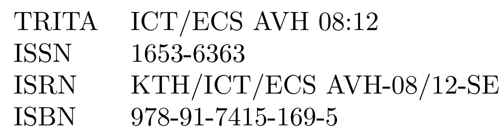
KTH Information and
Communication Technology

System Interconnection Design Trade-offs in Three-Dimensional (3-D) Integrated Circuits

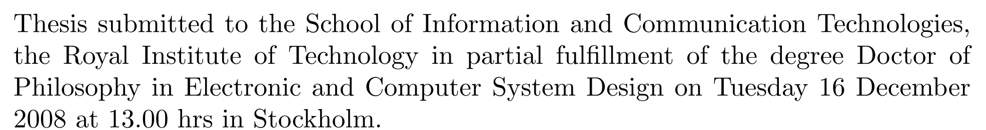
ROSHAN WEERASEKERA

Doctoral Thesis
Stockholm, Sweden 2008


TRITA ICT/ECS AVH 08:12
ISSN 1653-6363
ISRN KTH/ICT/ECS AVH-08/12-SE
ISBN 978-91-7415-169-5

© Roshan Weerasekera, December 2008

KTH School of Information and Communication Technologies
Department of Electronic, Computer, and Software Systems
ELECTRUM 229,
Kista SE-164 40,
Sweden.


Thesis submitted to the School of Information and Communication Technologies, the Royal Institute of Technology in partial fulfillment of the degree Doctor of Philosophy in Electronic and Computer System Design on Tuesday 16 December 2008 at 13.00 hrs in Stockholm.

Abstract

Continued technology scaling together with the integration of disparate technologies in a single chip means that device performance continues to outstrip interconnect and packaging capabilities, and hence there exist many difficult engineering challenges, most notably in power management, noise isolation, and intra and inter-chip communication. Significant research effort spanning many decades has been expended on traditional VLSI integration technologies, encompassing process, circuit and architectural issues to tackle these problems. Recently however, *three-dimensional* (3-D) integration has emerged as a leading contender in the challenge to meet performance, heterogeneous integration, cost, and size demands through this decade and beyond.

Through silicon via (TSV) based 3-D wafer-level integration is an emerging vertical interconnect methodology that is used to route the signal and power supply links through all chips in the stack vertically. Delay and signal integrity (SI) calculation for signal propagation through TSVs is a critical analysis step in the physical design of such systems. In order to reduce design time and mirror well established practices, it is desirable to carry this out in two stages, with the physical structures being modelled by parasitic parameters in equivalent circuits, and subsequent analysis of the equivalent circuits for the desired metric. This thesis addresses both these issues. Parasitic parameter extraction is carried out using a field solver to explore trends in typical technologies to gain an insight into the variation of resistive, capacitive and inductive parasitics including coupling effects.

A set of novel closed-form equations are proposed for TSV parasitics in terms of physical dimensions and material properties, allowing the electrical modelling of TSV bundles without the need for computationally expensive field-solvers. Suitable equivalent circuits including capacitive and inductive coupling are derived, and comparisons with field solver provided values are used to show the accuracy of the proposed parasitic parameter models for the purpose of performance and SI analysis.

The deep submicron era saw the interconnection delay rather than the gate delay become the major bottleneck in modern digital design. The nature of this problem in 3-D circuits is studied in detail in this thesis. The ubiquitous technique of repeater insertion for reducing propagation delay and signal degradation is examined for TSVs, and suitable strategies and analysis techniques are proposed. Further, a minimal power smart repeater suitable for global on-chip interconnects, which has the potential to reduce power consumption by as much as 20% with respect to a traditional inverter is proposed. A modeling and analysis methodology is also proposed, that makes the smart repeater easier to amalgamate in CAD flows at different levels of hierarchy from initial signal planning to detailed place and route when compared to alternatives proposed in the literature.

Finally, the topic of system-level performance estimation for massively integrated systems is discussed. As designers are presented with an extra spatial dimension in 3-D integration, the complexity of the layout and the architectural trade-offs also increase. Therefore, to obtain a true improvement in performance, a very careful analysis using detailed models at different hierarchical levels is crucial. This thesis presents a cohesive analysis of the technological, cost, and performance trade-offs for digital and mixed-mode systems, outlining the choices available at different points in the design and their ramifications.

Acknowledgements

This thesis comes at the end of four formative years, which has contributed enormously to my professional and personal growth, spent at the Department of Electronic, Computer, and Software Systems (ECS) of the School of Information and Communication technologies (ICT), the Royal Institute of Technology (KTH), Stockholm, Sweden. Here are the people to whom I want to express my deepest gratitude for what they have facilitated to this work.

First and foremost, I must thank my supervisors Prof. Hannu Tenhunen and Prof. Li-Rong Zheng for their guidance and support throughout my PhD studies. Their constant enthusiasm for learning, and finding new research avenues is highly appreciated, and both of them have a knack for explaining the most intricate theories with such a simple approach.

I owe a huge debt of thanks to Dinesh - Dr. Dinesh Pamunuwa of Lancaster University - for being a collaborator on almost all my publications and giving the necessary comments and criticism when I desperately needed them. Not only that I must greatly remember his endeavour on recruiting me as a research associate in Lancaster University under the EU-FP7 funded ELITE project. This stay was quite successful and effective on making series of new publications and also it opened up new research avenues.

I greatly appreciate the financial support for my research from SIDA under the auspices of the research capacity building project at the Department of Electrical and Electronic Engineering, University of Peradeniya (UPDN), Sri Lanka, without which this thesis would not have been possible. My thanks are also due to its coordinators, Dr. Sanath Alahakoon of UPDN and Tekn. Lic. Mats Leksell of KTH, not only for their exceptional administration capabilities but also for their humanism. Also, I must thank the SIDA/SAREC PhD student colleagues from UPDN in KTH and Chalmers University of Technology, Sweden.

I must thank the financial support of European Union research funding under grant FP7-ICT-215030 (ELITE) of the 7th framework programme for my internship at Lancaster University, UK. Also, the support received from the members in ELITE collaborative partner companies Qimonda Gmbh in Germany, LETI in France, and Hyperstone in Germany is highly regarded.

Many thanks are due to colleagues at ECS too numerous to mention here, especially in the iPack center, ESD, SAM and RaMSIS groups, for their friendship, support on learning various tools and help on understanding some theoretical aspects.

Also, a special word of thanks should go to the administrative staff in ECS and the IT support group in ICT.

Thanks are also due to the staff of the Centre for Microsystems Engineering at Lancaster University, UK, especially the humanism, great support, and friendship of Prof. Andrew Richardson, the Center Director.

The friendship of Sri Lankan postgraduate students community in Stockholm, and some other colleagues has made the time spent in Stockholm more pleasant.

I express my deepest gratitude to my parents for their restless efforts on growing and taking care of me through thick and thin; and to my brother and the sister, for being close associates and mentors on various other occasions.

I compliment most warmly my loving and caring wife, Indu, who had to endure and share the pain and the pleasures of my arduous professional life. Last but not least, I must thank our son, Savinu, who missed me sometimes from dawn to dusk for not being available to play with.

Roshan Weerasekera
December, 2008
Stockholm

List of Publications

1. **Roshan Weerasekera**, Dinesh Pamunuwa, Hannu Tenhunen, and Li-Rong Zheng, "Modelling Through-Silicon-Vias in 3D-ICs," *IET Electronic Letters*, September, 2008, Under Review.
2. **Roshan Weerasekera**, Dinesh Pamunuwa, Matt Grange, Hannu Tenhunen, and Li-Rong Zheng, "Parasitic Parameter Estimation and Electrical Modelling of Through-Silicon Vias in 3-D ICs," *IEEE International Symposium on Circuits and Systems*, Under Review.
3. Matt Grange, **Roshan Weerasekera**, Dinesh Pamunuwa, and Hannu Tenhunen, "Exploration of Through Silicon Via Interconnect Parasitics for 3-Dimensional Integrated Circuits," *IEEE International Symposium on Circuits and Systems*, Under Review.
4. Botao Shao, **Roshan Weerasekera**, Abraham Tareke Woldegiorgis, Li-Rong Zheng, Ran Liu, and Werner Zapka, "High Frequency Characterization and Modeling of Inkjet Printed Coplanar Strips on Flexible Substrate," *in Proceedings of the 2nd Electronics System-Integration Technology Conference*, London, UK, September, 2008, pp. 695-699.
5. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng, and Hannu Tenhunen, "2-D and 3-D Integration of Heterogeneous Electronic Systems under Cost, Performance and Technological Constraints," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, September, 2008, Under Review.
6. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng, and Hannu Tenhunen, "Minimal-Power, Delay-Balanced Smart Repeaters for Global Interconnects in the Nanometer Regime," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 5, pp. 589-593, May, 2008.
7. Botao Shao, **Roshan Weerasekera**, Li-Rong Zheng, Ran Liu, Werner Zapka, and Peter Lindberg, "High Frequency Characterization of Inkjet Printed Coplanar Waveguides," *12th IEEE Workshop on Signal Propagation on Interconnects*, May, 2008, pp.1-4.
8. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa, and Hannu Tenhunen, "Extending Systems-on-Chip to the Third Dimension: Performance, Cost and Technological Tradeoffs," *in Proceedings of the IEEE/ACM international conference on Computer-aided design*, IEEE Press, November, 2007, pp. 212-219.
9. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa, and Hannu Tenhunen, "Early selection of system implementation choice among SoC, SoP and 3-D Integration," *IEEE International System-on-Chip Conference*, Sep-

tember, 2007, pp.187-190.

10. Botao Shao, **Roshan Weerasekera**, Abraham Tareke Woldegiorgis, Li-Rong Zheng, Ran Liu, Werner Zapka, and Peter Lindberg, "Electrical Characterization of Inkjet Printed Interconnections on Flexible Substrates", *IEEE CPMT Symposium on Green Electronics*, Göteborg, Sweden, 2007.
11. **Roshan Weerasekera**, Abraham Tareke Woldegiorgis, Botao Shao, Saul Rodriguez Duenas, Li-Rong Zheng, Peter Lindberg, Werner Zapka, and Hannu Tenhunen, "Electrical Characterization of Ink-Jet Printed Interconnects on Plastic for Low-Cost RFID," *IEEE International Conference on Industrial and Information Systems*, Peradeniya, Sri Lanka, 2007.
12. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng, and Hannu Tenhunen, "Delay-Balanced Smart-Repeaters for on-chip Global Signaling," *20th International Conference on VLSI Design held jointly with 6th International Conference on Embedded Systems*, 2007, pp. 308-313.
13. Jian Liu, **Roshan Weerasekera**, Li-Rong Zheng, and Hannu Tenhunen, "Exploration of Autonomous Error-Tolerant (AET) Celluar Networks in System-on-a Package (SoP) for Future Nanoscale Electronic Systems," *IEEE International Conference on Industrial and Information Systems*, Peradeniya, Sri Lanka, 2006.
14. Dinesh Pamunuwa and **Roshan Weerasekera**, "Nanoelectronics: from novelty toys to functional Devices - an integration perspective," *IEEE International Conference on Industrial and Information Systems*, Peradeniya, Sri Lanka, 2006, Invited Paper.
15. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng, and Hannu Tenhunen, "Minimum-Power, Delay-Balanced Drivers for interconnects in the Nanometer Regime," *Proceedings of the international workshop on System-level interconnect prediction*, German, March, 2006, pp. 113-120.
16. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa, and Hannu Tenhunen, "Switching sensitive interconnect Driver to Combat Dynamic Delay in on-Chip Buses," *Lecture Notes in Computer Science (Proceedings of PATMOS)*, vol. LNCS 3728, pp. 277-285, 2005.
17. **Roshan Weerasekera**, Jian Liu, Li-Rong Zheng, and Hannu Tenhunen, "A Nanocore/ CMOS Hybrid System-on-Package (SoP) Architecture for Future Nanoelectronic Systems," *Conference on High Density Microsystem Design and Packaging and Component Failure Analysis*, June, 2005, pp.1-4.
18. Jian Liu, **Roshan Weerasekera**, Li-Rong Zheng, and Hannu Tenhunen. "Nanocore/ CMOS hybrid system-on-package(sop) architecture for autonomous error-tolerant (aet) cellular array network," *In 5th IEEE Conference on Nanotechnology*, Nagoya, Japan, 2005.
19. Jian Liu, **Roshan Weerasekera**, Li-Rong Zheng, and Hannu Tenhunen, "Nano scale autonomous error-tolerant (aet) cellular network," *In Technical Proceedings of the 2005 Nanotechnology Conference and Trade Show*, California, USA, May, 2005, pp. 748-751.
20. **Roshan Weerasekera**, Jian Liu, Li-Rong Zheng, and Hannu Tenhunen. "A nanocore/cmos hybrid system-on-package (sop) architecture for future nanoelectronic systems," *In Technical Proceedings of the 2005 Nanotechnology Conference and Trade Show*, California, USA, May, 2005, pp. 157-160.
21. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa, and Hannu Ten-

hunen, "Crosstalk Immune Interconnect Driver Design," *in the Proceedings of International Symposium System-on-Chip*, November, 2004, pp. 139 - 142.

Table of Contents

Acknowledgements **iii**

List of Publications **v**

Table of Contents **ix**

1 Introduction **1**
1.1 Evolution of Microelectronic Systems 1
1.1.1 Microelectronic 2
1.1.2 Packaging 3
1.2 Trends in Further Miniaturization 4
1.2.1 Device and Interconnect Scaling 4
1.2.2 Integrated Circuit Packaging 6
1.2.3 Dealing with Complexity: System-Level Integration 7
1.3 Interconnect Challenges and Strategic Solutions 9
1.3.1 Emerging Solutions - Alternatives for Cu/Low-κ 11
1.4 Scope of Thesis and Author's Contribution 12
1.4.1 Smart Repeaters for Interconnections in Nanometer Technologies 13
1.4.2 Cost and Performance Trade-offs for 2-D and 3-D Mixed-Signal ICs 14
1.4.3 Electrical Modelling of Through-Silicon Vias in 3-D Integrated Circuits 15
1.5 Thesis Organization 16

2 Interconnect Modelling and Analysis **17**
2.1 Introduction 17
2.1.1 Electromagnetic View of Interconnects 19
2.2 Parasitic Estimation and Extraction 20
2.2.1 Resistance 21
2.2.2 Inductance 25
2.2.3 Capacitance 29
2.2.4 Conduction 32
2.3 Electrical Level Modelling 33
2.3.1 Choosing a Wire Model 35
2.4 Interconnect Timing Analysis 39
2.4.1 Interconnect Driver Modelling 39

2.4.2 Interconnect Delay Modelling 43
2.4.3 Noise-on-Delay Effect 45
2.5 Interconnect Energy Dissipation Analysis 48
2.5.1 Switching Energy 48
2.5.2 Short Circuit Energy 49
2.5.3 Leakage or Static Energy 50
2.6 Summary 50

3 Electrical Modelling of Through-Silicon Vias 53
3.1 Introduction 53
3.2 TSV Specification and Physical Modelling 54
3.3 Trends in Parasitic Parameter Values 55
3.3.1 Isolated TSV 55
3.3.2 Two Parallel TSVs 58
3.3.3 TSV Bundle 62
3.4 Compact Modelling of TSV Parameters 67
3.4.1 RLC Extraction of an Isolated TSV 67
3.4.2 RLC Extraction of a TSV Bundle 70
3.5 Summary 77

4 Signalling Techniques for On-Chip Global Interconnects 79
4.1 Introduction 79
4.2 Design Methodologies for On-Chip Interconnects 80
4.2.1 Layout and Routing Level 80
4.2.2 Circuit Level 81
4.2.3 Architectural and System Level 83
4.3 SMART Driver Circuit 84
4.3.1 Limitations in Existing Driver circuits 85
4.3.2 The Concept 85
4.3.3 Circuit Realization 86
4.3.4 Noise Resiliency of the SMART Driver 90
4.3.5 Energy Saving of the SMART Driver 91
4.3.6 Design Methodology 92
4.3.7 Energy and Delay Model Validation 96
4.3.8 Impact of Technology Scaling 98
4.4 Vertical Signal Transmission Methodologies 99
4.4.1 Signal Transmission Characteristics of TSV interconnects . 100
4.4.2 Signalling Link Design for Layer-to-Layer Communication . 105
4.5 Summary 107

5 IC Cost Modelling at the System Conceptual Level 109
5.1 Introduction 109
5.2 Cost Analysis and Modelling 109
5.2.1 Rent's Rule 110
5.2.2 Wire-Length Modelling 112
5.3 Bare Die/Packaged Chip Cost Analysis 113
5.3.1 Packing Density and Area 114
5.3.2 Chip Size 117

5.3.3 Die Yield Analysis 118
5.3.4 Chip Cost Model 124
5.4 Board- or Package-Level Model 128
5.4.1 Number of Pins per Chip 128
5.4.2 Module Level Average Interconnection Length 128
5.4.3 Chip Footprint 129
5.4.4 Yield and Cost Analysis 130
5.5 Summary 132

6 Heterogeneous System-on-Chip Integration: 2-D or 3-D ? 133
6.1 Introduction 133
6.2 Three-Dimensional Integration 134
6.2.1 Benefits of 3-D Integration over 2-D Planar 134
6.2.2 Challenges for 3-D integration 137
6.2.3 Three-Dimensional Integration Options 139
6.3 Early Estimation of Cost and Performance 141
6.3.1 Models for Trade-off Analysis 142
6.4 Tradeoff Analysis for SoC, SoP and 3-D Implementations 147
6.4.1 Monolithic SoC 150
6.4.2 2D-SoP 150
6.4.3 3D-SiP 151
6.4.4 3D-WLI 152
6.5 Discussion 153
6.6 Summary 156

7 Conclusions 159
7.1 Summary 159
7.2 Future Work 160

References 163

1

Introduction

'I believe the best is yet to come'

Jack S. Kilby [1]

1.1 Evolution of Microelectronic Systems

Over the last fifty years, the synergistic interaction between solid-state physics, electrical engineering, and materials science has fueled the growth of the solid-state circuits industry from infancy to become one of the largest industries in the world. The technologies behind almost all modern electronic products, which touch every aspect of human life, from computers to communication equipments, toys, food, medical technology and the automobile industry are all based on microelectronic devices and packaging technologies.

The miniaturization of electronic systems goes back to the days of World War II. During this time, there was a increased demand for small size, light weight, low power, and reliable military electronic systems because of the increased use of these systems and ease in carrying them especially in aircrafts or for infantry personnel who carried equipment in combat. Due to this demand, electronic systems has moved from room-sized products toward hand-held devices with considerably greater computational horsepower; the functions that a today's chip performs are essentially no different from those earlier products. Even for today's applications the performance metrics remain the same, but cost the constraint has come into picture as a major design requirement especially in consumer electronic systems.

Consequently the process of device miniaturization evolved from few micrometers to nanometers today, and circuit complexity has advanced from Small-Scale Integration (SSI) in 1960s, to Medium/Large Scale Integration (MSI/LSI) in 1970s, to Very Large Scale Integration (VLSI) 1990s, and to Giga-Scale Integration (GSI) in 2000s. This tighter integration continues at a break-neck speed toward a trillion transistors per chip, Tera-Scale Integration (TSI) era, in 2020s. With the passage of time, not only digital devices and memory, but also analog/mixed-signal blocks,

MEMS based sensors, biological functions are also being integrated on the same die or package to build a complete system. In reconciling with feature size miniaturization and technology divergence, and achieving smaller, faster, and cheaper products, there exists many unprecedented difficult technological challenges at different hierarchy levels in electronic system design process [*2*, *3*].

1.1.1 Microelectronic

The electronics industry was launched by the invention of the Vacuum Tube, and its basic usage was to amplify signals for radio and other audio devices. But, Vacuum Tubes steadily spread into other devices, and the first tube was used as a switch in calculating machines in 1939. The ENIAC of 1947, intended for computing artillery firing tables, was the first electronic computer developed by John W. Mauchly and J. Presper Eckert, Jr. During that time ENIAC was the fastest computer, but it contained around 18,000 vacuum tubes which failed at the rate of one in every 7 minutes, occupied 16,200 cubic-feet, weighted 60,000 pounds, and consumed 174 kW (= 233 horsepower) of electricity. The reliability problems with the vacuum tubes and the excessive power consumption made the implementation of larger engines economically and practically infeasible. These problems were visible to many in the industry and hence momentum on research into miniaturization of electronic systems grew. One can clearly see that there are three inventions in the 20th century which has been greatly instrumental in the evolution of the Integrated Circuit (IC) and thus in Information Technology (IT): *the invention of the transistor, monolithic concept and the planar process.*

While searching for switches and amplifiers to replace mechanical relays and the valves that so troubled ENIAC, J. Bardeen, W.H. Brattain and W. Shockley of Bell labs, USA, invented the first point contact transistor. Thereafter, in 1950, Shockley invented a new device called a bipolar junction transistor (BJT), which was more reliable, easier and cheaper to build than the point contact devices.

On July 24, 1958, Jack Kilby of Texas Instruments scribed in his note book what has come to be known as the idea of monolithic circuits, that circuit elements such as resistors, capacitors, distributed capacitors and transistors - if all made of

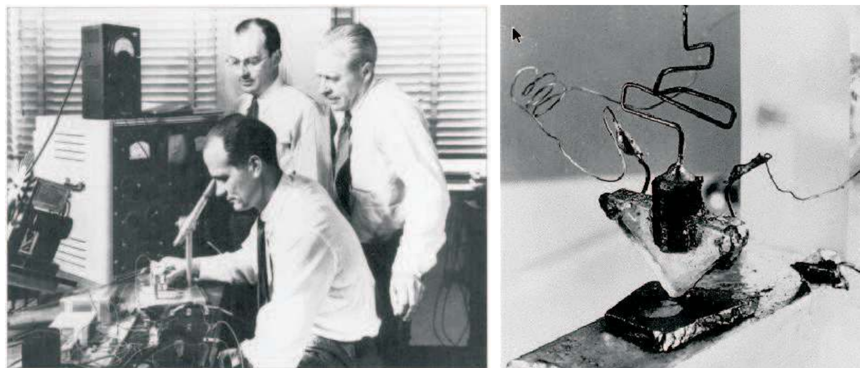


Figure 1.1: *First Point Contact Transistor and it's co-inventors.*

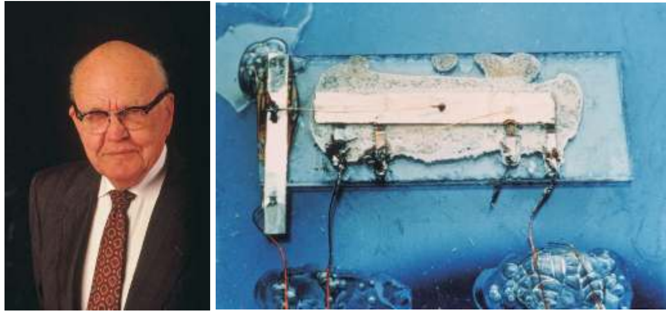


Figure 1.2: *Jack Kilby and the first IC (source: Texas instruments).*

the same material - could be included in a single chip. On September 12, 1958, he succeeded in fabricating a phase shift oscillator on a single piece of semiconductor, which earned him the Nobel Price for Physics in 2000, for the invention which revolutionized the modern electronic industry [*1*].

In 1959, the swiss physicist Jean Hoerni at Fairchild Semiconductor invented the planar process, in which optical lithographic techniques were used to diffuse the base into the collector and then diffuse the emitter into the base to produced a transistor. This process consists of three basic steps: Oxidization, Photolithography, and Etching. One of Hoerni's colleagues, Robert Noyce, invented a technique for growing an insulating layer of silicon dioxide over the transistor, leaving small areas over the base and emitter exposed and diffusing thin layers of aluminium into these areas to create wires, which led directly to modern ICs.

Gordon E. Moore, a co-founder of Intel, while working at Fairchild Semiconductor in 1964, foresaw that the number of components, as well as the functionality, that could be integrated on a single die would grow exponentially with time [*4*]. He also observed that the microprocessor performance (clock frequency $\times$ instructions per clock) also doubles every 1.5 to 2 years. As years went by, it turned out that Moore was right, and it became less of a prediction and more of a self-fulfilling prophecy known as Moore's Law. It has been a goal and key performance indicator of successful leading-edge semiconductor products and companies for the past four decades.

1.1.2 Packaging

Packaging is an essential and integral part of semiconductor products. According to [*5*], packaging serves major functions at the IC or device level, and at the system-level. At the IC or device level, it serves four purposes: interconnection of electrical signals, mechanical and environmental protection of circuits, distribution of power (i.e., electrical energy), and dissipation of heat generated by the semiconductor devices [*5*, *6*, *7*]. The approach for packaging must be selected based on the application, because system requirements for computer, handheld, automotive, medical, and bioapplications are all different. For example, miniaturization is more important for handheld devices than automotive applications. However, regardless of the

application, the functionality and complexity of the IC have been increasing and driving the development of microelectronic packaging over the decades [5, 6, 7].

The microelectronics packaging technology started with the discovery of the transistor in the late 1940s, and has evolved to serve the increasing complexity and performance of the IC with the passage of time. Early transistors were housed in plastic packages providing just the protection for the device. Once the military became interested in highly reliable applications, the need for hermetic packages were incorporated to prevent transistor gain degradation and junction leakage current due to contamination and moisture. This led to the development of the metal Transistor Outline (TO) packages [7]. With the development of silicon planar technology, electronic packages were developed to fulfill the requirements of high performance ICs containing large numbers of devices as it affects the operating frequency, power, complexity, reliability, and cost of semiconductor products.

