The



geometry factor,  $\xi$ , is also different because the return path resistance should be added to the signal line resistance.

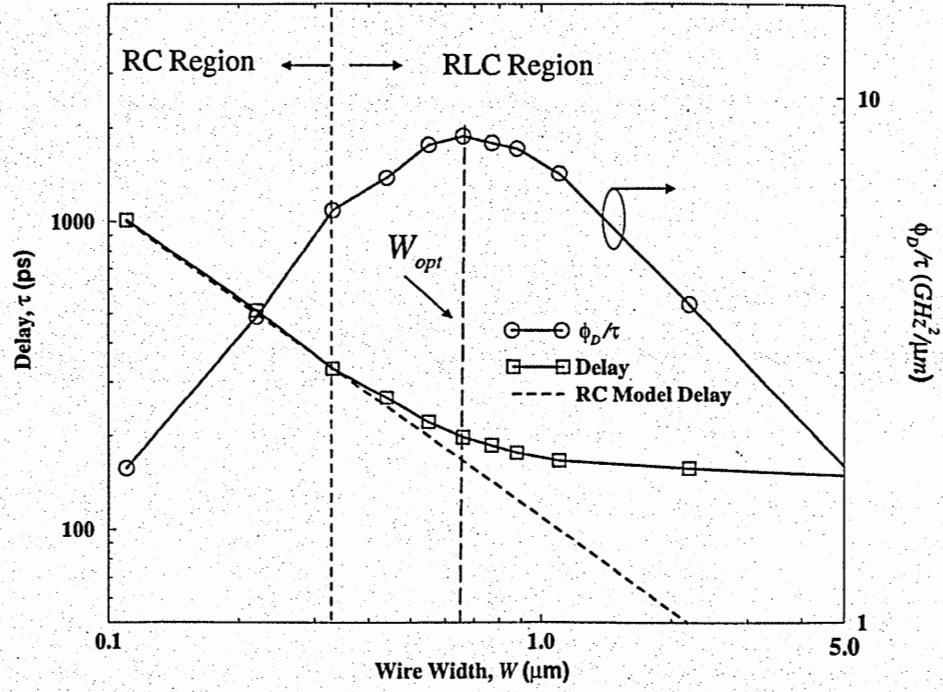


Figure 5.7: Data flux density-reciprocal latency product versus wire width for an interconnect 24 mm long implemented at the 45 nm technology node. In the shallow RLC region, the data flux density-reciprocal product attains its maximum.

### 5.3.2 Optimal Wire Width for Two Signal Lines

The same methodology can be used when there are two signal lines between power/ground lines to find the optimal wire width for common and differential modes. The optimal wire width for each mode can be identified by (5.11). Latency and data flux density of interconnects, however, are different for in-phase and anti-phase switching cases because the modes have different modified speeds of light and geometry factors.

Hence, the optimal wire width is different for common and differential modes. The design of interconnects is taken for the worst case which means that the optimal wire width is the width at which the lowest data flux density-reciprocal latency product is maximized. Data flux density-reciprocal latency product for both in-phase and anti-phase switching cases are plotted in Figure 5.8.

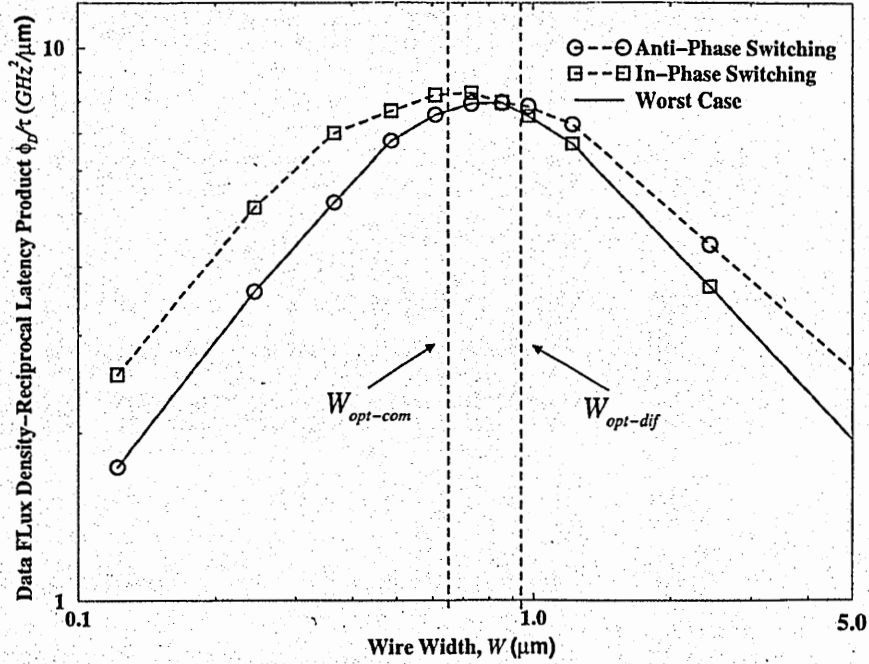


Figure 5.8: Data flux density-reciprocal latency product for in-phase and anti-phase switching cases. The optimal wire width for common and differential modes are different. The worst case  $\Phi_D/\tau$  is maximized when the difference between in-phase and anti-phase switching latencies is minimum.

It can be inferred from Figure 5.8 that the optimal wire width is the width at which  $\Phi_D/\tau$  for the two modes are equal, or in other words, the two modes have equal latencies. As it will be shown, a wire width equal to

$$W_{opt} = \sqrt{W_{opt-com} W_{opt-dif}} , \quad (5.12)$$

results in a small delay difference in most typical cases (less than 10% for  $0.3 < c_{orth}/c_g$ ) and therefore, can be used as the optimal wire width. Using (5.5), (5.7) and (5.11), the ratio of the optimal wire width for common and differential modes can be written as

$$\frac{W_{opt-dif}}{W_{opt-com}} = \frac{\sqrt{c_{dif} - c_{orth}}}{\sqrt{c_{com} - c_{orth}}} , \quad (5.13)$$

where  $c_{dif}$  and  $c_{com}$  are the equivalent capacitance per unit length for the differential and common modes, respectively. By substituting

$$c_{com} = c_g + c_m , \quad (5.14)$$

and (5.1) in (5.13) and assuming that the optimal spacing ratio is used ( $c_m = 0.45c_g$ ), (5.13) can be simplified to

$$\frac{W_{opt-dif}}{W_{opt-com}} = \sqrt{1.9} , \quad (5.15)$$

and (5.12) can be rewritten as

$$W_{opt} = \sqrt[4]{1.9} W_{opt-com} = \frac{1}{\sqrt[4]{1.9}} W_{opt-dif} . \quad (5.16)$$

To show that the delay variation at the optimal wire width is small, the ratio of the two latencies can be identified using the fact that the interconnect latency for each mode can be written in terms of the optimal wire width for that mode as

$$\tau_{mode} = (1 + 0.33 \frac{W_{opt-mode}^2}{W^2}) ToF_{mode} . \quad (5.17)$$

Using (5.17) and (5.12), the ratio of common and differential mode latencies at the optimal wire width can be written as

$$\frac{\tau_{dif}}{\tau_{com}} = \frac{(1 + 0.33 \frac{W_{opt-dif}^2}{W_{opt-com} W_{opt-dif}}) ToF_{dif}}{(1 + 0.33 \frac{W_{opt-com}^2}{W_{opt-com} W_{opt-dif}}) ToF_{com}}, \quad (5.18)$$

which can be simplified to

$$\frac{\tau_{dif}}{\tau_{com}} = 1.17 \frac{\sqrt{1 + \frac{c_{orth}}{1.88c_g}}}{\sqrt{1 + \frac{c_{orig}}{c_g}}}. \quad (5.19)$$

Equation (5.19) shows that the delay variation at the optimal wire width is a function of  $c_{orth}/c_g$  ratio. The normalized delay variation is plotted versus  $c_{orth}/c_g$  ratio in Figure 5.9, and it shows that for a wide range of  $c_{orth}/c_g$  ( $0.3 < c_{orth}/c_g$ ), the delay variation is less than 10%, considerably smaller than those of the deep RC or RLC cases.

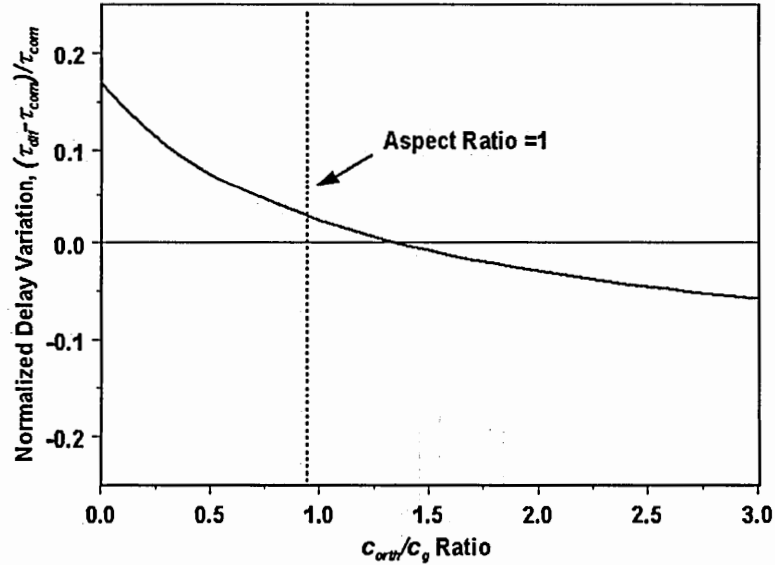


Figure 5.9: Normalized delay variation versus  $c_{orth}/c_g$  ratio when the optimal wire width is used. For a wide range of  $c_{orth}/c_g$  ratios, common and differential mode delays differ less than 10%.

The fact that using the optimal wire width results in a small delay variation is shown in Figure 5.10, wherein the normalized delay variation is plotted versus wire width. At the optimal wire width, the delay variation is less than 3%. This can be explained qualitatively. The differential mode has a larger capacitance which makes the characteristic impedance smaller and hence it has a larger attenuation in comparison with the common mode. Its propagation speed, however, is larger because of a smaller percentage increase in capacitance due to orthogonal lines. In the lossless case, the propagation speed determines the latency, whereas in the highly lossy case, the attenuation determines the latency. At the optimal wire width, however, the attenuation and propagation speed are equally important, and the common and differential mode latencies become almost equal.

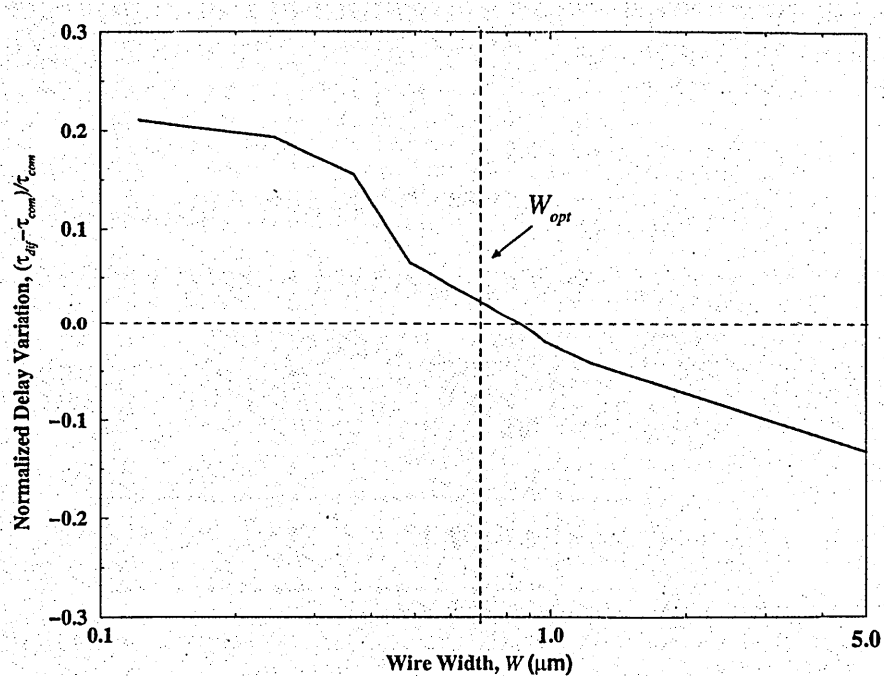


Figure 5.10: Normalized delay variation versus wire width for an optimally buffered interconnect implemented at the 45 nm technology node. At the optimal design point, delay variation is less than 3%.

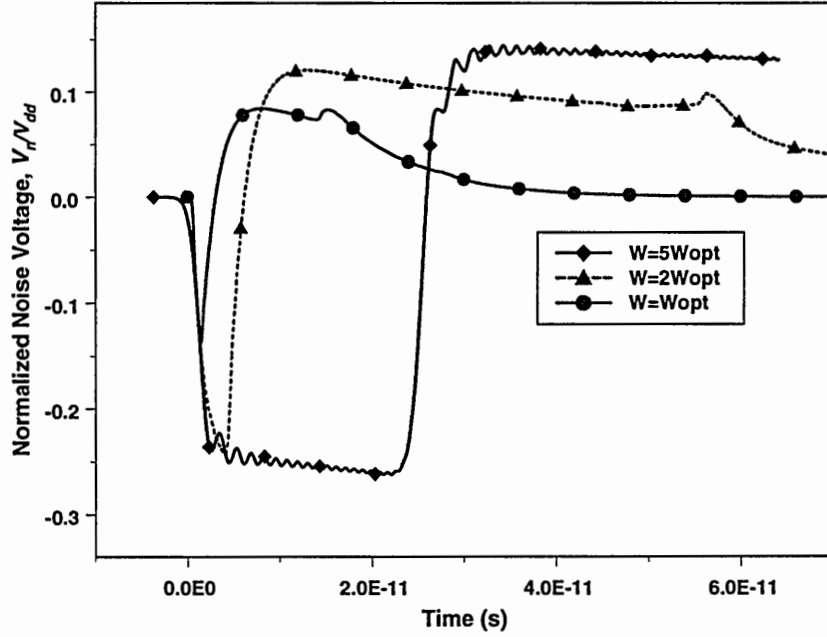


Figure 5.11: HSPICE simulations showing normalized crosstalk at the end of a quiet line when its adjacent line switches from low to high. Time zero corresponds to one time-of-flight delay. Peak and duration of out-of-phase noise match very well with what (5.25) and (5.31) predict with less than 3% error.

Another impact of wire width optimization is that it keeps crosstalk small and constant for all technology generations. Figure 5.11 shows the crosstalk voltage at the end of a quiet victim line when its adjacent line switches from low to high. It shows that the out-of-phase noise is dominant. The duration of the out-of-phase noise is

$$t_{out-phase} = \frac{\ell}{v_{com}} - \frac{\ell}{v_{dif}}, \quad (5.20)$$

where  $v_{com}$  and  $v_{dif}$  are the wave propagation speeds for the common and differential modes, and  $\ell$  is the interconnect length. Using (5.7) for each mode, (5.20) can be written as

$$t_{out-phase} = \frac{l_{seg} \sqrt{\epsilon_r}}{c_0} \left( \frac{1}{\sqrt{1 - \frac{c_{orth}}{c_{com}}}} - \frac{1}{\sqrt{1 - \frac{c_{orth}}{c_{dif}}}} \right), \quad (5.21)$$

where  $l_{seg}$  is the length of an interconnect segment between two repeaters, assuming that crosstalk for each segment should be modeled separately because a small noise voltage at the input of a repeater is not transferred to its output. The optimal number of repeaters in the RLC regime is [18]

$$k_{opt} = 0.95 R_{int} / \sqrt{Z_{com} Z_{dif}}, \quad (5.22)$$

which means that the resistance of each interconnect segment between two repeaters is

$$r\ell_{seg} = R_{int} / k_{opt} = 1.15 \sqrt{Z_{com} Z_{dif}}. \quad (5.23)$$

Duration of out-of-phase noise can, therefore, be written as

$$t_{out-phase} = 5R_0 C_0 \frac{W^2}{W_{opt}^2} \left[ \sqrt[4]{\frac{1 + \frac{c_{orth}}{c_g}}{1 + \frac{c_{orth}}{1.88c_g}}} - \sqrt[4]{\frac{1 + \frac{c_{orth}}{1.88c_g}}{1 + \frac{c_{orth}}{c_g}}} \right], \quad (5.24)$$

which can be approximated by

$$t_{out-phase} \approx 0.65 R_0 C_0 \left( \frac{c_{orth}}{c_g} \right)^{0.57} \frac{W^2}{W_{opt}^2}, \quad (5.25)$$

with less than 3% error for most practical cases ( $0.3 < c_{orth}/c_g < 3$ ). The open-ended out-of-phase noise voltage is

$$V_Q = -\frac{Z_{dif} V_{dd}}{Z_{dif} + R_{tr}} e^{-\frac{r\ell}{2Z_{dif}}}. \quad (5.26)$$

Assuming that optimal repeaters are used, the driver resistance is [18]

$$R_{tr} = \frac{R_0}{h_{opt}} = \frac{\sqrt{Z_{dif} Z_{com}}}{1.15}. \quad (5.27)$$

By substituting (5.23) and (5.27) in (5.26), the open-ended noise voltage can be written as

$$V_{open} = \frac{1}{1 + \frac{1}{1.15} \sqrt{\frac{Z_{com}}{Z_{dif}}}} \exp(-0.525 \sqrt{\frac{Z_{com}}{Z_{dif}}}). \quad (5.28)$$

which is approximately equal to

$$V_{open} = 0.245 V_{dd}, \quad (5.29)$$

with less than 4% error. The load capacitance is charged with the time-constant of  $Z_0 C_L$ , and the peak noise voltage considering the load capacitance is, therefore,

$$V_n = V_{open} (1 - \exp(-\frac{t_{neg}}{Z_{dif} C_L})), \quad (5.30)$$

which can be rewritten as

$$V_n = 0.245 V_{dd} \left[ 1 - \exp(-0.845 (\frac{c_{orth}}{c_g})^{0.57} \frac{W^2}{W_{opt}^2}) \right]. \quad (5.31)$$

Equation (5.31) shows that the peak crosstalk voltage is small (less than  $0.14V_{dd}$  for  $c_{orth}/c_g = 1$  and less than  $0.17V_{dd}$  for  $c_{orth}/c_g = 2$ ), and constant for all generations of technology if the optimal wire width is used.

## 5.4 Optimal Metal Thickness and Spacing

In Sections 5.2 and 5.3, the optimal wire width and the optimal spacing ratio are determined. The optimal values for the metal and dielectric thicknesses and the total spacing between interconnects ( $S_T = 2S_g + S_m$ ) are yet to be identified. A thin inter-level dielectric results in a large capacitance to orthogonal lines,  $c_{orth}$ , which makes the optimal wire width large, and therefore, reduces the data flux density. A large  $c_{orth}$  also makes the



difference between common and differential mode propagation speeds large that increases the peak and duration of out-of-phase crosstalk as (5.25) and (5.31) show. On the other hand, a thick inter-level dielectric is not technologically feasible because it requires vias with very large aspect ratios. Hence, for this analysis, it is assumed that the inter-level dielectric thickness is equal to the wire width. A slight change in the value of inter-level dielectric thickness has a small impact on the results of this section.

Metal thickness and the total spacing between interconnects are the two parameters that are left to be optimized. Since in all cases, it is assumed that the optimal wire width is used, the normalized metal thickness and total spacing values with respect to the wire width are considered, and the impact of  $T/W$  and  $S_T/W$  ratios on three key parameters, latency, data flux density and energy per bit, is studied.

By using the optimal wire width, interconnect latency remains equal to  $1.33T_{oF}$ . Time-of-flight, however, slightly changes for different  $T/W$  and  $S_T/W$  ratios because the capacitive loading caused by orthogonal lines ( $c_{orth}/c_g$ ) depends on  $T/W$  and  $S_T/W$  ratios. Figure 5.12 shows the latency of a 10mm-long interconnect versus  $S_T/W$  ratio for metal thicknesses of 0.5, 1, 2, and 3 times  $W_{opt}$ .

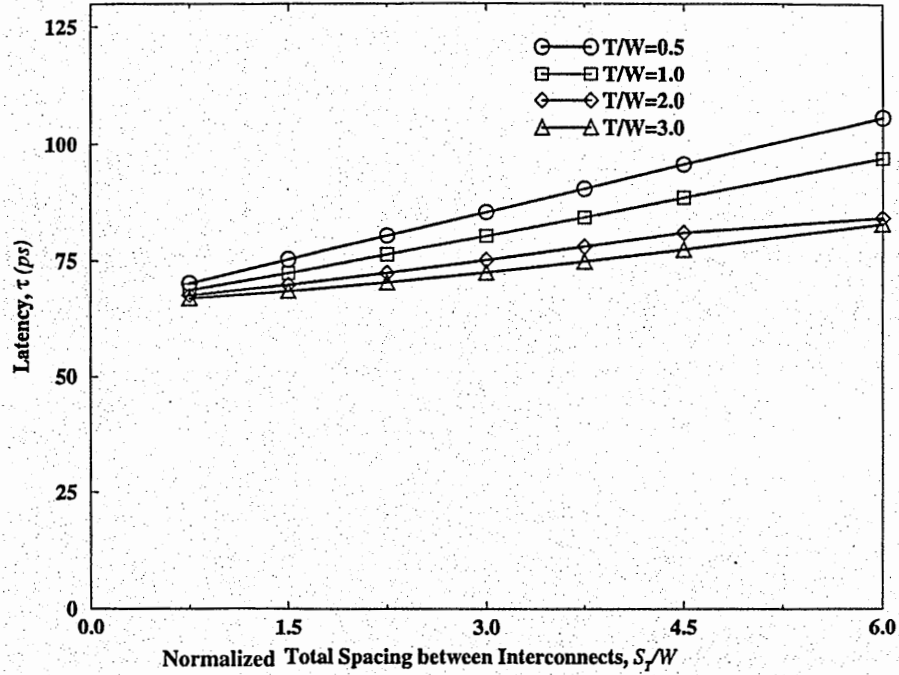


Figure 5.12: Interconnect latency versus total spacing for the four aspect ratios of 0.5, 1, 2, and 3. The total spacing ( $S_T = S_g + 2S_m$ ) is normalized to the wire width, and it is assumed that the optimal spacing ratio ( $S_m/S_g = 0.45$ ) and the optimal wire width are used.

Figure 5.12 shows that as the spacing between interconnects decreases or metal thickness increases, the interconnect latency becomes smaller. The reason is that the wave propagation speed increases as  $c_{orth}/c_g$  ratio decreases as (5.7) shows. The variation in latency, however, is not large. For instance, increasing  $T/W$  from 1 to 3 at  $S_T/W=1$ , results in only 13% reduction in latency.

The data flux density for the same interconnect is plotted in Figure 5.13 versus the total spacing for four different aspect ratios. For each aspect ratio, as the spacing between interconnects decreases, the optimal wire width increases because of larger capacitance per unit length which increases the geometry factor. For each aspect ratio, therefore, there

is a spacing which maximizes the data flux density. The curves, however, are flat around the peak values, which suggest using super- or sub-optimal values if necessary.