In much of the literature [5, 7] the evolution of microelectronic packaging is described starting with the Dual-In-line Package (DIP) of the 1970s. The DIP contained a single chip connected with wire-bonds to interconnections on the package, and the connections from package to system board were made with pins located on both sides of the package. As the Input/Output (I/O) count in chips increased with the passage of time, more connections were needed. Then, in the 1980s, the whole package area was filled with pins forming a Pin-Grid Array (PGA) package. Also at the same time, a Surface Mount Technology (SMT) was adopted for electronic production. SMT facilitated the assembly process, and the Quad-Flat Package (QFP) was introduced. Later, in the 1990s, area array SMT contacts provided by the Ball Grid Array (BGA) package started a new era in microelectronic packaging enabling much smaller package size. The evolution of packaging technology has increased the chip area to package area ratio, resulting in a much smaller, thinner, and lighter package with an increased number of I/O pins. This again has led to the Chip-Scale Package (CSP), which, by definition, is a package with an area of less than 1.2 times the area of the chip [8] and a pitch of a few hundred micrometers for the package I/O pads. At the beginning of the millennium, Wafer-Level Packages (WLP), Three-Dimensional (3-D) integration, stacked packages, and System-in-Packages (SiP) were adopted in the packaging industry to allow even higher packaging density. The latest packaging concept is the System-on-Package (SoP), which involves the integration of a whole system including passive components into a single package, leading to miniaturized systems [9].

1.2 Trends in Further Miniaturization

1.2.1 Device and Interconnect Scaling

Historical facts reveal that the reduction of feature size used to fabricate ICs continues at a rate of 0.7 per year [2] in compliance with Moore's Law [4]. It is quite interesting to note that the smallest dimension on a wafer has been reduced from several times the size of a red blood cell (6-8 μm) to that of the common cold virus (20 nm). Table 1.1 summarizes the effects of technology scaling to the transistor and interconnect performance metrics. As can be seen from the table, technology scaling basically achieves three goals [10]: doubles the gate density, reduces the energy per switching by 65%, and decreases the gate delay by 30%.

To meet the technology goals, according to the simple interconnect scaling theory, interconnect cross-sectional dimensions are scaled at the same rate as gates' dimensions. As a result, the resistance of a unit length wire increases at the rate of 104% per year. In general, die area should decrease by 50% per year in successive technologies, but new designs integrate more transistors and functionality per chip, resulting in die area increment instead, and die size has been increasing at 13% per year. Consequently global interconnect length increases at a rate of 6% per year, and it's RC time constant increases by approximately 130% per year! Delay of wires has dominated that of gates, and the ratio of wire delay to gate delay increases at a rate of 300% annually. Therefore, designers have had to pay attention to the interconnect delay bottleneck.

As transistor count per unit area increases, the current required per unit area increases by 43%. At the same time, the wire resistance rises rapidly, increasing IR drop over wires. The worst effect of this is, with decreasing voltage, the tolerable IR decreases proportionally. To account for total interconnect length, and to optimize signal and power distribution networks effectively across the chip, manufacturers have added more interconnection layers, which adds to the design complexity.

When the linear dimensions scale by a factor, the voltage must also be scaled by the same factor to keep the electric field within a certain limit. However, a higher supply voltage is necessary to provide a performance boost, because transistor drive current is proportional to gate over drive $(V_{dd} - V_t)^n$, where n is in the range of $1 - 2$. Recently, the International Technology Roadmap for Semiconductors

	Parameter	Symbol	Scaling/year
DEVICE	Dimensions	W, L, t_{ox}	0.7
	Supply Voltage	V	0.7
	Drain Current	$I_{DS} \Rightarrow \frac{W}{L}\frac{1}{t_{ox}}V^2$	0.7
	Gate Capacitance	$c_g \Rightarrow \frac{WL}{t_{ox}}$	0.7
	On Resistance	$R_{tr} \Rightarrow \frac{V}{I_{DS}}$	1
	Intrinsic Delay	$\tau \Rightarrow R_{tr}c_g$	0.7
	Power Dissipation	$P \Rightarrow VI_{DS}$	0.49
	Switching Energy	$E \Rightarrow P\tau$	0.34
	Gate Density	$n \Rightarrow \frac{1}{WL}$	2.04
CHIP	Chip Area	A	1.13
	Chip Edge	$y \Rightarrow \sqrt{A}$	1.06
	Current per unit area	$I_a \Rightarrow \frac{I_{DS}}{WL}$	1.43
	Total Chip Current	$I_a y^2$	1.61
WIRE	Cross-sectional Dimensions	w, h, s, t	0.7
	Resistance per unit length	$r \Rightarrow 1/wt$	2.04
	Capacitance per unit length	$c \Rightarrow w/h$	1
	RC Constant	rc	2.04
	RC delay/Gate Delay	rc/τ	2.9
	Local Interconnection Length	L_l	0.7
	Local Interconnection RC delay	rcL_l^2	1
	Global Interconnection Length	$L_g \Rightarrow y$	1.06
	Global Interconnection RC delay	rcL_g^2	2.29

Table 1.1: *Constant Field Scaling of Device and Wire Properties [2, 6, 11].*

(ITRS) predicts a slow down for scaling maximum supply voltage in the nanometer regime due to the inability to further reduce threshold voltage due to leakage power consumption and process variations. Moreover, lowering supply voltage on one hand reduces the dynamic power consumption, and on the other hand naturally increases operation current, which in turn requires thicker metal layers in order to reduce IR drop. Total chip current increases at an annual rate of 61%, thus creating challenges in power distribution system design and in package level thermal management [2].

Transistor switching energy is reducing at an annual rate of 66%, and it is reaching the minimum energy that must be transfered in a single interconnect's binary transition, which is $E_s = kTln(2)$, where k is Boltsmann's constant and T is absolute temperature [3, 12]. In essence, in the regime of tera-scale integration innovative and radical changes in logic devices are essential to overcome the challenges that hinder the performance and reliability of electronic systems.

1.2.2 Integrated Circuit Packaging

As narrated, the continuous reduction of cost per function has been the key to exponential growth of electronic industry. But, the cost of assembling and packaging ICs has not kept pace with the cost reduction in wafer fabrication; packaging cost exceeds the wafer production cost. Other driving forces for the evolution of electronic packaging are performance, size and volume, time-to-market and reliability.

As the technology advances towards nanometer generations, density and performance of individual chips are continually enhanced. Unfortunately, today, not all of these merits can be translated to the system level due to the problem of electronic packaging, which has presented a bottleneck for increasing system speed, reducing power, and shrinking system size [5, 7].

When the complexity of chips expand, the number of I/O pins rises exponentially according to Rents rule [13], which consequently increases wiring demands for system level interconnections. Thus, in order to provide enough wires for system interconnections at a reduced substrate size, interconnect pitch has to be reduced providing stringent limits on signal integrity at the package-level.

Technology	Wire-bond	Solder bump	Adhesive bumps	Micro-via
Resistance ($m\Omega$)	$30-100$	1.0-3.0	15-30	0.2-1.0
Inductance (nH)	1.0-3.0	0.05-0.1	0.05-0.1	0.01-0.3
Capacitance (pF)	0.01-0.05	0.002-0.01	0.002-0.01	0.0002-0.001
Discontinuities	Severe	Moderate	Moderate	None

Table 1.2: *Typical values of parasitic components in different chip interconnection technologies[8].*

When chip speed is higher than several hundred MHz, the package exhibits very large parasitic effects. For example, in today's VLSI chips, the chip I/O pads are still quite large, which requires very large buffers and off-chip drivers for off-chip communications. In addition, the package itself and on-board interconnects have much larger dimensions than that of the on-chip's. They are hence large loads for the off-chip drivers. Besides the higher power consumption and larger chip size for these off-chip drivers, system performance is severely degraded. With higher

operating frequencies and signal rise time shorter than two and a half times the time-of-flight, transmission line effects become significant. Consequently, preservation of signal integrity and timing becomes a difficult challenge as signals move from chip to chip within the system. Table 1.2 summarizes typical values of chip to package interconnection parasitic parameters. Also, it is essential to minimize impedance discontinuities at chip-to-package and package-to-board interconnection junctions and reduce cross-talk noise between adjacent lines.

The dearth of I/O pins also places more restrictions on the power supply network design, as more gates per pin means longer current paths and increased current in each path, requiring more on-chip bypass capacitance [*5*, *11*]. With the area array bonding techniques where the pins are placed over the entire surface of the package, the number of pins grow with the square of the chip dimension. Also with the elimination of bond wire inductance, and the resistive drop over the on-chip power supply grid is much less, as the current paths are shorter. This eases the requirements on on-chip bypass capacitance, but there still exists a need for innovative off-chip signalling schemes. Multi-chip packaging techniques, where several chips exist in one package in a vertical stack (SiP), and the inter-chip links are implemented locally, are another option to System-on-Chip (SoC).

As per most of the predictions if the current technology scaling continues without particular low power design techniques, the power density of future microprocessors will be a main limiting factor. With the dimension of chip and package scaling down and clock frequency scaling up, electronic products have experienced a dramatic increase in power density. The task of dissipating heat from ICs while maintaining acceptable junction temperature has been a significant challenge for semiconductor and system manufactures. With low power circuit and system architectures, it is projected that in 2013, power dissipation of a high-performance CMOS chip will be around 0.64 W/mm^2 while its area is around 750 mm^2, and the maximum allowable junction temperature is about 90 ^oC [*2*]. This will result in a thermal resistance budget at 0.19 $^oC/W$ for the whole module, indicating a great challenge for heat removal in high performance products even with power efficiency circuit and system architectures.

1.2.3 Dealing with Complexity: System-Level Integration

As the functionality and the number of gates in a chip has increased, the chip complexity has also increased. Hence, a modular based approach is used. Today, for example, an IC performs very different functionalities and uses diverse implementation styles such as processor cores, DSP blocks, FPGA blocks, analog/RF circuits and memory, and it is designed from blocks of such interconnected resources - the overall system is designed at a higher abstraction level. Usually such building blocks can be shared and also re-uses as Intellectual Property (IP) blocks, which further improves the productivity and reduces the time-to-market. This methodology has been termed as *System-on-a-Chip*.

Many of today's systems consists of complex SoCs with embedded processors, significant amounts of memory and FPGAs, but they do not provide the total system solution for real world systems. Such electronic systems digital and storage blocks coexists with many other functional devices such as analog/RF, passive components, sensors, and biological functions. These sensors and biological functions

can be non-CMOS and non-silicon technologies, with different design and implementation styles. In integrating such disparate technologies in to a single chip, designers are confronted with many technical and economic barriers [2, 14], namely huge initial investment for masks and their development, process dependent memory blocks, high precision analog blocks and the management of substrate coupled noise, and process incompatibility with non-Si materials and/or MEMS. The mask count increases as much if different types of technologies merge together to form a single die. For example, dozens of smaller chips with different functionalities are interconnected on a substrate using chip-to-chip interconnections, and packaged as a single module. System level integration methodologies are preferred to overcome most of the above mentioned barriers, reduce time-to-market, and offers greater flexibility than single chip solutions.

Multi-chip modules (MCMs) were introduced by IBM in 1970s to enable high-performance systems and in such a system dozens of smaller chips have been interconnected. The MCM technology allows the chips to be spaced more tightly with less volume and weight than individually packaged ICs. There are three major variations of MCM implementations: MCM-D, a multilayer, thin-film structures on semiconductor or ceramic base layers, with deposited metal conductors and dielectrics; MCM-C, a thick-film or co-fired ceramic technology; and MCM-L, organic laminated multilayer boards.

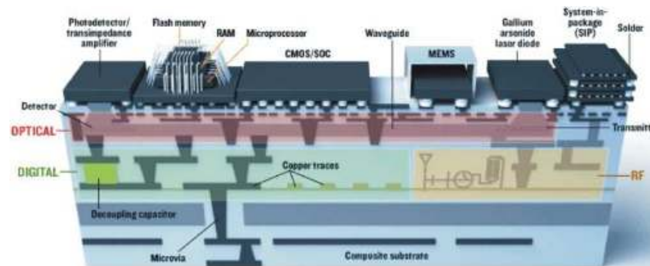


Figure 1.3: *System-on-Package, latest packaging concept, is involved in integrating a whole system into a single package [15].*

Moreover, due to ever increasing demands for low cost, smaller chips with more functionality, and smaller time-to-market for portable systems, vertical integration found to be an attractive option. Interconnecting bare or packaged chips in the vertical dimension, known as three-dimensional integration.

By contrast, stacking packaged dies sometimes known as Package-on-Package (PoP) - has its own advantages, including the ability to integrate chips from multiple suppliers and different IC technologies, such as analog, digital, mixed-signal, RF, and optoelectronic. In addition, packaged dies can be tested and burned in before being stacked. Stacked packaging and wafer-level packaging methods are both rising in popularity. After 3-D designs, the next highest efficiency can be achieved by wafer-level packaging, especially CSP designs. These provide a footprint that is just barely larger than the size of the die.

Multiple-die packages must address key logistics issues, such as being able to accommodate incompatible die shrinks; simplify management of multiple IC vendors; enable package-level test and burn-in; enable the combination of high- and low-yield devices; contribute to product quality and reliability; maximize configuration flexibility and minimize time to market; and risk, because time is our most precious commodity.

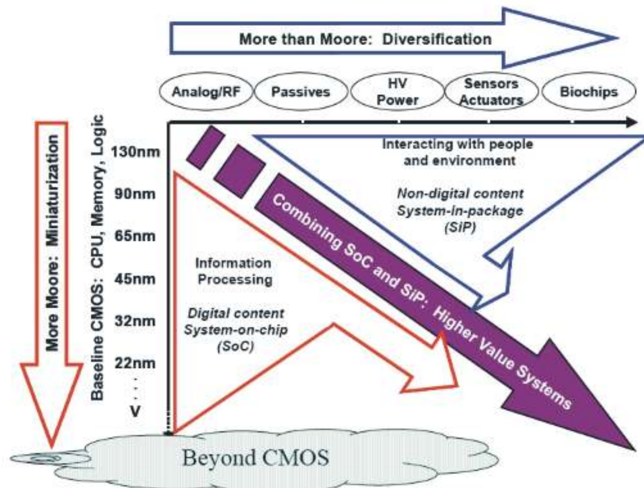


Figure 1.4: *Moore's Law and More - all the functions in a Electronic System does not scale with Moores Law [2].*

1.3 Interconnect Challenges and Strategic Solutions

Continued transistor scaling will not be as straightforward in the future as it has been in the past because fundamental material and process limits are rapidly being approached. Meindel *et.al.* in [*3*] derived five key fundamental limits for Tera-scale integration from thermodynamics, quantum mechanics, and electromagnetism. These limits are independent of material, device structure, circuit configuration, or system architecture.

The problem of interconnect delay and the possible solution domain can be explained simply by the RC time constant of a wire [*16*, *17*], which is:

$$\tau \quad = \quad rcL^2 = \frac{\rho}{wt}\frac{\epsilon w}{t} = [\rho\epsilon]\left[\frac{1}{ht}\right][L^2]. \tag{1.1}$$

As (1.1) expresses, the growing interconnect delay issue can be addressed by material processes, reverse scaling (reduce the width while maintaining the same aspect ratio) and reducing interconnect length.

Copper (Cu) interconnects perform better than aluminium (Al) because resistivity of Cu is approximately 40% lower than that of aluminium [*18*]. Also, Cu has a higher resistance to electromigration effects. Electromigration lifetime of Cu is

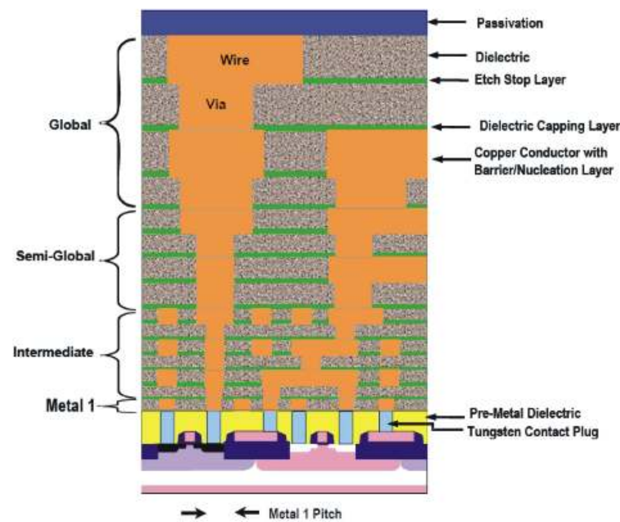


Figure 1.5: *Cross-section of Hierarchical Wiring with steadily increasing pitch and thickness at each conductor levels to alleviate the impact of interconnect delay on performance [2].*

about 100 times longer than Al at the same current density. However, a drawback with Cu as interconnect material is that Cu readily diffuses in most dielectrics and acts as a recombination center in Silicon. Hence, a metallic (such as Ta, TaN) or dielectric (such as SiN, SiC) diffusion barrier is generally needed to encapsulate a Cu line to prevent electrical leakage and degradation of transistor performance. These barrier films have much higher resistivity than Cu and approximately 20% of the wire width is consumed by the film. Also, the cross-sectional dimesion of the wire is close to the electron mean free path hence the electron scattering effect at the conductor surface as well as the grain boundaries result in increased resistance.

In order to reduce the rate of resistance increase, the thickness is increased to achieve a larger cross-section, which leads to tall and thin wires. Contemporary technologies use wires with an Aspect Ratio (AR) approximately equal to $2-2.5$. By doing so, the annual increment of RC constant can be maintained at a constant rate, and the delay over a global wire increases at a rate of 13% per year. Nonetheless, the continuous increment of aspect ratio will not bring similar benefits because the reduction in resistance will be offset by the dominance of inter-wire capacitance to the total wire capacitance. The impact of this effect is a rise in coupling noise, which is in two forms: cross-talk and signal integrity, and dynamic delay. Since the reduction in packing density is not an option, the only way of reducing capacitance is to use low permitivity (low-κ) dielectrics instead of SiO_2 whose dielectric constant is about 3.9. The low-κ and porous SiO_2 currently being proposed are not robust enough to withstand assembly and packaging process such as wire bonding. IBM, who introduced the Cu/low-κ interconnect technology has announced recently their new strategy to use air gaps as dielectric for the power hungry advanced technology nodes.

As is evident from (1.1), another key technique in reducing interconnect delay is

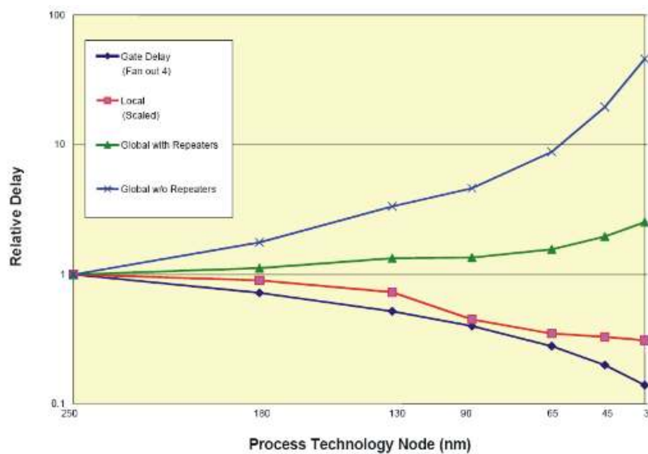


Figure 1.6: *Delay for Metal 1 and Global Wiring versus Feature Size. From 180nm down to 15 nm, the delay of scaled wires increases by approximately 10 ps while that of fixed length wires increases by about 2000 ps. If these wires are modified with repeaters, these delays reduce to approximately 3 ps for scaled wires and 40 ps for fixed length wires. [2].*

to reduce its length. The most simple way to do is to insert repeaters by breaking the wire into several sections [*19*, *20*]. Usually these wire sections are highly capacitive and high strength repeaters are needed. The adverse effect of this is increased power consumption; it has been estimated that over 50% of the power in a high performance microprocessor is dissipated by repeaters charging and discharging interconnects [*21*, *22*, *23*]. Further, over 90% of this power is concentrated in only 10% of the interconnects; *i.e.* those which are classed as global and run for a significant fraction of the die length. The length of global interconnects can effectively be reduced by integrating blocks in a stack and hence power consumption can be reduced significantly.

1.3.1 Emerging Solutions - Alternatives for Cu/Low-κ

For some of the manufacturing challenges and limitations in Cu/low-κ interconnect systems, the strategic solutions for the technology node beyond 45 nm is not shown. Alternative to that some predominant options for interconnect design are of greater importance for further miniaturization because: there are no metals with conductivity significantly higher than that of copper; Dielectric constants cannot go below 1, and to achieve dielectric constants below 2.5 porosity needs to be incorporated into the material, which weakens it; and unlike transistors, scaling or shrinking deteriorate the performance of interconnects, and that deterioration will (already has) become a significant limiter in overall circuit performance.

The performance limitation of interconnects and packaging shows clearly the inadequacy of conventional solutions to meet the overall performance requirements in high performance electronic systems in the nanometer regime. Traditionally, the interconnect requirements have been met through distinctly separate functions of

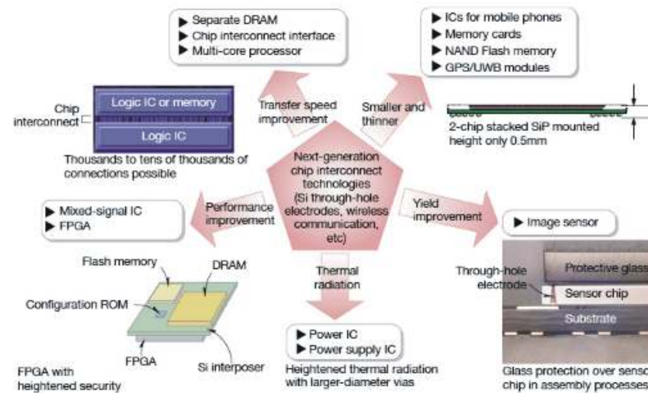


Figure 1.7: *Next generation chip interconnect technologies, such as Si through-hole vias. They boost chip-to-chip data transfer tares, contribute to smaller and thinner chips with higher performance, and improve heat radiation and yield [24].*

on-chip interconnects, package, silicon chip, and board-level technologies, but integrated system level solutions are required for growing interconnect and packaging challenges.

ITRS [2] suggest some of the possible solutions which can and have already been proven to be effective in alleviating tyranny of interconnects. They are:

- **Use different signaling methods**: Circuit techniques, signaling techniques such as multi-level, near speed of light, signal coding techniques. (Uses the available technology with innovative approaches to reduce delay, crosstalk etc.)
- **Innovative design and package options**: interconnect-centric design, Package intermediated interconnect, Chip-Package co-design
- **Use geometry**: Three-Dimensional integration (reduces the wire length)
- **Use different physics**: optics, RF microwaves, Tera-hertz photonics (introduce different information carriers other than charge.)
- **Radical Solutions**: Nanowires/nanotubes, Molecules, Spin, Quantum wave functions

1.4 Scope of Thesis and Author's Contribution

During the last decade KTH has a track record of research in response to interconnect-centric design [25, 26], chip-package co-design [27], and different signaling and interconnect optimization strategies [25, 26]. The scope of this thesis is the design, modelling and analysis of system interconnections and their effects in massively integrated 3-D ICs under cost, performance, and other technological constraints. The technical contributions of this thesis are three-fold: signalling techniques for global on-chip interconnects; cost, performance and technological trade-offs for 2-D and 3-D mixed-signal ICs; and electrical modelling of Through-Silicon Vias (TSV)

in 3-D ICs. The author's contributions are discussed in the next few sections, with a brief summary and key publications.

1.4.1 Smart Repeaters for Interconnections in Nanometer Technologies

Summary: Smart repeaters exploit the fact that in a parallel wire structure, the effective capacitance of a given wire is dynamic; i.e. it is a function of not only the physical geometry, but also the relative switching pattern described by the bits on the wire in question (the victim) and the adjacent wires (aggressors). With a traditional repeater, since the drive strength is static, the result is a spread of the propagation delay, with the repeater strength being essentially too much for every bit pattern other than the worst-case pattern. In the proposed repeater, the drive strength is dynamically altered depending on the relative bit pattern, by partitioning it into a Main Driver and Assistant Driver. For a higher effective load capacitance both drivers switch, while for a lower effective capacitance the assistant driver is quiet. By disconnecting part of the repeater when it is not needed, the total load capacitance to the previous stage is reduced, resulting in reduced energy consumption for those instances. It is experimentally shown that for a UMC 0.18 μm technology the potential energy saving is 10% over a traditional repeater for typical global wire lengths. Also, with the technology scaling the potential average saving in energy can be as much 20%-30% for typical global wire lengths in nanometre technologies.

Author's Contribution: *The first author came up with the concept, carried out analytical work on dynamic energy saving and the timing model, designed the circuit, carried out the simulations, and wrote the manuscripts of all of the following publications.*

Related Publications:

1. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa and Hannu Tenhunen, "Switching sensitive interconnect Driver to Combat Dynamic Delay in on-Chip Buses," in *Lecture Notes in Computer Science (Proceedings of PATMOS)*, vol. 3728, pp. 277-285, 2005.

   Technical Contribution in the paper: *This paper proposes a switching pattern dependent-driver and provides circuit-level proof of concept.*

2. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng and Hannu Tenhunen, "Minimum-Power, Delay-Balanced Drivers for interconnects in the Nanometer Regime," in *Proceedings of the international workshop on System-Level Interconnect Prediction*, German, March, 2006, pp. 113-120.

   Technical Contribution in the paper: *A methodology for design of the SMART repeater is proposed and a high-level analysis of the energy saving is presented.*

3. **Roshan Weerasekera**,Dinesh Pamunuwa, Li-Rong Zheng and Hannu Tenhunen, "Delay-Balanced Smart-Repeaters for on-chip Global Signaling", in P*Proceedings of the 20th International Conference on VLSI Design held jointly with 6th International Conference on Embedded Systems*, 2007, pp. 308-313.

   Technical Contribution in the paper: *The circuit level implementation of the smart repeater, a first order delay model, power and delay comparison, crosstalk, and sizing of assistant and main driver for different coupling capacitances are presented.*

4. **Roshan Weerasekera**,Dinesh Pamunuwa, Li-Rong Zheng and Hannu Tenhunen, "Minimal-Power, Delay-Balanced Smart Repeaters for Global Interconnects in the Nanometer Regime", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 5, pp. 589-593, May, 2008.

   Technical Contribution in the paper: *Energy and timing models, driver circuit design, and experimental verification of the models have been revisited.*

1.4.2 Cost and Performance Trade-offs for 2-D and 3-D Mixed-Signal ICs

Summary: Because of today's market demand for high-performance, high-density portable hand-held applications, electronic system design technology has shifted the focus from 2-D planar single-chip solutions to alternative options such as tiled silicon and single-level embedded modules as well as 3-D integration. Among the various choices, finding an optimal solution for system implementation deals usually with cost, performance and other technological trade-off analysis at the system conceptual level. It has been identified that decisions made within the first 20% of the total design cycle time will ultimately affect upto 80% of the final product cost. In this work, we discuss appropriate and realistic metrics for performance and cost trade-off analyses both at system conceptual level (up-front in the design phase) and at the implementation phase for verification in the 3-D integration. In order to validate the methodology, two ubiquitous electronic systems are analyzed under various implementation schemes and the pros and cons of each of them are discussed.

Author's Contribution: *The author came up with the idea, derived all the models, carried out the analyses, and wrote the manuscripts of all the following publications.*

1. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa and Hannu Tenhunen, Weerasekera, Roshan; Li-Rong Zheng,; Pamunuwa, Dinesh; Tenhunen, Hannu, "Early selection of system implementation choice among SoC, SoP and 3-D Integration," in *IEEE International System-on-Chip Conference*, September, 2007, pp.187-190.

   Technical Contribution in the paper: *A preliminary description on yield, cost, and performance models for SoC, SoP, and 3-D integration for trade-off analyses are presented in this paper.*

2. **Roshan Weerasekera**, Li-Rong Zheng, Dinesh Pamunuwa and Hannu Tenhunen, " Extending Systems-on-Chip to the Third Dimension: Performance, Cost and Technological Tradeoffs," in *Proceedings of the IEEE/ACM international conference on Computer-aided design*, IEEE Press, November, 2007, pp. 212-219.

   Technical Contribution in the paper: *An extensive discussion on the yield, cost, thermal issues and performance of 2-D and 3-D integration options are carried out in this paper.*

3. **Roshan Weerasekera**, Dinesh Pamunuwa, Li-Rong Zheng and Hannu Tenhunen, "2-D and 3-D Integration of Heterogeneous Electronic Systems under Cost, Performance and Technological Constraints," in *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems,* September, 2008, Under Review.

   Technical Contribution in the paper: *Yield, cost and performance models presented in first and second papers have been refined introducing the effects of thermal-via insertion to reduce excess heat generated in 3D-ICs and the testing cost.*

1.4.3 Electrical Modelling of Through-Silicon Vias in 3-D Integrated Circuits

Summary: Parasitic parameter (resistance, capacitance, and inductance) extraction of TSV structures is a critical step towards a successful physical design of three-dimensional integrated circuits. Various TSV structures starting from a lone TSV and going up to a 3×3 bundle have been simulated in a field solver with varying geometrical parameters, and its electrical parameters have been extracted. Then, a set of novel closed-form equations are proposed for TSV parasitics in terms of physical dimensions and material properties, allowing the electrical modelling of TSV bundles without the need for computationally expensive field-solvers. Finally, suitable equivalent circuits including capacitive and inductive coupling are derived, and comparisons with field solver provided values are used to show the accuracy of the proposed parasitic parameter models for the purpose of performance and SI analysis.

Author's Contribution: *The author came up with the idea, built the test structures, analysis methodology, derived the empirical formulae, and wrote the manuscripts of the following publications.*

1. **Roshan Weerasekera**, Dinesh Pamunuwa, Hannu Tenhunen, and Li-Rong Zheng, "Modelling Through-Silicon-Vias in 3D-ICs," *IET Electronic Letters*, September, 2008, Under Review.

   Technical Contribution in the paper: *This paper proposes novel compact closed-form equations for TSV parasitics in terms of physical dimensions and mate-*

rial properties allowing electrical modelling of TSV bundles without the need for computationally expensive field-solvers.

2. **Roshan Weerasekera**, Dinesh Pamunuwa, Matt Grange, Hannu Tenhunen, and Li-Rong Zheng, "Parasitic Parameter Estimation and Electrical Modelling of Through-Silicon Vias in 3-D ICs," *IEEE International Symposium on Circuits and Systems 2009*, Under Review.

   Technical Contribution in the paper: *This is a detailed discussion of the compact closed-form equations proposed in paper 1 for TSV parasitics with delay and noise amplitude estimations using extracted and predicted parasitics.*

3. Matt Grange, **Roshan Weerasekera**, Dinesh Pamunuwa and Hannu Tenhunen, "Exploration of Through Silicon Via Interconnect Parasitics for 3-Dimensional Integrated Circuits" *IEEE International Symposium on Circuits and Systems 2009*, Under Review.

   Technical Contribution in the paper: *Trends in TSV bundle parasitics, signal integrity issues and related metrics are discussed.*

4. **Roshan Weerasekera**, Matt Grange, Dinesh Pamunuwa, Hannu Tenhunen, and Li-Rong Zheng "Modelling and Analysis of Through Silicon Via Interconnects in 3-Dimensional Integrated Circuits" In submission to IEEE transactions of VLSI.

   Technical Contribution in the paper: *Trends in TSV parasitics, their empirical models, and signal integrity issues are throughly discussed in this paper.*

1.5 Thesis Organization

This thesis is organized into seven chapters. Chapter one is the introduction to the thesis where the historical evolution and trends in electronic system design, research overview and author's contribution are discussed. The second chapter provides basic theoretical background for interconnect modelling and analysis. Trends in TSV parasitics and novel compact closed-form equations for them are proposed in Chapter 3. In Chapter 4, signalling techniques suitable for on-chip global interconnects for 2-D and 3-D integrated circuits are presented. Based on the process and gate level data available, a system conceptual level chip/die parameter estimation methodology is discussed in Chapter 5. These models are not originally from the author's research work, but a comprehensive collection and some changes being made in compliance with the current needs. Using these system level parameters, yield and cost models for various packaging options are derived. Chapter 6 discusses the cost and performance trade-off analysis methodology for 2-D and 3-D integration are presented with two case studies. In Chapter 7, the conclusions and future work is elaborated.

2

Interconnect Modelling and Analysis

This chapter serves as a general introduction to the research issues discussed in this thesis. It starts by discussing non-idealities in wires, and carries out a comprehensive review on established parasitic estimation techniques and electrical modelling from low to high frequencies for wires. The chapter ends with a discussion of delay estimation models.

2.1 Introduction

The significant difference between any two electronic system, for example a personal computer and a washing machine controller is the pattern of interconnections between various active and passive components. These interconnections or wires carry signals from one place to another and make up the different functionality that the user expects. Thus, realizing the interconnections between various devices and modules make up an electronic circuit or system.

Electronic Systems are packaged in a hierarchy of chips, carriers, circuit boards, chassis and cabinets (Refer Figure 2.1). At each level of hierarchy, signals are transported on different kinds of interconnections. On-chip wires constitute the lowest level in a hierarchy that spans chip- to package-level connections (such as bond wires, package vias and solder balls, and package traces), circuit-board level connections (thick film wires), backplane-level wires (thick film metal layers or cables), chassis-level connections (more cables) and finally rack-level connections (such as bus bars made of solid metal straps or rods for power connections).

The designer of an electronic system/circuit has multiple choices in realizing the interconnections, which appear in the schematic diagrams as simple lines without apparent impact on the overall circuit performance. These are ideal wires assumed to be equi-potential regions, voltage variations at the near end of the wire are assumed to appear at the far end of the wire at exactly the same point of time, *i.e.* propagation speed is infinitely high. A real wire, however, is not an ideal conductor with zero resistance, capacitance and inductance, but rather an unintended or parasitic circuit element. Also, interconnect structures in state-of-the-art ICs and packages form a complex geometry with capacitive, resistive and inductive coupling between neighbouring systems. All these non-idealities have an effect on the desired circuit behaviour.

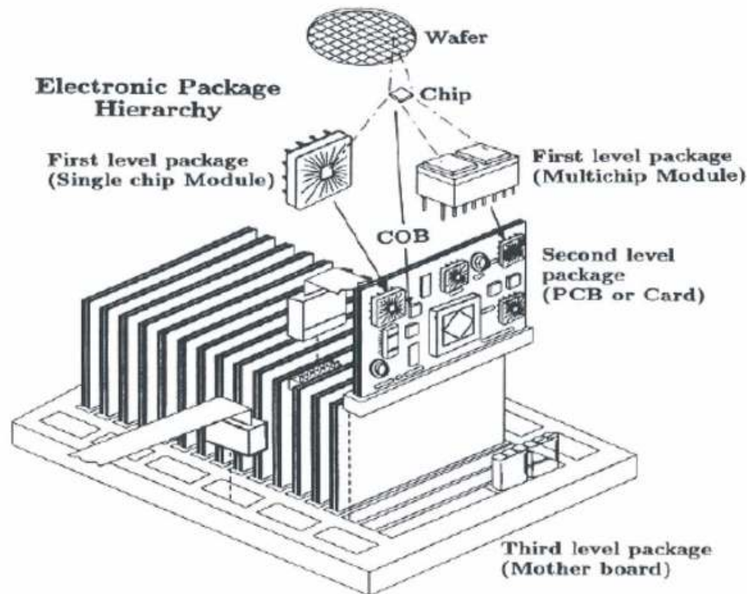


Figure 2.1: *Electronic Packaging hierarchy in Electronic Systems [5].*

With the increase in circuit performance, complexity, density and levels of integration in nanometer technologies, it is essential to include all parasitic effects in the circuit analysis and optimization process. But this approach is not very constructive due to a plethora of design variables in the optimization process, and the complexity of the overall circuit with millions of nodes require an unacceptably high computational time. Furthermore, this approach has the disadvantage of potentially masking the true problem, because at a given circuit node, only few dominant parameters affect the overall performance. Therefore, complete circuit optimization represents a trial and error or heuristic approach rather than a methodological approach to the design process. Thus, usually designer are compelled to have a clear insight into the parasitic wiring effects, their relative importance, and their reduced-order models. By identifying the critical portions of a system, it can be analyzed more effectively for relevant parasitic effects.

Furthermore, in order to compare different interconnect schemes the most important metrics that are usually used as figures of merits in interconnect performance are: [*28*, *29*] propagation delay, or equivalently, performance; power consumption; and noise coupling, which impacts the reliability. Evaluating above mentioned figure of merits require estimating wire parasitics because all figures of merits are functions of parasitics. This chapter discusses the electromagnetic view of wires, methods and basis for estimating wire parasitics, and interconnect delay estimation techniques.

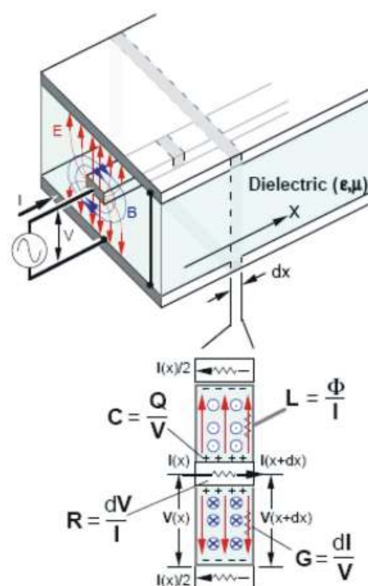


Figure 2.2: *Electromagnetic View of a Wire [30]. If I,V change at the drive point, B,E change as well; disturbance propagates away from the drive point at the speed of light* $v = \frac{c}{\sqrt{\epsilon_r}}$.

2.1.1 Electromagnetic View of Interconnects

When a wire is excited with an electrical signal it will travel down the wire forming an electric field due to its potential and a magnetic field due to the current flowing in the wire. Ideally the electromagnetic field components in a wire, shown in Figure 2.2, are perpendicular to each other and to the direction of wave propagation, called Transverse Electro-Magnetic (TEM) mode waves. All electromagnetic behaviours can ultimately be explained by Maxwells four basic equations shown in Table 2.1. Maxwell's equations for complex geometries are usually solvable by numerical methods with the aid of a field solver [*31*], requiring prohibitive amounts of computation time for large ICs. Field solvers use numerical techniques to solve Maxwell's equations by one of two classes of methods. The first uses the differential form of the governing equations and are called Finite Difference (FD) and Finite Element Method (FE) methods. The other methods use integral equation approaches such as the Method of Moment (MoM) and the Boundary Element Methods (BEM).

For the most part, the use of field solvers is restricted to critical portions of the chip due to the complexity and the resultant high computational time. Hence, based on the frequency range of interest, length of the line, and the rise time of the signal, these equations are simplified to achieve faster computation.

To perform timing and signal integrity analysis, it is necessary to translate layout information such as wire width and length, the geometry of surrounding wires and substrate parameters into electrical parameters. Then, they can be combined with other circuit elements to estimate the overall system performance, and also

optimization. The most general wire model is the transmission line, but most wires do not behave as transmission lines, for example on-chip wires are more resistive than off-chip wires and do not show transmission line properties. Generally wires may be modeled as Capacitive (C), Resistive (R), RC, LC, RLC, $R(f)L(f)C$ lines with capacitive and/or inductively coupling. The next section describes how these parameters are extracted in order to complete an electrical model of the wire.

	Differential Form	Integral Form
Gauss's Law	$\nabla \cdot \vec{D} = \rho$	$\oint_S \vec{D} \cdot d\vec{A} = \int_V \rho d\vec{V}$
Gauss's Law	$\nabla \cdot \vec{B} = 0$	$\oint \vec{B} \cdot d\vec{A} = 0$
Faraday's Law	$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$	$\oint_c \vec{E} \cdot d\vec{l} = -\frac{d}{dt} \int_S \vec{B}.d\vec{A}$
Ampere's Law*	$\nabla \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t}$	$\oint_c \vec{H} \cdot d\vec{l} = \int_S \vec{J}.d\vec{A} + \frac{d}{dt} \int_S \vec{D}.d\vec{A}$

Table 2.1: *Maxwell Equations: Maxwell's four equations express how electric charges produce electric fields (Gauss' Law), how currents and changing electric fields produce magnetic fields, and how changing magnetic fields produce electric fields. In non-dispersive, isotropic media, the field vectors are related as:* $\vec{B} = \mu\vec{H}, \vec{D} = \mu\vec{E}$ *Also, from the law of charge conservation, we can write* $\nabla \times \vec{J} = -\frac{\partial \rho}{\partial t}$

2.2 Parasitic Estimation and Extraction

Wire parasitic extraction is usually carried out by representing complex structures as a collection of simple geometric elements, and then each parasitic value is combined using superposition or introducing scale factors to obtain the parasitics of the complex structure. There are commonly used industrial tools which simply extract the wire parameters for given any complex structures such as ANSOFT Q3D Extractor [*32*], FastHenry [*33*] and FastCap [*34*]. Many of these commercial tools assume that the electromagnetic field through interconnects is quasi-static; they ignore the displacement current in Maxwell's equations. With such a simplification, electrical fields remain static outside conductors, but magnetic fields retain frequency dependency inside conductors so that the skin effect can be accounted for properly. Capacitance and conductance of a structure are determined by electrical fields only; resistance and inductance are determined only by magnetic fields. In other words, by ignoring the displacement current, magnetic and electrical fields are decoupled in the quasi-static theory, and can be solved independently. Because of the decoupling a quasi-static field solver is quicker and it can solve much bigger problems in less time than a full-wave solver. Many modern quasi-static solvers can perform whole-package RLGC extraction of a complicated package design in a few hours. However, it would be far too inefficient to embed multiple field solver calls during the course of an iterative optimization involving circuit simulation, and

the methods explained below are proven to be accurate within 1-10%, and used in general for the vast majority of calculations.

2.2.1 Resistance

By definition, from the fundamental laws of electrostatics, the resistance is the ratio of potential difference of the two ends of a wire to the total current flowing into it:

$$R \equiv \frac{\Phi_{12}}{I} = \frac{-\int_L \vec{E}.\vec{dl}}{\int_A \sigma \vec{E}.\vec{dl}} \tag{2.1}$$

Resistance is dominated by the cross sectional area and the resistivity (inverse of conductivity) of the signal conductor. The resistance of a uniform wire with width w, thickness t, and resistivity ρ, is:

$$r_{dc} = \underbrace{\frac{\rho}{t}}_{=R_\square} \frac{l}{w} = R_\square \frac{l}{w} \tag{2.2}$$

Since the thickness is usually a constant for a given technology, it is customary to incorporate it with the resistivity and form a single constant called *sheet resistance* of the material.

(A) Frequency dependency: The Skin Effect

At DC or low frequencies, the current flowing in a conductor will spread out uniformly as much as possible over the cross-section. As the frequency increases, the current density inside is not uniform, but drops away exponentially with depth into the conductor. This phenomenon is known as the *skin effect*. This leads to current crowding primarily on the surface and the effective cross-section where current flows reduces. As a consequence, wire resistance increases with the frequency.

Skin effect is defined as the depth below the surface of the conductor at which the current density decays to $1/e$ (about 0.37) of the current density at the surface [*35*], and is given by:

$$\delta_e = \sqrt{\frac{\rho}{\mu \pi f}}. \tag{2.3}$$

Skin effect onset occurs generally close to the frequency (cut-off frequency, f_s) where $\delta \leq 0.3t$ and is fully developed when $\delta \ll t$ (as a guideline $\delta \leq 0.1t$) [*36*]. For

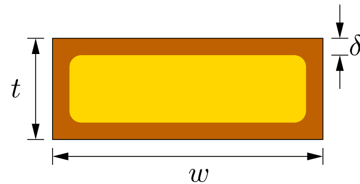


Figure 2.3: *Wire Cross Section.*

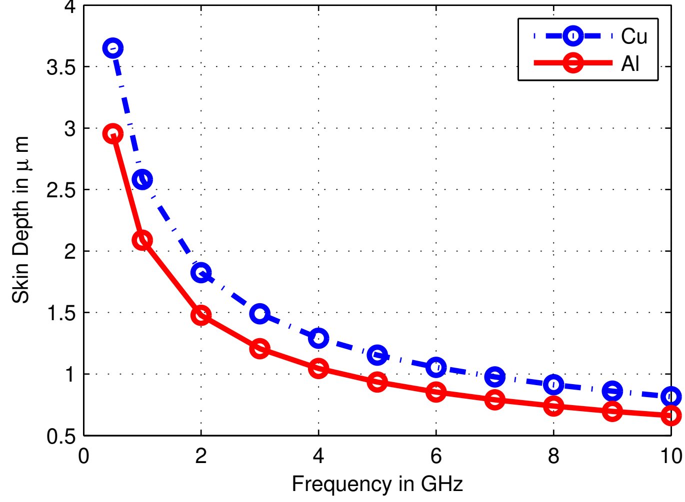


Figure 2.4: *Skin Depth versus frequency of Al and Cu On-Chip Interconnects [37].*

typical on-chip wires, δ_e is found to be equal to $1.5tw/(t+w)$ with relative error less than 5% for $0.25 < t/w < 10$ [*25*].

It is straightforward to define an effective resistance by dividing the product of resistivity and wire length by the effective area that the total current passes through. The effective area is now limited to $wt - (w-2\delta)(t-2\delta) \approx 2\delta(w+t)$, and the expression for frequency dependant resistance at high frequencies is:

$$R(f) = \frac{l\sqrt{\pi\mu\rho f}}{2(w+t)} \tag{2.4}$$

Additionally to that there is an empirical formula which is widely used to describe the frequency dependent behaviour of a wire over a ground plane:

$$R(f) = \begin{cases} r_{dc} & f \leq f_0 \\ r_{dc}\sqrt{\frac{f}{f_0}} & f \geq f_0 \end{cases} \tag{2.5}$$

where $f_0 = \frac{\rho}{\mu\pi\delta_e^2}$ is referred to as the break frequency at which this phenomenon begins to dominate.

Furthermore, the accurate frequency dependent modelling of wire parameters includes resistance and inductance. A thorough discussion is included in Section 2.2.2.

(B) Diffusion Barrier Effect

Another factor responsible for increased resistivity - effective Cu wire resistivity of 2.2×10^{-8} Ωm compared to 1.7×10^{-8} Ωm for bulk Cu - is the presence of a finite cross-sectional area consumed by the higher resistivity metal barrier material which

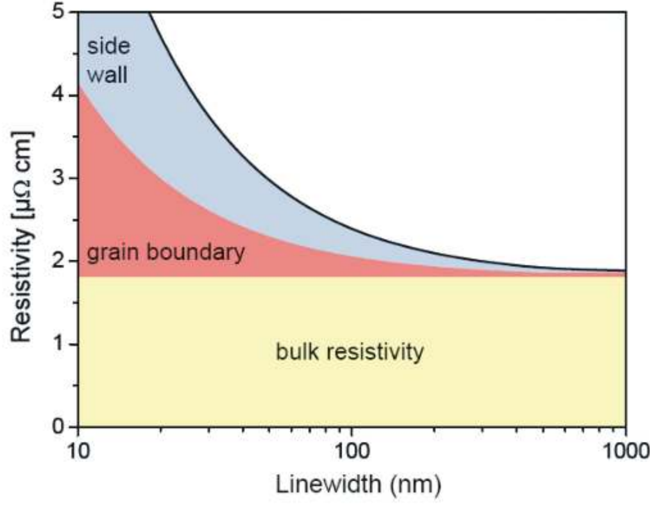


Figure 2.5: *Effect of Grain Boundary and Surface scattering to Cu Resistivity [2].*

encapsulates the Cu interconnect. Barrier material is usually a refractory material such as Titanium (Ti) or Tantalum (Ta) or their Nitrides [*38*]. This metal barrier prevents the diffusion of Cu into the surrounding dielectric. Since the resistivity of the barrier material is extremely high, it is reasonable to assume that Cu carries all the current, and therefore, the effective area through which current conducts is reduced. As the barrier thickness cannot scale as rapidly as the interconnects, it increasingly occupies a higher fraction of the interconnect cross sectional area while restricting the current flow within the material with lower resistivity. The effect on resistivity because of the barrier is given by [*39*]:

$$\rho_b = \frac{\rho_o}{1 - \frac{A_b}{wt}}, \tag{2.6}$$

where ρ_o is the bulk resistivity at a given reference temperature, A_b is the area occupied by the barrier, and w and t are the wire width and thickness respectively.

(C) Surface and grain boundary scattering Effect

In addition to that, resistivity of on-chip metal interconnects begins to increase as the minimum dimension of the metal line becomes comparable to the mean free path of the electrons due to the fact that surface scattering has a significant contribution to resistivity compared to the contribution from bulk scattering. The modelling of this effect dates back to 1938 by Fuchs [*40*] for 1-dimension, which was later extended to 2-D [*41*] in 1952. Fuchs's scattering governed expression for resistivity of a thin film metal is in terms of bulk resistivity:

$$\rho = \frac{\rho_o}{\left[1 - \frac{3(1-p)}{2k} \int_1^\infty \left(\frac{1}{x^3} - \frac{1}{x^5}\right) \frac{1-e^{-kx}}{1-pe^{-kx}} dx\right]}. \tag{2.7}$$

Here $k = \frac{d}{\lambda_{mfp}}$; d is the smallest dimension of the film, λ_{mfp} the bulk mean free path of electrons, p the fraction of electrons which are elastically reflected at the surface, and ρ_o the bulk resistivity of Cu equal to 1.7 $\mu\Omega cm$. The dominance of the surface effect depends on the parameter k. For Cu, $p = 0.47$ and $\lambda_{mfp} = 421$ $^oA@0^oC$ [42]. When $k \gg 1$, scattering governed resistivity can be expressed as [42]:

$$\rho = \frac{\rho_o}{1 - \frac{3(1-p)}{8k}} \tag{2.8}$$

Grain boundaries in polycrystalline interconnect act like partially reflecting planes. When the grain size is comparable to the electron mean-free path, the electrons suffer greater gain boundary effect further increasing the resistivity. That can be expressed mathematically [43] as:

$$\rho_g = \frac{\rho_o}{3\left[\frac{1}{3} - \frac{\alpha_g}{2} + \alpha_g^2 - \alpha_g^3 ln\left(1 + \frac{1}{\alpha_g}\right)\right]}, \tag{2.9}$$

where, $\alpha_g = \frac{\lambda p_g}{d_g(1-p_g)}$. Here, d_g grain diameter, and p_g the grain boundary reelection coefficient ($0 < p_g < 1$). In the limits of very small and very large α, (2.9) takes the simple forms:

$$\rho_g \approx \begin{cases} \left(1 + \frac{3}{2}\alpha\right)\rho_o & \text{for } \alpha \ll 1, \\ \frac{4}{3}\alpha\rho_o & \text{for } \alpha \gg 1. \end{cases} \tag{2.10}$$

Resistivity equation for a thin wire combined with surface scattering and gain boundary effects is rather complex to use in simple calculations and hence, a reduced-form expression is highly desirable and useful in interconnect analysis. Based on an empirical study on surface and grain boundary scattering models proposed in [40, 41, 43] and with the aid of curve fitting techniques, [44] has presented a simple closed form resistivity model given by:

$$\rho(w) = \rho_B + \frac{K_p}{w}, \tag{2.11}$$

where the fitting parameters ρ_B and K_p are 2.202×10^{-8} Ωm and 1.030×10^{-15} Ωm^2, respectively. Notably, ρ_B is almost the same as bulk resistivity of Cu. The notable absence in this formula is a term which describes the dependency of wire thickness on resistivity in relation to the scattering effect. [45] has presented an experimentally validated model including wire thickness too. That is:

$$\rho = \beta + \alpha\frac{1}{wt}, \tag{2.12}$$

where, $\alpha = 0.0072$ and $\beta = 1.9357$ $\mu\Omega cm^{-1}$.