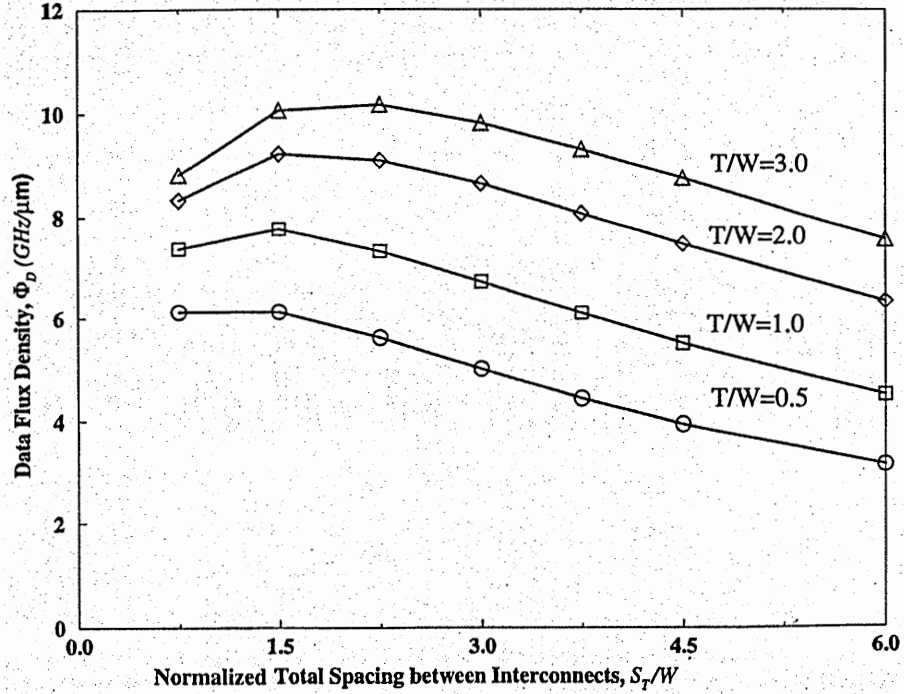


Figure 5.13: Data flux density versus total spacing for the aspect ratios of 0.5, 1, 2, and 3. Interconnects are 10mm long and are implemented at the 45 nm technology node.

Energy per bit is also plotted in Figure 5.14 which shows that as the spacing between interconnects decreases or the aspect ratio increases, the energy per bit increases rapidly. For instance, for the unity aspect ratio, reducing the  $S_T/W$  ratio increases the energy per bit by 41%, while increases the data flux density by only 9%. This shows that a super-optimal spacing is more energy efficient. To identify the spacing which offers the best trade-off between the data flux density and the energy per bit, the data flux density-reciprocal energy per bit product is plotted versus the total spacing for four aspect ratios in Figure 5.15. It can be seen that the peak values for  $\Phi_D/E_b$  for different aspect ratios are

about the same. However, the total spacing at which  $\Phi_D/E_b$  attains its maximum is larger for larger aspect ratios. Hence, to have a good trade-off between data flux density and energy per bit, a larger spacing should be used for larger aspect ratios. The spacing which results in maximum  $\Phi_D/E_b$  is plotted versus the aspect ratio in Figure 5.16 from which it can be acquired that the energy-efficient spacing can be approximately written as

$$S_{T-\Phi_D/E_b-\max} \approx 2 + \frac{T}{W}. \quad (5.32)$$

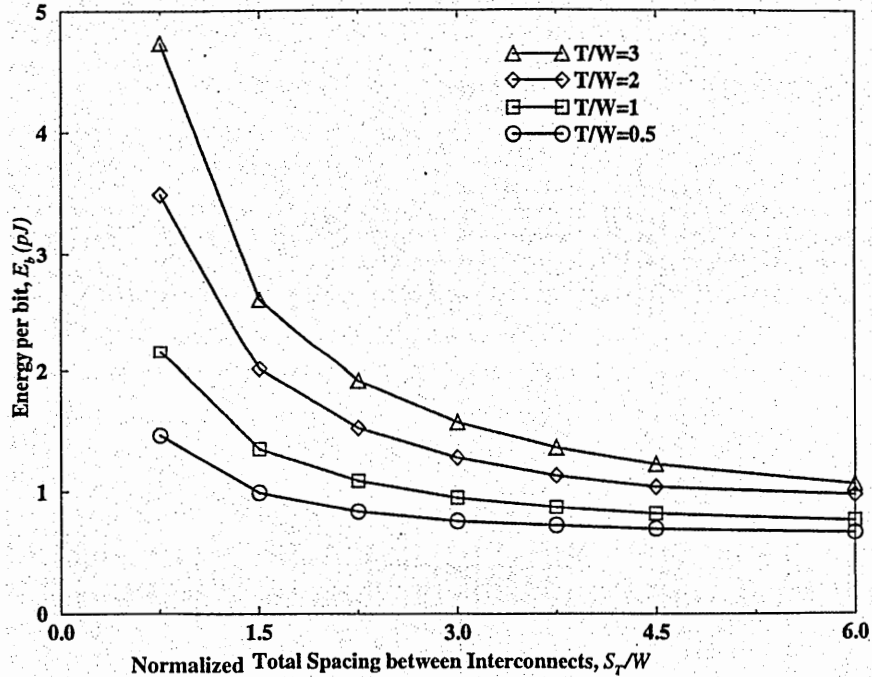


Figure 5.14: Energy per bit versus total spacing between interconnects for the aspect ratios of 0.5, 1, 2, and 3. Interconnects are assumed to be 10 mm long.

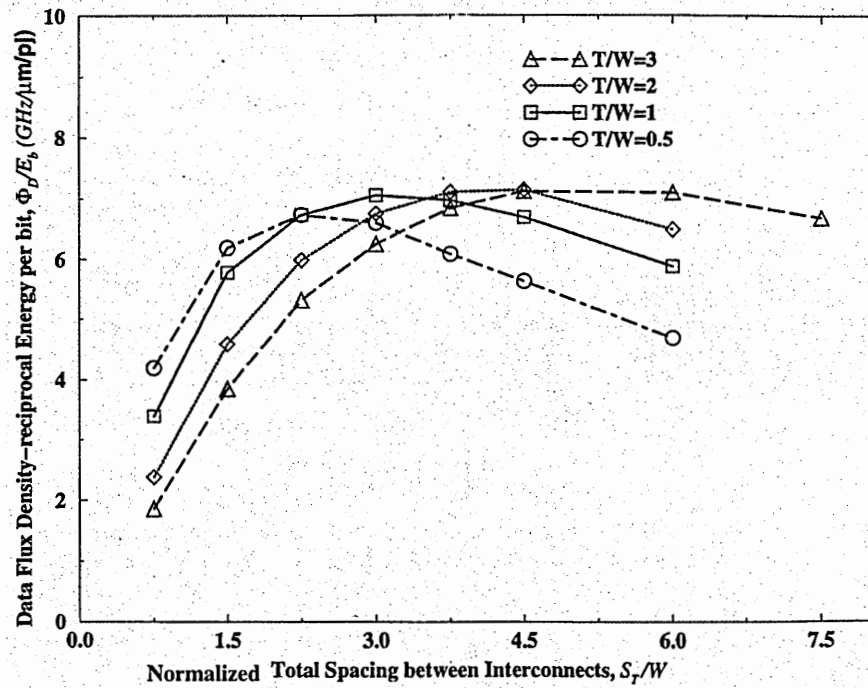


Figure 5.15: Data flux density-reciprocal energy per bit product versus total spacing between interconnects for aspect ratios of 0.5, 1, 2, and 3. Interconnects are 10mm long and are implemented at the 45nm technology node.

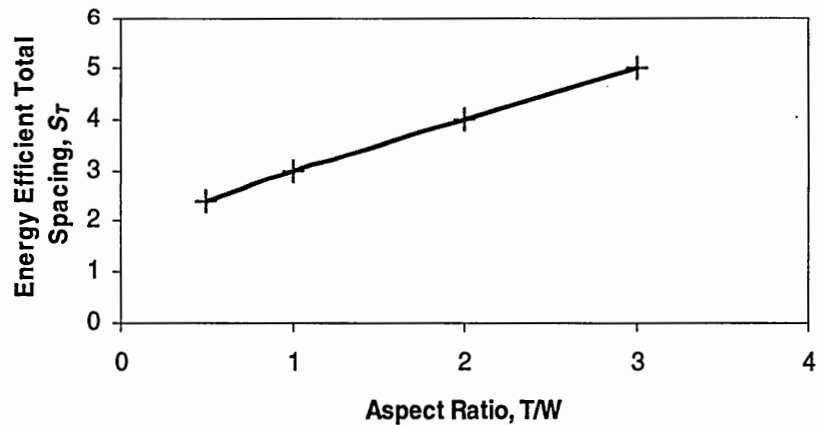


Figure 5.16: The energy efficient spacing versus aspect ratio. The energy efficient spacing, which maximizes  $\Phi_D/E_b$ , increases as the aspect ratio increases.

It should be noted that by using a larger aspect ratio, a larger data flux density can be achieved at the price of a larger energy per bit. For instance, by using an aspect ratio of 3 instead of 1, data flux density increases by 24% and energy per bit increases by the same percentage. Hence, for high-performance applications, a larger aspect ratio can be used while for low-power applications, a smaller aspect ratio is more appropriate.

## 5.5 Conclusions

Compact physical models for the delay and crosstalk of on-chip co-planar transmission lines are used to optimize the design of global interconnects. For the case that there are two signal lines between power and ground lines, it is proven that the latency is minimized when the ratio of signal-ground spacing to signal-signal spacing is 0.45 independent of the interconnect aspect ratio or the absolute values of the spacing between interconnects. The optimal wire width is also identified that maximizes the data flux density-reciprocal latency product. The data flux density-reciprocal latency product is maximized so that the global interconnects can transfer as many bits as possible with low latency. The value of the optimal width is independent of the interconnect length and is determined by the resistivity of the metal, the intrinsic delay of the repeaters, and the wire geometry. Therefore, a unique optimal wire width can be used for virtually all global interconnects regardless of their lengths. Using the optimal wire width results in a 42% smaller latency, 30% smaller energy-per-bit, and 84% smaller repeater area at a cost of only a 14% decrease in data flux density, compared to using half the optimal wire width (sub-optimal design). On the other hand, using twice the optimal wire width (super-optimal design) results in only a 14% decrease in latency at the cost of a 35% decrease in

data flux density, compared to the optimal wire width design. By simultaneously using the optimal spacing ratio and optimal wire width, the dynamic delay variation and crosstalk are limited to 10% and  $0.2V_{dd}$ , respectively. Finally, the trade-off between the data flux density and energy per bit is presented which suggests using larger spacing between interconnects for larger aspect ratios to maximize the data flux density-reciprocal energy per bit product.

While in this chapter it is assumed that power and ground lines are wide enough to isolate signal lines from far aggressors, the next chapter focuses on deriving compact physical models for far inductive noise and its impact on the global interconnect optimization.

## Chapter 6

### Multilevel Interconnect Crosstalk Modeling

#### 6.1 Introduction

Rigorous models for the crosstalk of co-planar distributed RLC lines above orthogonal lines have been presented in Chapter 4. For these models, it has been assumed that power and ground lines are wide enough that they isolate signal lines from far lines. This assumption, however, may not be true if power and ground lines are not wide enough, and in fact, a large noise may be induced by far aggressors. In this chapter, new compact physical models for the total crosstalk of co-planar RLC lines are presented that consider near and far aggressors. Far aggressors that are two metal levels below the victim line also are taken into account.

In Section 6.2, the problem and the methodology of solving it are defined. Far inductive crosstalk is modeled in Section 6.3 for the case that aggressor and victim lines are identical. Impact of far aggressors that are not identical to a victim line is then modeled in Section 6.4. In Section 6.5, the noise voltage-time integral is introduced and the interconnect length at which this integral attains its maximum is identified. Superposition theorem is used in Section 6.6 to calculate crosstalk noise when a victim line is attacked by near and far lines simultaneously. An integrated crosstalk model is proposed in Section 6.7 to calculate the crosstalk noise voltage induced by virtually all



near and far aggressors. Finally, in Section 6.8 it is proved that the optimal wire dimensions make crosstalk small and constant in various technology generations.

## 6.2 Methodology

Figure 6.1 shows a cross-sectional view of four top metal levels in a GSI chip. Wires in the top two metal levels are relatively thick and wide, and have prominent inductive effects. Hence, each signal line has a nearby power/ground line to have a nearby return path. Wires in lower metal levels are more resistive, and, therefore, not every signal line has a nearby power/ground line.

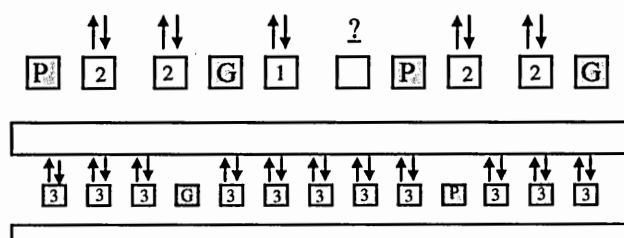


Figure 6.1: A cross-sectional view of 4 top metal levels. Top two levels are relatively fat and due to inductive effects each signal line has a nearby power/ground line as a return path. The spaces between signal and ground lines are 0.45 times smaller than signal to signal spaces to reduce crosstalk and minimize worst-case delay.

A victim line on the top most level can be affected by three types of aggressors, (1) a nearby aggressor, (2) far aggressors in the same metal level, and (3) far aggressors that are two metal levels below the victim line. The near aggressor is inductively and capacitively coupled to the victim line whereas far aggressors are only inductively

coupled to the victim line. Aggressors that are within the same metal level as the victim line have the same inductance, capacitance, and resistance per unit length as the victim line does. Far aggressors in a lower metal level, however, have different resistance, inductance, and capacitance values compared with the victim line. Hence, the impact of different kinds of aggressors can be quite different. The mutual inductance between the victim and orthogonal lines is zero; therefore, they do not contribute to noise. One way of solving this complicated problem is to solve the set of differential equations described by

$$\frac{\partial^2}{\partial x^2}[V(x,t)] = [L][C]\frac{\partial^2}{\partial t^2}[V(x,t)] + [R][C]\frac{\partial}{\partial t}[V(x,t)], \quad (6.1)$$

where  $[R]$ ,  $[L]$  and  $[C]$  are resistance, inductance and capacitance matrices, respectively. Since, different aggressors have different characteristics, a direct analytical solution for (6.1) may not be feasible. A numerical solution can also be very time consuming, and offers little physical insight.

To find compact physical models for this problem, the near aggressor is ignored initially. Two cases are then solved. In the first case, a victim line is attacked by some aggressors that are identical to the victim. In the second case, a victim line is affected by aggressors that are not identical to the victim line. By using the superposition theorem, the total noise caused by all near and far aggressors is modeled.

### 6.3 Identical Victim and Aggressor Lines

In the case that far lines and a victim line are identical, far inductive noise is found by solving (6.1), where  $[R]$  is substituted by scalar  $r$ , the loop resistance per unit length of each line. The loop resistance includes the resistance of return paths which are the power

and ground lines. Since capacitance is a local effect, mutual capacitances between the far lines and the quiet line are negligible. The worst case crosstalk occurs when all far aggressors switch in phase, and therefore, the mutual capacitance between the far lines can be ignored, too. Hence,  $[C]$  can be substituted by scalar  $c$ , the distributed self capacitance of the lines (Figure 6.2).

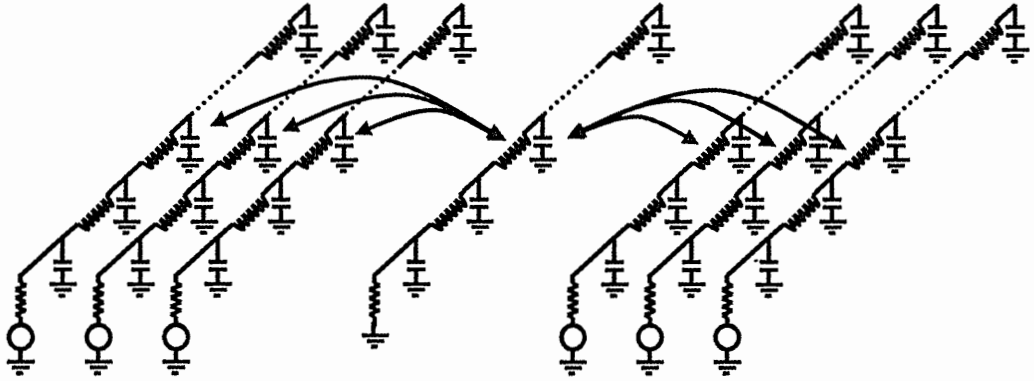


Figure 6.2: A quiet victim line is attacked by some far aggressors. It has been assumed that there is no nearby aggressor and far lines are inductively coupled to the victim line. The resistance and capacitance matrices can be substituted by scalar resistance and capacitance values.

By finding the eigen-vectors of the inductance matrix, the set of differential equations described by (6.1) can be decoupled. For the case that all aggressors switch in-phase (the worst case for crosstalk), the following two modes of propagation describe the system:

$$V^- = \sum_{\text{All Aggressors}} \left( \frac{l_{vi}}{l_s} V_i \right) + \sqrt{\sum_{\text{All Aggressors}} \frac{l_{vi}^2}{l_s^2}} V_v, \quad (6.2)$$

and

$$V^+ = \sum_{All \text{ Aggressors}} \left( \frac{l_{vi}}{l_s} V_i \right) - \sqrt{\sum_{All \text{ Aggressors}} \frac{l_{vi}^2}{l_s^2}} V_v, \quad (6.3)$$

where  $l_{vi}$  is the mutual inductance between the victim line and the  $i^{th}$  aggressor,  $l_s$  is the self inductance of lines,  $V_i$  is the voltage of the  $i^{th}$  aggressor, and  $V_v$  is the voltage of the victim line. In the two-line case, the plus mode represents the aggressor and victim lines switching in the same direction, and the minus mode represents switching in opposite directions. The equivalent inductance and capacitance per unit length for each mode are given by

$$l^+ = l_s + \sqrt{\sum_{All \text{ Aggressors}} l_{vi}^2} \quad ; \quad c^+ = c, \quad (6.4)$$

and

$$l^- = l_s - \sqrt{\sum_{All \text{ Aggressors}} l_{vi}^2} \quad ; \quad c^- = c. \quad (6.5)$$

The propagation speed for the plus and minus modes is given by

$$v = 1 / \sqrt{lc}, \quad (6.6)$$

where  $l$  is the equivalent inductance per unit length corresponding to each mode. Voltage of each mode can be solved by either exact or low-loss approximate solutions for distributed RLC lines [26]. Low-loss approximation is used here because it results in simple expressions that are insightful. The error is also small, especially for the peak noise voltage. The differential equation for a single RLC line in Laplace domain is

$$\frac{\partial}{\partial x^2} V(x, s) = V(x, s) lcs^2 \left( 1 + \frac{r}{sl} \right), \quad (6.7)$$

and the solution of (6.7) for an infinite long line is

$$V_{\text{inf}}(x, s) = V_{\text{in}}(s) \frac{Z(s)}{Z(s) + R_{\text{tr}}} e^{-xs\sqrt{lc}\sqrt{1+\frac{r}{sl}}}, \quad (6.8)$$

where  $R_{\text{tr}}$  is the driver resistance,  $V_{\text{in}}$  is the input voltage, and  $Z(s)$  is the lossy characteristic impedance defined as

$$Z(s) = Z_0 \sqrt{\frac{s+r/l}{s}}, \quad (6.9)$$

where the loss-less characteristic impedance is  $Z_0 = \sqrt{\frac{l}{c}}$ .

Assuming that

$$\left| \frac{r}{ls} \right| \ll 1, \quad (6.10)$$

which is the low-loss approximation, (6.8) can be approximated by

$$V_{\text{inf}}(x, s) = V_{\text{in}}(s) \frac{Z_0}{Z_0 + R_{\text{tr}}} e^{-x\sqrt{lc}(s+\frac{r}{2l})}, \quad (6.11)$$

and for a step input, (6.11) in time domain can be written as

$$V_{\text{inf}}(x, t) = \frac{Z_0 V_{\text{dd}}}{Z_0 + R_{\text{tr}}} e^{-\frac{rx}{2Z_0}} u(t - x\sqrt{lc}), \quad (6.12)$$

which is the low-loss approximate solution for a single distributed RLC line. Note that even for lossy lines (6.10) is valid at high frequencies ( $s = j\omega$ ), which means that the low-loss approximation accurately describes the voltage of the line for  $t \approx x\sqrt{lc}$ . As it will be shown, the peak noise occurs at a time close to the time-of-flight (ToF). The low-loss approximation can, therefore, accurately model the peak noise voltage. For a further accurate analysis, one can use the rigorous solution for a single RLC line.

Voltage of the victim line can be written in terms of the two modes as

$$V_v = (V^+ - V^-) / \sqrt{\sum_{All \text{ Aggressors}} (l_{vi}^2 / l_s^2)}. \quad (6.13)$$

As (6.4) and (6.5) show, the minus mode has a smaller equivalent inductance. Hence, the propagation speed of the minus mode is larger, and as (6.13) shows, an out-of-phase noise appears at the end of the victim line. Assuming that aggressors are excited by step inputs and using the low-loss approximation, the noise voltage at the end of a quiet victim line is

$$V_{identical}(t) = - \frac{\sum_{All \text{ Aggressors}} l_{vi}}{\sqrt{\sum_{All \text{ Aggressors}} l_{vi}^2}} \frac{Z_0}{R_{tr} + Z_0} V_{F-in} e^{-\frac{r\ell}{2Z_0}}, \quad (6.14)$$

for  $\sqrt{c(l - \sum_{All \text{ Aggressors}} l_{vi})\ell} < t < \sqrt{c(l + \sum_{All \text{ Aggressors}} l_{vi})\ell}$ , which is the time that the minus mode has arrived at the end of victim line and the plus mode has not arrived yet. For  $t > \sqrt{c(l + \sum_{All \text{ Aggressors}} l_{vi})\ell}$ , the plus and minus modes cancel impact of each other and the noise voltage can be ignored.  $V_{F-in}$  in (6.14) is the input voltage of the aggressor lines and  $\ell$  is the length of interconnects. If aggressors switch from logic 0 to 1,  $V_{F-in}$  is  $+V_{dd}$ , and if they switch from logic 1 to 0 it is  $-V_{dd}$ . The noise duration is

$$t_n = \frac{\ell}{v^+} - \frac{\ell}{v^-} = (ToF) \sqrt{\sum_{All \text{ Aggressors}} (l_{vi} / l_s)^2}, \quad (6.15)$$

where  $ToF$  is the time-of-flight. In the case that there is a large capacitance at the end of the victim line, the load capacitance is charged through the characteristic impedance of the line, and the noise voltage is

$$V_{load} = \begin{cases} V_{open} (1 - e^{-t/Z_0 C_L}) & t \leq t_n \\ V_{open} (1 - e^{-t_n/Z_0 C_L}) e^{-t/Z_0 C_L} & t > t_n \end{cases}, \quad (6.16)$$

where  $V_{open}$  is given by (6.14). As an example, the induced noise on a middle quiet line is plotted versus time in Figure 6.3 when all signal lines are shielded from both sides by power/ground lines. All inductance and capacitance values are extracted by RAPHAEL [17].