(D) Temperature Effect

A qualitative view of the temperature dependence for resistance may be obtained by examining the effects of temperature on carrier concentration; in basic terms, conductivity relates to the carrier concentration (q) and mobility of the carriers (μ) as given by: $\sigma = q\mu$. Carriers are created by the ionization of atoms within the

lattice comprising the solid, and conductors are easily ionized by nature. They all have a surplus of free electrons. At the temperature of interest, essentially all the atoms in a conductor are ionized and the supply of electrons is virtually constant with temperature. However, the carriers usually do not move in a straight line when they traverse through a material. This movement is influenced by defects in the lattice, impurities, grain boundaries, and fixed ions. As temperature increases, the carriers are more active and suffer more collisions, thereby reducing the mobility.

In the case of conductors, the loss of mobility is entirely due to ionic scattering and depends on the characteristic of the particular material and can be usually characterize using the traditional relationship [46]:

$$\rho(T) = \rho_o(T_o)\left[1 + t_{cr}(T - T_o)\right] \tag{2.13}$$

where $\rho(T)$ is the wire resistivity at any given temperature T, $\rho(T_0)$ is the wire resistivity at the reference temperature T_0, t_{cr} is the temperature coefficient of resistance (TCR) of the bulk material. Mathematically, the TCR is the slope of $\rho(T)$ vs. T curve normalized to $\rho(T)$, and for the cases where the TCR is nonlinear, a linearized average over a range of temperature may be derived. For bulk Cu, t_{cr} = 0.39-0.43 $\%^oC^{-1}$ at 20 ^oC [46, 47].

A study on Cu wires in 65 nm technology has been carried out by Lu *et.al.* [48] of IBM corporation. They proposed an experimentally validated empirical equation which described the dependence of wire resistance with surface and grain boundary scattering together with the temperature :

$$\rho_{sg} = \rho_0\left[1 + t_{cr_bulk}(T - T_o) + \frac{\alpha}{w} + \frac{\beta}{h}\right] \tag{2.14}$$

where ρ_0 is bulk wire resistivity, w and h are wire width and height of the Cu portion, and the model parameters α and β are positive constants, which are functions of a surface scattering coefficient and gain boundary scattering coefficient. α has been extracted for each metal level (*i.e.* for each wire thickness h) as $\alpha = a + \frac{b}{h}$. The coefficients are: $a = 0.021\ \mu m$, $\beta = 0.016\ \mu m$, $b = 0.0014\ \mu m^2$ [48]. They have also found that t_{cr} is equal to 0.43 $\%^oC^{-1}$ at 20 ^oC.

2.2.2 Inductance

Inductance is a measure of the distribution of the magnetic field near and inside a current-carrying conductor. This measure is a property of the physical layout of the conductor, and is a measure of the ability of that conductor to link magnetic flux, or store magnetic energy. The fundamental definition for inductance is

$$L = \frac{\oint \vec{B} \cdot d\vec{A}}{I} \tag{2.15}$$

The definition of inductance follows a loop property: *i.e.* in order to determine the inductance accurately the current return path should be known. In modern interconnect structures return current is over a range, and an exact return path cannot easily be identified. However prior work established that the current return path is primarily in the power distribution network, and other adjacent wires [49]. The loop formed by the signal wire and the return path can potentially extend to

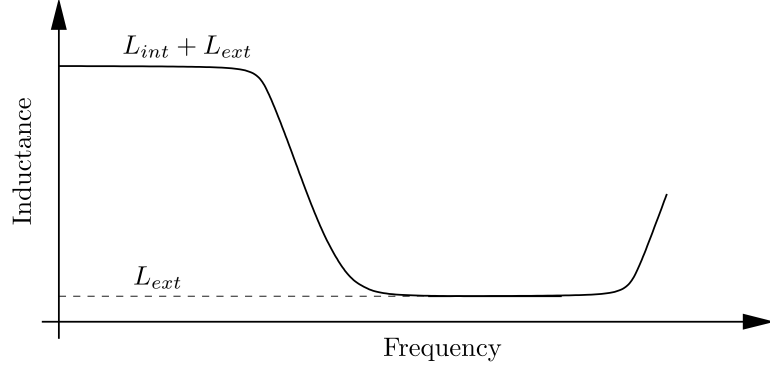


Figure 2.6: *Frequency Dependence of Inductance. Low frequency inductance of a wire is the sum of internal and external inductances, while at high frequencies, internal inductance vanishes.*

several hundred micrometers away from the wire under consideration. This vastly complicates the extraction of parasitic inductance of a given wire, as it depends not only on the characteristics of a particular wire, but also potentially on the characteristics of several thousand other wires. In order to estimate the inductance, the induced current is assumed to return at infinity. This method was proposed by Rosa [*50*] in the early part if the 20th century, and was further introduced for circuit analysis by Ruheli [*51*].

(A) Partial and loop-based inductance

A simple approach that can be used for inductive parasitic extraction is to use the free space relationship, which relates loop inductance (L) of a wire to its capacitance ($C_{\epsilon_r=1}$) assuming a dielectric of air in the medium [*52*], given by:

$$L = \frac{\epsilon_0 \mu_0}{C_{\epsilon_r=1}} \tag{2.16}$$

This method is used in the tool Raphael RC2 [*37*], which is a two dimensional parasitic extraction tool. Considering the middle conductor in a three parallel conductor system, the self and mutual inductance equations become:

$$L_s = \frac{\epsilon_0 \mu_0}{2} \left(\frac{1}{C_s} + \frac{1}{C_s + 2C_c} \right) \tag{2.17}$$

$$L_m = \frac{\epsilon_0 \mu_0}{2} \left(\frac{1}{C_s} - \frac{1}{C_s + 2C_c} \right) \tag{2.18}$$

where C_s can either be C_{smid} or C_{scorn} based on the wire in consideration. Unfortunately in a IC, this assumption does not hold up and more sophisticated methods need to be used.

When the return paths are not known a priori, the widely accepted method requires partial inductance elements be calculated for the whole loop. This is defined as the flux created by the current of one segment through the virtual loop

which another segment forms with infinity. The partial inductance can be found for two coupled segments, a and b by solving the integral [*53*]:

$$L_{ab,partial} = \frac{\mu}{4\pi}\frac{1}{A_a A_b}\int_{A_a}\int_{l_a}\int_{A_b}\int_{l_b}\frac{dl_a \cdot dl_b}{|r_a - r_b|}dA_a dA_b \tag{2.19}$$

where A_a and A_b are the cross-sections of the segments, and l_a and l_b their lengths.

To calculate the partial inductances of rectangular cross-sectional wires, closed-form equations presented in [*50*] are used. The formulae for self and mutual inductances of a rectangular wire with $l \gg w + t$ are shown in (2.20) and (2.21).

$$L_{self} = \frac{\mu_o l}{2\pi}\left[\ln\left(\frac{2l}{w+t}\right) + \frac{0.2235(w+t)}{l} + \frac{1}{2}\right] \tag{2.20}$$

$$L_{mutual} = \frac{\mu_o l}{2\pi}\left[\ln\left(\frac{2l}{s}\right) - 1 + \frac{s}{l}\right] \tag{2.21}$$

Here, μ_0 is the permeability of air equal to $4\pi \times 10^{-7}\frac{H}{m}$.

Loop inductance is the sum of partial self and mutual inductances of the segments which form all loops in the system. This can be expressed mathematically as:

$$L_{loop} = \sum_i \sum_j s_{ij} L_{p,ij} \qquad \text{with} \tag{2.22}$$

$$s_{ij} = \begin{cases} -1 & \text{when current flows in opposite directions} \\ +1 & \text{when current flows in the same direction} \end{cases}$$

where $L_{p,ij}$ is the partial inductance of segment l_i due to current i_j on segment l_j. If $i = j$, $L_{p,ij}$ is the partial self inductance, else the partial mutual inductance. In order to find $L_{p,ij}$ of each branch either (2.19) and (2.21) or (2.20) may be used.

(B) Internal and External Inductance

For a wire with a finite conductivity, the magnetic flux exists both inside and outside the conductor. Therefore inductance of a wire can be subdivided into two components: internal inductance (L_i), for the inductance of the wire due to magnetic flux inside the wire; external inductance (L_{ext}), for the inductance of the wire due to magnetic flux outside the wire (Loop or partial inductance is external to the wire). Then, the total inductance of a wire is the summation of external and internal inductances. Typically, internal inductance accounts for less than 10% of the total low-frequency inductance of a single wire, or open loop. For closed loops, the internal inductance may be a significant portion of the loop inductance, due to the cancellation of self- and mutual inductances.

When modeling the internal inductance, the high frequency effect of the current distribution has to be considered, because when the skin effect is well developed, current resides on the surface of the wire. For a wire with a circular cross section, the internal inductance can be found using the formula:

$$L_i = \frac{\mu l}{8\pi}. \tag{2.23}$$

For a wire with rectangular cross-section, the widely used formula to find the internal inductance is:

$$L_i = \begin{cases} \frac{r_{dc}}{2\pi f_0} & f \leq f_0 \\ \frac{r_{dc}}{2\pi\sqrt{f f_0}} & f \geq f_0 \end{cases} \quad (2.24)$$

where f_0 is the break frequency. In addition to this, Choudhury et al in [*54*] described a modelling methodology for internal inductance, but no expressions were presented as a function of wire geometries, which can easily be used early in the design phase.

(C) High frequency and proximity effects

In addition to the skin effect mentioned in Section 2.2.1, the current distribution inside a conductor also changes with frequency due to the proximity effect. If the current in these two wires flows in opposite directions, the currents concentrate towards each other; otherwise, the two currents shift away from each other. Both the skin effect and the proximity effect are essentially due to the same mechanism - the current tends to concentrate closer to the current return path in order to minimize the inductance. Note that at high frequencies, the resistance of a conductor also depends on the surrounding signal activities due to the proximity effect.

Another effect of frequency on the inductance is due to multi-path current re-distribution. In an integrated circuit, there are many possible current return paths, e.g., the power/ground network, nearby signal lines, and the substrate. The distribution of the return current among these possible paths is determined by the impedance of the individual paths. At different frequencies, the relationship among the impedances of different paths will change, as well as the distribution of the return current. The return current is distributed in those paths so as to minimize the total impedance at a specific frequency .

If the frequency dependent effects are very important to consider in a desired frequency range, the cross-sections are subdivided into sections smaller than the skin depth at the maximum frequency of interest. Then, the current distribution in each filament can be regarded as uniform. To calculate the partial inductances of rectangular cross-sectional wires, closed-form equations proposed by Rosa [*50*] are used. In this manner, an inductively coupled RL circuit can be formed for the conductor. By solving currents in this circuit at several points in the frequency domain, the frequency dependent resistance and inductance can be obtained [*55*]. This technique, which is known as partial element equivalent circuit (PEEC), is the foundation for frequency dependent parasitic extraction tools such as FastHenry [*33*], and was first proposed in [*56*].

To capture the proximity effect, it is reasonable to consider the inductance between different RL circuits formed for the conductors in the neighbourhood [*57*]. It is quite obvious that such an inductively coupled RL circuit is computationally inefficient to solve in SPICE, and therefore, wires can be modelled with frequency independent lumped-element circuit models up to an arbitrary maximum frequency using simple ladder networks [*35*]. Some works that present such ladder circuits are [*58*, *59*, *60*].

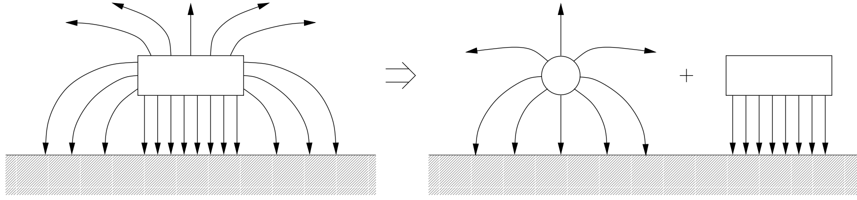


Figure 2.7: *Interconnection Line facing substrate and its Electrostatic Field. Capacitance of a conductor over a ground plane has two components: due to it direct vertically below electrostatic field, and fringing field. As wire width scales down, the parallel plate capacitance estimation underestimates the capacitance in several orders of magnitude due to the significant fringing field.*

2.2.3 Capacitance

Capacitance relates to the electric field as represented by the ratio of voltage to charge, and for a two conductor system, the wire capacitance can be defined as

$$C \equiv \frac{Q}{\phi_{12}} = \frac{\oint_S \vec{D} \cdot d\vec{A}}{-\int_A \sigma \vec{E}.d\vec{l}} \tag{2.25}$$

Here, ϕ_{12} is the voltage between the two conductors, A is any surface enclosing the positively charged conductor, and L is any path going from the negative conductor to the positive conductor. A physical approach requires the analytical solution of Poisson's equation, which often results in lengthy and complicated equations, often nonsolvable. To extract the capacitance of multi-conductor systems, electric field solvers such as Ansoft Q3D Extractor [*32*], FastCap [*34*] provide an accurate but computationally expensive solution.

In the design phase, there is a compelling need for fast, but accurate formulae for the estimation of capacitance values. Such equations necessarily represents an approximation. Several such analytical equations have been proposed in the literature, which are accurate enough to estimate interconnect capacitance for optimization purposes.

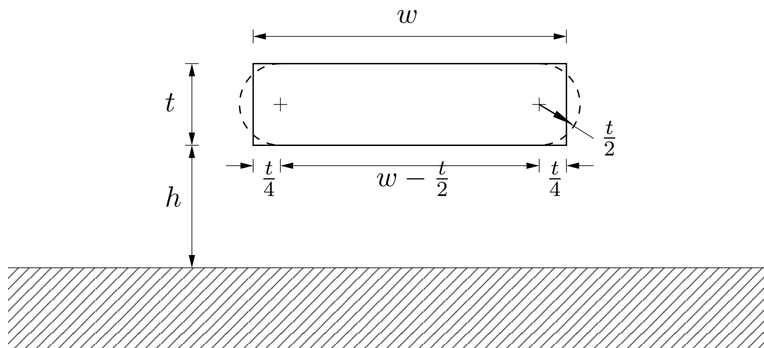


Figure 2.8: *Replace the rectangular line profile with an oval one, composed of a rectangle and two half cylinders.*

Single Wire over a Ground Plane: The per unit length parallel plate capacitance of the micro-strip line structure is given by:

$$C_p = \epsilon_k \frac{w}{h}, \tag{2.26}$$

where ϵ_k is the permit

The above simple parallel plate approximation underestimates the capacitance of a wire by several orders of magnitude if applied to the very high aspect ratio (height /width) wires in DSM technologies because the fringing field contributes significantly to the wire-to-ground capacitance. It is essential that the contribution of the fringe components of the E-field to the capacitance is taken into account. [*61*] presents an equation including fringing component for infinitesimal thin wire where width (w) is much greater than the distance from the ground plane (h):

$$C = \frac{\varepsilon w}{h}\left\{1 + \frac{2h}{\pi w}\left[1 + \ln\left(\frac{\pi w}{h}\right)\right]\right\} \qquad \text{for } w \gg h \tag{2.27}$$

This formula underestimates the capacitance because the derivation does not take into account the thickness of the conductor.

Another method to capture this fringing field effect is that a wire is decomposed into a rectangular cross section with a width (w) and a circular cross section with a diameter equal to wire thickness (t). Wire capacitance is then calculated as the sum of a parallel plate capacitor and a cylindrical wire over a ground plane. However, smooth surfaces such as an ellipsoid will have less charge accumulated near the ground plane than a square, which has sharp corners for congregation of charges. Also, circle has a perimeter πt compared to a perimeter of $4t$ for a square. These two factors may underestimate the total capacitance and therefore, to compensate, Yuan *et.al.* suggested in [*62*] to consider a parallel plate capacitor with width $w - \frac{t}{2}$ and a cylinder with radius $\frac{t}{2}$, giving a total capacitance of:

$$C = \varepsilon\left[\frac{w - \frac{t}{2}}{h} + \frac{2\pi}{\ln\left(1 + \frac{2h}{t} + \sqrt{\frac{2h}{t}\left(\frac{2h}{t} + 2\right)}\right)}\right] \tag{2.28}$$

However, this formula determines capacitance accurately only when $w \gg \frac{t}{2}$ and $t \approx h$. As this ratio drops in the region of $w < \frac{t}{2}$, this formula underestimates the capacitance, and therefore, the physically motivated approach was abandoned and an empirical formula suggested in [*62*]. Since then, several such capacitance estimation empirical formulae depending solely on curve-fitting techniques have been proposed in the literature, and a comparison has been carried out in [*63*]. [*64*] presents an equation which has a better accuracy than previously presented models when width/height ration drops below 2-3.

$$C_{sak} = \varepsilon\left[\frac{w}{h} + \frac{0.15w}{h} + 2.8\left(\frac{t}{h}\right)^{0.222}\right] \tag{2.29}$$

Another formula which is slightly more complex is reported in [*65*]:

$$C_{meij} = \varepsilon\left[\frac{w}{h} + 0.77 + 1.06\left(\frac{w}{h}\right)^{0.25} + 1.06\left(\frac{t}{h}\right)^{0.5}\right] \tag{2.30}$$

A Wire in a Multi Layered Structure: In contemporary ICs, multiple metal layers are in use, and these 3-D interconnects have been simplified to two-dimensional or quasi-three-dimensional structures, based on the layout pattern. If the layers above or below a set of wires in consideration are routed densely, they can be approximated as a ground plane, reducing the structure to a two-dimensional model. Under this condition, capacitive parasitics shown in Figure 2.9 are scalable functions of wire cross-sectional dimensions. Considering a single wire in a multilayer interconnect system, capacitance can be decomposed into two components: self capacitance (C_s) and mutual or line-to-line capacitance (C_c). In a multilevel interconnect structure, two capacitance structures can be identified: parallel lines on one plate, and parallel lines between two plates. The first structure emulates lines without top wiring, and the second structure emulates lines with top wiring.

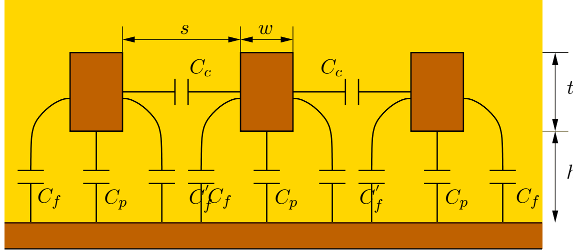


Figure 2.9: *Wire geometries and related capacitive parameters of a top-Layer.*

In such an environment since the presence of adjacent conductors significantly alters the electric field around the central conductor, the effect of the wire spacing, s, must be taken into account in the expression for wire capacitances. [*64*] presents self and mutual capacitance formulae for a wire, which can be used to estimate capacitance of the middle wire by $C_s + 2C_c$. Sakurai's mutual capacitance formula is:

$$C_c = \varepsilon \left[0.03\frac{w}{h} + 0.83\left(\frac{t}{h}\right) - 0.07\left(\frac{t}{h}\right)^{0.222} \right] \left(\frac{s}{h}\right)^{-1.34} \tag{2.31}$$

Total capacitance given by $C_s + 2C_c$ is in good agreement with the values predicted by a field solver, but individual components are not intended to provide accurate results. Such accurate formulae are proposed in [*66*, *67*].

A complete set of such equations partitioning the wire capacitances into ground and coupling components have been proposed in [*68*, *69*]. These equations allow a self capacitance to be defined both for a conductor sandwiched between two other conductors in (2.34), and also a conductor which has just one adjacent conductor

in (2.35).

$$C_f = \epsilon_k \left[0.0075\frac{w}{h} + 1.4\left(\frac{t}{h}\right)^{0.222}\right] \quad (2.32)$$

$$C_f^{'} = C_f \left[1 + \left(\frac{h}{s}\right)^{\beta}\right] \quad (2.33)$$

$$C_{smid} = C_p + 2C_f^{'} \quad (2.34)$$

$$C_{scorn} = C_p + C_f + C_f^{'} \quad (2.35)$$

The per unit inter-wire capacitance or the coupling capacitance is given by:

$$C_c = C_f - C_f^{'} + \epsilon_k \left[0.03\left(\frac{w}{h}\right) + 0.83\left(\frac{t}{h}\right) - 0.007\left(\frac{w}{h}\right)^{0.222}\right]\left(\frac{h}{s}\right)^{1.34} \quad (2.36)$$

The above set of equations are accurate to within 90% only when the wire geometries satisfy the following inequalities:

$$0.3 < \frac{w}{h} < 30, \qquad 0.3 < \frac{t}{h} < 10, \qquad 0.3 < \frac{s}{h} < 30.$$

When the wire geometry is out of that range, it is possible to treat the rectangular conductors as equivalent round wires if $w \leq 2H$, where $H = h + \frac{t}{2}$. There is a mutual coupling capacitance formula given in [70] for two round conductors over a ground plane. Zheng [25] presents an equivalent radius of square conductors to find self and mutual capacitance. The radius of the equivalent round conductors is then $R = 0.25w + 0.335t$, and the self and mutual capacitance terms are given by :

$$C_s = \frac{\pi\varepsilon_k}{\ln\left[\frac{2H\sqrt{4H^2+(s+w)^2}}{R(s+w)}\right]} - \frac{w\varepsilon_k}{2H} \quad (2.37)$$

$$C_c = \frac{2\pi\varepsilon_k \ln\left[\frac{\sqrt{4H^2+(s+w)^2}}{R(s+w)}\right]}{\ln\left[\frac{2H\sqrt{4H^2+(s+w)^2}}{R(s+w)}\right]\ln\left[\frac{2H(s+w)}{R\sqrt{4H^2+(s+w)^2}}\right]} \quad (2.38)$$

The relative error of the above two equations is less than 12% for most of VLSI interconnect geometries when $\frac{t}{w} < 2$ and $\frac{s}{w} < 2$ [69].

2.2.4 Conduction

The typical interconnect materials such as Cu and Al shows a considerably constant capacitance with the frequency variation. Since the DC leakage and the time-varying fields deposit negligible charge, the charge remains constant, and under quasi-static conditions, electric field remains constant with the frequency, which leads to a constant potential difference. Hence, the ratio of charge to potential difference is also a constant. However, when the dielectrics are inhomogeneous or

lossy, the capacitance will be strongly dependant on frequency. Then the dielectric constant can be defined as:

$$\epsilon = \epsilon^{'} - j\epsilon^{''} \quad (2.39)$$

$$= \epsilon^{'} - j(\epsilon_b + \frac{\sigma}{\omega}) = \epsilon^{'}(1 - jtan\delta) \quad (2.40)$$

Then the conduction is:

$$G = \omega C tan\delta \quad (2.41)$$

2.3 Electrical Level Modelling

As long as the wire cross-sectional dimension is much smaller than the wavelength, the signal propagation in the medium can be assumed as TEM or quasi-TEM mode [*71*]. This requirement is generally satisfied for on-chip interconnects. For example, the wave length of a 10 GHz frequency signal is around 30 *cm*, which is several orders of magnitude greater than the cross-sectional dimension of interconnects in nanometer regime. The basic physical structure and electrical and magnetic fields of a general electronic system interconnection were shown in Figure 2.2. Under the TEM mode of propagation, all wires can be generalized as transmission lines which have series resistance and inductance, and parallel capacitance and conductance. The corresponding electrical model for an infinitesimal length (Δx) is depicted in Figure 2.10. In the limit that $\Delta x \to 0$, the model in Figure 2.10 renders the differential equation relating the spatial and time dependence of currents and voltages [*25*] as:

$$-\frac{\partial V}{\partial x} = r_w I + l_w \frac{\partial I}{\partial t} \quad \text{and} \quad (2.42)$$

$$-\frac{\partial I}{\partial x} = g_w I + c_w \frac{\partial V}{\partial t} \quad (2.43)$$

where x is the length, t the time, V the voltage, I the current, and r_w, l_w, g_w and c_w the per unit values. In many practical applications, it is reasonable to assume that

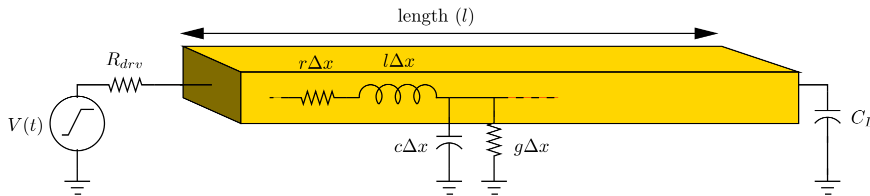


Figure 2.10: *A typical on-chip interconnect with its driver, which is represented by a saturated ramped input voltage source ($V(t)$) and its internal resistance (R_{drv}), and the load capacitance (C_L). Per-unit parasitic parameters r, l, g and c for a infinitesimal section is also shown. Note: the term "wire" refers to the wire on its own, while "interconnect" refers to the wire with its load and driver.*

g_w is negligible, because dielectrics exhibit relatively lossless behaviour in many cases. Combining (2.42) and (2.43) results in:

$$\frac{\partial^2 V}{\partial x^2} = (r_w c_w + l_w g_w)\frac{\partial V}{\partial t} + l_w c_w \frac{\partial^2 V}{\partial x^2} \tag{2.44}$$

The general solution to (2.44) can be written as:

$$v(x,t) = Ae^{-\gamma x} e^{j\omega t} \tag{2.45}$$

where γ is the propagation constant equal to $\sqrt{(g + j\omega c)(r + j\omega l)}$, which provides information about the wire characteristics. The propagation constant can be broken down to its real and imaginary parts; the real part, α is the attenuation constant, and β, the phase velocity. From a practical point of view α describes the way that the signal attenuates when travels along the wire while β represents how fast the signal propagates. The velocity of propagation is $v_p = \frac{\omega}{\beta}$. Then (2.45) can be described in the following form:

$$v(x,t) = Ae^{-\alpha x} e^{j(\omega t + \beta x)} \tag{2.46}$$

In the low frequency RC regime, the velocity of signal propagation is $\sqrt{\frac{\omega}{RC}}$; *i.e.* phase velocity is proportional to $\sqrt{f}$. At high frequencies, $r << \omega l$ and the phase velocity approaches to the speed of light - *i.e.* wire is acting as a waveguide.

This analysis gives us a different physical insight to the wave propagation nature in a wire. In the low frequency RC regime, signals do not attenuate significantly but travel more slowly along the line. As the frequency increases, the velocity increases while the amplitude of the signal decreases due to attenuation; in the high frequency domain (LC-regime), signals travel near their maximum velocity, with severe attenuation.

Each frequency component travels at a different speed along the wire. For short wires, the delay between the fastest and slowest components is negligible compared to the transition time itself. Hence, no discernible difference can be seen at the output. For longer wires, the high frequency components arrive first, though they have undergone severe attenuation. Eventually the lower frequency components, which contain a significant portion of power, catch up later. Hence, we do not get sharp edges toward the end of the wire. It is worth to note that inserting repeaters or boosters amplify the high frequency components of a signal, but does not change the frequency characteristics of the interconnect at all [*72*].

There are two special cases of the above wave equation, which typically arise in electronic systems:

1. *Resistive Interconnections*: In the case of highly resistive wires such as on-chip interconnects, where $r_w >> l_w$, and assuming a perfect dielectric ($g_w \approx 0$), the wave equation reduces to the well-known diffusion equation:

$$\frac{\partial^2 V}{\partial x^2} = rc\frac{\partial V}{\partial t} \tag{2.47}$$

   Sakurai rigorously derives the solution to this condition for a single distributed RC line in [*73*].

2. *Ideal Transmission Line*: When the series inductance dominates the series resistance (i.e. $l_w >> r_w$) and ignoring g_w, the two equations in (2.44) become the standard wave equation:

$$\frac{\partial^2 V}{\partial t^2} = v_p^2 \frac{\partial^2 V}{\partial t^2} \tag{2.48}$$

where v_p is the propagation velocity given by $v_p = \frac{1}{l_w c_w}$. Also, The velocity of propagation along a line under TEM is $v_p = \sqrt{\frac{1}{\mu\epsilon}} = \frac{C}{\sqrt{\epsilon_r}}$, where μ is the permeability of the medium, ϵ_r the relative permit

2.3.1 Choosing a Wire Model

(A) Lumped and Distributed Wire Models

A dimensionless ratio of the physical length of a wire to the signal wavelength, $\frac{l}{\lambda}$, which is referred as the *electrical length* is used to determine whether to model the wire as a lumped or distributed model. A wire is considered to be electrically short if the electrical length is less than unity. These electrically short wires belong to classical circuit analysis and it is quite safe to approximate the entire line as a lumped RC or RLC segment because the signal level along the entire length of the wire is almost constant. A rule of thumb to determine whether a wire can be represented by a lumped circuit or not is to test its length against the following criterion:

$$l \leq \frac{\lambda}{10} \tag{2.49}$$

Alternatively stated, the wire length should be significantly smaller than the shortest wave length, which is equal to $\frac{v_c}{f_T}$. In general, f_T, the highest operating frequency (the cut-off frequency or the corner frequency) is determined by the rise and fall times of the propagated signal. In the case of a simple RC circuit representation of a wire, the cutoff frequency occurs at $\frac{1}{2\pi RC}$. For a input signal with rise time t_r, the rise-time measured between 10% and 90% is $t_r = 2.2RC$, and substituting for RC, the cut off frequency reduces to [*55*, *71*, *74*]:

$$f_T = \frac{0.35}{t_r}. \tag{2.50}$$

That is however from the signaling medium perspective. Even though the frequency spectrum of a trapezoidal pulse is infinite, the energy of the signal is concentrated in the lower part of the spectrum and rapidly decreases with increasing frequencies. To be more specific, approximately 15% of its frequency components are at higher than f_{3dB}, and the magnitude of the pulse sepctrum at the frequencies higher than f_{3dB} is less than 10% of its maximum value [*37*]. For example, for a trapezoidal waveform f_{3dB} is found to be equal to $\frac{0.885}{t_r}$, where t_r is the rise time. Hence, 3 dB bandwidth is not adequate to reconstruct signals and the bandwidths of $\frac{1}{t_r}$ or more is used in accurate interconnect simulations [*55*, *71*].

Nevertheless, due to faster rise times and increasing interconnect lengths, the electrical length of interconnects becomes a significant fraction of the operating wavelength, and transmission line effects must be taken into account. Important

effects like resistive shielding cannot be ignored anymore and lumped models become inadequate because they cannot accurately predict crosstalk, rise time, or delay. Moreover, when dimensions are electrically large, the structure can be broken into a set of electrically small substructures. Each of these substructures is equivalent to a lumped model based on the so-called per unit length parameters.

A frequently used rule of thumb to determine the number of lumped segments is theoretically derived in [*75*] based on the f_T is, the propagation delay caused by a single segment should be smaller than one fifth of the shortest rise time, which is in mathematically:

$$n \geq \frac{5l\sqrt{LC}}{t_{r_{10\%-20\%}}} \tag{2.51}$$

where n is the number of segments. However, usually a five section π ($\pi 5$) model is 99% accurate to the response of a true distributed line [*6*].

(B) When to Consider the Effect of Inductance

At low clock speeds, on-chip interconnects are usually modeled only with lumped or distributed RC elements, whereas for off-chip lines the inductance is very important. When the clock frequency enters the gigahertz regime the contribution to the wire impedance from the inductance (ωL) becomes comparable to the line resistance (R). Inductance and inductive coupling have become important not only in the signal delay estimation but also in the noise analysis of a growing number of on-chip signal lines. Moreover, inductive coupling, along with capacitive coupling, can be a significant source of noise on quiet nets due to the switching of nearby aggressors. On the other hand, the introduction of the inductance to the wire models usually requires complex analysis, it is quite desirable to consider it only when it is important.

A growing body of literature exists which attempts to precisely quantify when inductance effects are important. These simple relations apply to quasi-TEM propagation in a lossy transmission line. While they differ slightly in formulation, the general result is best expressed as one of two equivalent expressions.

The first of these is stipulated by Deutsch *et.al.* in [*76*], which is the error in delay prediction between RC and RLC modelling of a wire exceeds 15% if

$$C_L \ll c_w l \tag{2.52}$$

$$\frac{r_w l}{2Z_0} \leq 1 \tag{2.53}$$

$$Z_{drv} < nZ_0 \tag{2.54}$$

where n is between 0.5 and 1. The acceptable error limit will affect (2.53) and (2.54). A different variant is used for crosstalk prediction; when inductive coupling has to be taken into account in order for crosstalk prediction difference to be greater than 20% between RC and RLC modeling:

$$C_L \ll c_w l \tag{2.55}$$

$$\frac{r_w l}{2Z_0} \leq 1.5 \tag{2.56}$$

$$Z_{drv} < nZ_0 \tag{2.57}$$

with n between 1 and 1.5.

The second, which is very widely used, is stipulated as a combination of two conditions [77]:

1. Is the rise/fall time of the input signal smaller than the time required for the signal round trip from the driver to the end of a line? This condition implies that when the switching is fast enough, the signal transmission is affected by the reflection.
$$t_r < t_{tof} \tag{2.58}$$
Substituting $t_{tof} = \frac{l}{v_c} = l\sqrt{l_w c_w}$, the condition becomes
$$\frac{t_r}{2\sqrt{l_w c_w}} < l \tag{2.59}$$

2. Is the time-of-flight greater than Elmore delay for an RC line, *i.e.* $t_{tof} > \frac{r_w c_w l^2}{2}$. This condition is also described as the wire resistance being smaller than the characteristic impedance.
$$t_{tof} > \frac{r_w c_w l^2}{2} \Rightarrow l < \frac{2}{r_w}\sqrt{\frac{l_w}{c_w}} \tag{2.60}$$
Evidently, this condition gives another view that $\frac{2}{r_w}\sqrt{\frac{l_w}{c_w}}$ represents the damping factor, usually denoted by ξ, of a single section approximation of a wire. If $\xi > 1$, the circuit is overdamped and has small inductance effects. The greater the value of ξ, the more accurate the RC model. However, as ξ becomes less than one, the circuit becomes underdamped and oscillations occur, where inductance cannot be neglected.

Combining the two conditions (2.59) and (2.60), we obtain:
$$\frac{t_r}{2\sqrt{l_w c_w}} < l < \frac{2}{r_w}\sqrt{\frac{l_w}{c_w}} \tag{2.61}$$
However, in the case the constraint on the left-hand side is larger than that on right-hand side, the relation may not exist: $t_r > \frac{4L}{R}$. To elaborate, the combination of rise time and loss is such that short wires have a t_{tof} much less than the rise time (t_r), and long wires have far too much loss for inductance to be important. In such a case, the inductance effect can be ignored regardless of the line length.

Alternatively, the double inequality (2.61) can be interpreted as a bound on the total line inductance L_t. As indicated in [49], the interconnect exhibits non-negligible inductive characteristics if the following two conditions hold :
$$L_w > \frac{1}{4}\frac{t_r^2}{C_w}, \text{and} \tag{2.62}$$
$$L_w > \frac{1}{4}R_w^2 C_w \tag{2.63}$$

The penetrating nature of the magnetic fields causes all on-chip non-orthogonal conductors to be magnetically coupled, and the wire may not be a uniform lossy transmission line. Thus, the conditions used to predict when the inductive effect

is important, involving the characteristic impedance and time-of-flight are not applicable for on-chip interconnects where nonuniform transmission line properties exist [37]. Also, [37] claims that inductive effects become important only when the inductive reactance, $j\omega l_w l$, becomes a significant portion of the total reactance including driver's resistance, $(r_w + j\omega l_w)l + Z_{drv}$. As frequency increases, the inductive reactance becomes increasingly dominant, and larger inductive effects show up. Therefore, the third rule of thumb is designers need to consider the inductance of an interconnect when

$$C_L < \frac{1}{8} c_w l \tag{2.64}$$

$$l \leq \frac{2}{r_w}\sqrt{\frac{l_w}{c_w}} \tag{2.65}$$

$$2\pi f_T l_w l > \frac{r_w l + Z_{drv}}{2} \tag{2.66}$$

Here, f_T is defined in (2.50). Under this criteria, the inductive reactance occupies more than one-third of the total reactance, and the delay and crosstalk errors, without considering inductance, exceed 25%. When the values of C_L and $c_w l$ are very close, the delay and crosstalk errors may exceed 25%, ignoring inductance if.

$$C_L > \frac{1}{8} c_w l \tag{2.67}$$

$$l \leq \frac{2}{r_w}\sqrt{\frac{l_w}{c_w}} \tag{2.68}$$

$$2\pi f_T l_w l > \frac{r_w l + Z_{drv}}{4} \tag{2.69}$$

These conditions apply for on-chip interconnects.

(C) When to Consider Frequency Dependent Effects

The frequency dependant models are of interest for off-chip and microwave circuits due to their larger wire sizes. But, since the chip operating frequency has been approaching to GHz range, these effects have migrated to on-chip interconnects as well. Therefore, modelling wires using constant R and L may not be accurate enough, but modelling them with frequency dependent electrical parameters is complex and requires a lot of computational time. To treat frequency-dependency of interconnect parameters, several circuit models have been proposed. Tsuchiya *et.al.* in [78] propose a representative frequency to extract the parameters of an interconnect. This frequency is based on the rise time of input signal is: $f_{sig} = 0.34/t_r$.

In contrast, [60] claims that it is sufficient to consider DC resistance and inductance values for delay analysis. The reason as they claim is when the operating frequency reaches GHz values, $\omega l_w \gg r_w$, and the skin effect becomes more prominent, and thereby wire resistance increases exponentially and inductance decreases slightly. Also, $R(f)$ and $L(f)$ have opposing dependencies. Therefore, in this regime, delay is more sensitive to wire inductance than the resistance.

2.4 Interconnect Timing Analysis

After the physical information of wires have converted into their electrical representation either as RC or RLC components, performance of the interconnects can be analyzed analytically or using generic circuit simulators. However, due to the prohibitive number of nets and the complex nature of interconnects, it impractical to simulate at the SPICE level to perform timing analysis on an IC. The most practical and widely used method is to represent gate delay using table look-up methods and analyse interconnects using model order reduction techniques, such as the Elmore delay model [*79*], Asymptotic Waveform Evaluation (AWE) [*80*], PRIMA [*81*], or Krylov-subspace based techniques [*82*, *83*]. However, the nature of the problem dictates what is more efficient in determining accurate result.

Signal delay is identified as the primary design parameter in synchronous design since it directly deals with the performance. It is usually measured at the 50% point of signal swing, from the input of the driver to the input of the receiver, and is a function of driver strength and wire loading. With technology scaling, the majority of the wire loading has shifted from metal-to-ground to the coupling capacitance, and hence, it is also a function of the switching activities of the neighbouring wires. For simplicity in analysis the signal delay is decoupled into two parts: gate delay and wire delay. The major benefit of such an approach is that it isolates the nonlinearities from the linearities, because the parasitics that are associated with MOS transistors show significant nonlinearities and the wire parasitics are linear. Each part is usually analysed individually and summed up to obtained the overall timing [†]. The performance of local circuits is usually dominated by the gate delay due to short interconnects, but for global signalling, both the line delay and the gate delay are important for overal timing. It is very important to optimize interconnects with drivers and also repeaters to minimize path delay. For this purpose, proper driver models are paramount for efficient optimization and analysis.

2.4.1 Interconnect Driver Modelling

In order to determine the equivalent delay of a buffer, the complete downstream network is abstracted as an equivalent load, and the delay is then a function of the input transition time and the equivalent load. Key elements of this methodology are the estimation of the equivalent (or effective) load and the delay of the wire. For instance, in typical gate delay analysis, RC or RLC interconnects are usually approximated as an equivalent capacitance or single section pi network, using model order reduction techniques such as AWE. In many libraries, gate and cell delays are usually pre-characterized for static timing analysis to shorten design cycles and reduce costs. In general, gate and cell delays as well as slews are expressed as an empirical function of load and input slew or a look-up table [*84*].

Basically, Interconnect driver (or gate) modeling has two widespread accepted approaches [*84*]: empirically derived expressions or look-up tables for delay and output signal transition as a function of input-signal transition time; and a switch-resistor method or the Thevenin equivalent model comprised of a linear resistor

[†] note that the receiver is usually modeled as a loading capacitor at the end of the line.

and a time-variant voltage source[‡]. The main benefit of the second method is the inherent modeling of the coupling with an RC interconnect. However, the drawback comes from the fact that only a single resistor is used for capturing the gate switching behavior. This leads to inaccuracies in the prediction of the slew rate of gate output signal, especially when its input slew rate and loading capacitance vary significantly over a wide range. In order to cope with this issue, more complex models consisting of time and slew dependent non-linear resistances have been proposed [85, 86]. In practical RC analysis, the values of R_{drv} and C_{eff} are found for two points of the gate output waveform (e.g. 50% and 90%). The following sections outlines a methodology used in this thesis to extract a Thevenin driver model.

(A) Device Resistance

The device resistance of a MOS transistor is a nonlinear function of the supply voltage and gate-source voltage, and different approximations have been used to determine it in the literature. In most digital designs, the transistor is assumed to be a switch with an infinite off-resistance, and a finite on-resistance R_{dev}. The first-order approximation for R_{dev} is [6]:

$$R_{dev} = \frac{1}{\mu C_{ox}\left(\frac{W}{L}\right)(V_{dd} - |V_t|)}$$

However, R_{dev} is time-variant and non-linear depending upon the operating point of the transistor and therefore, this approximation is valid only when the transistor is in the active region, before saturation occurs. In [28] a method is stipulated to find a more accurate value for R_{dev}, assuming a constant and linear value for R_{dev} while switching between different logic states. A reasonable approach is to use the average value of the resistance over the operating region of interest; simply, take the average value of the resistances at the end-points of the transition.

$$R_{dev} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{V_{DS}(t)}{I_D(t)} \tag{2.70}$$

This assumption works well only if the resistance does not experience any non-linearities over the range of interest. Simply according to that method, R_{dev} is the average value of R_{mid} and R_o from the simulated I-V curves for a MOS transistor, as shown in Figure 2.11. Alternatively it can be estimated with the aid of (2.70) as:

$$R_{dev} = \frac{1}{0.5V_{DD}}\int_{0.5V_{DD}}^{V_{DD}} \frac{V}{I_{dsat}(1+\lambda V)}dV \tag{2.71}$$

$$\approx \frac{3}{4}\frac{V_{DD}}{I_{dsat}}\left(1 - \frac{7}{9}\lambda V_{DD}\right), \tag{2.72}$$

where I_{dsat} is the saturation current.

[‡]More accurate waveform analysis can be performed by using a time-varying current source model which captures the gates behaviour over the entire input range.

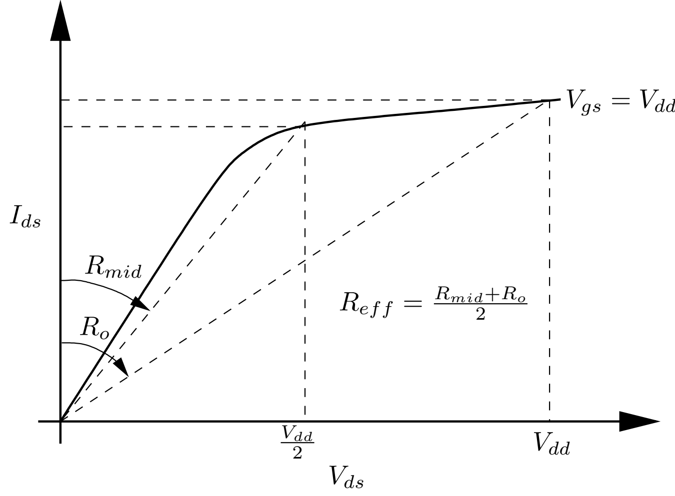


Figure 2.11: *Estimating Device Resistance.*

Another method of estimating R_{dev} is by considering the equivalence between a lumped RC network, where the 50% delay is defined as $0.69RC$, and the actual driver as a current source charging/discharging the load capacitance, in which the 50% delay point is set by [87]

$$C_{load}\frac{V_{swing}}{2I_{supply}}.$$

If we assume that the driver supplies αI_{dsat}, where I_{dsat} is the saturation current throughout the voltage swing of interest, the above two equations are equal and gives us the effective device resistance. (The parameter α represents the fact that the saturation current will not flow throughout the voltage swing of interest. However, due to velocity saturation effects this value α is close to 1.)

$$R_{dev} = \frac{0.5C_L V_{dd}}{0.69C_L \alpha I_{dsat}}$$

From the simulated $I_{ds} - V_{ds}$ curves for 0.18 μm technology it has been found that the constant α is approximately 0.9, and the device resistance for future technologies is found using the relation:

$$R_{dev} = 0.805\frac{V_{dd}}{I_{sat}} \tag{2.73}$$

For the same technology the device resistance is experimentally estimated by Spectre (or Spice) simulations by loading an inverter with a capacitor, C_L and driving it by a step input. As is mentioned in this section, the 50% delay of this inverter, t_d, is equal to the lumped RC network delay ($0.69R_{dev}C_L$).

$$t_d = t_{self} + 0.69R_{dev}C_L$$

For two different loadings, C_{L1} and C_{L2}, t_{d1} and t_{d2} can be measured and then the R_{dev} is derivable as follows:

$$R_{dev} = \frac{t_{d1} - t_{d2}}{0.69(C_{L1} - C_{L2})} \tag{2.74}$$

(B) Device Capacitance

Estimating both input and junction capacitances is straightforward from the layout geometries.

Device Input Capacitance: Ideal MOSFET input capacitance or gate oxide capacitance, C_{ox}, is defined as:

$$C_{ox} = \frac{\epsilon_{ox} W L_{eff}}{T_o x}$$

where W and L_{eff} are the width and effective channel length of the transistor, and ϵ_{ox} and T_{ox} represent the dielectric constant for the gate dielectric and gate oxide thickness, respectively. Because of the overlap of the source and drain with the gate, the effective gate input capacitance, C_{in}, is given by [87]:

$$\begin{aligned} C_{in} &= C_{ox} + C_{overlap} \\ &= \frac{\epsilon_{sio2}\epsilon_0 W L_{eff}}{T_{ox}} + (C_{GD0} + C_{GS0})W \end{aligned} \tag{2.75}$$

Junction (Parasitic Source/Drain Diffusion) Capacitance: The junction capacitance originates from the ionized dopants in the vicinity of the source and drain junctions. The source and drain diffusion regions have a capacitance to the substrate that depends on the voltage between the diffusion regions and substrate or well and the "base" area and "perimeter" of these regions. The model generally expresses the total diffusion capacitance for a source or drain area at zero DC bias across the junction as:

$$C_j = C_{jbase} A_{base} + C_{jperiphery} P_{junction}$$

where C_{jbase} is the junction capacitance per unit area and $C_{jperiphery}$ the periphery capacitance per unit length, while A_{base} and $P_{junction}$ are the area of the base and the perimeter of the region excluding the gate side, respectively. According to typical layout practices, the drain/source must contain a square contact which has side length of L_{eff} and ensuring a spacing of $L_{eff}/2$ on either side of the contact. Overall, these design rules result in a drain/source length of $2.5L_{eff}$.

In reality C_{jbase} and $C_{jperiphery}$ are a function of the junction voltage, V_j, which determines the actual thickness of the junction depletion layer. This is generally expressed as:

$$C_j = \frac{C_{j0}}{\left(1 + \frac{V_R}{\phi_{o,m}}\right)^m}$$

where C_{j0} represents the zero bias junction capacitance ($V_R = 0V$) and $\phi_{o,m}$ is the built-in junction potential, which is typically in the range of $0.5 - 0.7V$. The value of grading coefficient m depends on the junction doping profile and usually $m < 1$. For abrupt junctions, such as the bottom area of the diffusion region, $m = 0.5$ corresponding to a square root dependence, while a graded junctions, such as sidewall areas of the diffusion regions are described by a cubed root dependence

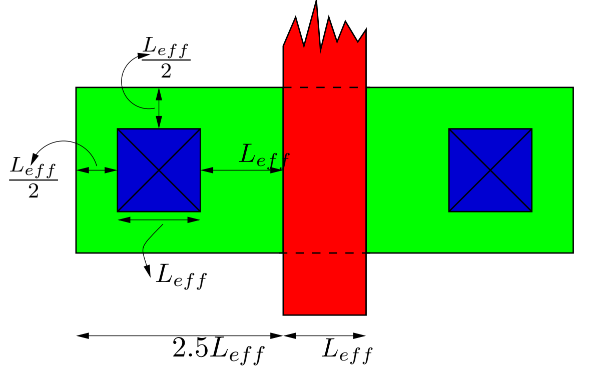


Figure 2.12: *MOS transistor Layout and typical minimum distances.*

with $m = 1/3$. Back of the envelope calculations, it is quite common to approximate the bottom junction as being step-like and the sidewalls linearly graded. The average junction capacitance for source/drain is usually approximated by setting $V_R = V_{dd}/2$ and is of the form [87]:

$$C_j = \frac{C_{j0}A_d}{\left(1 + \frac{V_R}{\phi_o}\right)^{m_j}} + \frac{C_{j0sw}P_d}{\left(1 + \frac{V_{dd}}{\phi_{osw}}\right)^{m_{jsw}}} \tag{2.76}$$

where

$$A_d = 2.5L_{eff}W \tag{2.77}$$

$$P_d = 5L_{eff} + W \tag{2.78}$$

The junction capacitance values C_{j0} and C_{j0sw} are a function of the substrate doping. However, the junction values for PMOS and NMOS can be assumed to be the same in this analysis.

2.4.2 Interconnect Delay Modelling

First Order Delay Models: In the early days of IC design, interconnect was treated as a lumped capacitor and the RC constant of the interconnect delay is estimated as:

$$\tau = R_{drv}(cl + C_L), \tag{2.79}$$

which is valid only when the driver resistance overwhelms the wire resistance. This is still valid for short local on-chip interconnects.

When the driver resistance and wire resistance are comparable, resistive shielding causes the delay at the driver output to be equivalent to a situation where it drives a lumped load that is less than the total capacitance of the interconnect.

Closed-form Elmore Delay Metric: Typically, global on-chip wires are becoming highly resistive with feature size reduction. Hence, obviously the lumped capacitor model

ignores the resistive shielding effect of the interconnect resistance. When the inductance is negligible, signal propagation obeys the diffusion equation, which does not lend itself to a simple closed from solution and require approximate solutions [*6*, *73*]. The Elmore delay provides a useful technique for estimating the delay of circuits whose response is well-captured by a dominant time constant. As the complexity of interconnect structures increase the accuracy of the Elmore delay drops. Therefore models such as AWE [*80*] , PRIMA [*81*] surfaced in the interconnect modelling domain. However, Elmore metric corresponds to a first order AWE approximation of a circuit, where multiple time constants have been used to capture behaviour of RC network, and provides an upperbound for the delay.