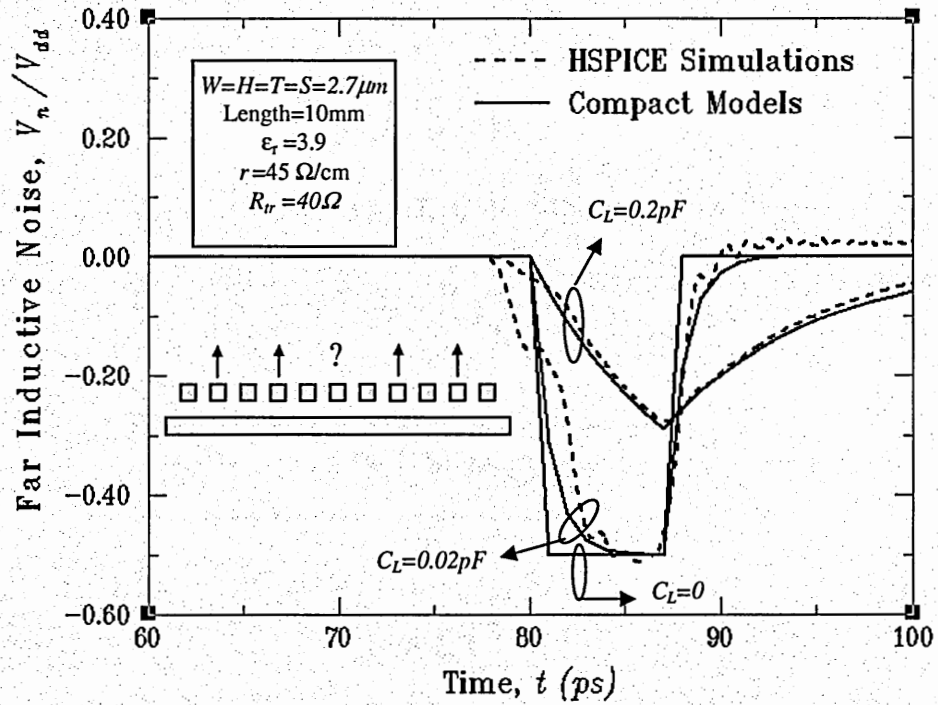


Figure 6.3: The induced noise on the quiet line when all signal lines are shielded by two power/ground lines. HSPICE simulations are compared with compact models for different load capacitances. The input of the aggressors is a positive step voltage,  $V_{dd} u_o(t)$ .

Equation (6.14) shows that the peak noise voltage of an open-ended line is independent of the absolute values of the inductive couplings. This means that if all mutual couplings are made smaller by the same factor, e.g. by making power/ground lines wider, the peak noise does not change. The reason is that the aggressor and victim lines are identical, and their natural frequencies are the same. Hence, the aggressors and

the victim line resonate, and even a very weak coupling can cause a considerable noise voltage. This is explained in more detailed in Appendix B. Figure 6.4 shows noise voltage and duration versus ground line width for the structure shown in Figure 6.3. Making ground lines wider and not changing other dimensions reduces all mutual inductances almost equally.

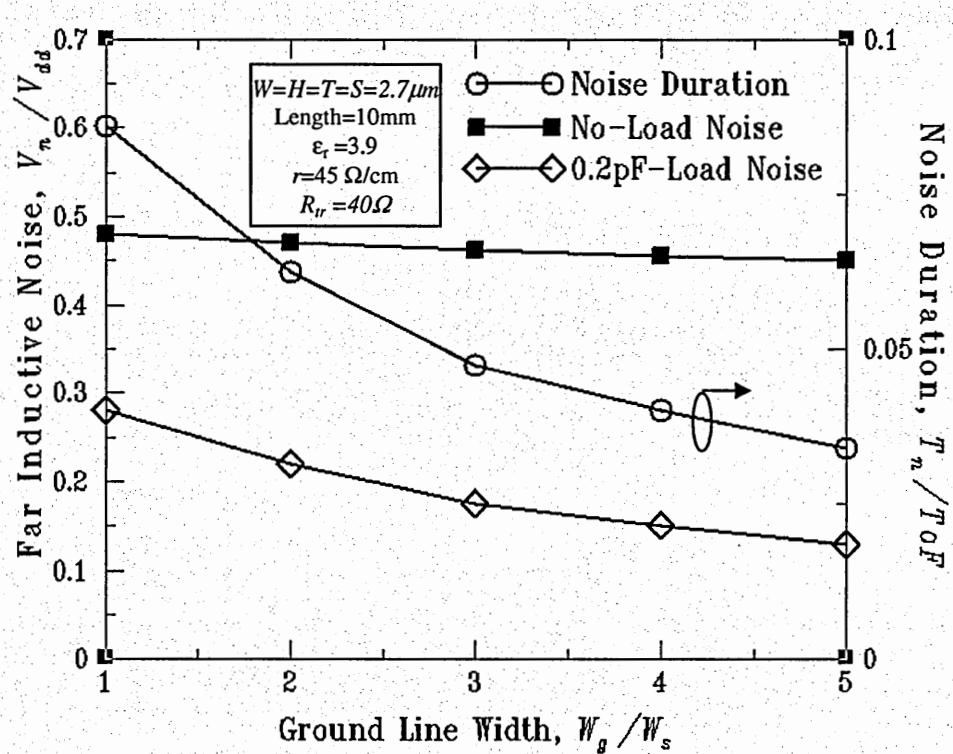


Figure 6.4: The peak and duration of far inductive noise versus the ground line width for the structure shown in Figure 6.3. Increasing ground line width reduces all mutual inductances approximately by the same ratio. The peak noise of an open-ended line is independent of ground line width.

## 6.4 Non-Identical Victim and Aggressor Lines

Equations (6.14) and (6.15) do not hold if aggressors and a quiet line have different capacitance and inductance values. In this case, far aggressors can be modeled by the single low-loss RLC line model given by



$$V_f(x, t) = \frac{Z_{0f}}{Z_{0f} + R_{lr}} V_{F-in} e^{\frac{r_f \ell}{2Z_{0f}}} u(t - x\sqrt{l_f c_f}), \quad (6.17)$$

where  $l_f$  and  $c_f$  are the far lines' inductance and capacitance per unit length, respectively,  $Z_{0f}$  is the far lines' characteristic impedance, and  $u(t)$  is a unit step function. This is based on a loosely coupled assumption that the impact of the quiet line on the aggressors can be neglected [36]. If the coupling between the aggressors is not negligible, their common mode can be used. The differential equation for the victim line is given by

$$\frac{\partial^2 V(x, t)}{\partial x^2} = l_v c_v \frac{\partial^2 V(x, t)}{\partial t^2} + r_v c_v \frac{\partial V(x, t)}{\partial t} + \sum_{\text{all Aggressors}} l_{vi} c_{fi} \frac{\partial^2 V_f(x, t)}{\partial t^2}, \quad (6.18)$$

where  $r_v$ ,  $l_v$  and  $c_v$  are resistance, inductance and capacitance per unit length of the victim line, respectively. The voltage at the end of the victim line for  $x\sqrt{l_v c_v} < t < x\sqrt{l_f c_f}$  is

$$V_{\text{non-identical}}(t) = \frac{-2Z_{0f} V_{F-in}}{Z_{0f} + R_{lr}} \frac{\sum_{\text{All Aggressors}} l_{vi} c_{fi}}{(l_f c_f - l_v c_v)} \left[ \frac{\ell \sqrt{l_f c_f} - t}{\ell \sqrt{l_f c_f} - \ell \sqrt{l_v c_v}} e^{\frac{r_v \ell}{2Z_{0v}}} + \frac{t - \ell \sqrt{l_v c_v}}{\ell \sqrt{l_f c_f} - \ell \sqrt{l_v c_v}} e^{-1.15 \frac{r_f \ell}{2Z_{0f}}} \right], \quad (6.19)$$

and at all other times is approximately zero. In most cases, the two exponents inside the bracket are close to each other, and (6.19) can be approximated by

$$V_{\text{non-identical}}(t) = \frac{-2Z_{0f}}{Z_{0f} + R_{lr}} V_{F-in} \frac{\sum_{\text{All Aggressors}} l_{vi} c_{fi}}{(l_f c_f - l_v c_v)} \left[ \exp\left(-\frac{r_v \ell}{2Z_{0v}}\right) \right]. \quad (6.20)$$

By substituting (6.20) in (6.16), the peak noise voltage can be found when there is a load capacitance. An example of this case is when a victim line that is shielded from both sides is attacked by aggressors that are shielded only from one side. The noise voltage at the end of the victim line is plotted versus time in Figure 6.5 wherein HSPICE simulations verify compact expressions.

Unlike the previous case, the induced noise in the victim line travels faster or slower than the signal in aggressors, and far and victim lines do not resonate. *The peak noise of an open-ended line is, therefore, proportional to the total mutual inductance, and the noise duration is independent of the mutual couplings.*

It should be noted that using the loosely coupled assumption for the identical victim and aggressor lines results in an impulse function in the solution. The reason is that for the identical lines, aggressor and victim lines resonate and ignoring the impact of the victim line on aggressors makes the solution unrealistic. Hence, the solutions that are presented for identical and non-identical lines cannot be used interchangeably.

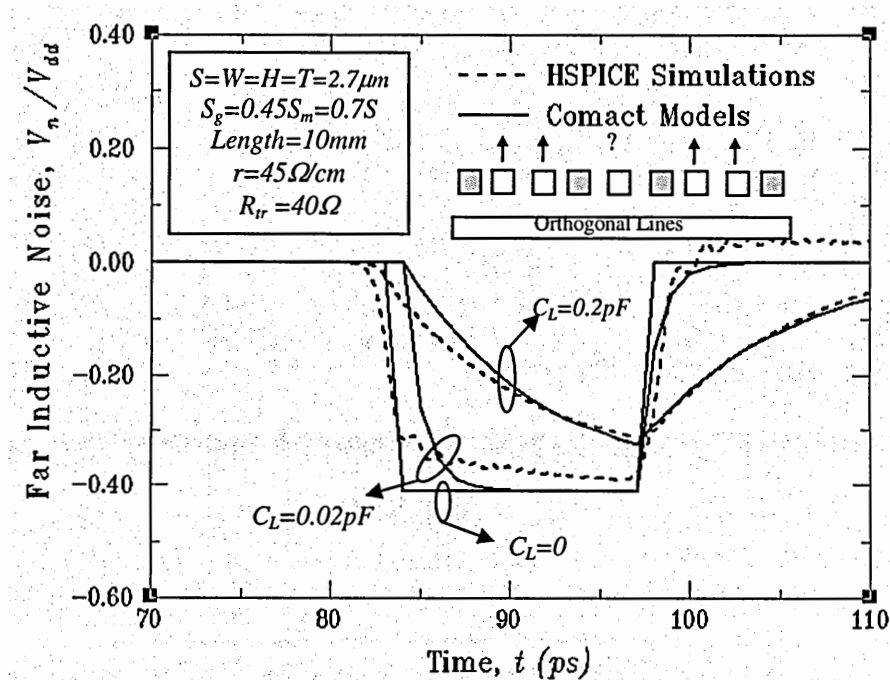


Figure 6.5: The induced noise at the end of a quiet line for three different load capacitances. HSPICE simulations are compared with compact models.  $S_g$  and  $S_m$  are optimized so that the worst-case delay is minimized.

Another application for (6.20) is when a quiet line is attacked by far lines which are two metal levels below the victim line. Lines in lower metal levels usually have a smaller cross-sectional area and are more resistive. Number of signal lines between power/ground lines is typically larger than that of the top most metal level. Having more than two signal lines between power/ground lines makes it more difficult to model far lines in lower metal levels because different signal lines have different inductance values based on their distance from power/ground lines. The worst case far inductive noise occurs when all aggressors switch simultaneously in the same direction because if one of them switches in an opposite direction it cancels the impact of others. Since far lines that are surrounded by power/ground lines are highly coupled, both inductively and capacitively, their voltages are close as they switch, and therefore with a good approximation, all of them can be modeled by a single equivalent line. The equivalent line's capacitance is equal to the total capacitance of the lines to lower and upper orthogonal lines and power/ground lines. The inductance and resistance of the equivalent line is the inductance and resistance of the lines in parallel.

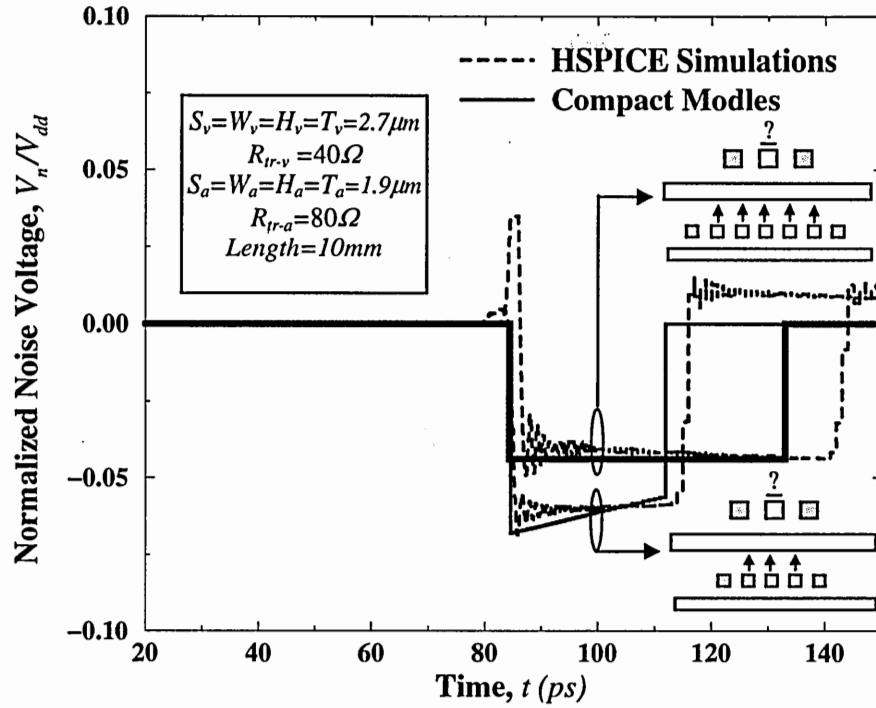


Figure 6.6: Voltage at the end of a quiet victim line when it is affected by either three or five far aggressors that are two metal levels below the victim line. Line resistance for the two top levels is  $45 \Omega/\text{cm}$  and for the lower metal levels is  $90 \Omega/\text{cm}$ . The parameters with index of “v” correspond to the victim line, and the parameters with index of “a” correspond to the aggressors.

The above mentioned method is verified against HSPICE simulations in Figure 6.6 where a quiet victim line is attacked by either three or five far aggressors. Although there is a larger mutual coupling between the aggressors and the victim line in the five-aggressor case, the peak noise voltage for the five-aggressor case is smaller. The reason is that the  $(l_v c_v - l_c f)$  term in (6.20) for the three-aggressor case is smaller than that of the five aggressor case. A simple physical interpretation is that the natural frequencies of the aggressors and the victim are closer in the three-aggressor case, and the aggressors and the victim line are closer to resonance.

In both cases, the errors in the peak and duration of noise are less than 10% and 18%, respectively. It should be noted that there is a narrow noise pulse in the noise graph obtained by the HSPICE simulations for the five-aggressor case that is not predicted by the compact expressions. This is because of the fact that the far aggressors have been modeled with a single equivalent line assuming that they have equal currents. However, in reality, their currents are slightly different and that causes the narrow noise pulse. Since this noise pulse is very narrow, it is filtered out by the load capacitance and can be neglected.

## 6.5 Noise Voltage-Time Integral

Both peak and duration of noise are important because even a large noise voltage with a very small duration cannot cause false switching and it will be filtered out by the load capacitance. The noise voltage-time integral can be, therefore, defined as a figure of merit. Using (6.16), it can be shown that the noise voltage-time integral is

$$\int_0^{\infty} V_n(t)dt = V_{open} t_{neg}, \quad (6.21)$$

which is independent of the load capacitance. For both identical and non-identical line cases, (6.21) is equal to

$$\int_0^{\infty} V_n(t)dt = \frac{2Z_{0f}}{Z_{0f} + R_{tr}} \frac{\sum_{\text{All Aggressors}} l_{vi} c_f}{(\sqrt{l_f c_f} + \sqrt{l_v c_v})} V_{dd} \ell e^{-\frac{r_v \ell}{2Z_{0v}}}. \quad (6.22)$$

It should be noted that for the case that lines are identical parameters corresponding to the victim and far lines are equal. By taking the derivative of (6.22) with respect to interconnect length, the length at which the noise voltage-time integral attains its

maximum can be identified. *The noise voltage-time is maximized when interconnect resistance is equal to twice characteristic impedance:*

$$r\ell = 2Z_0. \quad (6.23)$$

It is therefore imperative to avoid interconnect lengths close to  $2Z_0/r$  (Figure 6.7). Knowing this fact can be especially useful for repeater insertion algorithms.

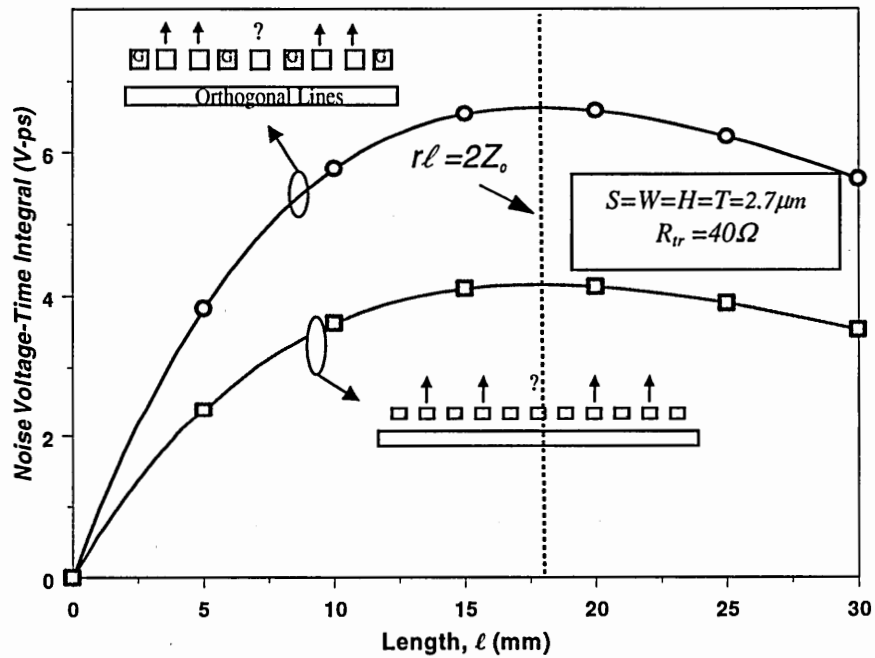


Figure 6.7: Noise voltage-time integral versus interconnect length.  $V_{dd}$  is assumed to be 1V. The value of this integral is independent of the load capacitance.

## 6.6 Near and Far Aggressors

When a victim line is affected by both near and far aggressors, total crosstalk can be found by using superposition theorem. The following subsections show how models that are derived in the previous sections can be used to find total crosstalk caused by near and far aggressors. First, the crosstalk induced by intra-level aggressors is modeled, and later, the impact of inter-level far lines is modeled.

### 6.6.1 Near and Intra-Level Far Aggressors

To use the superposition theorem, crosstalk should be modeled for two cases: far lines switch and the adjacent line stays quiet, and vice versa.

#### Far Lines Switch and The Near Line Stays Quiet

In this case, each two adjacent signal lines have equal voltages, and can be treated as a single line. The equivalent inductance of each pair is

$$l_p = (l_s + l_m) / 2, \quad (6.24)$$

and its equivalent capacitance is

$$c_p = 2c_g + 2c_{orth}, \quad (6.25)$$

where  $l_m$  is the mutual inductance between each two adjacent signal lines,  $c_g$  is the capacitance between a signal line and its adjacent power/ground line, and  $c_{orth}$  is the capacitance to lower orthogonal lines. The equivalent driver resistance and line resistance are  $R_{tr}/2$  and  $r/2$ , respectively. The voltage induced on the victim line can be found by the rigorous solution found in Section 6.3 as

$$V_{f\text{-int}r\text{alevel}} = V_{\text{identical}}(t, l_{vi} = l_{Pi}, l = l_p, c = c_p, R_{tr} = R_{tr}/2, r = r/2), \quad (6.26)$$

where  $l_{Pi}$  is the mutual inductance between the  $i^{\text{th}}$  pair of far aggressors and the victim-near aggressor pair and its value can be written as

$$l = (l_{vf1} + l_{vf2} + l_{Nf1} + l_{Nf2})/2. \quad (6.27)$$

The results are verified against HSPICE simulations and plotted in Figure 6.8.

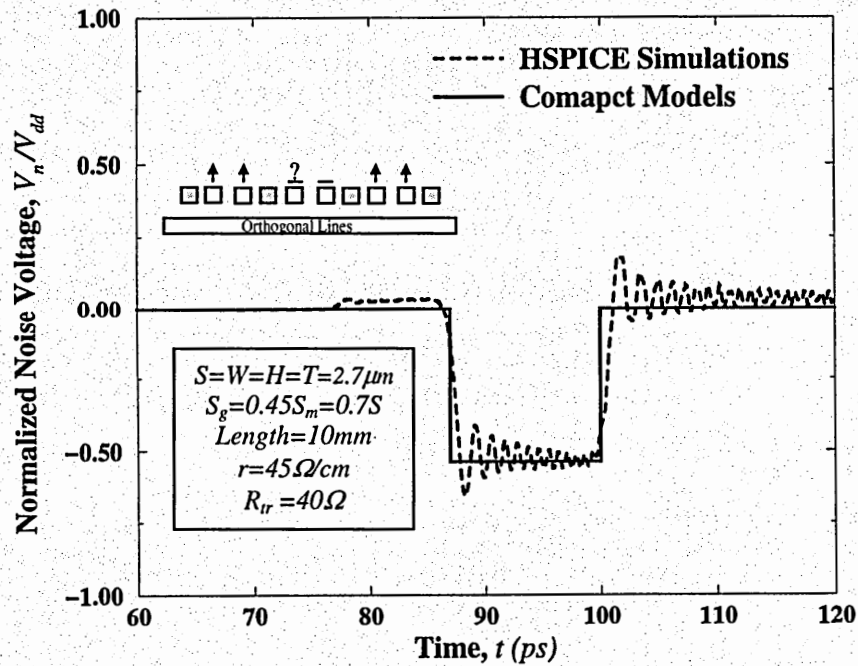


Figure 6.8: Noise voltage at the end of a quiet victim line when intra-level far lines switch upward and a near aggressor stays quiet.



## The Near Line Switches and Far Lines Stay Quiet

In this case, although the far aggressors are quiet, they cannot be neglected because the far and near aggressors are identical, and they hence resonate. This means that as the near aggressor switches, a considerable current is induced in the far aggressors, which in turn, can affect the noise voltage in the victim line.

This case can be solved by superposition of in-phase and anti-phase switching of the victim line and the near aggressor. For the in-phase switching case, voltage of each two adjacent lines are equal, and they can be treated as a single line. Likewise, using (6.13), the voltage of the victim line can be calculated as

$$V_{com}(t) = \frac{Z_{0P}}{R_{tr}/2 + Z_{0P}} V_{N-in} e^{\frac{r\ell/2}{2Z_{0P}}} \left[ u(t - \sqrt{c_p(l_p + \sqrt{\sum_{All\ Far\ Pairs} l_{pi}} \ell)}) + u(t - \sqrt{c_p(l_p - \sqrt{\sum_{All\ Far\ Pairs} l_{pi}} \ell)}) \right]. \quad (6.28)$$

For the anti-phase switching case, far lines can be ignored because they receive a negligible magnetic flux change, and no current is induced in them to affect the victim line. In this manner, the voltage at the end of the victim line is

$$V_{dif}(t) = -\frac{Z_{dif}}{R_{tr} + Z_{dif}} V_{N-in} e^{\frac{r\ell}{2Z_{dif}}} \left[ u(t - \ell \sqrt{c_{dif} l_{dif}}) \right], \quad (6.29)$$

where

$$l_{dif} = l_s - l_m, \quad (6.30)$$

and

$$c_{dif} = c_g + c_{orth} + 2c_m. \quad (6.31)$$

The induced voltage on the victim line when the nearby aggressor switches and far lines are quiet is

$$V_{Near} = V_{com}(t) + V_{dif}(t), \quad (6.32)$$

which is verified in Figure 6.9.

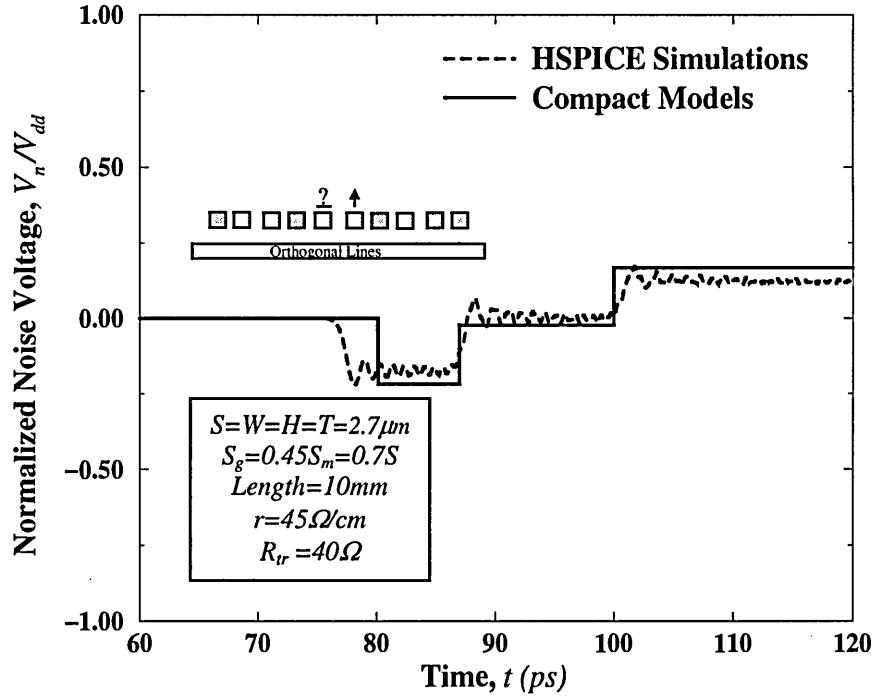


Figure 6.9: Noise voltage at the end of a quiet victim line when intra-level far lines are quiet and a near aggressor switches upward.

### Both Near and Far Lines Switch

By adding (6.26) and (6.32), the total noise caused by near and far aggressors can be found. Figure 6.10 shows noise voltage at the end of a quiet victim line when all aggressors switch from low to high exactly at the same time. This might happen if all lines are synchronized by a clock. Otherwise, there might be a phase shift between different drivers. Also, they may switch in different directions. Hence, Figure 6.10 does not show the worst case crosstalk. The worst case scenario is when far aggressors switch

in the opposite direction that the near aggressor switches with a time shift larger than the pulse width of the noise caused by far lines. Hence, the worst case peak noise voltage and its duration are

$$V_{peak} = V_{dd} \left[ \left( \frac{\sum_{All\ FarPairs} l_{pi}}{\sqrt{\sum_{All\ FarPairs} l_{pi}^2}} + 1 \right) \frac{Z_{0p}}{R_{ir}/2 + Z_{0p}} e^{-\frac{r\ell/2}{2Z_{0p}}} - \frac{Z_{dif}}{R_{ir} + Z_{dif}} e^{-\frac{r\ell}{2Z_{dif}}} \right], \quad (6.33)$$

and

$$t_{peak} = ToF_{com} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)}. \quad (6.34)$$

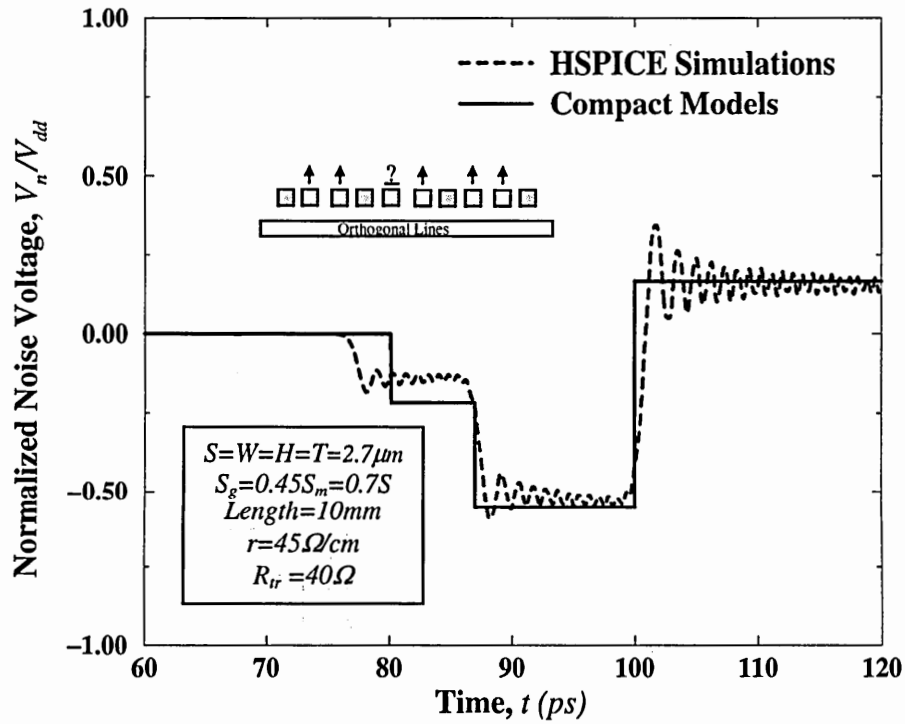


Figure 6.10: Noise voltage at the end of a quiet victim line when all intra-level near and far aggressors switch simultaneously.

## 6.6.2 Near and Inter-Level Far Aggressors

The same methodology can be used for the far lines which are two metal levels below the victim line. The noise voltage when far lines switch and the near line is quiet can be found by the models derived in Section 6.4. The victim and near aggressor have equal voltages and their common mode can be used to find the noise caused by intra-level far lines as

$$V_{f\text{-interlevel}} = V_{non\text{-identical}}(t, l_{vi} = l_{pi}, l_v = l_p, c_v = c_p, l_f = l_{eqv}, c_f = c_{eqv}). \quad (6.35)$$

To find the noise voltage when the near aggressor switches and far lines stay quiet, impact of far lines can be ignored because the inductances and capacitances of victim and far lines are different and they do not resonate. This makes the analysis simpler. The noise voltage caused by a near aggressor is

$$V_{near}(t) = \frac{Z_{com}}{R_{tr} + Z_{com}} V_{N-in} e^{\frac{r\ell}{2Z_{com}}} \left[ u(t - \ell\sqrt{c_{com}l_{com}}) \right] - \frac{Z_{dif}}{R_{tr} + Z_{dif}} V_{N-in} e^{\frac{r\ell}{2Z_{dif}}} \left[ u(t - \ell\sqrt{c_{dif}l_{dif}}) \right], \quad (6.36)$$

where

$$l_{com} = l_s + l_m, \quad (6.37)$$

and

$$c_{com} = c_g + c_{orth}. \quad (6.38)$$

Total noise when both near and intra-level far aggressors switch is equal to the summation of (6.35) and (6.36), which is verified in Figure 6.11.

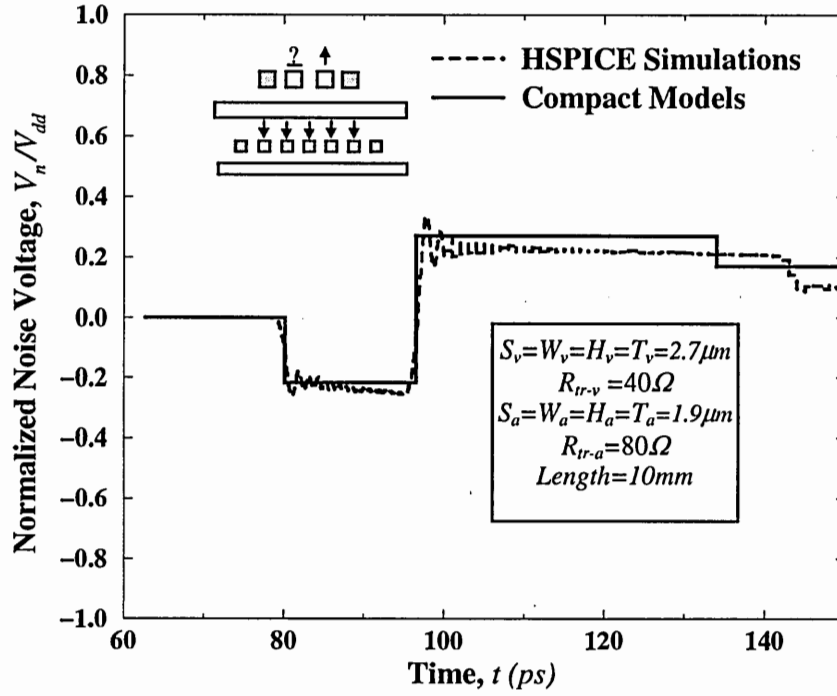


Figure 6.11: Noise voltage at the end of quiet victim line when near and intra-level far aggressors switch simultaneously. Line resistance for the two top levels is 45  $\Omega/\text{cm}$  and for the lower metal levels is 90  $\Omega/\text{cm}$ . The parameters with index of “v” correspond to the victim line and the parameters with index of “a” correspond to the aggressors.

## 6.7 Integrated Crosstalk Model

The results found in the previous section can be incorporated in an integrated expression for crosstalk caused by all near, intra- and inter-level far aggressors given by

$$V_{total} = V_{f-interlevel}(t - t_{f-interlevel}) + [V_{com}(t - t_{near}) + V_{dif}(t - t_{near})] + V_{f-int ralevel}(t - t_{f-int ralevel}), \quad (6.39)$$

where  $t_{f-interlevel}$ ,  $t_{f-intralevel}$  and  $t_{near}$  are times at which different sets of aggressors switch. If there is a large load capacitance at the end of the victim line, the RC charge-up equation should be used for each time period that the open-ended noise voltage is constant to find the peak noise voltage.

The integrated model is compared against HSPICE simulations in Figure 6.12 where noise at the end of a victim line is plotted versus time. This graph shows the worst case scenario for crosstalk when all aggressors are affecting the victim line constructively. The worst case is when far aggressors switch anti-phase compared to the near aggressor. Also, as Figure 6.8 and Figure 6.9 show, the intra-level far aggressors and the near aggressor constructively affect the victim line if the intra-level far aggressors switch 20 ps (far noise duration in this example) later than the near aggressor.

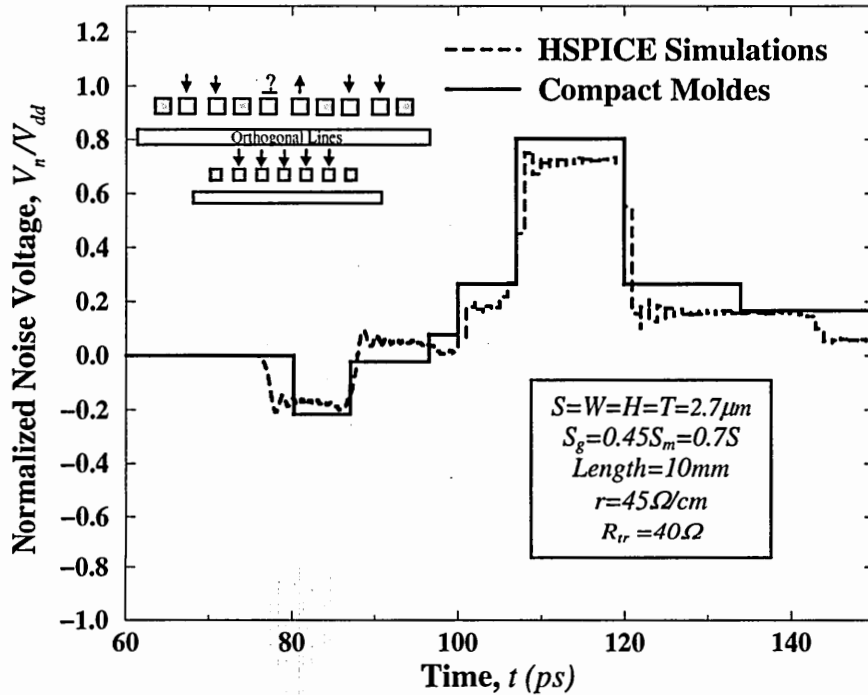


Figure 6.12: Total noise caused by near, far intra- and inter-level aggressors. Far Lines switch in the opposite direction and intra-level far aggressors switch 20 ps (the duration of intra-level far noise in this example) after near and inter-level far aggressors. This shows the worst case scenario for crosstalk.

Figure 6.12 shows that far aggressors can cause a prohibitively large crosstalk. The major way to decrease crosstalk is to avoid very small line resistance. For instance, for

the case shown in Figure 6.12, by increasing resistance per unit length from  $45 \Omega/cm$  ( $W=T=2.7\mu m$ ) to  $270 \Omega/cm$  ( $W=T=1.1\mu m$ ), the peak crosstalk voltage reduces from  $0.8V_{dd}$  to  $0.25V_{dd}$ . Increasing line resistance, however, increases interconnect latency. Hence, there is a trade-off between crosstalk and latency. Figure 6.13 has plotted crosstalk and latency versus line resistance for the same structure shown in Figure 6.12.

Inserting repeaters is an alternative way to reduce crosstalk and avoid large latency. This is shown in Figure 6.14 where the worst-case crosstalk and latency are plotted versus line resistance assuming that optimal repeaters are inserted [18]. Latency increases with line resistance more slowly compared to the latency of non-buffered interconnects shown in Figure 6.13.

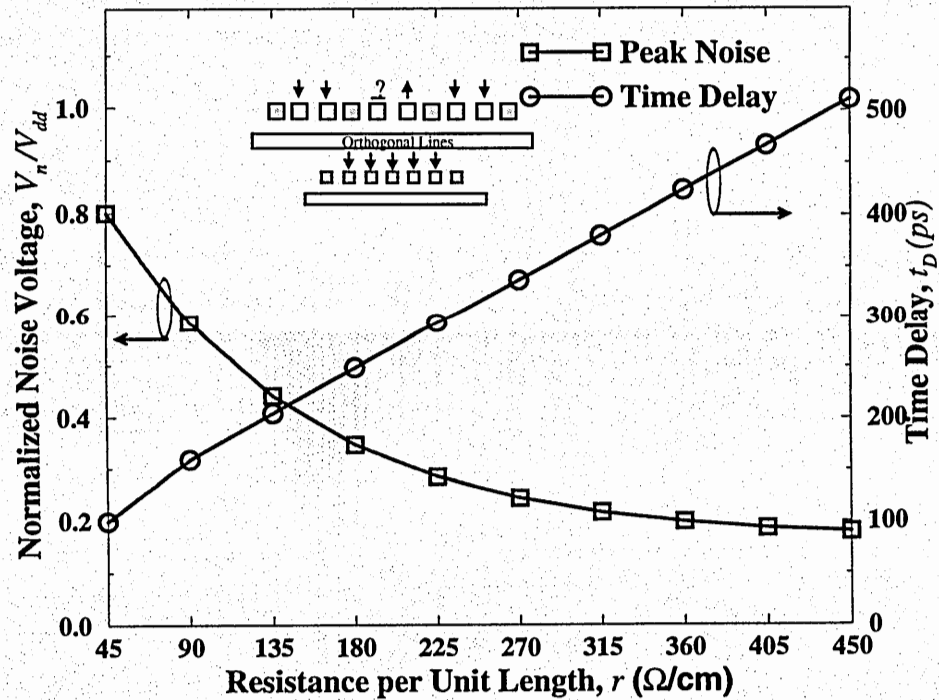


Figure 6.13: Worst case noise voltage and interconnect latency versus resistance per unit length of interconnects in the top metal level. It has been assumed that interconnects in the two lower metal levels have a resistance per unit length two times larger than that of the top two metal levels.

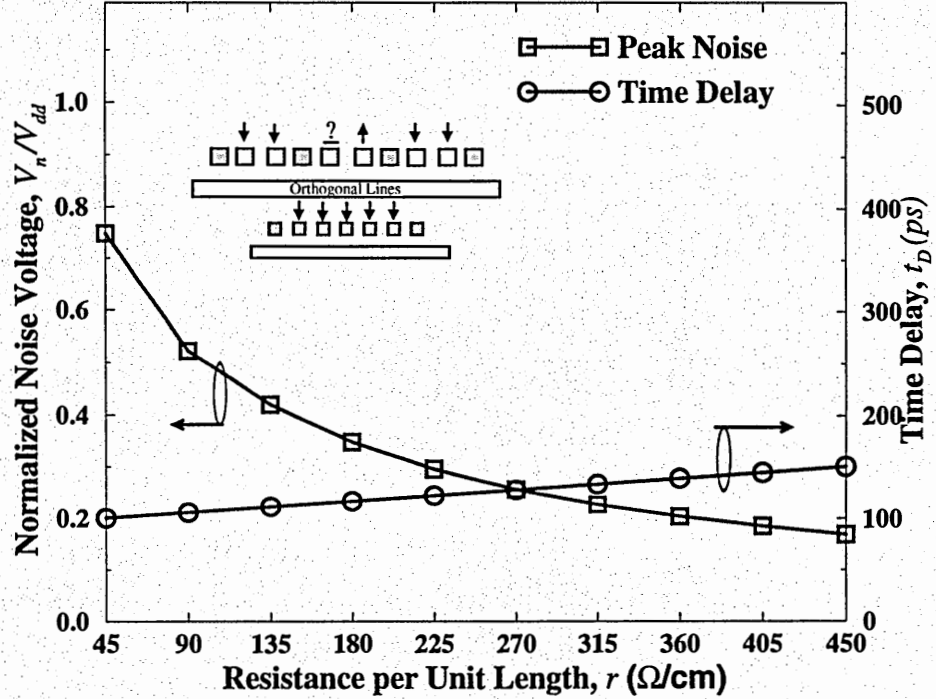


Figure 6.14: Worst case noise voltage and interconnect latency versus resistance per unit length of interconnects in the top metal level when optimal repeaters are used. All parameters are the same as those in Figure 6.13.

It is worthwhile to note that the peak crosstalk voltage for the no-repeater and optimal repeater cases are close as Figure 6.13 and Figure 6.14 show. The reason is that repeaters make interconnect segments shorter, and reduce the attenuation along the line. The open-ended crosstalk voltage, therefore, increases. The input capacitance of repeaters, however, filters the crosstalk noise. For the optimal repeater case, these two effects are equally important, and inserting repeaters does not change the peak crosstalk voltage considerably.



## 6.8 Impact of Wire Width Optimization

The compact crosstalk models show that crosstalk can be prohibitively large if resistance per unit length of interconnects is small. In Chapters 2 and 4, the optimal wire width has been identified that maximizes data flux density-reciprocal latency product. In Chapter 4, it is also shown that using optimal wire width makes the near crosstalk small and constant. In this section, the total crosstalk caused by near and far aggressors is calculated as a function of  $W/W_{opt}$  to illustrate the impact of wire width optimization on crosstalk.

Using a similar method used in Chapter 4, the peak noise voltage can be written as

$$V_{load} = V_{dd} \left[ 0.39 \frac{\sum_{All\ FarPairs} l_{pi}}{\sqrt{\sum_{All\ FarPairs} l_{pi}^2}} + 0.1 + 0.78 \frac{\sum_{All\ Aggressors} l_{vi} c_{f-eq}}{(l_{f-eq} c_{f-eq} - l_p c_p)} \right] \left[ 1 - \exp \left( -3.39 \frac{W^2}{W_{opt}^2} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)} \right) \right] \quad (6.40)$$

Equation (6.40) is proved in Appendix C. The first bracket in (6.40) represents the open-ended noise voltage and is independent of wire width assuming that all cross-sectional dimensions scale proportionally. The second bracket represents the impact of input capacitance of repeaters. As wire width decreases, the length of each segment between two repeaters decreases; hence the noise duration decreases and the peak noise voltage becomes smaller.

Equation (6.40) proves that crosstalk remains constant in various technology generations if the ratio of wire width over the optimal wire width remains constant. Peak crosstalk voltage is plotted versus  $W/W_{opt}$  ratio in Figure 6.15 for ground lines width of  $W$ ,  $2W$  and  $3W$ . It is evident that a  $W/W_{opt}$  of larger than 1 can result in a substantially larger peak crosstalk voltage. It also shows that using wider ground line widths can slightly lower crosstalk. Hence, by using the optimal wire width for signal lines and twice of it for power/ground lines, crosstalk will always be limited to  $0.25V_{dd}$ .

The analysis presented in this section is independent of interconnect length because by optimal repeater insertion, the segment length between two repeaters is the same for interconnects with different lengths.

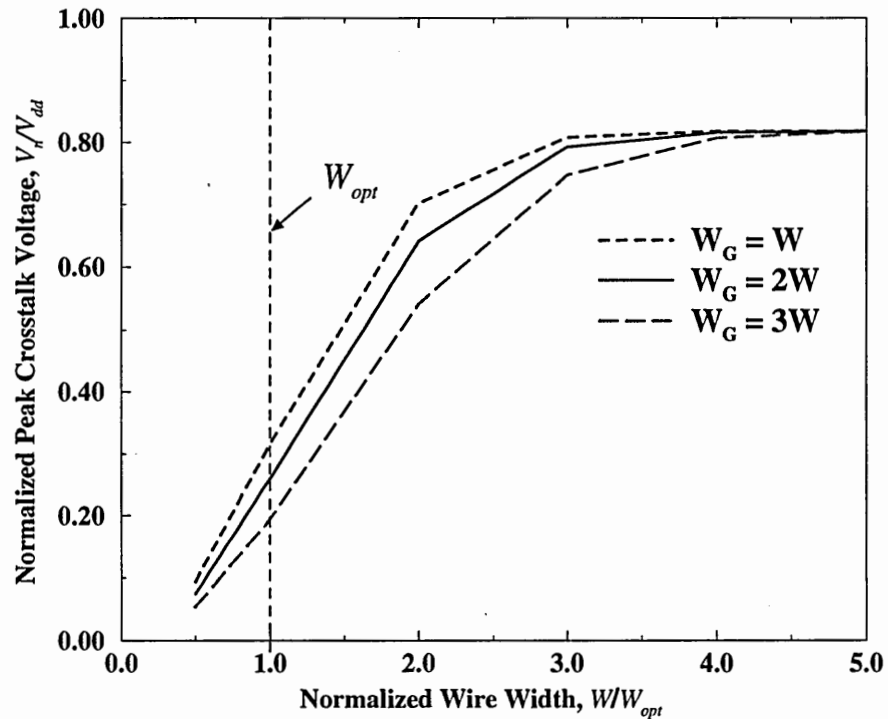


Figure 6.15: Normalized peak crosstalk voltage versus normalized wire width for three power/ground line widths. It has been assumed that optimal repeaters are inserted.

## 6.9 Conclusions

For the first time, compact physical models are derived for total noise caused by virtually all near and far aggressors. Far lines that are two metal levels below a victim line are also considered. Far inductive noise is modeled for two basic cases that aggressors and a victim line are identical or non-identical. For both cases it is shown that noise voltage-time integral is maximized when line resistance becomes equal to twice characteristic impedance of the lines, which can be quite useful for repeater insertion algorithms. Having the solutions for identical and non-identical lines and using superposition theorem, an integrated model is derived for total noise caused by all near and far aggressors. It is shown that the worst case scenario is when far aggressors switch anti-phase compared to a near aggressor. It is also shown that crosstalk can be prohibitively large if lines have a small resistance. Hence, there is a trade-off between latency and crosstalk. For instance, for 10 mm long interconnects it is shown that the worst case crosstalk can be reduced from  $0.80V_{dd}$  to  $0.25V_{dd}$  by increasing line resistance from  $45 \Omega/\text{cm}$  to  $270 \Omega/\text{cm}$ , which increases latency by 230% and 30% for no-repeater and optimal-repeater cases, respectively. Finally, it has been proved that by using optimal wire width for signal lines and twice of that for power/ground lines, crosstalk remains less than  $0.25V_{dd}$  in all generations of technology.