The RC time constant of circuit with a cascaded N-stage RC chain can be approximated by the Elmore delay [*79*].

$$\tau = \sum_{i=1}^{N} R_i \sum_{j=1}^{N} C_j = \sum_{i=1}^{N} C_i \sum_{j=1}^{N} R_j \tag{2.80}$$

The Elmore time constant for an interconnect with a sufficiently large number of distributed sections and driver resistance R_{drv} and a load capacitance C_L is:

$$\tau_r c = (R_{drv} c_w + r_w C_L)\, l + R_{drv} C_L + 0.5 r_w c_w l^2 \tag{2.81}$$

The factor 0.5 in the last term is from the distributed nature of the wire's resistance and capacitance. Some scale factors to the Elmore time delay constant have been proposed to predict the delay as accurately as possible [*6*]. Under a step voltage excitation, the times required for the output voltage at the far end of lumped and distributed RC networks to rise from 0 to 50% is $0.7RC$ and $0.4RC$, repectively. The 50% delay for a wire with driver resistance R_{drv} and load capacitance C_L can be written as:

$$\tau = 0.7 R_{drv}(cl + C_L) + 0.7 rlC_L + 0.4 rcl^2 \tag{2.82}$$

Expressing the solution for (2.47) in series expanded form and approximating

Output potential range	Time Elapsed	
	Distributed RC	Lumped RC
0% to 90%	1.0RC	2.3RC
10% to 90% (rise time)	0.9RC	2.2RC
0% to 63% (time constant)	0.5RC	1.0RC
0 to 50% (delay)	0.4RC	0.7RC
0% to 10%	0.1RC	0.1RC

Table 2.2: *The time delays between commonly used reference points in the output potential [6].*

a single-exponent, [*73*] provides a closed-form solution for (2.47) for the far-end voltage. That solution could be used to estimate the time elapsed to reach any voltage level. ate the time elapsed to reach any voltage level.

$$\frac{V(l,t)}{V_{DD}} = 1 - exp\left(-\frac{\frac{t}{RC} - 0.1}{R_T C_T + R_T + C_T + 0.4}\right) \tag{2.83}$$

where $R = rl$, $C = cl$, $R_T = \frac{R_{drv}}{R}$, and $C_T = \frac{C_L}{C}$. By solving t in terms of v, a delay expression can be obtained for the delay from $t = 0$ to the time when the normalized voltage at the receiving end reaches $v(= V/V_{DD})$. t_v can be expressed as

$$t = 0.1rcl^2 + ln\left(\frac{1}{1-v}\right)\left[R_{drv}(cl + C_L) + rlC_L + 0.4rcl^2\right] \tag{2.84}$$

For special values of v, that is, for 0.9 and 0.5, the following formulae can be obtained:

$$t_{0.9} = 1.02rcl^2 + 2.3(R_{drv}(cl + C_L) + rlC_L) \tag{2.85}$$

$$t_{0.5} = 0.377rcl^2 + 0.693(R_{drv}(cl + C_L) + rlC_L) \tag{2.86}$$

These two expressions are identical to the expressions given in (2.82), and Table 2.2.

Second and Higher order LRC Delay models: Inter-chip wires on a typical package substrate are characterized by low-loss dielectrics and by conductors with low resistivity and a large cross section, making losses due to shunt conductance negligible - *i.e.* obeys transmission line behaviour. The RC model is inadequate to accurately model delay of these RLC transmission lines, because the RC model is more suitable for modeling higher-order under-damped systems. In lossy transmission lines, both RC and LC delays co-exist. For LC dominated wires, the signal propagation delay is equal to it's time-of-flight, representing the time required for a signal traveling from one place to another at the wave velocity.

$$t_{LC} = t_{tof} = L\sqrt{l_w c_w} \tag{2.87}$$

If a wire is a very resistive transmission line, the following empirical formula for adding time-of-flight (t_{tof}) and conventional RC delay (t_{rc}) was found to predict the total wire delay well [*88*].

$$t_{RLC} = (t_{tof}^{1.6} + t_{rc}^{1.6})^{\frac{1}{1.6}} \tag{2.88}$$

Alternatively, [*89*] has proposed an empirical formula for the propagation delay of a RLC transmission line:

$$t_{RLC} = \frac{e^{-2.9\zeta^{1.35}}}{\omega_n} + 0.74r_w c_w L^2(R_T + C_T + R_T C_T + 0.5) \tag{2.89}$$

where $\zeta = \frac{r_w L}{2}\sqrt{\frac{c_w}{l_w}}\frac{R_T + C_T + R_T C_T + 0.5}{\sqrt{1+C_T}}$, $\omega_n = \frac{1}{\sqrt{l_w L(c_w L + C_L)}}$,$C_T = \frac{C_L}{cl}$, and $R_T = \frac{R_{tr}}{r_w L}$.

2.4.3 Noise-on-Delay Effect

In nanometer technologies due to the scaling of line widths, increasing aspect ratios, tight integration of wires, and larger die sizes increase the coupling between wires. This leads to two effects in terms of interconnect or circuit performance: Crosstalk

Noise, and Noise-on-Delay Effect (Dynamic Delay). Crosstalk noise corresponds to noise bumps that are injected to an adjacent silent wire from switching wire, or wires. The terminology that is frequently used in the context of crosstalk analysis is to label the wire under consideration as the *victim*, and label any wires that couple to it as *aggressors*.

Crosstalk noise analysis is very important due to shrinking noise margins, and have to be considered at nearly every stage of high-speed circuit design in order to reduce the number of expensive design iterations and ensure a successful design. Therefore, computationally efficient and accurate crosstalk models are predominantly important to quickly identify the nets that violate noise margins because a full-chip analysis scenario consists of a prohibitive amount of aggressor/victim combinations. Therefore, a detailed simulation of crosstalk noise on a victim is highly inefficient and time consuming.

Two major metrics are typically used to evaluate the impact of noise: Noise peak (V_{peak}) and noise width. V_{peak} describes the maximum amount of crosstalk noise between two nets, and its value depends on the coupling capacitance, other loading capacitances and parasitic resistances, and slew rate of the aggressor, and the victim driver strength. The noise width represents the length of time that the value of the noise is larger than a given threshold. However, the most basic model to estimate the crosstalk voltage is the charge-sharing model presented in much of the literature [*28*]. It is of the form:

$$V_{peak} = \frac{C_c}{C_c + C_{gv}} V_{DD}, \tag{2.90}$$

which is the upper bound of charge sharing, and valid only when the victim line is highly resistive and the aggressor is switching very fast.

Even if V_{peak} exceeds a certain threshold, the receiver may still be immune to noise when the noise bump has a very narrow width and the receiver capacitance is large. For this reason, the noise width is of a paramount importance in capturing the overall performance effect.

To effectively capture the effect of coupling noise from adjacent wires on delay, the most accurate method is to solve the coupled wire differential equation using model decomposition. However, a switch-factor based decoupled model is significantly more computationally efficient and widely used in delay estimation early in the design cycle of state-of-the-art VLSI circuits. For capacitively coupled nets, a coupling capacitance between two wires can be modeled as an effective capacitance to the ground together with the wire's self capacitance. This is a widely used theorem in electronics known as Miller's Effect [*90*]. Mathematically, the effective wire capacitance can be represented as $C_s + \lambda C_c$, where λ is a switch-factor. λ is totally dependent on the signal pattern on the neighbouring conductors (*i.e.* in-phase, quiet, or out-of-phase); generally it is between $(0, 2)$ for a victim with one aggressor, and $(0, 4)$ with two aggressors. However, Kahng *et.al.* in [*91*] show that the effective capacitance depends on the delay offset between signals and the their slew rates, and hence the switch factor takes values between $(-1, 3)$ for two coupled nets. This approximation is inaccurate since it assumes a constant slope for the voltage waveforms. Ghoneima and Ismail, in [*92*], obtain Miller coupling factors based on exponential waveforms. In addition to that [*20*] gives some empirical switch-factors claimed to be accurate within 3-4%.

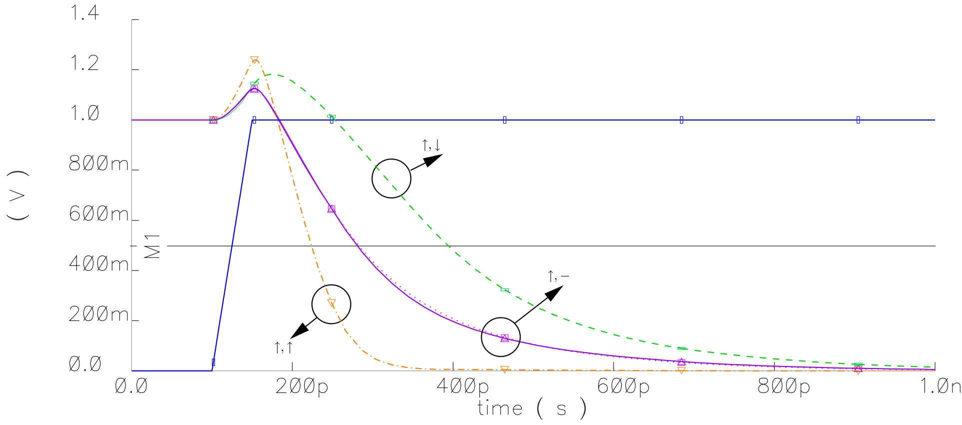


Figure 2.13: *Crosstalk noise on delay effect, or the dynamic delay for a 5mm long interconnect. In this case only one aggressor has considered. Victim input is going on upward transition and cases are shown when the aggressor is switching from 0 to 1 ($\uparrow,\uparrow$), quiet at 1 or 0 ($\uparrow,-$), and 1 to 0 ($\uparrow,\downarrow$).*

	Switching Pattern		
	$(\uparrow\uparrow),(\downarrow\downarrow)$	$(\uparrow -),(\downarrow -)$	$(\uparrow\downarrow),(\downarrow\uparrow)$
Traditional	0	1	2
Kahng *et.al.* [*91*]	-1	1	3
Pamunuwa *et.al.*[*20*]	0	0.65	2.2
Ghoneima *et.al.*[*92*]	-1.885		3.885

Table 2.3: *Switch Factor Comparison*

Moreover, since the electric field is effectively shielded by metal lines, when there are multiple lines on the same layer, the capacitive coupling rapidly decays increasing neighbours. In order to make the analysis simple while maintaining sufficient accuracy, just the nearest neighbours can be considered [*93*]. Introducing shielding lines effectively reduces the capacitive coupling noise.

It is worthwhile to discuss the nature of inductive coupling noise as well. The fundamental difference between inductive and capacitive coupling noise is that capacitive crosstalk noise always occurs in the same direction as the aggressor switches, whereas inductive coupling noise is induced through the return current, which opposes the direction of the aggressor switching and occurs more instantaneously than capacitive coupling noise. Moreover, since the return current induced by inductive coupling spreads over a long range, even farther wires may suffer from inductive crosstalk.

To accurately capture the mutual inductance coupling effect on delay, the tradition is to use switch factors -1 and +1 for two wires switching in opposite and the same direction respectively. Hence, the effective inductance term can be written as $L_s \pm L_m$. Though it is simple to estimate switch factors for capacitive coupling and inductive coupling for a victim with one or two aggressors, to capture the effect

of many neighbours switching with long range inductive coupling needs a complex analysis. The work in [94] presents a more physical methodology to estimate an empirical switch-factor which is used to solve for loop RLC parameters. They generalize the switch factor based decoupling approach to multiple RLC line conditions based on circuit theory.

2.5 Interconnect Energy Dissipation Analysis

Usually repeaters are inserted along on-chip global interconnects to reduce delay which otherwise is proportional to the square of the wire's length. As mentioned earlier, these repeaters consume a large amount of the total power consumption of a chip. In this section the modelling of energy consumption in a repeater is discussed. Typically energy dissipation in a CMOS circuit is categorized into three major components: dynamic, short-circuit and leakage. Dynamic and short-circuit energy dissipation occurs only during switching events whereas leakage energy is static.

2.5.1 Switching Energy

Each time a wire is driven from 0 to V_{DD}, an energy amounting to $C_{eff}V_{dd}^2$ is drawn from the power supply, where C_{eff} is the total effective load capacitance which includes downstream wire capacitance. Half of this is stored in the load capacitance while the rest is dissipated in the pull-up network of the driver. During a V_{DD} to 0 transition, the energy stored in the capacitance is dissipated in the pull-down network of the driver.

The average dynamic energy dissipation for a switching event is given by:

$$E_{dyn} = \frac{1}{2}C_{eff}V_{DD}^2 \qquad (2.91)$$

The energy dissipation per cycle depends on whether or not switching transitions occur, and on the relative switching pattern of neighbouring wires as well. Switching energy component is at about 70-90% of total energy consumption and most of the time designers pay attension to reducing this component.

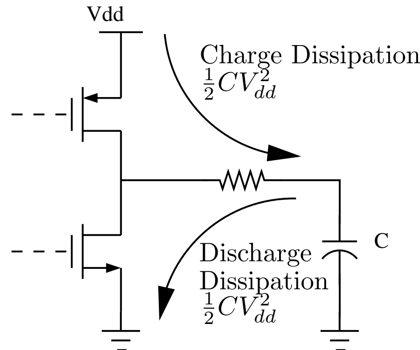


Figure 2.14: *Charge and Discharge Dissipation paths of a CMOS inverter.*

2.5.2 Short Circuit Energy

Due to the finite slew rate of signals, a brief period exists when both the PMOS and NMOS devices in the inverter structures are simultaneously on, resulting in the flow of a short-circuit current. With reference to Figure 2.15, when an upward transition is applied to the input of an inverter, the NMOS transistor starts conducting as soon as the input signal passes V_{thn}, but the PMOS transistor continues to conduct until the input signal passes the value of $V_{DD} - V_{thp}$. Hence a direct conducting path exists from V_{DD} to ground during this period. There is also a similar flow of short-circuit current for a downward transition at the input.

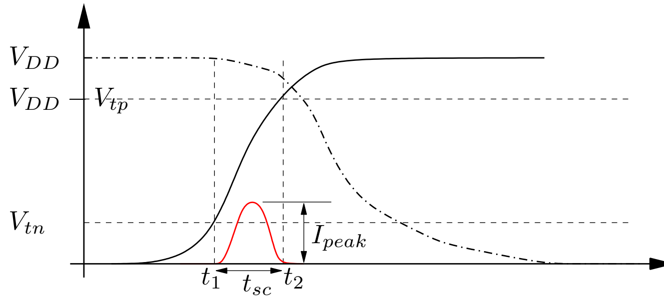


Figure 2.15: *Voltage and current waveforms of a CMOS gate.*

Many analytical models for short circuit energy estimation have been proposed in the literature, and the most suitable model for our purpose is adopted here. The current spike is assumed to be a triangle with a peak I_{peak}, and a base t_{sc}. Hence the total short-circuit charge is $I_{peak}t_{sc}$, and the short-circuit energy is given by:

$$E_{SC_{l\rightarrow h}} = \frac{1}{2} I_{peak} t_{sc} V_{DD} \tag{2.92}$$

The peak short-circuit current is found using the Alpha-Power law [*95*], with $\alpha = 1.3$, which is a typical value for current technologies [*87*]. Hence the peak current is:

$$I_{peak} = I_{dsat} \left(\frac{V_{gs} - V_t}{V_{DD} - V_t} \right)^{1.3} \tag{2.93}$$

The saturation voltage of short-channel MOS transistors is much smaller than that of long-channel devices $V_{gs} - V_t$. Here it is assumed that when $V_{gs} = V_{DD}/2$, the drain-saturated voltage is approximately $V_{DD}/8$ which is enough to saturate the device [*87*]. Similarly as is discussed in [*87*] t_{sc}, the time that the short-circuit current flows is found using Sakurai's delay formula described in Section 2.4.2 [*95*]. This equation can be used to calculate t_1 and t_2 in Figure 2.15 by substituting $v = V_{tn}$ and $v = V_{DD} - V_{tp}$ respectively. Then t_{sc} is simply $t_2 - t_1$ and results in:

$$t_{sc} = ln \left(\frac{V_{DD} - V_{tn}}{V_{tp}} \right) [R_d(C_d + C_g + C_w) + R_w C_g + 0.4 R_w C_w] \tag{2.94}$$

For current and future technologies V_{th} values are usually estimated as 20%-30% of Vdd [*87*]. Hence in this analysis it is assumed that $V_{th} = V_{DD}/4$, and $ln(V_{DD} -$

$V_{tn})/V_{tp}$ is equal to 1.09. The device resistance (R_d) is estimated by equating the delays over the required switching range of the transistor and a lumped RC model as described in Section 2.4.1(A).

2.5.3 Leakage or Static Energy

In the absence of switching activity (*i.e.* in the steady-state), a leakage current flows through the reverse-biased diode junctions of the transistors, located between the source or drain and the substrate. This contribution is very small compared to the switching current but the junction leakage currents are caused by thermally generated carriers, and it increases exponentially with increasing temperature [*96*].

Leakage energy is expected to dominate the overall energy consumption as the technology scales; as [*97*] predicts, leakage will be a significant portion of the total energy consumption which increases approximately 5× in each technology generation. If not properly addressed, this is going to make a major impact in nanoscale IC design.

The average leakage energy of a MOS transistor is given by

$$E_{leakage} = \frac{V_{DD} I_{leakage}}{f_{clk}} = \frac{V_{DD}(I_{offn} W_n + I_{offp} W_p)}{2 f_{clk}} \tag{2.95}$$

where $I_{leakage}$ is the subthreshold current, f_{clk} clock frequency, I_{offn} and I_{offp} are leakage current in NMOS and PMOS transistors respectively, and W_n and W_p are the sizes of NMOS and PMOS transistors. The subthreshold current can be computed when $V_{gs} \approx 0$ by [*98*]:

$$I_{leakage} = I_0 e^{\frac{(V_{gs} - V_{th})}{m V_T}} (1 - e^{\frac{-V_{ds}}{V_T}}) \tag{2.96}$$

where $I_0 = \mu_0 C_{ox} \frac{W}{L} V_T e^{1.8}$, V_T is the thermal voltage, and m the subthreshold slope coefficient.

The expression (2.95) is multiplied by a factor of $\frac{1}{2}$ to consider the effect of switching, because on average, half the drivers will have an input of logic high, while the other half will have logic low. The logic high signal turns on the NMOS network of the driver and the leakage current is determined by the PMOS network; the opposite happens for the drivers with a low input signal. In CMOS logic circuit design usually the width of PMOS and NMOS transistors are adjusted so that there is an equal amount of charging and discharging current at the load. Hence $I_{offn} \approx I_{offp}$.

2.6 Summary

Wires are not ideal as drawn in schematic diagrams but a parasitic element which exhibits undesired effects that hinders the performance of electronic systems. These non-idealities are usually captured by computing the electromagnetic bahaviour of a wire using a tool set known as field solvers. This requires expensive simulations, and requires lot of computational time. One way to reduce the complexity is to partition the problem into a set of geometry dependent parasitics, and solving a discrete electrical network made up of parasitic elements. The basic requirement in this partitioning is to allow both efficiency in simulation and the required accuracy.

Parasitic extraction basically pertains to calculating equivalent resistance, capacitance, and inductance for a given structure to build the electrical network. Extracting resistance is straightforward, as DC resistance is quite adequate for many cases. In contrast to that capacitance and inductance extraction have a high geometry dependence. Usually inductive coupling is long range and the coupling matrix for a multi-conductor system is fully populated whereas capacitive coupling is short range and the matrix is sparse. Although based on many assumptions and simplifications, closed-form formulae are sufficient for most occasions.

Electrical models of wires take different forms depending on accuracy and computational complexity. Major questions in selecting an electrical model is when to consider inductance and the frequency dependency in the models. Nevertheless a distributed RC model with DC parasitic parameters are adequate for many on-chip wires, and a lossy transmission model is suitable for off-chip wires.

Interconnect performance analysis methodologies are of utmost important in successful physical design optimization and therefore, efficient yet accurate models are required to estimate performance metrics such as delay, crosstalk, energy and bandwidth.

3

Electrical Modelling of Through-Silicon Vias

This chapter discusses general methodology that can be used to obtain closed-form equations for TSV parasitics in terms of physical dimensions and material properties. The proposed equations allow electrical modelling of TSV bundles without the need for computationally expensive field-solvers, within an error margin that makes them suitable for the system-conceptual studies in typical 3-D IC design flows.

3.1 Introduction

Stacking multiple processed chips or dies on top of each other into a vertical structure provides opportunities for improving performance, for heterogeneous integration, and for reducing form factor [*39*]. Stacking for example processor and memory blocks vertically in the same neighbourhood, the performance of the overall system can greatly be improved.

Wire bonding has been used as an interconnection between the stacked devices and the circuit board. However, this technique is not appropriate for high performance applications, because it causes several disadvantages such as limitation of chip size reduction, and deterioration of signal integrity and high frequency characteristics, and lower density. Therefore, a new vertical interconnect methodology, through silicon via (TSV) based wafer-level integration (WLI) has been proposed [*99*, *100*, *101*, *102*]. TSVs route the signal and power supply links through all chips in the stack vertically (TSV based WLI process and alternative layer-to-layer signal transmission methods are discussed in Section 4.4.).

In order to analyze the electrical characteristics such as delay, Signal Integrity (SI), and Power Integrity (PI) of a 3-D stacked chip, simulating the entire structure in a field solver would take an unacceptable amount of computational time. To reduce design time and mirror well established practices, it is desirable to model the whole physical structure as a collection of parasitic parameters in equivalent circuits. Chapter 2 has reviewed the already existing parasitic estimation techniques for horizontal wire structures, but a comprehensive set of self-consistent compact models for capacitance and inductance extraction in a TSV bundle do not currently exist. Therefore, compact models to obtain TSV parasitics more efficiently than with the use of a computationally expensive field solver are essential

in simulation-based explorations at the system-conceptual level in 3-D IC design. This chapter outlines the trends in TSV parasitics, a methodology for generation of compact models to estimate the parasitics and proposes a set of analytic equations for estimation of the various parameters.

Recently, parasitic modelling of TSVs and investigation of signal transmission characteristics in a TSV has received some attention in the literature. Alam et.al. in [*103*] have used closed-form equations to estimate TSV resistance and capacitance values as functions of their geometric parameters. These models have not been validated thorough field-solver based simulation or experimental results. There are a few works which provide high frequency parasitic models for TSVs [*102*, *104*, *105*], but these do not report any DC parameter models. A recent work [*106*] proposed an empirical delay model, but no explicit formulae are given for parasitic parameter estimation. Although the propagation delay of a TSV is a function of its physical dimensions, the more useful formulation would be to describe the TSV using its equivalent circuit. This provides circuit intuition that allows not only propagation delay calculation, but also power and energy calculations and signal integrity analyses. Another work [*107*], published very recently, discusses trends in TSV parasitics for a specific structure in the MIT Lincoln Lab 3-D integration process. It presents a thorough study, articulating capacitive and inductive time constants and loop inductance behaviour, but does not in general give a clear methodology to estimate parasitics without the help of a field solver.

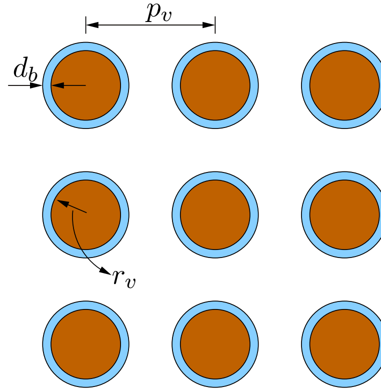


Figure 3.1: *Top view of a TSV bundle. Here, r_v is the radius of TSV, d_b is the thickness of dielectric barrier, and p_v is the pitch.*

3.2 TSV Specification and Physical Modelling

Various forms of TSV processes have been proposed by the academia and the industry through out the past few years, but still a well established technology is not known for the designers. However, the general structure of a TSV is assumed to have a uniform circular cross-section. The material used for TSVs is Cu, with an annular dielectric barrier typically of silicon dioxide (SiO_2) or Silicon Nitride (Si_3N_4) surrounding the copper cylinder. Further, a thin annular Titanium Nitride (TiN) layer is usually deposited between the Cu and SiO_2 layers, which acts as

an adhesion layer [*108*]. This TiN barrier layer has been neglected for the sake of simplicity and to reduce computational time in the field solver, since its inclusion has an apparently negligible effect on the parasitic parameters. Also, the high resistivity in TiN region will staunch the current flow in it, and concentrated in copper bar. The notation used to represent the TSV physical dimensions, as well as their simulated ranges where relevant is shown in Table 3.1. The geometrical quantities in Table 3.1 refer to those specified in Figure 3.1.

In a 3-D chip stack, the likely configuration for TSVs is in a regular matrix, for which a representative unit is a 3×3 bundle. Such a structure has been simulated in a 3-D/2-D quasi-static electromagnetic-field solver specifically used for parasitic extraction of electronic components [*32*]. This tool utilizes the Finite Element Method (FEM) and the Method of Moments (MoM) to solve Maxwell's equations and estimate RLGC parameters of a structure.

Notation	Description	Simulated range
r_v	TSV radius	$10\mu m - 40\mu m$
l_v	TSV length or height	$20\mu m - 140\mu m$
d_b	SiO_2 dielectric barrier thickness	$0\mu m - 1\mu m$
s_v	Separation of two TSVs	$40\mu m - 200\mu m$
p_v	Pitch of TSVs ($p_v = s_v + 2(r_v + d_b)$)	
σ	Conductivity of Bulk Copper	$58 \times 10^{-6}\ S/m$
ϵ_{SiO2}	Relative permittivity of SiO_2	3.9
ϵ_{Si}	Relative permittivity of Si	11.9
ϵ_0	Permittivity of air	$\frac{1}{36\pi} \times 10^{-9}\ F/m$
μ_0	Permeability of air	$4\pi \times 10^{-7}\ H/m$
R_{tsv}	Resistance of a TSV	
C_{tsv}	Capacitance of an isolated TSV	
L_{tsv}	Inductance (self) of an isolated TSV	

Table 3.1: *Notations and simulated ranges of physical dimensions*

3.3 Trends in Parasitic Parameter Values

In this section, we discuss trends in TSV parasitic parameter values for three different configurations: for an isolated TSV, two parallel TSVs, and a 3×3 TSV bundle. The simulated data for selected points are plotted to give the reader a qualitative sense of the variation of the parasitic parameters with relevant variables.

3.3.1 Isolated TSV

The equivalent electrical circuit diagram for an isolated TSV is a conventional T-model wire segment including parasitic resistance, inductance and capacitance to ground, as shown in Figure 3.2.