## Chapter 7

### Chip-Package Co-Design

#### 7.1 Introduction

Today's chips and packages are so related that they can no longer be designed independently. Besides the important role of a package in providing mechanical support and facilitating heat removal, a package can enhance the electrical performance of a chip. Global power, clock and even signal interconnection can be partially pursued by package or printed wiring board (PWB) interconnects. Thick and wide wires that are available in a package or a PWB can add a new level to the hierarchy of GSI multi-level interconnection whereas implementing such a level on-chip can be prohibitively expensive. High quality low-loss transmission lines at the board level can be used for on-chip global signal interconnection if high-density low parasitics input/output interconnects are available. In this manner, the maximum interconnect latency decreases that results in a higher global clock frequency. Power and ground planes at the board or package level provide low impedance paths for power and ground distribution across the chip, and a sufficiently small IR drop can be achieved by using an adequate number of I/O pins. Clock can be distributed across the chip by the board-level electrical or optical interconnects to lower the clock jitter and skew or power dissipation [37, 38, 39].

In this chapter, it is shown how package can be optimally used to improve signal and power distribution. For signal interconnection, a partition length is identified which shows interconnects beyond which lengths should be “exterconnects,” (external

interconnects), and for a case study, it is shown that the global clock frequency can be increased by 33% using 4 additional PWB levels and 1800 more I/O pads. For power distribution, the required number of I/O pins is calculated to have a reasonable IR drop.

## 7.2 Signal Interconnection through PWB

Because of nearby ground planes which exist off-chip, fat off-chip interconnects can be considered as low-loss transmission lines. Hence, if they are properly driven, their delay is time-of-flight (ToF), which is the smallest possible delay. A new packaging technology such as sea-of-leads (SoL) can provide high-density I/O pin arrays with negligible parasitics [40]. Such a packaging technology facilitates routing long global interconnects through the printed wiring board (PWB), and in this way, the performance of the system can be improved. The idea of taking advantage of high quality off-chip interconnects has also been suggested by [41] where it has been shown that the delay of long global off-chip interconnects (longer than 6 mm) is time-of-flight limited. However, no clear boundary has ever been defined between on-chip and off-chip interconnects to determine which nets should be routed off the chip. Likewise, the trade off between the overall cost and performance improvement has not been evaluated. Off-chip interconnects have considerably lower wiring density (1/100-1/36 smaller than the typical on-chip interconnects) such that off-chip routing of global nets may require many layers. Hence, a careful partition is required between on-chip and off-chip interconnects to improve the performance at a reasonable cost.

In this section, an optimal partition between on-chip and off-chip interconnects is defined to achieve the highest performance with the minimum numbers of off-chip layers

and I/O pads. In Section 7.2.1, the maximum on-chip wire length is found to achieve the highest possible global clock frequency. In Section 7.2.2, the cost of achieving the highest global clock frequency is estimated in terms of PWB layers and I/O pads for a projected microprocessor in year 2011.

### 7.2.1 Maximum On-Chip Interconnect Length to Achieve the Highest Performance

The global clock frequency of a system-on-a-chip is determined by the largest delay of synchronized interconnects. The global clock frequency is typically selected such that the largest global interconnect delay is 90% of the clock cycle. In this way, 10% variation in signal delays or clock frequency can be tolerated [8]. The lowest possible latency that can be achieved is the time-of-flight which is determined by the speed of light. Having ToF latency for long interconnects requires thick and wide interconnects. For example, for a 24 mm long interconnect, which is projected to be one edge of the chip at the 45 nm technology node, the metal and dielectric thicknesses of around 5  $\mu\text{m}$  are needed to achieve ToF. Such thicknesses and widths are prohibitively large for on-chip interconnects. However, printed wiring board (PWB) wires usually have large cross-sectional areas, and therefore, if terminated properly, have ToF delay. If some of the long interconnects are routed externally through PWB “*exterconnects*”, the maximum on-chip interconnect length reduces, and the maximum delay reduces. In this way, the global clock frequency can be improved. *Use of more exterconnects results in a higher global clock frequency that can be achieved up to the point where the ToF delay of the longest exterconnect becomes dominant.* If the maximum on-chip interconnect length decreases

such that the maximum on-chip delay becomes equal to the ToF delay of the longest exterconnect,  $l_{max}$ , the maximum possible global clock frequency is achieved. In Chapter 2 it is shown that the optimal wire width that maximizes the product of data flux density and reciprocal latency is the width at which interconnect latency becomes  $1.33\text{ToF}$ . Hence, the maximum on-chip interconnect length can be found by solving

$$1.33 \frac{l_{par}}{c_0 / \sqrt{\epsilon_{r-on-chip}}} \equiv \frac{l_{max}}{c_0 / \sqrt{\epsilon_{r-PWB}}}, \quad (7.1)$$

for

$$l_{par} = 0.75 l_{max} \frac{\sqrt{\epsilon_{r-PWB}}}{\sqrt{\epsilon_{r-on-chip}}}, \quad (7.2)$$

where  $\epsilon_{r-on-chip}$  and  $\epsilon_{r-PWB}$  are the relative permittivity of on-chip and on-board dielectrics, respectively. All interconnects for  $\ell_{par} < \ell < \ell_{max}$  should be on-PWB to have the highest global clock frequency. This method reveals the optimal partition between on-chip and off-chip interconnects for the global net distribution of a GSI chip and maximizes the global clock frequency without compromising the on-chip wiring density. Assuming that the dielectric constant at the board and chip levels are equal, by reducing the maximum on-chip length to  $\ell_{par}$ , the global clock frequency can be improved by 33%.

As shown in Chapter 3, because of on-board ground planes, exterconnect crosstalk is minimum, local and well known. Exterconnects have no overshoot problem, and, unlike on-chip interconnects, have very small delay variation. Typically, the thickness to width ratio ( $T/W$ ) of the off-chip wires is 0.2-0.6 as compared with on-chip interconnects that have  $T/W \geq 1$ . Therefore, the capacitive coupling,  $K_c$  of the exterconnects is smaller than

that of the on-chip interconnects [11]. Consequently, exterconnect crosstalk is significantly lower.

### 7.2.2 Cost Estimation

By transferring all interconnects longer than the partition length to the board, the maximum possible global clock frequency can be achieved. The question that is yet to be answered is how many off-chip layers and I/O pads are needed. In other words, is it feasible to partition interconnects and exterconnects optimally?

#### 7.2.2.1. Extracting the Wiring Distribution

Increasing the number of PWB layers and I/O pads is the cost of utilizing exterconnects. Estimating the number of off-chip layers and I/Os requires stochastic models for the global net length distribution. Using these models, the total length of all nets which should be routed by exterconnects can be found.

The ITRS projections for high performance microprocessors in the year 2011 are used [12]. Key parameters are summarized in Table 7.1. Using a stochastic net-length distribution model for heterogeneous systems [42], the global net density function,  $i(\ell)$  of the projected chip is found as shown in Figure 7.1. The total number of nets with length between  $\ell = a$  and  $\ell = b$  is

$$I(a < \ell < b) = \int_a^b i(\ell) d\ell. \quad (7.3)$$

This stochastic model also specifies the net length distribution for each individual fan out separately.



Table 7.1: Key parameters for the projected microprocessor in the year 2011.

Chip Area	818mm <sup>2</sup>
Number of Gates	1080M
Memory Percentage	90%
Number of Mega cells	20

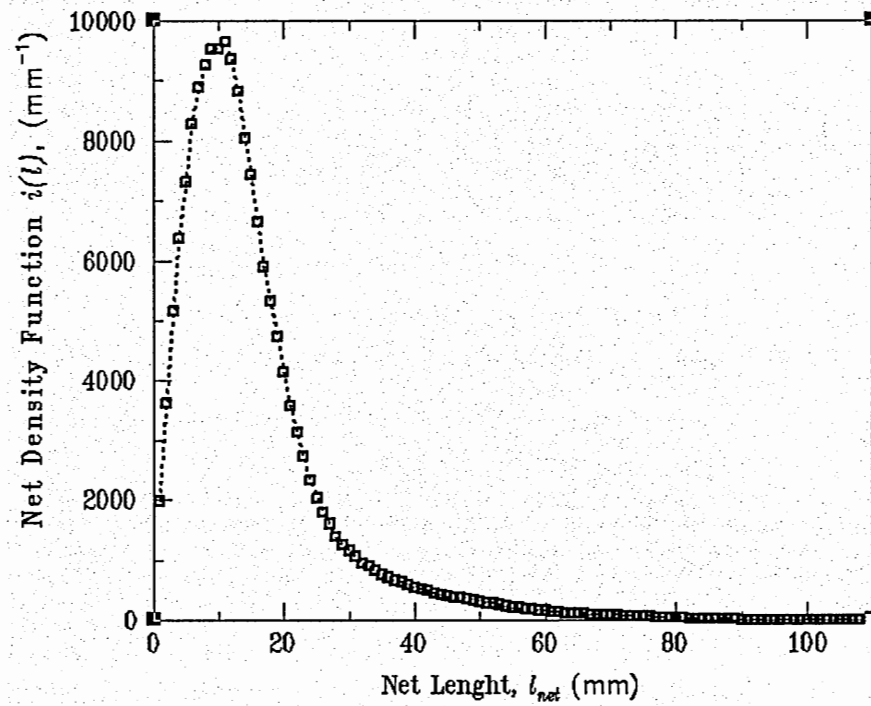


Figure 7.1: Net length distribution for the projected microprocessor in the year 2011.

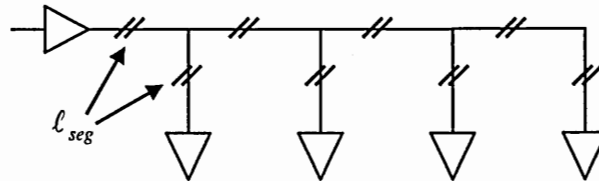


Figure 7.2: A typical net with fan out 4.

Each net consists of one driver and several receivers as shown in Figure 7.2. Within each net, the farthest receiver has the largest delay. If no repeater is used, the capacitance of the whole net contributes to the delay of the farthest receiver. However, since many repeaters are often inserted along the net, the path between the driver and the farthest receiver is isolated from the capacitance of the branches. Therefore, the length of the wire connecting the driver to the farthest receiver determines the largest delay of each net. Calling this length “*effective*” net length, it is necessary to find the total net length versus the *effective* net length. Assuming that the nets are in the form of a tree shown in Figure 7.2, the length of each segment in the net can be found by

$$\ell_{seg} = \frac{\ell_{net}}{2(f.o.)}, \quad (7.4)$$

where *f.o.* is the fan-out of the net and  $\ell_{net}$  is the net length. It can be assumed that each terminal has an equal probability to be served as the driver. For instance, in the case shown in Figure 7.3, the fan out is four. For the driver terminals corresponding to a, b, or e terminals, the effective length of the net will be  $5\ell_{seg}$ . If c or d terminals are the driver terminal, the effective length will be  $4\ell_{seg}$ . In this way, 60% of all nets with a fan-out equal to four have the effective length of  $5\ell_{seg}$  and 40% of them have the effective length of  $4\ell_{seg}$ . The same method can be applied to different fan-outs, and in this way, the effective net length density function,  $L(\ell_{eff})$  can be obtained as shown in Figure 7.4. The total length of all nets having effective length between  $\ell_{eff} = a$  and  $\ell_{eff} = b$  can be found by

$$L_t(a < \ell_{eff} < b) = \int_a^b L(\ell_{eff}) d\ell_{eff}. \quad (7.5)$$

The maximum effective net length is equal to twice the chip edge dimension.

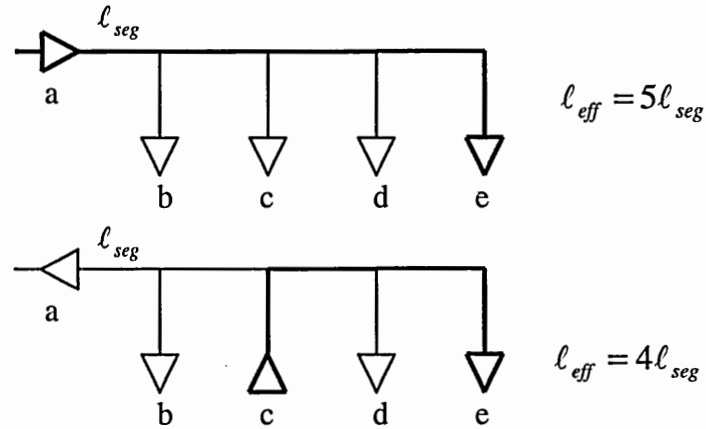


Figure 7.3: The effective length depends on the location of the driver

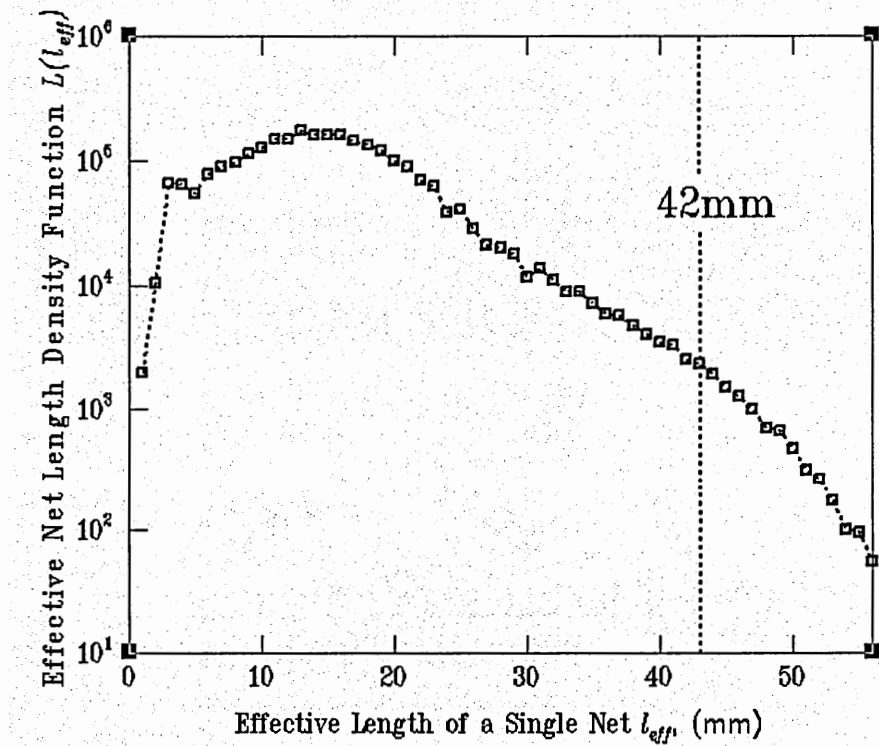


Figure 7.4: Effective global net-length distribution for the projected SoC in the year 2011.

### 7.2.2.2. Estimating the Required Area and I/O Pads

Having the effective global net-length distribution, the off-chip area required to route all on-chip interconnects longer than any maximum length can be found by

$$A_{off} = \frac{L_t(\ell_{eff} > \ell_{max}) \times P_{off}}{e_w}, \quad (7.6)$$

where  $P_{off}$  is the off-chip wire pitch, and  $e_w$  is the wiring efficiency (assumed to be 0.5).

The ITRS has projected wire pitch of  $72 \mu m$  for the printed wiring boards in the year 2011 [12].

In order to route exterconnects, additional I/O pads are required. Figure 7.5 shows the required number of pads versus maximum on-chip interconnect length in two cases. In the first case, all parts of the long nets are placed on board, and therefore, each net needs  $(f.o.+1)$  pads. In the second case, only long parts of the nets are routed through the PWB and the other parts remain on-chip. The second case requires fewer pads and also less off-chip area. Figure 7.6 shows these two cases more clearly.

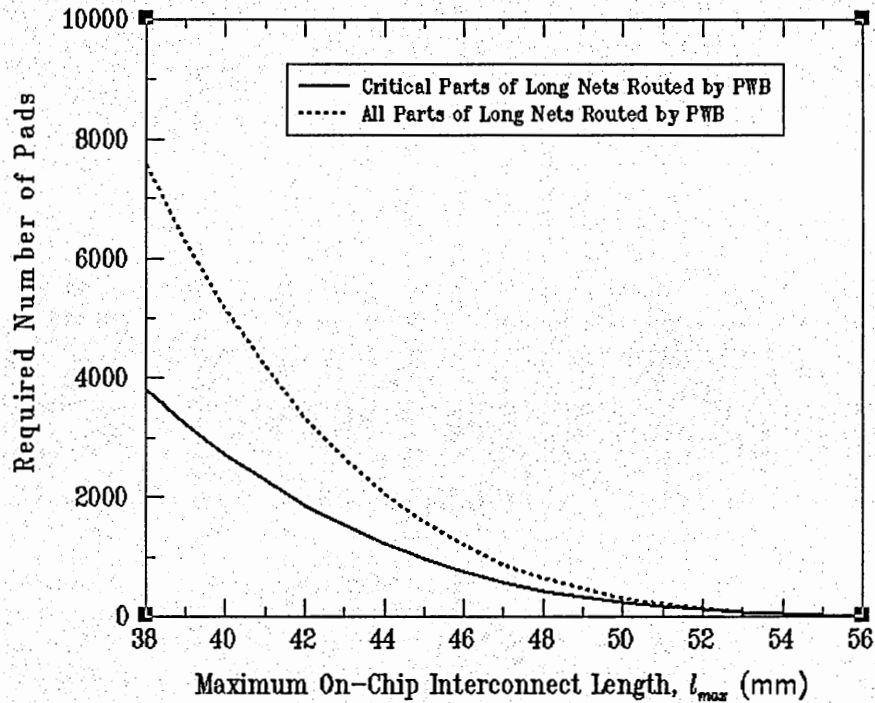


Figure 7.5: The required number of pads versus the maximum on-chip interconnect length.

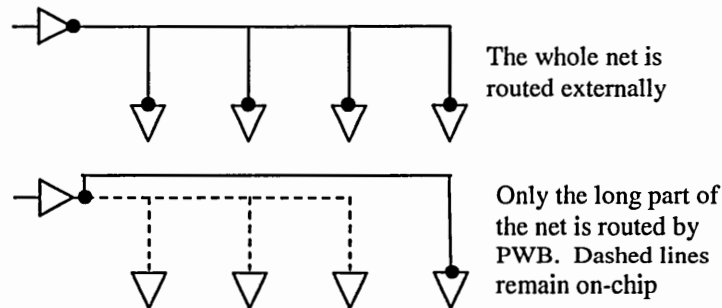


Figure 7.6: By transferring just the longest parts of the nets to PWB the required number of pads can be reduced.

### 7.2.2.3. Estimating Total PWB Area

After finding the off-chip area required for exterconnects, it is necessary to find the total number of layers that should be added to the PWB due to utilizing exterconnects. As

Figure 7.7 shows, three groups of PWB layers can be defined. The first group consists of power and ground layers. These layers distribute power and ground across the chip. The second group consists of PWB layers which route standard I/O pads to the other chips. These layers can be called the “*escape*” layers. Exterconnects are routed through the third group of PWB layers. The number of these layers can be found by using the required area in Section 7.2.2.1. The ground and power planes are also inserted between each two PWB layers and the number of them depends on the number of the other layers. The required number of layers to route the standard I/Os, however, depends on the number of exterconnect pads due to the via blockage that they cause. Hence, the number of escape layers should also be estimated considering the impact of exterconnects to find the total cost of utilizing exterconnects.

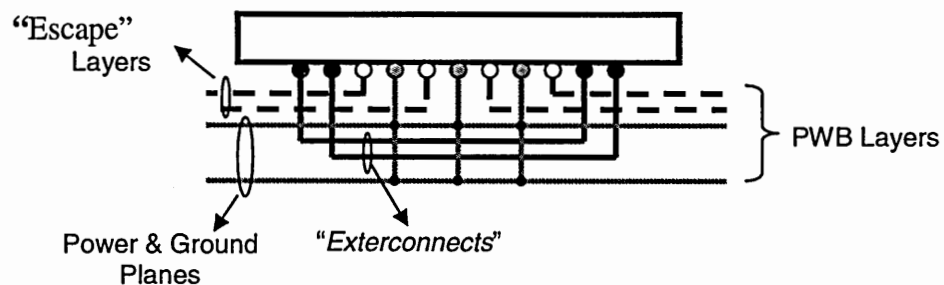


Figure 7.7: PWB layers are used for routing the standard I/Os to other chips, power and ground distribution, and exterconnects.

An algorithm to find the total number of PWB layers to route all I/O pads within the footprint of the package on the PWB has been developed [43]. A closed form expression, however, gives a better insight to the designers.

Ground and power I/Os are connected to the ground and power planes, while the standard signal I/Os should be routed to the outside area of the chip. To have a smaller via blockage for exterconnects, upper PWB layers are assigned for routing signal I/Os, and lower PWB layers are used to route exterconnects. Assuming that all pads are homogeneously distributed beneath the chip area, a closed-form model is developed to determine the number of escape layers.

Pad pitch, which is equal to the via pitch, is found as

$$P_p = \frac{\sqrt{A_{chip}}}{\sqrt{N_p}}, \quad (7.7)$$

where  $N_p$  is the total number of pads, and  $A_{chip}$  is the chip area. The number of wires that can be passed through each two adjacent vias, defined as lanes per channel is equal to (Figure 7.8)

$$N_l = \text{int} \left[ \frac{P_p - 2r_v - s}{w + s} \right], \quad (7.8)$$

where  $r_v$  is the radius of the via cross section,  $w$  is the wire thickness, and  $s$  is the minimum spacing between two wires or a wire and a via.

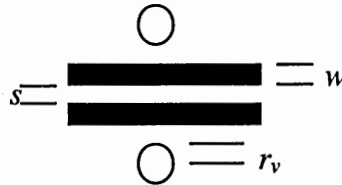


Figure 7.8: A channel with two lanes.

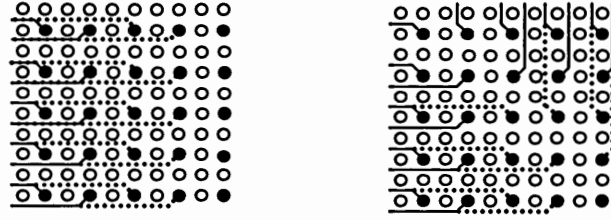


Figure 7.9: A quadrant of the standard I/Os routed to the other chips through PWB. They can be routed either through two edges (left) or four edges (right). The dashed lines show the wires in the next layer.

Figure 7.9 shows a quadrant of the pad array. Since the three other quadrants are routed similarly, they are not shown in Figure 7.9. The open circles are used for either power and ground pads or exterconnects, and therefore, are not routed to the outside. Standard signal I/Os are routed through either two (left picture in Figure 7.9) or four (right picture in Figure 7.9) edges. The required numbers of layers in both cases are the same since the center row (the lowest row in Figure 7.9) determines the required number of layers. The number of PWB layers required to route the standard signal I/O pads can be found by

$$N_{esc} = \frac{\frac{\sqrt{N_{stdIO}}}{2}}{\frac{\sqrt{N_p}}{\sqrt{N_{stdIO}}} N_l} = \frac{N_{stdIO}}{2N_l \sqrt{N_p}}, \quad (7.9)$$



where  $N_{stdIo}$  is the number of standard signal I/O pads. The numerator of the first fraction is the number of signal pad columns from center of the chip to an edge (5 columns in Figure 7.9), and the denominator represents the number of columns that can be routed in each PWB layer (2 column in Figure 7.9, where  $N_l=1$ ).