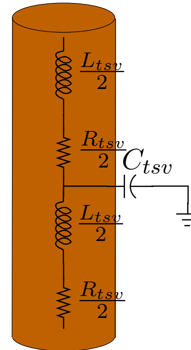


Figure 3.2: *Equivalent Circuit for a single TSV.*

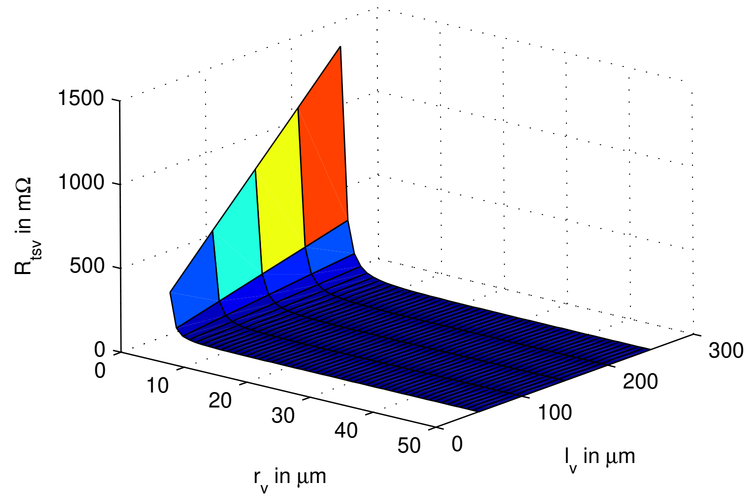


Figure 3.3: *Variation of TSV Resistance with its radius and length.*

(A) Resistance

The variation of resistance shows the expected linear inverse and proportional relationships with cross-section and length respectively (see Figure 3.3). This accounts for the quadratic dependence of resistance on radius, for a given length. For a given radius, the resistance increases linearly with length.

(B) Capacitance

Figure 3.4 shows the variation of TSV capacitance with radius and length when $d_b = 0.2\ \mu m$. As both radius and length increase, capacitance increases monotonically. This can be seen both in the surface plot and the line plots. This calculation of capacitance assumes a reference at infinity.

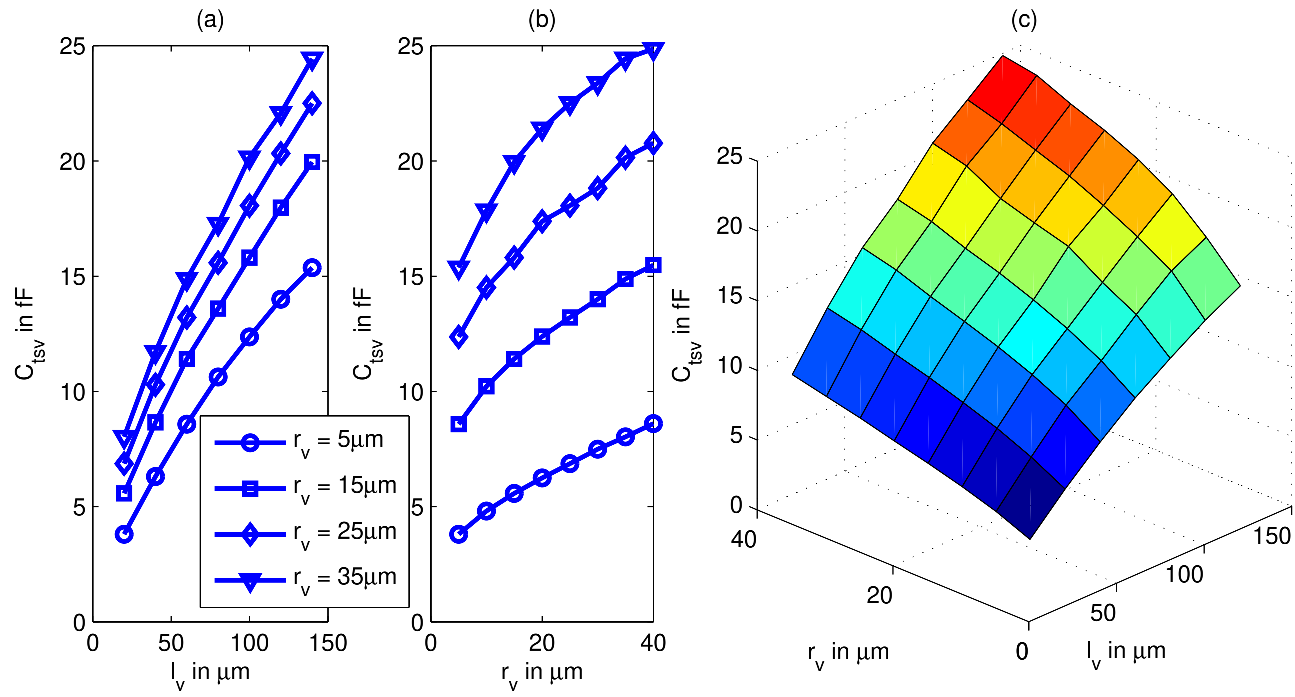


Figure 3.4: *Variation of TSV Capacitance with radius and length when* $d_b = 0.2\ \mu m$.

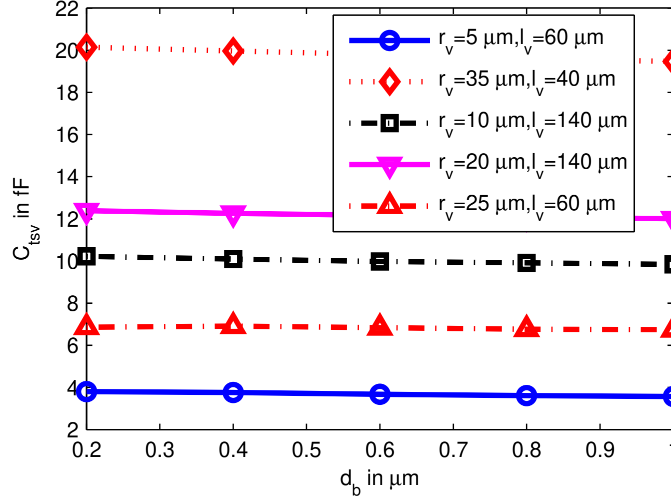


Figure 3.5: *Variation of TSV capacitance with d_b for various radii and lengths.*

In Figure **??**, for various length and radii combinations, TSV capacitance is plotted against dielectric barrier thickness. The variation of capacitance with d_b is not significant.

(C) Inductance

Variation of TSV inductance with radius and length is shown in Figure 3.6. As evident from the figure, TSV inductance increases with increasing length, but decreases with increasing radius, as predicted by the analytic formulation of inductance for an isolated conductor. It can be seen from the figure that as radius increases TSV inductance initially decreases and eventually levels off with further increases of radius.

3.3.2 Two Parallel TSVs

The electrical model of two parallel TSVs including coupling is shown in Figure 3.7. The subsequent plots are for the values of coupling or mutual capacitance C_c, self capacitance C_s, mutual inductance L_m and self inductance L_s. The resistance R_{tsv} is the same as for an isolated TSV.

(A) Capacitance

The coupling capacitance between two TSVs is a function of radius, length and inter-via spacing, as well as dielectric barrier thickness and permittivity. The higher the dielectric barrier thickness, the lower the parasitic coupling between two TSVs, as can be seen in Figure 3.8.

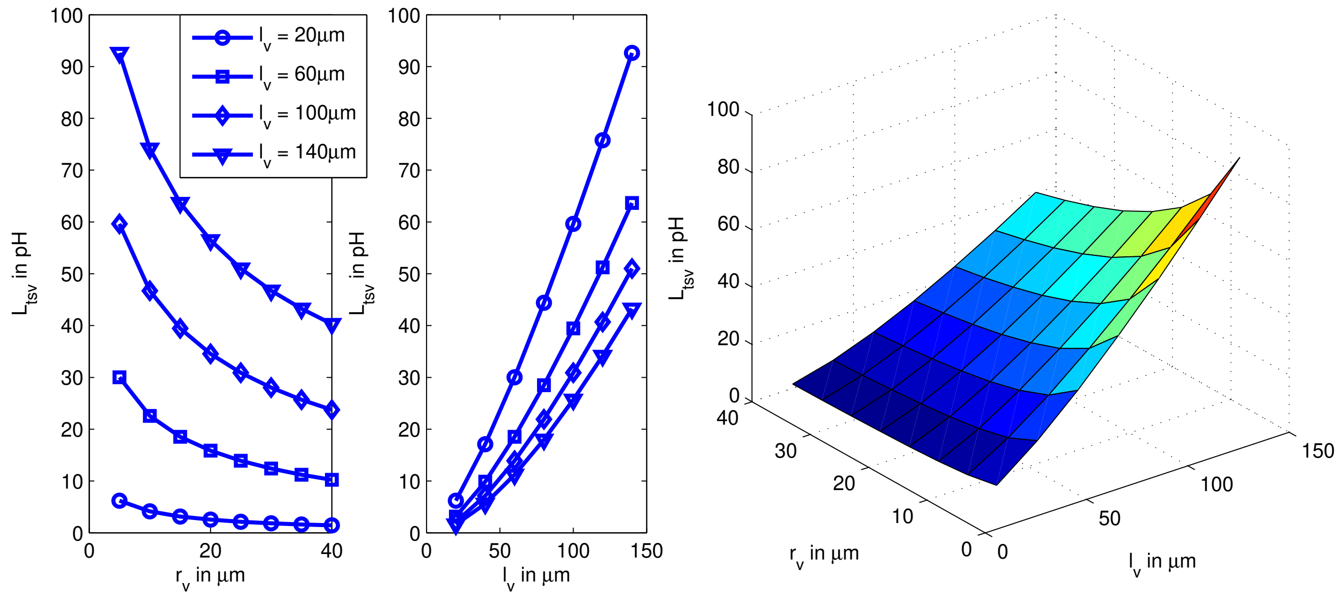


Figure 3.6: *Variation of TSV inductance with for various radii and lengths*

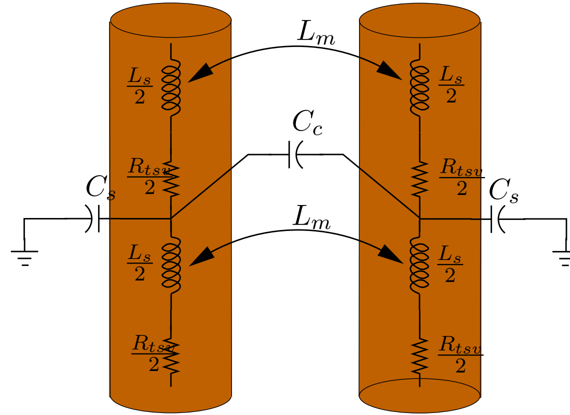


Figure 3.7: *Two coupled TSV model.*

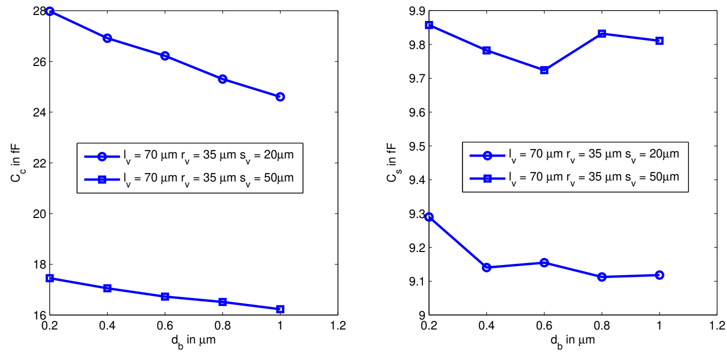


Figure 3.8: *Variation of coupling capacitance with dielectric barrier thickness.*

In order to achieve a high density of TSVs, a dielectric thickness that is as small as possible is desirable, and is generally a constant for a particular technology. The dielectric barrier thickness is assumed to be 0.2 μm and this value is used in further simulations.

Figure 3.9 shows the variation of coupling capacitance between two TSVs with spacing for different radii and lengths. For a higher radius and length the coupling capacitance is higher. As expected, coupling capacitance decreases with increasing spacing, becoming asymptotically zero as the spacing approaches infinity. The coupling capacitance of a TSV increases monotonically with increasing radius and decreasing spacing. As the TSV spacing increases, the self capacitance increases, and finally reaches a limit that is the capacitance of an isolated TSV for the given geometrical parameters. This variation is depicted in Figure 3.10.

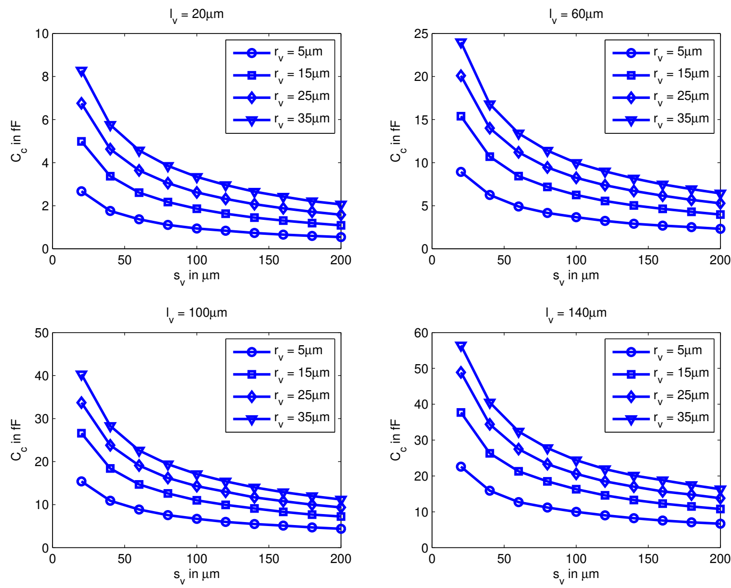


Figure 3.9: *Variation of TSV coupling capacitance with spacing for various lengths and radii.*

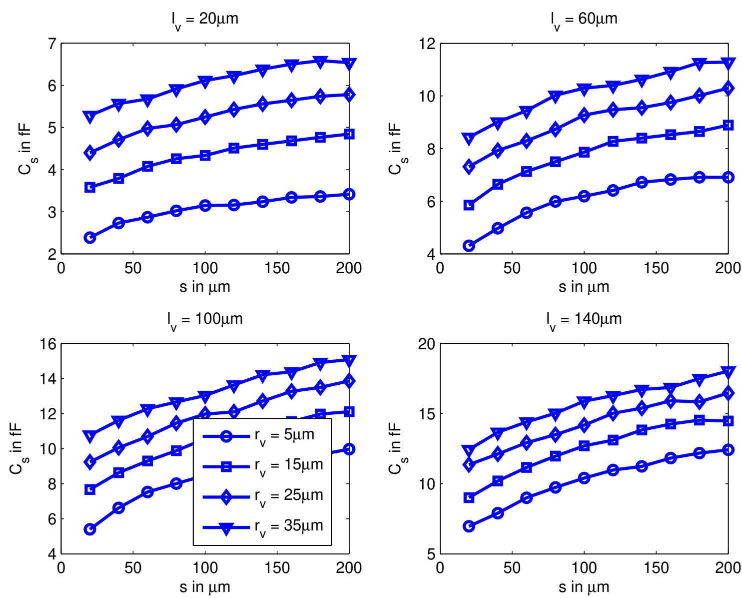


Figure 3.10: *Variation of TSV self capacitance with spacing for various lengths and radii.*

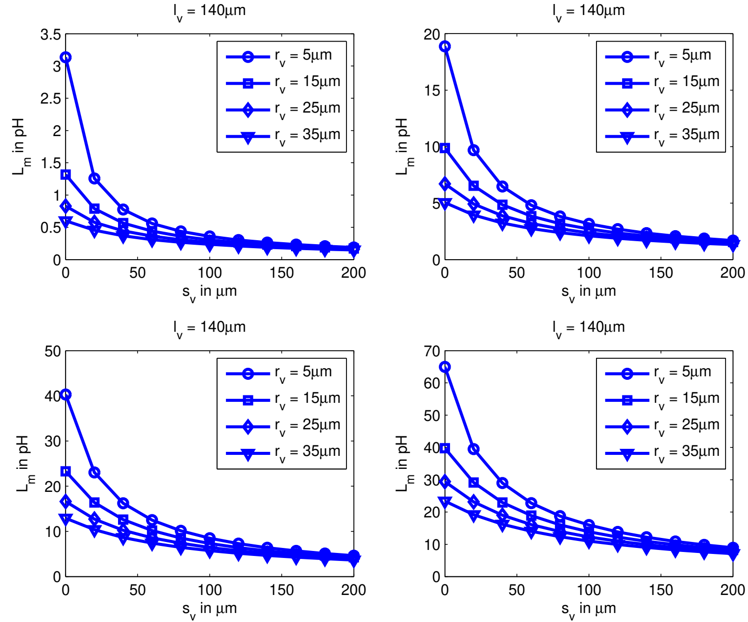


Figure 3.11: *Variation of mutual inductance of two TSVs with radius and spacing for different lengths.*

(B) Inductance

The self inductance of a TSV is not affected by the presence of a neighbouring TSV. Therefore, self inductance of a TSV in a two parallel configuration (or a TSV bundle) is the same as that of an isolated TSV. Figure 3.11 depicts the variation of TSV inductance with radius (r_v) and inter-via spacing (s_v). Subplots represent different TSV lengths. As can be expected, the mutual inductance levels off to the same value for different radii with increasing s_v. Intuitively, mutual inductance should asymptotically approach zero as the spacing goes to infinity. The mutual inductance also levels off with increasing radius.

3.3.3 TSV Bundle

In a TSV bundle there exists mutual coupling between any two TSVs. For convenience, the naming convention of TSVs in a bundle as illustrated in Figure 3.12 is adopted; the middle TSV is denoted as M for Middle and the others named in relation to their orientation with respect to the M TSV; for example, North (N), North East (NE), etc. Then the capacitance between TSV NE and TSV SW is denoted as $C_{NE,SW}$.

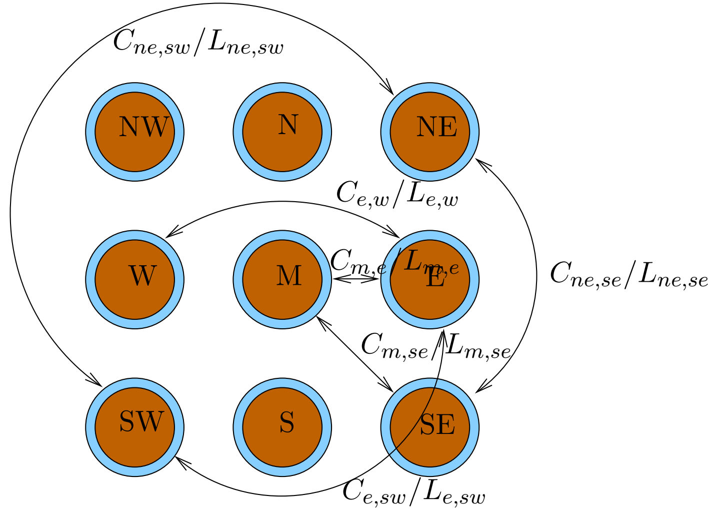


Figure 3.12: *Coupling terms in a 3x3 TSV bundle.*

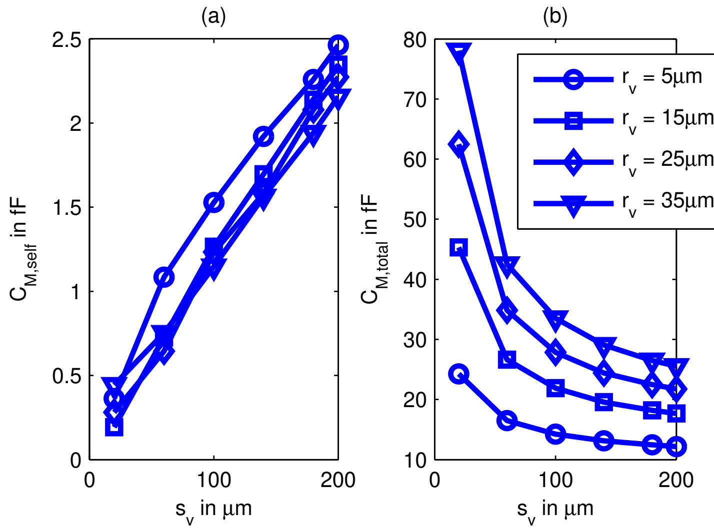


Figure 3.13: *Self (a) and total (b) Capacitance of middle TSV in 3×3 bundle.*

(A) Capacitance

The field-solver gives the total capacitance for a given TSV which is the summation of the self and all coupling capacitances to every other TSV. Figure 3.13 depicts the self-capacitance in (a) and the total capacitance in (b) of the middle TSV in a bundle. As the inter-via spacing increases the self capacitance increases and finally reaches the capacitance of an isolated TSV. Increasing spacing also results in decreasing total capacitance due to the diminishing contribution of coupling

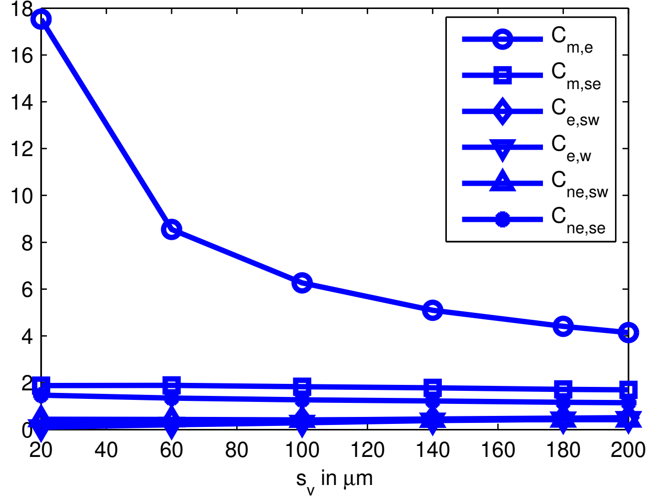


Figure 3.14: *Variation of coupling capacitance with spacing in a TSV bundle.*

capacitance.

Variation of the various TSV coupling capacitances shown in Figure 3.12 with inter-via spacing for a TSV with radius 35 μm is given in Figure 3.14. Among them, $C_{M,E}$ is most significant in comparison to all other terms. The terms $C_{NE,SW}$, $C_{NE,SE}$ and $C_{E,SW}$ are relatively small because the intervening TSV acts as a shield. In other words, electric field lines tend to terminate on the nearest conductor.

The nature of capacitive coupling of a TSV at the center in a 7×7 bundle and all the surrounding TSVs is depicted in Figure 3.15 for $s_v = 20\ \mu m$. As in the on-chip case, the capacitive coupling terms to nearest neighbors dominate over the coupling terms to nonadjacent lines, which are mostly insignificant. With reference to the naming convention proposed earlier, the distances from M TSV to N,E,S and W TSVs are the same, while the distances to NE, NW, SE and SW TSVs are also equal. Therefore, their coupling capacitances are also equal. Within the set of nearest neighbors the lateral terms ($C_{M,E}$) are more significant than the diagonal terms ($C_{M,SE}$). This is observable in Figure 3.15 and is due to the fact that the diagonal neighbors are partly shielded by the lateral conductors and the non-adjacent lines are almost completely shielded by the ring of adjacent lines.

(B) Inductance

The self and mutual inductance terms in a bundle exhibit markedly different characteristics to the capacitance, due to the fact that magnetic flux lines extend far more globally than electric field lines. The self inductance for example, shows a negligible variation with inter-via spacing (Figure 3.16).

Significantly, the coupling terms between non-adjacent lines are not negligible, as shown in Figure 3.17. The different terms decrease with increasing inter-via

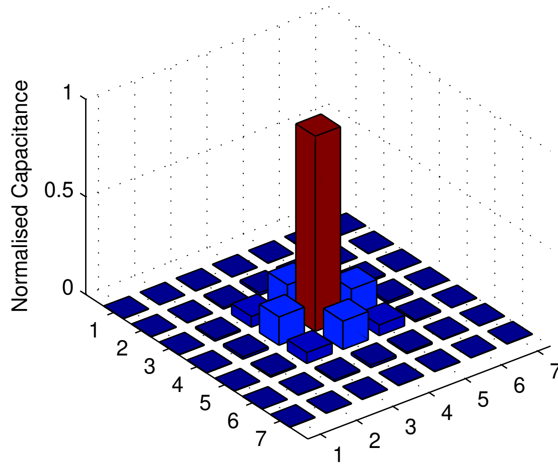


Figure 3.15: *Capacitive coupling between TSV in the middle of a 7×7 bundle with its surrounding TSVs. Values in the graph are normalized to middle TSV total capacitance.*

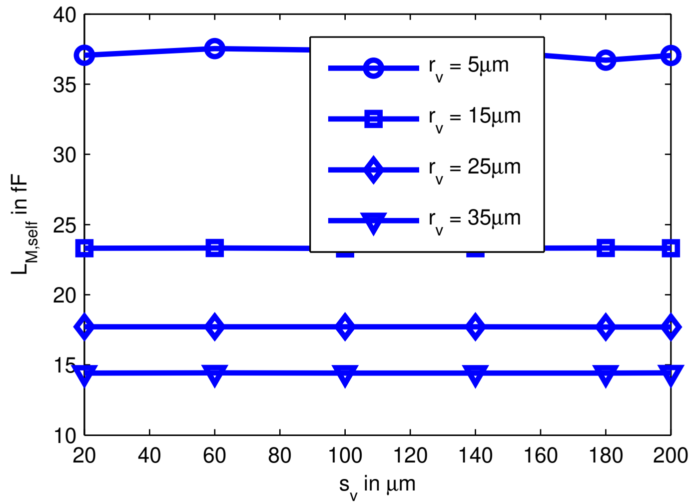


Figure 3.16: *Variation of self inductance with inter-via spacing for various radii in a TSV bundle*

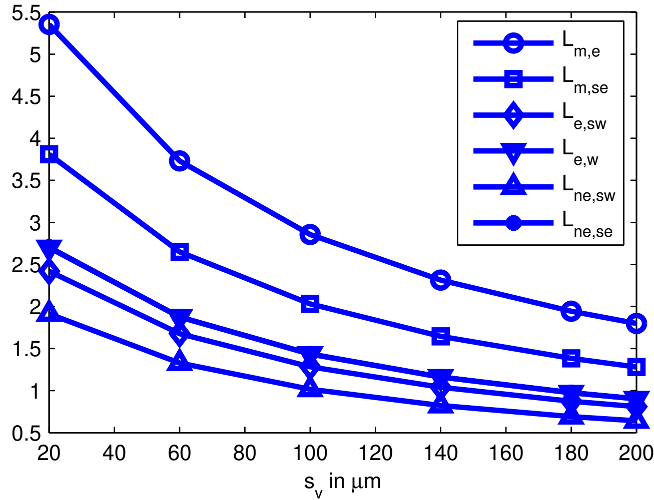


Figure 3.17: *Variation of Mutual inductance with inter-via spacing in a 3 times 3 TSV bundle*

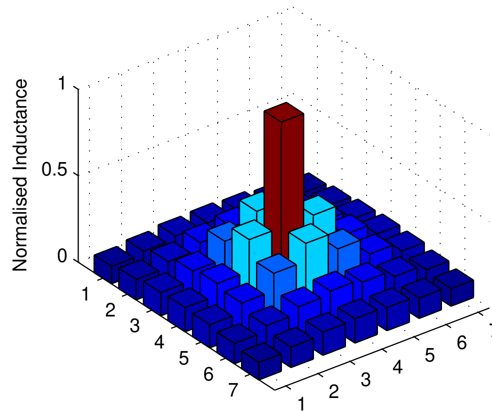


Figure 3.18: *Mutual inductance between C TSV of a 7×7 bundle normalized to self inductance C TSV for physical dimensions r_v = 20 μm, l_v = 80 μm and s_v = 20 μm.*

spacing, and gradually converge. These values should reach zero asymptotically, as the inter-via spacing tends to infinity.