The ITRS projects 1400 standard signal I/O pads for the year 2011 [12]. It is shown in [18] that by assigning more I/O pads for power and ground distribution, the IR drop and also the simultaneous switching noise will reduce considerably. Therefore, 4000 I/O pads are considered to distribute power supply and ground across the chip. The ITRS has also projected 36  $\mu m$  for  $w$  and  $s$  in that year [12] (although some manufacturers have already reached this resolution [44]). Via cross-section radius is often twice the wire width [43]. Figure 7.10 shows the required number of PWB layers to route standard signal I/O pads versus the number of exterconnect pads. A small increase in the number of exterconnect pads does not change the number of lanes per channel, and due to a larger number of channels, the number of required escape layers decreases. Adding more than 2300 pads, however, decreases the number of lanes per channel from three to two, and therefore, as Figure 7.10 shows, there is a jump in the number of layers and one more layer is required.

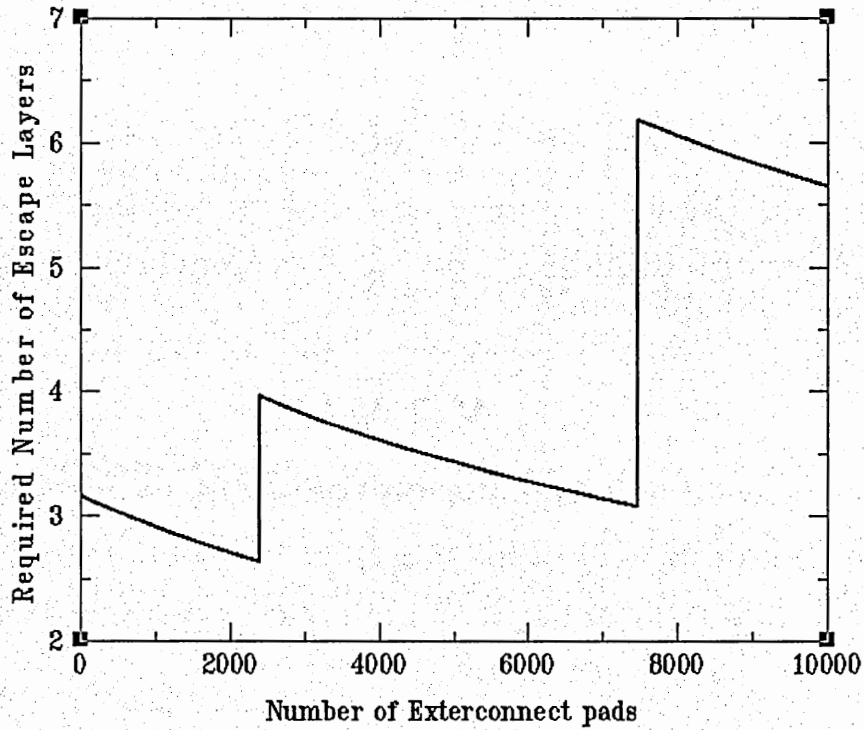


Figure 7.10: The required number of PWB layers to route standard I/Os versus the number of exterconnect pads.

Increasing the number of pads also has an impact on the effective area of PWB layers used for exterconnects. Each pad needs one via to be connected to one of the PWB layers. This via blocks two wires on each layer that it passes through as shown in Figure 7.11. Again, the cross-section of each via passing through a layer is assumed to be twice the wire width. The blocked wires cannot be used when the distance between the vias is small [20]. Therefore, each row of passing vias wastes the space of two wires. Hence, the effective area of each layer is

$$A_{eff} = \left(1 - \frac{2\sqrt{N_v}}{\sqrt{A_{chip}}/2w}\right) A_{chip} = \left(1 - \frac{4w\sqrt{N_v}}{\sqrt{A_{chip}}}\right) A_{chip} \quad (7.10)$$

where  $N_V$  is the number of passing vias.  $\sqrt{N_V}$  represents the number of via rows and  $\sqrt{A_{chip}} / 2w$  is the total number of wires in x or y directions on each layer. For example, if 2000 vias pass through a PWB layer with a wire width of  $36 \mu m$  and chip area of  $817 mm^2$ , approximately 22% of that layer's area is wasted. Having the required off-chip area and knowing the effective area of each PWB layer, the number of PWB layers to route exterconnects can be found.

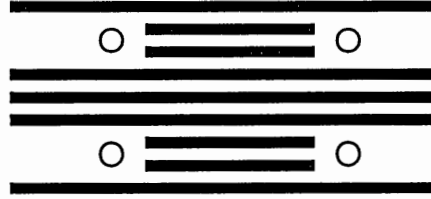


Figure 7.11: The via blockage.

The total number of PWB layers versus the maximum on-chip interconnect length is shown in Figure 7.12. The power and ground layers are assumed to be inserted after every signal layer such that each layer has nearby ground planes. In the case of not using exterconnects, about 8 PWB layers are required ( $\ell_{max} = 56 mm$ ). Reducing the maximum on-chip interconnect length to  $42 mm$ , requires 12 PWB layers. This means that by adding 4 layers to the PWB, the maximum possible global clock frequency can be achieved, which is about 33% increase in the global clock frequency. It is worth while to note that even today, boards with more than 12 layers are manufactured [44].

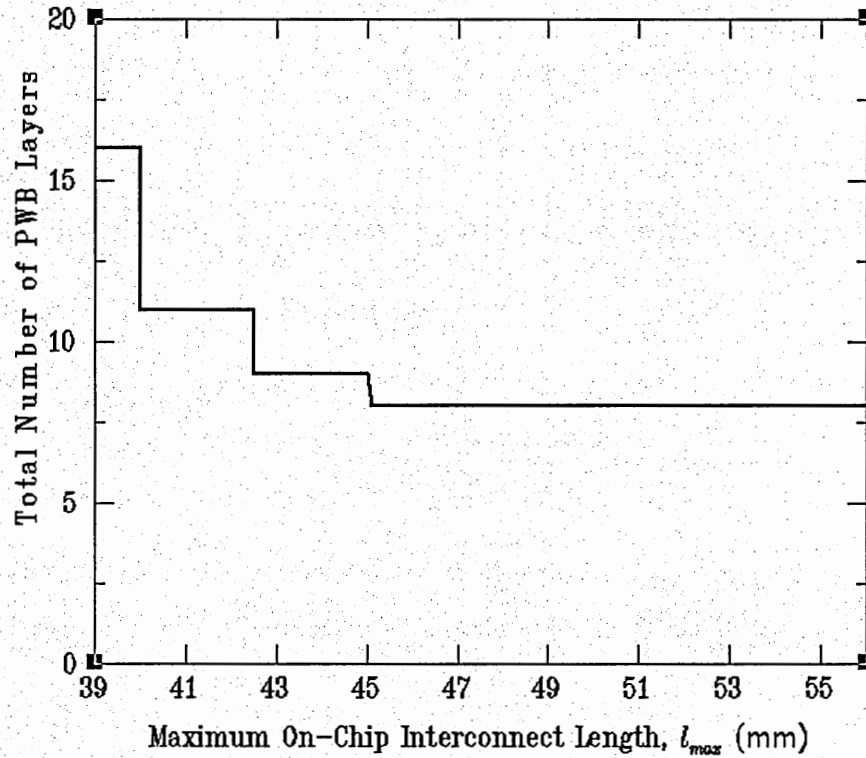


Figure 7.12: Total number of PWB layers versus the maximum on-chip interconnect length. Four additional layers are required to achieve the highest global clock frequency.

Table 7.2 summarizes the results and shows the advantages and costs of the optimal partition between exterconnects and interconnects. Silicon area required for repeaters and also exterconnect drivers is estimated using the models given by [45] for the size of inverters. It is worth while to note that the area required for exterconnect drivers is less than the saved repeater area. The required area for electrostatic discharge (ESD) protection circuits is estimated by using the model given in [46]. It has been projected that the size of ESD circuits scales down with  $s^{1/2}$ , where  $s$  is the channel length scaling factor.

Table 7.2: Costs and advantages of optimal partition between interconnects and exterconnects.

Parameters	No Exterconnect	Optimal Partition between Interconnects & Exterconnects
On-Chip Wire Width	0.85um	0.85um
Maximum On-Chip Interconnect Length	56mm	42mm
Global Clock Frequency	3GHz	<i>4GHz</i>
PWB Wiring Layers	8	12
# I/O Pads	5400	7200
Area for Global Interconnect Repeaters	11.38 mm <sup>2</sup>	11.11 mm <sup>2</sup>
Area of <i>Exterconnects</i> Drivers	0	0.035mm <sup>2</sup> (Output Drivers)
Area of ESD Protection Circuits	1.092 mm <sup>2</sup>	2.952 mm <sup>2</sup>

### 7.3 Power Distribution

The IR drop voltage across a chip is determined by the number of pins and the on-chip metal area that are dedicated to power and ground distribution. Hence, for a given IR drop budget, any reduction in the number of pins requires a larger on-chip metal area for power and ground distribution, and vice versa. In this section, it is shown that how on-chip wire width optimization dictates the number of power and ground pads.

Assuming that the power distribution network is mainly in the top two metal levels, the maximum IR drop voltage in a chip with area array pins is given by [47]

$$V_{IR} = 4\rho \frac{(k_G + 1)}{A_R \cdot (2W)} \frac{I_T}{N_G} \ln \left( 0.65 \frac{D_{chip}}{\sqrt{N_G M}} \right), \quad (7.11)$$

where  $\rho$  is metal resistivity,  $k_G$  is the number of signal lines between power and ground lines in the top metal levels,  $A_R$  is the aspect ratio,  $W$  is the wire width in the top levels,  $I_T$  is the chip total current,  $D_{chip}$  is the chip edge dimension,  $M$  is the size of pads, and  $N_G$  is the number of power pads.

For a given IR drop budget, (7.11) can be used to identify the required number of power and ground pins. In Chapter 3 it was shown that signal lines in the top metal levels need nearby power and ground lines to avoid large noise voltages. Hence, the number of signal lines between power and ground lines,  $k_G$ , is 2. Based on the ITRS projections, the aspect ratio can be in the range of 1-2.2. The optimal wire width that simultaneously maximizes data flux density and minimizes latency is also given in Chapter 4. By substituting these parameters and the ITRS projections for the chip size, chip current, chip supply voltage and pad size in (7.11), the required number of power and ground pins at various technology nodes can be identified. Table 7.3 has summarized the key

parameters and also the predicted number of power and ground pads in various technology nodes for the aspect ratios of 1 and 2. The target IR drop is assumed to be  $0.1V_{dd}$  ( $0.05V_{dd}$  for  $V_{dd}$  and  $0.05V_{dd}$  for  $V_{ss}$ ). Table 7.3 shows that because of rapid increases in the chip's supply currents, the number of power and ground pins should increase substantially as technology advances. It should be noted that these estimations are based on IR drop, and the simultaneous switching noise is not considered. If I/O pins have a large inductance, a larger number of pins might be required to have a small simultaneous switching or  $Ldi/dt$  noise.

Table 7.3: The key parameters and the required number of power/ground pins for various technology generations. Aspect ratios 1 and 2 are considered for the top global interconnects ( $A_r=1, 2$ ).  $I_T$  is the total current associated with power and ground I/O pins.

Year (Technology node)	2001 (130 nm)	2004 (90 nm)	2007 (65 nm)	2010 (45 nm)	2013 (32 nm)	2016 (22 nm)
$V_{dd}$ (V)	1.1	1	0.7	0.6	0.5	0.4
Power (W)	130	160	190	218	251	288
$I_T$ (A)= $2I_G$	236	320	542	726	1004	1440
$D_{chip}$ (mm)	17	17	17	17	17	17
$W_{opt}$ ( $\mu m$ )	1.4	1.1	0.9	0.67	0.54	0.42
$N_T = 2N_G$ ( $A_r=1$ )	312	542	1414	2854	5320	10720
$N_T = 2N_G$ ( $A_r=2$ )	174	306	808	1636	3084	6284

## 7.4 Conclusions

In this chapter, it is shown how package can be optimally used to enhance power and signal interconnection. The maximum on-chip interconnect length is found such that the maximum possible global clock frequency limited by the ToF delay can be achieved. Using a stochastic model for the net length distribution of a representative heterogeneous system-on-a-chip, the effective net length distribution of a projected system-on-a-chip in the year 2011 is found. Deriving new models for the required number of PWB layers to route standard signal I/O pads, the total cost of exploiting exterconnects is evaluated. By adding 4 layers to the PWB and 1800 I/O pads to the package, the global clock frequency can be increased from 3 *GHz* up to 4 *GHz* for a projected microprocessor in the year 2011.

Knowing the optimal wire width and density of power and ground lines at the chip-level, the optimal number of power and ground pins are projected for various technology generations to have a reasonable IR drop. While for current technology generations a few hundred pins are sufficient, more than 5000 pins are needed for 32 and 22 *nm* technology nodes.



## Chapter 8

### Optical Versus Electrical Interconnection

#### 8.1 Introduction

While optical interconnection has found mainstream application in fiber-optics telecommunications, it is still unclear at which technology generation optics will be used for on-chip or chip-to-chip communication in CMOS microelectronic systems. Optics can potentially improve interconnection in terms of energy dissipation, latency, and/or bandwidth. Energy dissipation of optical interconnects can be smaller than that of electrical interconnects with a voltage swing of 1V [48]. However, reducing the voltage swing of electrical interconnects can make the energy dissipation of electrical and optical interconnects comparable while their signal-to-noise ratios are roughly equal [49]. Also, although latency of optical interconnects may be smaller for long distances [50], the improvement that optics can offer in terms of latency is not significant because latency is limited by time-of-flight (ToF) and “fat” electrical wires that are available at the board-level also can offer latencies close to ToF [49]. The bandwidth of optical interconnects, however, can be substantially larger than that of electrical interconnects, and this fact can be a major motivation for using optics for chip-to-chip or even on-chip interconnection [51]. Svensson has concluded in [49] that the models that are used for bandwidth of electrical interconnects in [51] are too conservative, and for a case-study of 10 *cm* long interconnects he has shown that optical and electrical interconnects have comparable aggregate bandwidths [49]. However, the analyses presented in both [49] and [51] have

assumed that wire bandwidth is proportional to cross-sectional area, which is only true for wires with circular or square cross-sections [52]. Due to skin effect, current flows at the surface of a wire, and at high frequencies its resistance becomes inversely proportional to its circumference. Bandwidth of a wire is, therefore, proportional to the square of circumference not the cross-sectional area [52]. The assumption that wire bandwidth is proportional to its cross-sectional area using the relationship between circumference and area of a circle or square that is used in [49] and [51] can be a source of error for bandwidth of wires with rectangular cross-section.

In this chapter, a new partition length is identified that illustrates the length beyond which optical waveguides can offer a larger aggregate inter-chip bandwidth in comparison to electrical interconnects *when constrained by a fixed routing area*. In this analysis, it is assumed that optical drivers and receivers eventually will mature and become comparable with their electrical counterparts in terms of power, size, and cost. The comparison, therefore, emphasizes the interconnect media, or “wires versus waveguides,” rather than the interface circuits. Initial opportunities to use optical interconnects appear to be more promising for chip-to-chip interconnection rather than on-chip interconnection; therefore, board-level interconnects are considered. The partition lengths that are finally found confirm this assumption.

To consider bandwidth and wiring density simultaneously, data flux density is defined as the product of interconnect bandwidth and reciprocal pitch that represents bandwidth per unit width of an interconnect. It is desired to have a large data flux density to transfer as many bits per second as possible using a constant wiring area. In Section 8.2, data flux density is found for electrical interconnects, and the wire width that

maximizes data flux density is determined. In Section 8.3, the maximum data flux density for optical interconnects is identified, and the lengths beyond which optical interconnects offer larger data flux densities are specified in Section 8.4. The key results are summarized in Section 8.5.

## 8.2 Electrical Interconnects

Off-chip or on-board wires are usually fat and work in the RLC regime, where skin effect limits the bandwidth. The bit-rate limit of a distributed RLC interconnect is rigorously found [52] to be

$$B_{elect} = B_0 P^2 / kl^2, \quad (8.1)$$

where  $B_0$  is a constant determined by the conductor material,  $P$  is the cross-sectional perimeter,  $l$  is the wire length, and  $k$  is a factor between 60 to 120 depending on the “eye opening” that is desired.  $B_0$  is  $1.846 \times 10^{18}$  (bit/s) for copper, and  $k$  is chosen equal to 120, a conservative value. Equation (8.1) is re-derived and confirmed in Appendix C.

At the board level, wire thickness,  $T$ , is smaller than wire width  $W$  ( $T/W \sim 0.2-0.6$ ), and because of *proximity effect*, current flows mostly through the lower and upper regions of wires that are close to ground planes [53]. Hence, the effective perimeter of a wire can be taken as  $2W$ , and (8.1) can be written as

$$B_{elect} = K_0 W^2 / l^2, \quad (8.2)$$

where  $K_0 \equiv 4B_0/k = 6.152 \times 10^{16}$ . It should be noted that (8.2) is valid if the dielectric thickness scales according to wire width and the line’s characteristic impedance remains constant. Assuming that the spacings between interconnects are equal to their widths, data flux density, or bandwidth per unit width of an interconnect, is

$$\phi_{D-elect} \equiv B_{elect} / p = K_0 W / 2l^2, \quad (8.3)$$

where  $p$  is interconnect pitch. Equation (8.3) shows that data flux density increases linearly with increasing wire width. For instance, by doubling the width of a wire, the bandwidth of that wire becomes 4 times larger, but wiring density decreases by a factor of 2. In this manner, data flux density doubles. Assuming that interconnect bandwidth is proportional to cross-sectional area can be quite misleading because it results in a data flux density independent of wire width. Equation (8.3) suggests that on-board electrical interconnects should be made as wide as possible to maximize the aggregate bandwidth. The maximum frequency that a driver can switch, however, introduces a limit on maximum wire width. From (8.2), the optimal wire width, therefore, is the width at which interconnect bandwidth becomes equal to the maximum frequency that the driver can switch,  $f_{elect-max}$ :

$$W_{opt} = l \sqrt{f_{elect-max} / K_0}. \quad (8.4)$$

By substituting (8.4) into (8.3), the maximum data flux density that electrical interconnects can offer is

$$\phi_{D-elect-max} = \sqrt{K_0 f_{elect-max}} / 2l. \quad (8.5)$$

Equation (8.5) shows that the maximum data flux density is inversely proportional to  $l$ , not  $l^2$ . The reason is that for longer interconnects, wider wires can be used to enable a bandwidth equal to the maximum driver frequency. It also shows that as technology advances, transistors become faster and  $f_{elect-max}$  increases, increasing in turn the maximum data flux density that can be achieved by electrical interconnects. To estimate the maximum data flux density, we assume that  $f_{elect-max}$  will be equal to the International Technology Roadmap for Semiconductors (ITRS) [12] projections for the chip-to-board

clock frequency. The optimal wire width and the maximum data flux density are plotted in Figure 8.1 for various technology generations and interconnect lengths.

Figure 8.2 shows the maximum data flux density and the optimal wire width versus interconnect length at the 45 nm technology node. Figure 8.2 shows that as interconnect length decreases, the optimal wire width decreases, and in this manner, data flux density increases. The wire width, however, is limited by the minimum line width available on-board,  $W_{min}$ , implying that a minimum length exists below which data flux density remains constant. The value of minimum wire width is taken as 36  $\mu m$  as projected by the ITRS for the 45 nm technology node [12].

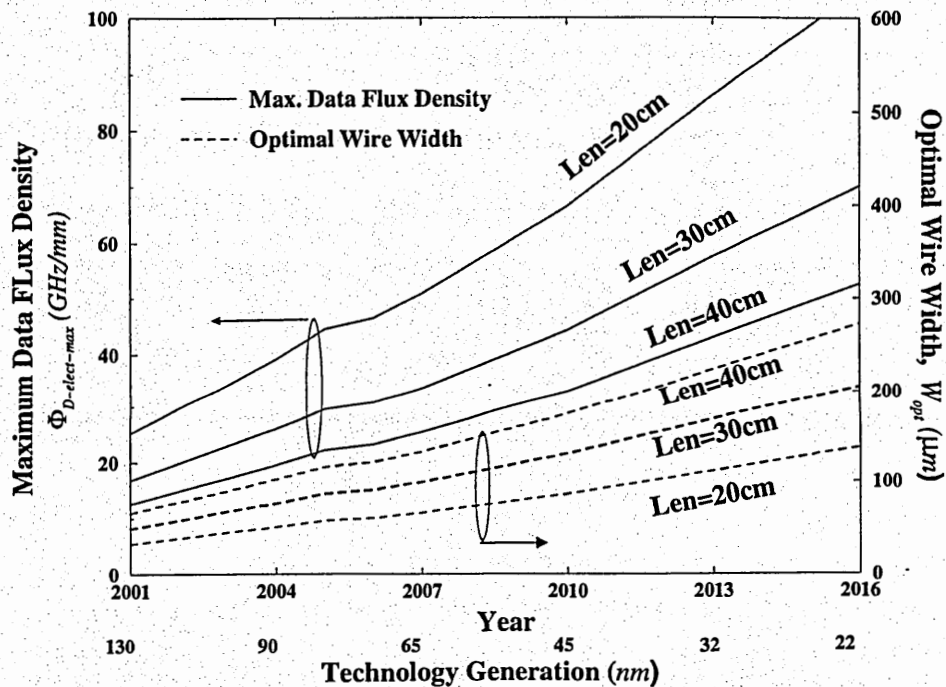


Figure 8.1: Maximum data flux density and optimal wire width for 20, 30 and 40 cm long interconnects implemented in different technology generations. Both maximum data flux density and optimal wire width increase in future generations because of faster transistors.

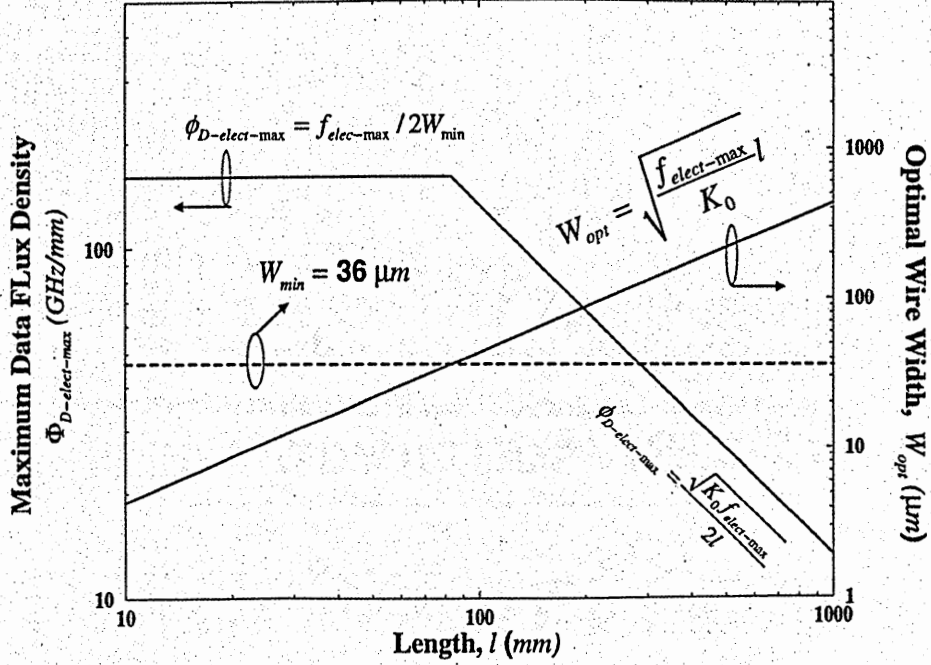


Figure 8.2: Maximum data flux density and optimal wire width versus interconnect length for the 45 nm technology node. The dashed line shows the minimum line width available on-board as projected by the ITRS [12].

Power consumption is a big concern for GSI chips and it is important to send bits of information in a manner that minimizes energy dissipation. The energy that should be dissipated to send one bit of data can be written approximately as

$$E_b = \frac{C_p V_{dd}^2}{2} + \frac{V_{dd}^2}{R_{tr} + Z_0} T_b, \quad (8.6)$$

where  $C_p$  is the parasitic capacitance associated with an I/O pad and its via,  $R_{tr}$  is the driver resistance,  $Z_0$  is the line's characteristic impedance, and  $T_b$  is the bit duration time. The first term in (8.6) is the energy required to charge the parasitic pad capacitance, and the second term is the energy dissipated to transfer a pulse through the transmission line. Typically, the first term is smaller than the second term, especially for the new packaging technologies with small I/O parasitics. Equation (8.6) shows that to have a small energy

per bit,  $T_b$  should be small, and to have valid data at the end of the line,  $T_b$  cannot be smaller than  $1/B_{elect}$ . Earlier, it was shown that increasing wire width increases the interconnect bandwidth until wire width becomes equal to  $W_{opt}$ , and interconnect bandwidth becomes equal to  $f_{elect-max}$ . Hence, optimal wire width not only maximizes data flux density, but also minimizes energy-per-bit. Assuming that driver resistance is equal to the line's characteristic impedance, the minimum energy per bit is

$$E_{b-min} = \frac{C_p V_{dd}^2}{2} + \frac{V_{dd}^2}{2Z_0} \frac{1}{f_{elect-max}}. \quad (8.7)$$

### 8.3 Optical Interconnects

While bandwidth of an electrical interconnect is determined by its length and cross-sectional dimensions, in practice, physical dimensions of an optical waveguide have no impact on data bandwidth. Indeed, the maximum number of bits per second that an optical link can transfer is determined by the driver and receiver [51, 54]. In this way, the maximum data flux density is

$$\phi_{D-optic} = f_{opt-max} / p_{opt}, \quad (8.8)$$

where  $f_{opt-max}$  is the maximum frequency at which optical drivers and receivers can switch, and  $p_{opt}$  is the pitch of optical waveguides. Although optical wavelength and waveguide technology influence optical crosstalk and hence the minimum value of  $p_{opt}$  [55],  $p_{opt}$  is tentatively set to  $2W_{min}$ . The maximum data flux density for the optical interconnects, therefore, is

$$\phi_{D-optic-max} = f_{opt-max} / 2W_{min}, \quad (8.9)$$

where  $W_{min}$  is the minimum line width that the available PWB technology permits. Figure 8.3 illustrates data flux density for both electrical and optical interconnects at 130 nm and 45 nm technology nodes assuming that maximum switching frequency is the same for optical and electrical transceivers.

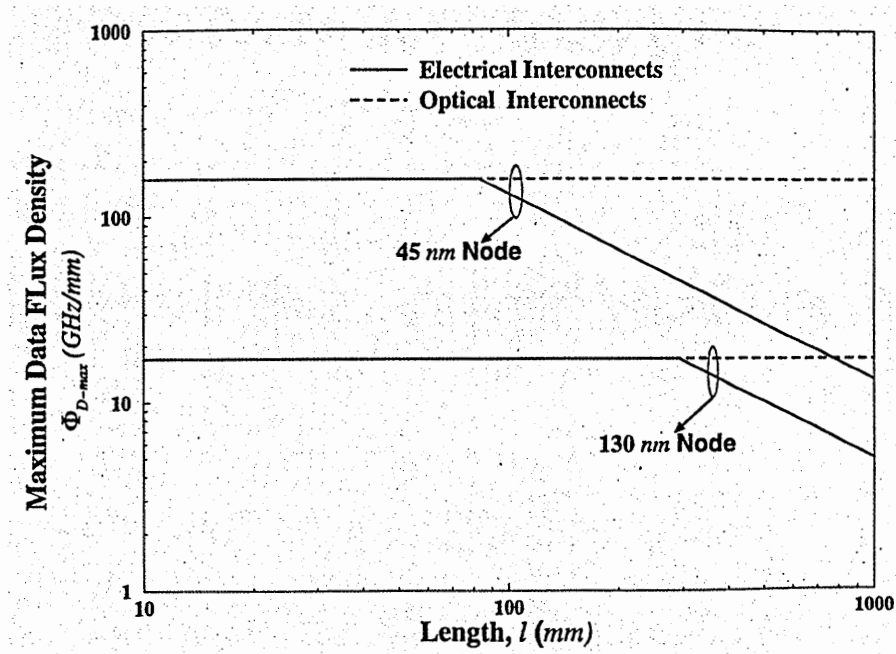


Figure 8.3: Maximum data flux density for electrical and optical interconnects implemented in 130 and 45 nm technology nodes. Performance of optical waveguides is length independent [51], [54].

Although area efficient routing is relatively easy for electrical interconnects because of their multilevel orthogonal routing capabilities, optical waveguides require less-efficient routing schemes to avoid lossy sharp bends and the need for inter-level communication. Although mirrors and grating couplers have been realized within a PWB [38], it is beneficial to minimize their use due to power budget constraints and cost. Hence, optical I/Os should be carefully placed such that board-level waveguides do not block one another.



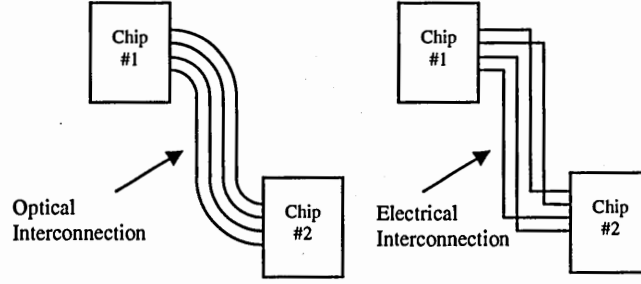


Figure 8.4: Unlike electrical interconnects, optical waveguides can not be routed arbitrarily because sharp bends and inter-level communication should be minimized due to power budget constraints.

## 8.4 The Partition Length between Electrical and Optical Interconnects

Although  $f_{opt-max}$  and  $f_{elec-max}$  can be different, they cannot be completely unrelated, since an optical driver is driven by an electrical driver and an optical receiver feeds an electrical driver. In this analysis, the maximum switching frequencies for optical and electrical transceivers are assumed to be equal. A small modification would be necessary if these two frequencies are different.

By comparing (8.5) and (8.9), a partition length for electrical and optical interconnects can be found as

$$l_{part} = W_{min} \sqrt{K_0 / f_{max}} , \quad (8.10)$$

which shows the interconnect length beyond which optical waveguides offer a larger data flux density compared to electrical interconnects. Figure 8.5 plots the partition length for different ITRS [12] technology nodes. From Figure 8.3, it is evident that the data flux density of optical waveguides is constant and larger than that of electrical interconnects for interconnects longer than  $l_{part}$ . As technology advances, the partition length decreases

due to decreases in available board-level line width *and* increases in clock frequency. For instance, the partition length decreases from 29 cm at the 130 nm to 8.3 cm at the 45 nm because of 7 times faster transistors and 25% smaller line width.

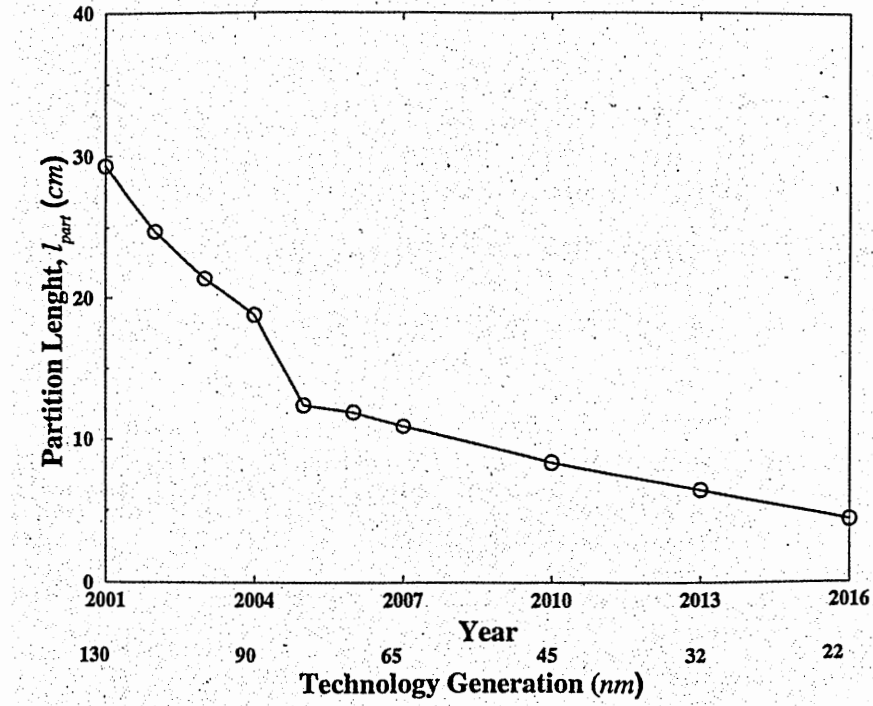


Figure 8.5: Partition length between optical and electrical interconnects for different generations of technology. Partition length decreases in future generations because of faster transistors and finer PWB line width.

By pushing the PWB technology and achieving minimum line widths smaller than those the ITRS has projected, smaller partition lengths can be attained and optical waveguides outperform electrical interconnects to an even greater extent. Hence, *if the minimum wire width is sufficiently small, it would be better to replace virtually all on-board wires with optical waveguides*. Figure 8.6 plots the partition length versus minimum board-level line width for four different maximum switching frequencies. From Figure 8.6 it can be seen that a minimum line width of 10  $\mu m$  makes the partition length

shorter than most board-level interconnects. The minimum waveguide pitch, however, is limited by the crosstalk between adjacent lines, and is different for various waveguide technologies. For instance, the minimum waveguide pitch for optical waveguides composed of  $\text{SiO}_2/\text{Air}/\text{MSQ}$  and operating at  $632\text{ nm}$  wavelength is  $5\mu\text{m}$  [55].

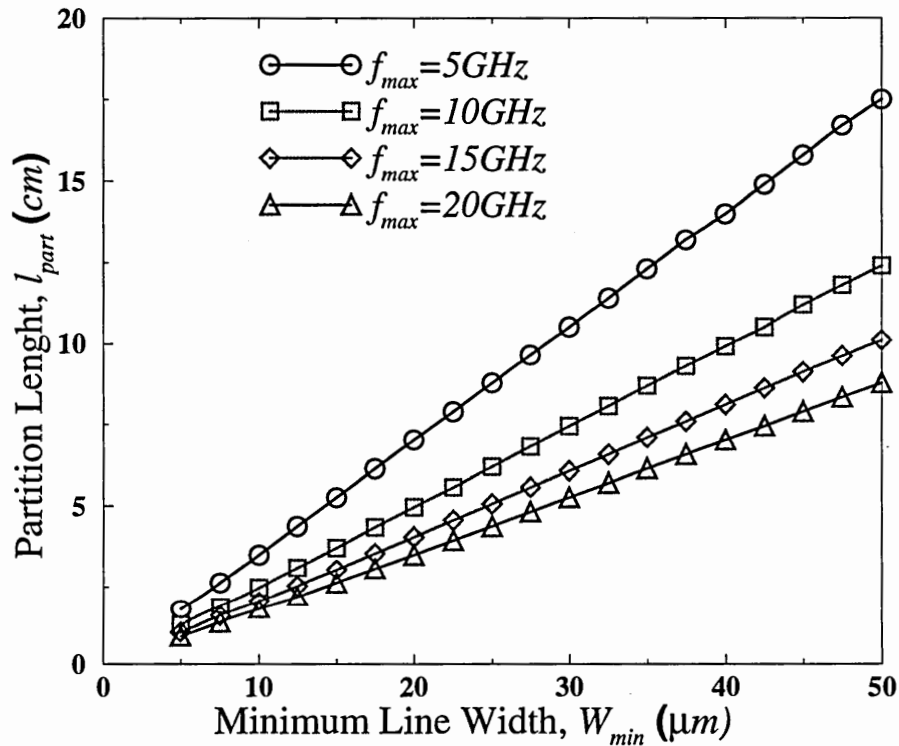


Figure 8.6: The partition length between optical and electrical interconnects versus minimum available board-level line width. Having a  $W_{min}$  of about  $10\mu\text{m}$  makes the partition length shorter than almost all board-level interconnects; hence, justifies replacing all board-level electrical interconnects with optical waveguides.

## 8.5 Conclusions

For various generations of technology, a partition length between electrical and optical interconnects is identified based on aggregate bandwidth constrained by a fixed routing area. Based on ITRS projections for chip-to-board clock frequency and minimum available PWB line width, this partition length is 29 *cm* for the 130 *nm* technology node which decreases to 8.3 *cm* at the 45 *nm* technology node. However, by pushing the PWB industry and achieving a minimum line width of around 10  $\mu m$ , the partition length will be shorter than almost all board-level interconnects.

## Chapter 9

### Future Work and Conclusion

In this chapter, some of the potential extensions to this thesis that can be perused for future research endeavors are briefly discussed. These extensions include advancing compact physical models, optimizing new interconnection techniques, developing stochastic models for crosstalk, extending chip-package co-design methodologies, and analyzing and optimizing clock distribution networks. At the end of this chapter, the key conclusions of this dissertation are summarized.

#### 9.1 Advancing Compact Physical Models

Novel physical models for latency and crosstalk of chip- and package-level interconnects have been presented in this thesis. However, more thoroughgoing models are required for enhancing the accuracy of the analyses. For instance, throughout this dissertation, it has been assumed that interconnects are driven by linear sources that are excited by step inputs. However, CMOS gates, which are the main drivers in GSI chips, are non-linear elements, and also have a relatively large output capacitance. Their input rise-time can be large, too especially if they are driven through long wires.

Skin effect, proximity effect, and surface scattering are also ignored in this thesis. As the switching frequency rises and the skin depth becomes comparable with wire cross-sectional dimensions, effective cross-sectional area of wires decreases and interconnect

resistance becomes frequency-dependent. Furthermore, at different frequencies, the distribution of return current changes, and that makes the inductance of interconnects a function of frequency. Accordingly, chip-level interconnects should be modeled as  $R(f)L(f)C$  lines. Surface scattering can also affect interconnect resistance if either wire dimensions or the skin depth are in the order of “electron mean free path [56].”

## 9.2 Optimizing Other Interconnecting Techniques

The optimal wire width that maximizes data flux density-reciprocal latency product is identified in this thesis. To calculate data flux density, which is bandwidth per unit width of interconnects, interconnect bandwidth is assumed to be equal to its reciprocal latency. This assumption is valid if each bit of data is lunched when the former bit has been detected by the receiver. However, as GSI chips become more and more interconnect-limited, more aggressive interconnecting techniques such as wave-pipelining are being investigated [57]. In the wave-pipelining technique, an interconnect is treated as a pipeline, and interconnect bandwidth is therefore not limited by its latency.

Global interconnects are good candidates for wave-pipelining because of two reasons. First, global interconnects are typically partitioned by several repeaters each of which provides one pipeline stage. Second, several pulses can be pipelined along the thick and wide global interconnects that work in the RLC regime. To design and optimize pipelined interconnects, interconnect bandwidth should be modeled rigorously. Since the existing optimal size and number of repeaters are based on latency minimization, new bandwidth-oriented repeater insertion algorithms should be developed. The cross-sectional

dimensions of interconnects should then be optimized for maximizing pipelined data flux density and minimizing latency.

### 9.3 Developing Stochastic Models for Crosstalk Noise

The crosstalk models presented in this thesis show that the number of aggressors affecting a victim line can be quite large. Since different aggressors can switch at any time and direction, the noise voltage strongly depends on the aggressors switching pattern. To have a reliable and robust design, crosstalk noise voltage is modeled for the worst-case scenario when all aggressors are switching at specific directions and at specific time phases. However, designing global interconnects based on these worst-case models might be too conservative. For instance, the probability of simultaneous switching of 10 aggressors at a specific direction might be very low. Stochastic models for crosstalk noise can, therefore, be very insightful. These stochastic models can predict the probability of various noise voltages on each node. Global interconnects can then be designed for error probabilities larger than zero along with error detection and error correction circuits. For instance, parity bits can be added to data-buses to detect and even correct possible errors. The number of parity bits needed depends on the error probability; hence, there will be a trade-off between error detection/correction overhead and the noise level. In this manner, the stochastic crosstalk models can be utilized to optimize the structure of global interconnects.

## 9.4 Extending Chip-Package Co-Design Methodologies

In Chapter 7, the impact of interconnect optimization on the package design has been studied, which illustrates how for a given IR drop budget, knowing the on-chip power/ground metal area determines the required number of power and ground pins. However, IR drop is not the only concern that the designers have in designing power and ground distribution network. Simultaneous switching noise (SSN) and electromigration (EM) are as important as IR drop. SSN, also called  $di/dt$  noise, is due to sudden current changes in an inductive power distribution network. Inductances of package I/O pins and on-chip power/ground grid both can contribute to SSN.

Electromigration, which is due to large current densities, can damage I/O pins and also on-chip vias and thereby determines the chip's life-time. Rigorous models are needed for SSN and EM to optimally design chip- and package-level power/ground distribution networks.

## 9.5 Analysis and Design of Clock Distribution Networks

While signal and power interconnects are analyzed and optimized in this thesis, clock distribution is not studied in detail. The reason is that clock distribution networks have become so sophisticated that modeling and optimizing them are beyond the time-frame and the scope of this thesis. Clock distribution networks in state-of-the-art high-performance systems consist of a hierarchy of H-trees, grids, and active jitter and skew compensation techniques [58, 59, 60]. In some cases, optical off-chip clock distribution networks are also employed [38]. The physical models that are derived for co-planar transmission lines can be directly applied to H-trees; clock skew can be modeled by



latency models, and crosstalk models can predict jitter. New models, however, are required for skew, jitter and power dissipation of grids and active compensations circuits. Taking advantage of these models, the hierarchy of clock distribution network can be optimized for minimizing area, power, skew and jitter simultaneously.

## 9.6 Conclusion of Dissertation

The main objectives of this thesis are: 1) to develop interconnect centric-methodologies to optimize global interconnects, 2) to derive compact physical models for delay and crosstalk of interconnects and incorporate them in optimization of global interconnect, and 3) to optimally utilize package to improve on-chip and chip-to-chip signal and power interconnection.

The main contributions of this thesis are as follows:

1. A new interconnect-centric approach is proposed to optimize global interconnects to simultaneously maximize data flux density and minimize latency. An optimal wire width is rigorously derived that is length independent and can be used for virtually all global interconnects.
2. The set of differential equations of  $n$ -coupled distributed RLC lines above an ideal ground plane is rigorously solved. Using the solution, it is proved that far inductive noise is negligible if a nearby ground plane exists.
3. Compact physical models are derived for the latency and crosstalk of co-planar transmission lines that are above orthogonal lines.
4. The compact physical models of co-planar transmission lines are used to optimize the cross-sectional dimensions of on-chip global interconnects. Through this

optimization, data flux density-reciprocal latency product is maximized, crosstalk and dynamic delay variation due to different switching patterns are substantially reduced, and the best trade-off between energy-per-bit and data flux density is achieved.

5. Compact physical models are derived to model crosstalk noise caused by virtually all near, inter- and intra-level far aggressors. The models show that a prohibitively large noise can be induced on a victim line if interconnect resistance is too small and emphasize the role of repeaters in reducing noise voltage with an inconsiderable latency penalty.
6. It is rigorously proved that by employing optimal wire width, the crosstalk caused by all near and far aggressors remains small and constant in all generations of technology.
7. An optimal partition is identified between exterconnects and interconnects to achieve the maximum possible global clock frequency with minimum numbers of board levels and I/Os. For a projected chip implemented at the 45 nm node, the global clock frequency can be improved by 33% using 4 additional board levels and 1800 I/Os.
8. The required number of power and ground I/Os dictated by IR-drop are calculated for various technology generations assuming that the optimal global wire dimensions are used.
9. The lengths beyond which optical waveguides can outperform electrical wires in terms of data flux density are identified for various technology generations. It is illustrated that if boards with line resolutions of around 10  $\mu\text{m}$  are available,

optical interconnects will offer larger data flux densities for most typical chip-to-chip interconnect lengths.

Some of the potential extensions to this thesis include advancing the compact physical models to consider non-linearity of the drivers as well as frequency-dependent inductance and resistance values, optimizing new interconnection techniques, developing stochastic models for crosstalk, extending chip-package co-design methodologies, and analyzing and optimizing clock distribution networks.

## Appendix A

### Derivation of Optimal Wire Width

It was shown that the optimal wire width is the wire width at which the data flux density-reciprocal latency product is maximized (or  $W\tau^2$  is minimized). In this appendix the value of this optimal wire width is rigorously derived. For optimal repeater insertion in the RLC region, physical models derived in [13] are used.

Delay of an interconnect in the RLC regime with optimal repeaters is [13]

$$\tau_{RLC} = (1 + 1.5 \frac{R_0 C_0}{Z_0^2} \frac{r_{int}}{c_{int}}) T_o F. \quad (A.1)$$

By using (2.18) and (A.1) it can be written

$$W\tau_{RLC}^2 = W[(1 + 1.5 R_0 C_0 r c \frac{c_0^2}{\epsilon_r}) T_o F]^2, \quad (A.2)$$

and by substituting (2.20) in (A.2)

$$W\tau_{RLC}^2 = W[(1 + 1.5 R_0 C_0 \frac{\xi \rho \epsilon_0 c_0^2}{W^2}) T_o F]^2. \quad (A.3)$$

In order to find the width at which  $W\tau^2$  is minimum, (A.4) should be solved:

$$\frac{\partial}{\partial W} (W\tau_{RLC}^2) \equiv 0, \quad (A.4)$$

which gives the optimal wire width as

$$W_{opt} = 2.12 c_0 \sqrt{R_0 C_0 \rho \xi \epsilon_0}. \quad (A.5)$$

By substituting the optimal wire width (A.5) into the RC model delay (2.1), it can be shown that at the optimal width:

$$\tau_{RC} = 2.5 \frac{l}{W_{opt}} \sqrt{\rho \xi \epsilon_r \epsilon_0 R_0 C_0} = 1.18 ToF. \quad (A.6)$$

Using (A.1), the RLC delay at the optimal wire width is

$$\tau_{RLC} = (1 + 1.5 R_0 C_0 \frac{\xi \rho \epsilon_0 c_0^2}{W_{opt}^2}) ToF = 1.33 ToF. \quad (A.7)$$

Comparing (A.6) and (A.7) shows that at the optimal wire width, the RC model underestimates the delay by 15%.

It is worthwhile to note that in this derivation, parasitic output capacitance of a repeater,  $C_{out}$  is neglected. The reason is that although  $C_{out}$  can be large in deep submicron technologies, its impact on interconnect latency and therefore on optimal wire width is small. The following analysis shows this more precisely.

Latency of each segment between two repeaters is

$$\tau_{seg} = \frac{\tau}{k_{opt}}, \quad (A.8)$$

and using (2.15) and (A.7), latency of each segment with optimal wire width can be written as

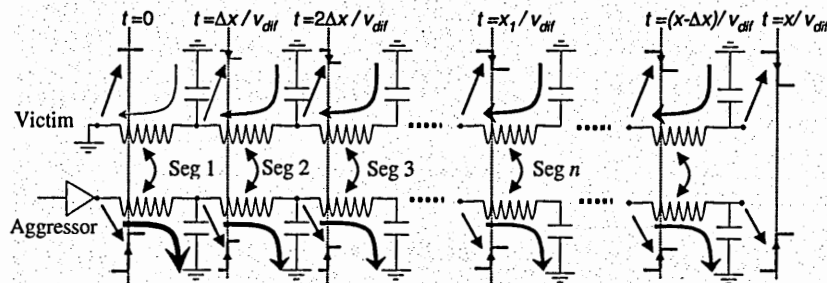
$$\tau_{seg} = 6.3 R_0 C_0. \quad (A.9)$$

Assuming that the value of  $C_{out}$  is equal to the input capacitance of the repeater [21], the ratio of the time-constant for the  $C_{out}$  over the segment latency would be constant and equal to 0.16 for all interconnect lengths and geometries, and all technology generations. Hence, if the error caused by neglecting  $C_{out}$  is identified for a specific case, it would be the same for all cases. HSPICE simulations show that neglecting  $C_{out}$  causes 10% error in the latency and less than 8% error in the optimal wire width. Hence, to have a compact expression for optimal wire width,  $C_{out}$  is neglected.

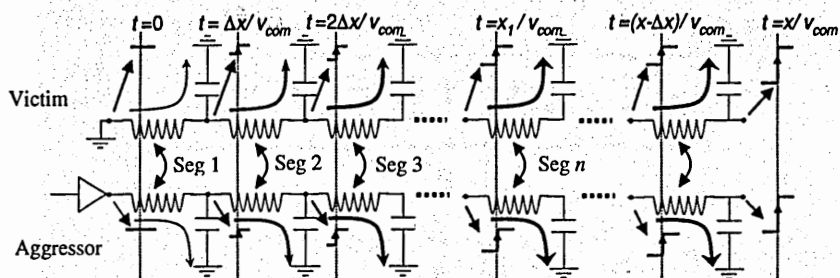
## Appendix B

### Physical Explanation for Crosstalk of Identical Lines

Equation (6.14) shows that the peak noise voltage of an open-ended line is independent of the absolute values of the inductive couplings. This means that if all mutual couplings are made smaller by the same factor, e.g. by making power/ground lines wider, the peak noise does not change. For instance, in a lossless two-line case, the open-ended noise voltage is  $-V_{dd}$ , independent of  $l_m$ . The reason is shown in Figure 71 and Figure 72. As the aggressor switches, a current is induced in the first infinitesimal segment of the victim line (Seg. 1) due to the inductive coupling. This current causes a small negative voltage swing on the input of the next segment (Seg. 2). The next segment (Seg. 2) is then excited by both inductive coupling and the voltage swing at its input, and a larger voltage appears on the third segment (Seg. 3). In this manner, as signal propagates, two lines resonate, and the aggressor transfers part of its energy to the victim line until their voltages become equal with opposite signs. This differential mode travels with the speed of  $((l_s - l_m)c)^{-0.5}$ ,



**Figure 71:** The differential mode caused in a lossless two-line case. Voltage and current changes corresponding to each infinitesimal segment are shown as a wave travels along the lines. It is shown how aggressor and victim lines resonate until the absolute value of their voltages become equal. The induced noise voltage is  $-V_{dd}$  regardless of the value of the mutual inductance.



**Figure 72: The common mode caused in a two-line lossless case for which the wave travels slower than the differential mode. The common mode discharges the noise voltage caused by the differential mode. The mutual inductance determines the noise duration.**

and its amplitude is doubled when the differential signal reaches the end of the lines. Meanwhile, two equal currents in the same directions are induced in both lines to charge them to their input voltages as shown in Figure 72. This common mode wave propagates with a lower speed  $((l_s + l_m)c)^{-0.5}$ .

## Appendix C

### Peak Crosstalk Voltage in Terms of $W/W_{opt}$

To study the impact of wire width optimization on crosstalk, the peak noise voltage should be found as a function of  $W/W_{opt}$  ratio. Assuming that optimal repeaters are utilized, the peak open-ended noise voltage is calculated in Section C.1 and in Section C.2 the impact of the load capacitance is taken into account.

#### C. 1 Open-Ended Noise Voltage

By substituting (5.27) and (5.23) in (6.33), the peak noise voltage induced by intra-level interconnects, can be written as

$$V_{\text{intra-level}} = V_{dd} \left[ \left( \frac{\sum_{\text{All FarPairs}} l_{pi}}{\sqrt{\sum_{\text{All FarPairs}} l_{pi}^2}} + 1 \right) \frac{1}{\frac{1}{1.15} \sqrt{\frac{Z_{dif}}{Z_{com}}} + 1} e^{-0.525 \sqrt{\frac{Z_{dif}}{Z_{com}}}} - \frac{1}{\frac{1}{1.15} \sqrt{\frac{Z_{com}}{Z_{dif}}} + 1} e^{-0.525 \sqrt{\frac{Z_{com}}{Z_{dif}}}} \right], \quad (\text{C.1})$$

which is independent of resistance per unit length of interconnects, and is determined by the  $\sqrt{Z_{com}/Z_{dif}}$  and the mutual inductances between far aggressors and the victim line.

The ratio of the common and differential mode characteristic impedances is given by (4.28), and is equal to

$$\frac{Z_{com}}{Z_{dif}} = \sqrt{\left(1 + \frac{2c_m}{c_g}\right) \left(1 + \frac{2c_m}{c_g + c_{orth}}\right)} \quad (\text{C.2})$$



where  $c_m$  is the mutual capacitance between two near signal lines and  $c_g$  is the capacitance between a signal line and a nearby ground line and  $c_{orth}$  is the capacitance between a signal line and the orthogonal lines. For the optimal spacing case,  $c_m=0.45c_g$ ,  $\sqrt{Z_{com}/Z_{dif}}$  can be approximated by 1.28 for a wide range of  $c_{orth}/c_g$  ( $0.3 < c_{orth}/c_g$ ) with less than 4% error. In this manner, (C.1) can be rewritten as

$$V_{intra-level} = V_{dd} \left[ 0.39 \frac{\sum_{All\ FarPairs} l_{pi}}{\sqrt{\sum_{All\ FarPairs} l_{pi}^2}} + 0.1 \right] \quad (C.3)$$

The same approach can be used for the crosstalk caused by inter-level aggressors. By substituting (5.27) and (5.23) in (6.20), the peak open-ended noise caused by inter-level aggressors can be rewritten as

$$V_{non-identical}(t) = 2 \frac{\sum_{All\ Aggressors} l_{vi} c_{f-eq}}{(l_{f-eq} c_{f-eq} - l_p c_p)} \frac{1}{1.15 \sqrt{\frac{Z_{dif}}{Z_{com}} + 1}} e^{-0.525 \sqrt{\frac{Z_{dif}}{Z_{com}}}}, \quad (C.4)$$

which can be similarly approximated by

$$V_{inter-level}(t) = 0.78 \frac{\sum_{All\ Aggressors} l_{vi} c_{f-eq}}{(l_{f-eq} c_{f-eq} - l_p c_p)}. \quad (C.5)$$

The peak open-ended noise voltage is the summation of (C.3) and (C.5):

$$V_{open} = V_{dd} \left[ 0.39 \frac{\sum_{All\ FarPairs} l_{pi}}{\sqrt{\sum_{All\ FarPairs} l_{pi}^2}} + 0.1 + 0.78 \frac{\sum_{All\ Aggressors} l_{vi} c_{f-eq}}{(l_{f-eq} c_{f-eq} - l_p c_p)} \right]. \quad (C.6)$$

## C. 2 Impact of Load Capacitance

To find the peak noise voltage, the impact of the load capacitance should be taken into account. The load capacitance is charged by the time constant of  $Z_0 C_L$ . The peak noise voltage is, therefore,

$$V_{load} = V_{open} [1 - \exp(-t_n / Z_0 C_L)]. \quad (C.7)$$

The duration of noise pulse is given by (6.34) and can be written as

$$t_n = \frac{\ell_{seg} \sqrt{\epsilon_r}}{c_0} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)}. \quad (C.8)$$

By substituting (5.23) in (C.8), the noise duration is

$$t_n = \frac{1.05 \sqrt{\epsilon_r Z_{com} Z_{dif}}}{rc_0} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)}. \quad (C.9)$$

Using the equation for the optimal wire width, (5.12), the noise duration can be written in terms of  $W/W_{opt}$ :

$$t_n = 5R_0 C_0 \frac{W^2}{W_{opt}^2} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)}. \quad (C.10)$$

The last parameter in (C.7) that should be identified is the load capacitance,  $C_L$ . Assuming that optimal repeaters are inserted, the load capacitance would be the input capacitance of an optimal repeaters given by [18]

$$C_L = C_0 h_{opt} = \frac{1.15 R_0 C_0}{\sqrt{Z_{dif} Z_{com}}}. \quad (C.11)$$

By substituting (C.6), (C.10), and (C.11) in (C.7), the peak noise voltage considering the load capacitance is

$$V_{load} = V_{dd} \left[ 0.39 \frac{\sum_{All\ FarPairs} l_{pi}}{\sqrt{\sum_{All\ FarPairs} l_{pi}^2}} + 0.1 + 0.78 \frac{\sum_{All\ Aggressors} l_{vi} c_{f-eq}}{(l_{f-eq} c_{f-eq} - l_p c_p)} \right] \left[ 1 - \exp \left( -3.39 \frac{W^2}{W_{opt}^2} \sqrt{\sum_{All\ Far\ Pairs} (l_{pi}^2 / l_p^2)} \right) \right]. \quad (C.12)$$

## Appendix D

### Derivation of Bit-Rate Limit of RLC Lines <sup>[52]</sup>

The transfer function of a transmission line in Laplace domain can be written as [61]

$$V(x=l)/V_{in} = \exp(-sl\sqrt{LC}) \exp(-l\sqrt{s}R_{sk}/2Z_0), \quad (D.1)$$

where  $Z_0$  is the line's characteristic impedance, and  $L$  and  $C$  are the line's inductance and capacitance per unit length, respectively.  $R_{sk} = \sqrt{\rho\mu}/P$  is the skin effect resistance [61], where  $\mu$  is the magnetic permeability,  $\rho$  is metal resistivity, and  $P$  is the circumference. By taking the inverse Laplace transform of (D.1), the step response of a transmission line can be obtained as

$$f(t) = \text{erfc}\sqrt{\beta l(t - l\sqrt{LC})} \quad (D.2)$$

for  $t \geq \sqrt{LC}$ , and zero for  $t < \sqrt{LC}$ , where  $\text{erfc}()$  is the complementary error function, and  $\beta \equiv (\mu\rho/16Z_0^2)(P/l)^2$ . Equation (D.2) shows that the signal rise time at the end of the transmission line, which determines the opening dimension in the eye diagram [49], is determined by  $\beta$ . Hence, the interconnect bit rate limit can be written as

$$B = 1/k\beta. \quad (D.3)$$

Constant  $k$  determines the desired eye-opening and is typically chosen between 60 to 120 [49].

## References

- [1] G. E. Moore, "Progress in digital integrated circuits," *IEEE IEDM Tech. Digst.*, Dec. 1975, pp. 11-13.
- [2] J. D. Meindl, R. Venkatesan, J. A. Davis, J. Joiner, A. Naeemi, P. Zarkesh-Ha, M. S. Bakir, T. M. Mule, P. A. Kohl, K. P. Martin, "Interconnecting device opportunities for gigascale integration (GSI)," *IEEE IEDM Tech. Digst.*, December 2001, pp. 525-528.
- [3] D. Sylvester and K. Keutzer, "Impact of small process geometry on microarchitectures in systems on a chip," *Proc. IEEE*, vol. 89, no. 5, pp. 467-489, April 2001.
- [4] R. H. Havemann and J. A. Hutchby, "High-performance interconnects: an integration overview," *Proc. IEEE*, vol. 89, no. 5, pp. 586-601, May 2001.
- [5] G. A. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, pp. 20-36, Jan. 1995.
- [6] D. Sylvester and C. Hu, "Analytical modeling and characterization of deep-submicron interconnect," *Proc. IEEE*, vol. 89, no. 5, pp. 634-664, May 2001.
- [7] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI)—Parts I and II," *IEEE Trans. Electron Devices*, vol. 45, pp. 580-597, March 1998.
- [8] R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal n-tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Trans. VLSI Syst.*, vol. 9, no. 6, pp. 899-912, December 2001.
- [9] A. B. Kahng, D. Stroobant, "Wiring layer assignment with consistent stage delays," in *Proc. Int. Workshop on System-Level Interconnect Prediction*, April 2000, pp. 115-122.
- [10] P. Zarkesh-Ha and J. D. Meindl, "An integrated architecture for global; interconnects in a gigascale system-on-a-chip," *IEEE Proc. VLSI Symp.*, June 2000, pp. 194-195.
- [11] A. Deutsch et al., "On-chip wiring design challenges for gigahertz operation," *Proc. IEEE*, vol. 89, no. 4, pp. 529-554, April 2001.

- [12] Semiconductor Industry Association, "International Technology Roadmap for Semiconductors (ITRS)," Edition 2001.
- [13] H. B. Bakoglu and J. D. Meindl, "Optimal interconnect circuits for VLSI," *IEEE Trans. Electron Devices*, pp. 903-909, May 1985.
- [14] D. Bradley, P. Mahoney, B. Stackhouse, "The 16kB single-cycle read access cache on a next-generation 64b Itanium microprocessor," *IEEE ISSCC Dig. Tech. Papers*, February 2002, pp. 110-111.
- [15] R. Riedlinger, T. Grukowski, "The high-bandwidth 256kB 2<sup>nd</sup> level cache on an Itanium microprocessor," *IEEE ISSCC Dig. Tech. Papers*, February 2002, pp. 418-419.
- [16] J. Cong, "An interconnect-centric design flow for nanometer technologies," *Proc. IEEE*, vol. 89, no. 5, pp. 505-528, April 2001.
- [17] RAPHAEL: *Interconnect Analysis Program*, TMA Inc, 1996.
- [18] R. Venkatesan, J. A. Davis, and J. D. Meindl, "A physical model for the transient response of capacitively loaded distributed *rlc* interconnects," *Proc. DAC*, June 2002, pp. 763-766.
- [19] A. Deutsch, G. V. Kopcsay, P. W. Coteus, C. W. Surovic, P. Dahlenz, D. L. Heckmann, D. W. Duan, "Bandwidth prediction for high-performance interconnects," *Proc. of Electronic Components and Technology Conference*, pp. 256-66, May 2000.
- [20] J. A. Davis and J. D. Meindl., "Compact distributed RLC interconnect models. I. Single line transient, time delay, and overshoot expressions," *IEEE Trans. Electron Devices*, vol. 47, no. 11, pp. 2068-2087, Nov 2000.
- [21] J. C. Eble, V. K De, D. S. Wills, J. D. Meindl, " Minimum repeater count, size, and energy dissipation for gigascale integration interconnects," *IEEE International Interconnect Technology Conference*, June 1998, pp. 56-58.
- [22] A. Naeemi and J. D. Meindl, "An optimal partition between on-chip and on-board interconnects," *IEEE International Interconnect Technology Conference*, June 2001, pp. 131-133.
- [23] J. C. Eble, "A generic system simulator with novel on-chip cache and throughput models for gigascale integration," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, November 1998.

- [24] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A compact physical via blockage model," *IEEE Trans. VLSI Syst.*, vol. 8, no. 2, pp. 689-692, December 2000.
- [25] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in VLSI's," *IEEE Trans. Electron Devices*, vol. 40, pp. 118-124, Jan 1993.
- [26] J. A. Davis and J. D. Meindl., "Compact distributed RLC interconnect models. I. Single line transient, time delay, and overshoot expressions," *IEEE Trans. Electron Devices*, vol. 47, no. 11, pp. 2068-2087, Nov 2000.
- [27] J. A. Davis and J. D. Meindl., "Compact distributed RLC interconnect models. II. Single line transient, time delay, and overshoot expressions," *IEEE Trans. Electron Devices*, vol. 47, no. 11, pp. 2068-2087, Nov 2000.
- [28] Clayton R. Paul, *Analysis of Multiconductor Transmission Lines*. New York: John Wiley and Sons, 1994, pp. 46-63.
- [29] A. Deutsch, et al, "On-chip wiring design challenges for gigahertz operation," *Proc. of IEEE*, vol. 89, no. 4, April 2001, pp. 529-554.
- [30] A. E Ruehli, "Inductance calculations in a complex integrated circuit environment," *IBM J. Res. Develop.*, vol. 16, no. 5, pp 470-481, September 1972.
- [31] M. W. Beattie, and L. T. Pilleggie, "On-chip induction modeling: basics and advanced methods," *IEEE Trans. VLSI Syst.*, vol. 10, no. 6, pp. 712-729, December 2002.
- [32] G. V. Kopcsay, B. Krauter, D. Widiger, A. Deutsch, B. J. Rubin, and H. H. Smith, "A comprehensive 2-D inductance modeling approach for VLSI interconnects: frequency-dependent extraction and compact circuit model synthesis," *IEEE Trans. VLSI*, vol. 10, no. 6, pp. 665-711, December 2002.
- [33] A. Kowalczyk et al., "First-generation MAJC dual microprocessor," *ISSCC Tech Digst.*, February 2001, pp. 236-237.
- [34] S. Rusu et al., "The first IA-64 Microprocessor," *IEEE J. Solid-State Circuits*, pp. 1539-1544, November 2000.
- 35 R. E. Collin, *Foundations for Microwave Engineering*. New York: IEEE Press, 2001, pp. 550-590.
- [36] D. B. Jarvis, "The effects of interconnections on high-speed logic circuits," *IEEE Trans. On Electronic Computers*, pp. 476-487, October 1963.

- [37] Q. Zhu and S. Tam, "Package clock distribution design optimization for high-speed low-power VLSI's," *IEEE Trans. Components, Packaging, and Manufacturing Technol.* pp.56-63, February 1997.
- [38] R. T. Chen et al, "Fully embedded board-level guided-wave optoelectronic interconnects," *Proc. IEEE*, vol. 88, pp. 780-793, June 2000.
- [39] A. V. Mule', E. N. Glytsis, T. K. Gaylord, and J. D. Meindl, "Electrical and optical clock distribution networks for high performance microprocessors," *IEEE Trans. VLSI Syst.*, vol. 10, no. 5, pp. 582-594, October 2002.
- [40] A. Naeemi, C. S. Patel, M. S. Bakir, P. Zarkesh-Ha, K. P. Martin and J. D. Meindl, "Sea of Leads: A disruptive paradigm for a system-on-a-chip (SoC)", *IEEE International Solid-State Circuits Conference*, February 2001, pp 280-281.
- [41] S. Afonso, L. W Schaper, J. P. Parkerson, W. D. Brown, S. S. Ang, and H. A. Naseem, "Modeling and electrical analysis of seamless high off-chip connectivity (SHOCC) interconnects," *IEEE Trans. Advanced Packaging*, pp. 309-320, August 1999.
- [42] P. Zarkesh-Ha, J. A. Davis, J. D. Meindl, "Prediction of net-length distribution for global interconnects in a heterogeneous system-on-a-chip," *IEEE Trans. VLSI Syst*, pp. 649-659, December 2000.
- [43] C. S. Patel, et al., "Optimal printed wiring board design for high I/O density chip size packages," *Circuit World, Journal of the Institute of Circuit Technology*, pp. 25-27, August 1999.
- [44] C. S. Patel, et al., "An analysis of the gap between PWB technology and chip I/O interconnect technology, and a new wafer-level batch packaging concept," *32<sup>nd</sup> International Symposium on Microelectronics*, October 1999, pp. 611-618.
- [45] J. C. Eble, "A generic system simulator with novel on-chip cache and throughput models for gigascale integration," Ph.D. Thesis, Georgia Institute of Technology, Atlanta, 1998.
- [46] S. Voldman, V. Gross, "Scaling, optimization and design considerations of electrostatic discharge protection circuit in CMOS technology," *Proc. of Electrical Overstress/Electrostatic Discharge Symposium*, September 1993 pp. 251-260.



- [47] K. Shakeri and J. D. Meindl, "Compact physical IR-drop models for GSI power distribution networks," *IEEE International Interconnect Technology Conference*, June 2003, pp. 54-56.
- [48] D. A. B. Miller, "Optics for low-energy communication inside digital processors," *Opt. Lett.*, pp. 146-149, January 1989.
- [49] C. Svensson, "Electrical interconnects revitalized," *IEEE Trans. VLSI Syst.*, vol. 10, pp. 777-789, December 2002.
- [50] E. D. Kyriakis-Bitxaros et al., "Realistic end-to-end simulation of the optoelectronic links and comparison with the electrical interconnects for system-on-a-chip applications," *J. Lightwave Technol.*, pp. 1532-1542, October 2001.
- [51] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728-749, June 2000.
- [52] Personal communication with Dr. Jianping Xu from Intel, email: Jianping.xu@intel.com.
- [53] M. Mizuno et al., "Clock distribution networks with on-chip Transmission lines," *IEEE Intl. Interconnect Technol. Conf.*, June 2000, pp. 3-5.
- [54] A. Z. Shang and F. A. P. Tooley, "Digital optical interconnects for networks and computing systems," *J. Lightwave Technol.*, vol. 18, pp. 2086-2094, Dec. 2000.
- [55] A. V. Mule et al., "Towards a comparison between chip-level optical interconnection and board-level interconnection," *IEEE Intl. Interconnect Technol. Conf.*, June 2002, pp. 92-94.
- [56] R. Sarvari and J.D. Meindl, "On the study of anomalous skin effect for GSI interconnections," *IEEE Intl. Interconnect Technol. Conf.*, June 2003, pp. 42-44.
- [57] H. Shah, P. Shiu, B. Bell, M. Aldredge, N. Sopory, and J. A. Davis, "Repeater insertion and wire sizing opportunities for throughput-centric VLSI global interconnects," *IEEE/ACM International Conf. Computer Aided Design*, Nov. 2002, pp. 280-284.
- [58] P. J. Restle and A. Deutsch, "Designing the best clock distribution network," *IEEE Symposium on VLSI Circuits*, pp. 2-5, 1998.
- [59] S. Rusu and G. Singer, "The first IA-64 Microprocessor," *IEEE J. Solid-State Circuits*, pp. 1539-1544, December 2000.

- [60] F. Anderson, J. Steve Wells, and E. Z. Berta, "The core clock distribution on the next generation Itanium microprocessor," *IEEE Int. Solid-State Circuit Conf.* pp. 146-147, February 2002.
- [61] R. Matick, *Transmission Lines for Digital and Communication Networks*. New York: IEEE Press, 1995, pp. 192-204.

## Vita

Azad Jafari Naeemi was born in Kerman, Iran in August 1972. In 1989, he was selected as one of the seven finalists in the National Iranian Physics Olympiad, and moved to Tehran to attend the special preparation program for the International Physics Olympiad. In 1994, he received the bachelor's degree in electrical engineering at Sharif University of Technology. From 1994 to 1999, he taught at the Khazra'i Institute and worked as a design engineer at Partban and KCR Companies in Tehran, Iran. In 1999, he started his graduate studies at Georgia Institute of Technology and has the privilege of working at Professor Meindl's research group. In spring 2000, Mr. Naeemi was awarded the Colonel Oscar Cleaver Prize, which recognized him as the outstanding graduate student in the school of Electrical and Computer Engineering at the Georgia Institute of Technology. Over the past three years, he has authored more than 12 papers in refereed journals and international conferences. His current research interests are in the areas of interconnect modeling and optimization, impact of inductance on GSI interconnects, far inductive noise modeling, and chip-package co-design methodologies.