In the case of inductance, the coupling is significant within the entire bundle because magnetic field lines tend to permeate the length and breadth of the global structure, again analogous to the on-chip case. This relationship can be observed in Figure 3.18.

3.4 Compact Modelling of TSV Parameters

The parasitic parameters of TSVs in a bundle have complex field dependence predicated on the physical geometry, and the material constants. Identifying these dependencies and creating a simple compact model is a challenging task. The data to be matched consists of electrical quantities extracted from a field solver for a range of physical dimensions.

Two methods can be identified for derivation of formulae for the prediction of these TSV parasitics. One method is the response surface method [*109*], where a least-squares approach is used to estimate formulae in terms of linear, square, and product terms of all independent variables. Though it is quite simple to come up with an equation using this method, the outcome is a long and unwieldy set of equations, providing no physical insight, or portability for different boundary conditions.

The second method, which is used in this work, is to use dimensional analysis [*110*, *111*]. Dimensional analysis enables the number of independent variables in a function to be reduced, through the combination of two or more variables into a single variable, such that the resulting variable is dimension neutral. The combination of the variables has to be carried out in such a way that the resultant variable has a meaningful interpretation.

A 3×3 TSV configuration (see Figure 3.1) is a general representative unit of a bundle. In such a structure, the TSV in the middle experiences lateral as well as diagonal coupling. In order to get a clear insight into the self and mutual components of capacitance and inductance, the full model is built up by studying an isolated TSV first, followed by a 3×3 bundle.

3.4.1 RLC Extraction of an Isolated TSV

Figure 3.2 shows the equivalent circuit for the TSV structure, while Table 3.2 indicates the resistive, inductive and capacitive parasitics as well as the (L/R) and (RC) time constants for various TSV geometry combinations. It is found that the L/R time constant is several orders of magnitude greater than the RC time constant. This reveals that in a TSV inductive effects dominate, and that it acts like a transmission line. In analyzing TSV delays an RLC transmission line model appears to be more accurate than an RC line as widely used in on-chip wire delay models.

(A) Resistance

Resistance can be described accurately as a function of its conductivity and cross sectional area. For a TSV with its radius r_v, conductivity of material σ, and length l_v, the resistance is:

$$R_{tsv} = f(l_v, r_v, \sigma) \tag{3.1}$$

$$R_{via} = \frac{l_v}{\sigma \pi r_v^2} \tag{3.2}$$

l_v (μm)	r_v (μm)	d_b (μm)	R_{tsv} ($m\Omega$)	L_{tsv} (pH)	C_{tsv} (fF)	$\frac{L_{tsv}}{R_{tsv}}$ (ns)	$R_{tsv}C_{tsv}$ (fF)
20	5	0.2	4.44	6.18	3.81	1.39	0.017
20	10	0.4	1.11	4.13	4.75	3.72	0.015
140	20	0.2	1.94	56.48	21.32	9.16	0.041
40	20	0.2	5.55	8.27	9.52	1.49	0.053
40	40	0.4	1.39	5.08	12.25	3.66	0.017
140	40	0.2	0.49	40.31	9.04	82.98	0.004
70	70	0.2	0.08	8.89	8.14	12.01	0.001
140	70	0.2	0.16	28.94	10.79	182.45	0.001

Table 3.2: *RC and $\frac{L}{R}$ time constants of an isolated TSV*

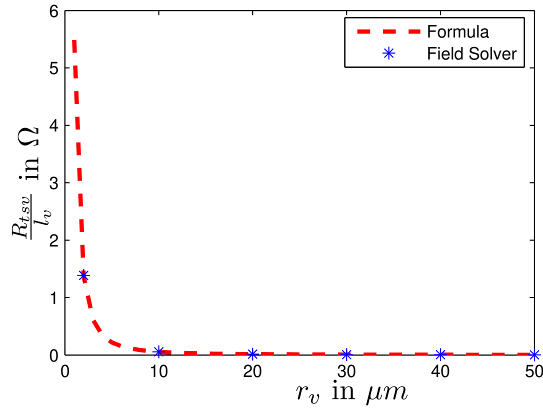


Figure 3.19: *Variation of TSV resistance with its radius.*

As usual, TSV resistance can also be expressed in unit length format, which is:

$$r_{via} = \frac{R_{via}}{l_v} \Rightarrow r_{via} = \frac{1}{\sigma\pi}\frac{1}{r_v^2} \qquad [\Omega m^{-1}] \tag{3.3}$$

The simulated values are accurate within 98% of those found from the analytical equation given in (3.2); see Figure 3.19.

(B) Capacitance

The capacitance C_{tsv} of an isolated TSV is a function of its geometry, *i.e.* radius r_v, length l_v, and thickness of SiO_2 barrier d_b, as well as the effective permittivity of the surrounding dielectrics, ϵ. Using the principles of dimensional analysis, C_{tsv} may be expressed as a function of dimensionless variables as follows:

$$\frac{C_{tsv}}{\epsilon_0 l_v} = f\left(\frac{l_v}{r_v}, \frac{d_b}{r_v}\right) \tag{3.4}$$

Note that the independent dimensionless variables have been selected so that they represent a meaningful physical quantity, such as aspect ratio (height to diameter ratio).

For a given technology, since the dielectric barrier thickness is a constant, the capacitance model is characterised under the assumption that its dielectric thickness is held constant at 0.2 μm. This assumption is logical given that d_b is a constant related to the technology, as well as the considerable reduction in complexity afforded by not treating this parameter as an independent variable. It is possible to recalibrate the equation constants for different d_b values. Should there be a need to treat d_b as an independent variable the model could perhaps be revisited.

Therefore, C_{tsv} is now reduced to a function of one independent variable. For various combinations of l_v and r_v, the plot of $\frac{C_{tsv}}{\epsilon l_v}$ against the aspect ratio, $\frac{l_v}{r_v}$, is shown in Figure 3.20 .

$$\frac{C_{tsv}}{\epsilon l_v} = f\left(\frac{l_v}{r_v}, \frac{d_b}{r_v}\right) \tag{3.5}$$

The relationship between $\frac{C_{tsv}}{\epsilon l_v}$ and $\frac{l_v}{r_v}$ can be expressed in the form of a compact model as:

$$\frac{C_{tsv}}{\epsilon_0 l_v} = \frac{63.34}{\log\left(1 + 5.26\frac{l_v}{r_v}\right)} \tag{3.6}$$

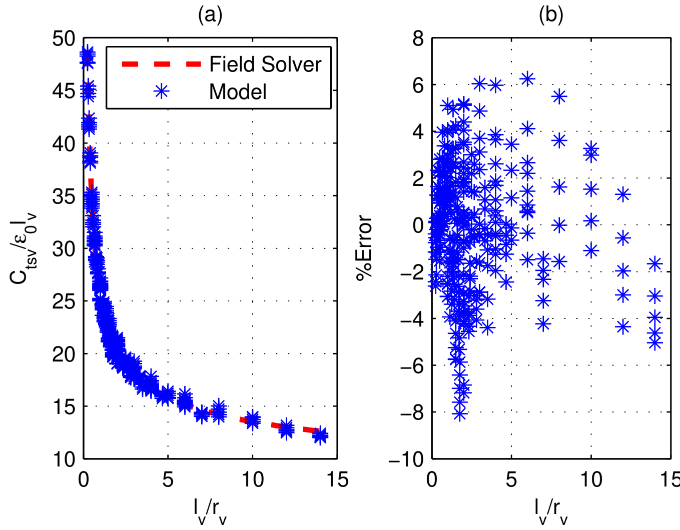


Figure 3.20: *(a) Variation of* $\frac{C_{tsv}}{\epsilon_0 l_v}$ *vs* $\frac{l_v}{r_v}$ *ratio, and (b) percentage error in predicted and extracted values.*

Even though Equation (3.6) does not contain a term that represents dielectric barrier thickness d_b, since the variation of capacitance with d_b is not significant for typical ranges, the proposed empirical self-capacitance formula has a maximum error contained to within 8% for the simulated range of $0 < d_b < 1$ μm. The form of the equation is derived from analytical insight given by field theory.

(C) Inductance

The inductance of an isolated TSV is a function of the geometrical parameters of radius r_v and length l_v and permeability μ of the surrounding medium. Due to the nature of the electromagnetic field, the dependence of inductance on d_b is negligible. Again using dimensional analysis, the TSV inductance can be expressed as:

$$\frac{L_{tsv}}{\mu l_v} = f\left(\frac{r_v}{l_v}\right) \tag{3.7}$$

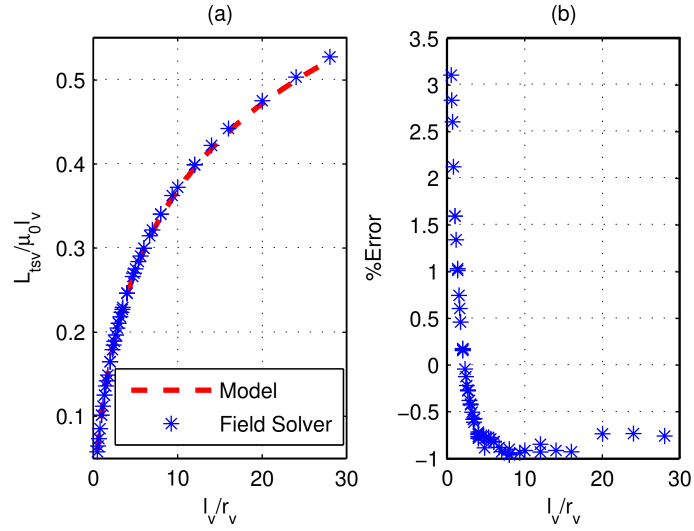


Figure 3.21: *(a) Variation of $\frac{L_{tsv}}{\mu_0 l_v}$ vs $\frac{l_v}{r_v}$ ratio, and (b) percentage error in predicted and extracted values.*

The variation of $\frac{L}{\mu l_v}$ versus $\frac{l_v}{r_v}$ is depicted in Figure 3.21. The empirical formula in (3.8) can be formulated for the self-inductance of an isolated TSV. The maximum error in this model is contained to within 3%. As with the capacitance, the form of the function is suggested by field theory.

$$\frac{L_{tsv}}{\mu l_v} = 0.16 ln\left(1 + 0.9\frac{l_v}{r_v}\right) \tag{3.8}$$

3.4.2 RLC Extraction of a TSV Bundle

(A) Capacitance

In a TSV bundle, the self and coupling capacitance between each and every TSV can be defined by:

$$C_{bundle} = \begin{bmatrix} C_{1,1} & -C_{1,2} & \cdots & -C_{1,n} \\ -C_{2,1} & C_{2,2} & \cdots & -C_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{n,1} & -C_{n,2} & \cdots & C_{n,n} \end{bmatrix} \tag{3.9}$$

In (3.9), the diagonal element $C_{i,i}$ represents the sum of the self and inter-via coupling capacitances $C_{i,j}$ as given in (3.10).

$$C_{i,i} = C_{i,0} + \sum_{j=1}^{n} C_{i,j} \tag{3.10}$$

As we have discussed in section 3.3.3(A), the capacitive matrix is sparse; the main diagonal and adjacent diagonals are populated while the other entries vanishingly small compared to the coupling capacitances to nearest neighbours (Figure 3.15). Therefore, it is reasonable to model the coupling between nearest neighbours. Intuitively a 3×3 bundle is the representative unit for a any size TSV bundle, and in a 3×3 bundle. the symmetry in the structure is exploited to reduce the number of terms to be investigated. Referring the naming convention given in Figure 3.22 the distances from M TSV to N, E, S and W TSVs are the same, as are the distances to NE, NW, SE and SW TSVs. Therefore, the closed-form capacitance formulae for the total capacitance of C, N, and NE TSVs ($C_{i,i}$), and their coupling terms to the nearest neighbours ($C_{i,j}$) as defined in Figure 3.22 is proposed for early signal and power integrity estimation of 3-D ICs.

The formula for the total capacitance C_t for N, NE and C TSVs is of the form:

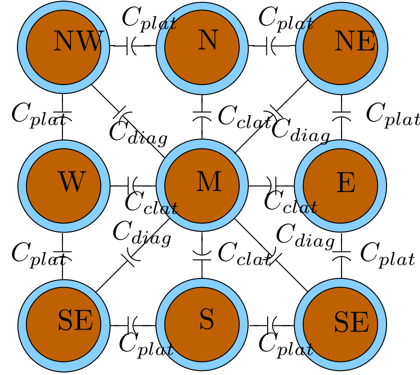


Figure 3.22: *TSV bundle nearest neighbour coupling capacitances.*

$$C_t = C_{tsv} + \frac{\epsilon_0 l_v}{ln\left(k_5 \frac{pv}{rv}\right)} \left[k_1 \left(\frac{p_v}{r_v}\right)^{k_2} + k_3 \left(\frac{p_v}{l_v}\right)^{k_4} \right], \tag{3.11}$$

and the constants are defined in Table 3.3.

As TSV pitch approaches to infinity, the total capacitance of a TSV in a bundle should approach to the capacitance of an isolated TSV (C_{tsv}). In (3.11) the constants k_2 and k_4 are negative and therefore, as p_v approaches to infinity C_t approaches to C_{tsv}. The isolated TSV capacitance formula C_{tsv} has a maximum error contained within 6% for the simulated range.

The formula for coupling capacitance C_c terms which are defined in Figure 3.22 of a TSVs in a bundle is of the form:

$$C_c = \frac{k_1 \epsilon_0 l_v}{ln\left(k_2 \frac{pv}{rv}\right)} \left[1 + k_3 \left(\frac{p_v}{r_v}\right)^{k_4} + k_5 \left(\frac{p_v}{l_v}\right)^{k_6} + k_7 \left(\frac{p_v}{l_v}\right)^{k_8} \right], \tag{3.12}$$

		k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	Max. % Error	Average % Error
(a)	C_{t_M}	34.0156	-0.8350	20.3800	-0.150	0.55	-	-	-	6.7	2.2
(b)	C_{t_N}	37.5900	-1.0830	19.3406	-0.156	0.57	-	-	-	5.5	2.0
(c)	C_{t_NE}	33.5413	-1.3317	17.1720	-0.168	0.59	-	-	-	4.5	1.6
(d)	C_{c_l}	10.191	0.5490	-0.014	0.796	0.054	-1.157	-0.018	-0.600	8.7	1.9
(e)	C_{c_p}	3.180	0.5440	-0.199	0.586	0.122	0.540	2.176	0.110	10.9	1.8
(f)	C_{c_d}	18.117	28.457	-1.734	-2.178	0.600	-0.518	-0.470	0.188	8.0	1.4
(g)	C_{s_M}	0.1505	-0.0071	-0.091	0.1849	-1.9371	6.9577	-0.0131	-0.0354	48.0	7.8
(h)	C_{s_N}	0.6876	-0.0390	-0.0583	1.8076	-0.2229	11.3537	0.0402	-13.1813	10.2	1.9
(i)	C_{s_NE}	0.3406	-0.0345	-0.0686	5.0708	-0.1530	-5.6346	-0.3859	-0.7643	13.3	2.0

Table 3.3: *Constants for Total, Coupling and Self Capacitances of* 3×3 *TSV bundle.*

and the constants $k_1, \cdots, k_8$ are given in Table 3.3.

However, the well establish practice is to distribute the total capacitance of a wire into its self ($C_{i,0}$) and capacitive components ($C_{i,j}$) such that those can be used in a circuit simulator for signal integrity analysis. Coupling capacitance terms have already been presented in (3.12), and the self capacitance component can be estimated using:

$$C_s = C_{tsv}\left\{1 - k_1 e^{\left(k_2 \frac{p_v}{r_v} + k_3 \frac{p_v}{l_v}\right)}\left[k_4\left(\frac{l_v}{r_v}\right)^{k_5} + k_6\left(\frac{p_v}{r_v}\right)^{k_7} + k_8\right]\right\} \quad (3.13)$$

where the constants are given in the last three rows of Table 3.3.

Also shown in Table 3.3 are the absolute *maximum* errors. As can be seen, all models have a *minimum* accuracy over the full simulated range of approximately 90%, except in the case of the self capacitance term C_{s_M}, which has a maximum error of approximately 50%. However, the comparison between the calculated C_t values from the proposed equations and extracted values for M,N, and NE TSVs have maximum absolute errors 2.3%,3.6%, and 2.9% respectively. It may seem that such an error renders this particular model unusable, but all large errors are contributed by capacitance values that are negligible for any meaningful delay, SI or PI analysis, because for those geometries, the self capacitance is a very small fraction of the total capacitance which is dominated by the coupling terms. This is borne out in Figure 3.24, where all large errors are for self capacitance values that fall within 5% of the total capacitance.

In this range, the self capacitance values are indistinguishable from numerical noise in the field solver. Hence the error of this model for meaningful capacitance values (greater than roughly 5% of the total capacitance) is no more than approximately 15%. For example, comparisons between the calculated and extracted C_t values for M, N, and NE TSVs in this range have maximum absolute errors of 2.3%, 3.6% and 2.9% respectively.

Further, circuit simulations were carried out for a structure with a representative geometry within the high error range to investigate the worst-case delay and coupled noise of the M TSV in a 3×3 bundle using values for C_{s_M} that vary $+50\%$ and -50% from the nominal field solver extracted values (Figure 3.25). The simulations verify that the errors in delay and noise are restricted to 1% and 4% respectively; *i.e.* the large errors in the insignificant C_{s_M} values are not reflected in high-level metrics. The average absolute error over all values is also shown in the table, which emphasises that the maximum errors are for a few pathological cases, and the overall fit is within a few percentage points of the simulated values.

(B) Inductance

The self and mutual inductance terms for a TSV bundle is defined by:

$$L_{bundle} = \begin{bmatrix} L_{1,1} & L_{1,2} & \cdots & L_{1,n} \\ L_{2,1} & L_{2,2} & \cdots & L_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & L_{n,n} \end{bmatrix}, \quad (3.14)$$

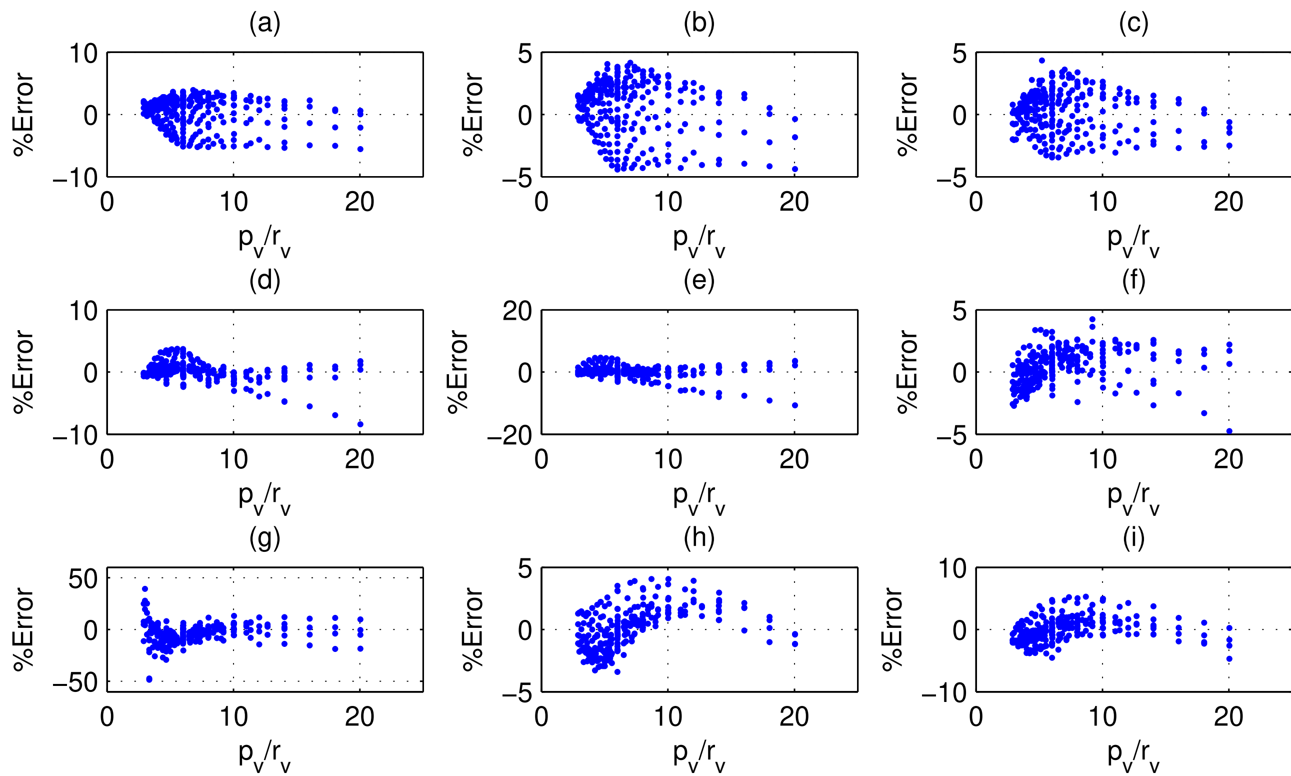


Figure 3.23: *Percentage Error in parasitics estimated from formulae* (3.11), (3.12), *and* (3.13) *for the order (a)-(i) given in Table 3.3.*

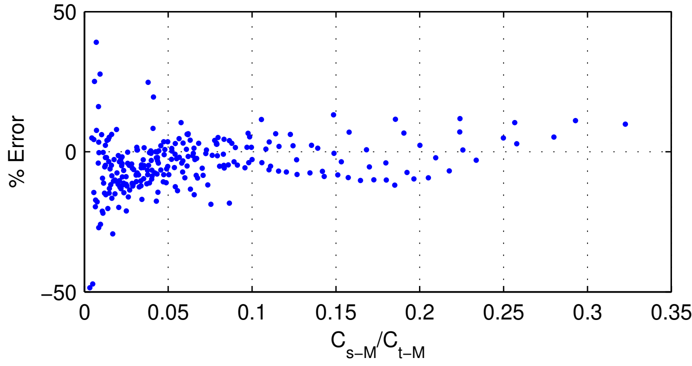
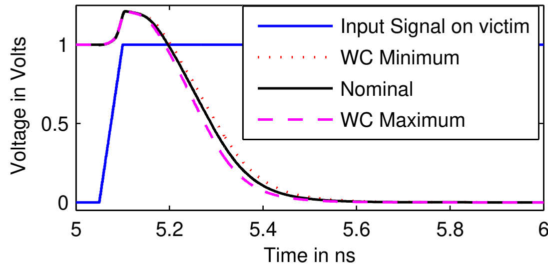
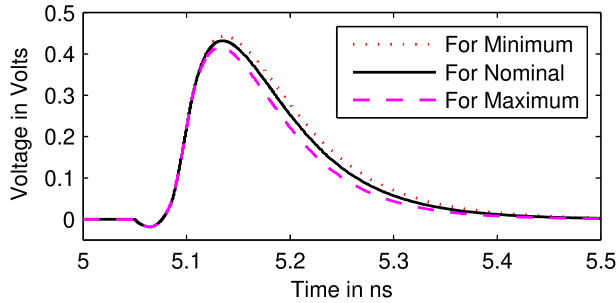


Figure 3.24: *Error plot of C_{s_M} as a fraction of C_{t_M}.*



(a) Output for delay with victim switching from '0' to '1'



(b) Output for noise with victim quiet

Figure 3.25: *Normalised output signal on the middle conductor in a* 3×3 *TSV bundle using nominal parasitic values obtained from the field solver, as well as worst-case minimum and maximum error combinations when all adjacent TSVs switch from '1' to '0'.*

where diagonal elements represent the self inductance (L_s) of the TSVs in a bundle, and off diagonal terms the mutual inductance between TSVs in the bundle. Inductive coupling is long range and therefore the inductance matrix is well populated,

with all elements being non-negligible. Inductive coupling is long range and therefore the inductance matrix is well populated, with all elements being non-negligible (Refer Figure 3.18).

The self inductance of a wire is not affected due to the presence of neighbouring wires, an observation borne out by simulations of two parallel TSVs. Therefore, the self inductance L_s can be estimated from the equation derived for an isolated conductor. However, mutual inductance between lines is a function of the effective permeability of the material, as well as the geometrical parameters of TSV length l_v, radius r_v, and the distance between the lines (the center to center distance, say d_v).

$$L_m = f(\mu, r_v, l_v, d_v) \tag{3.15}$$

Following principles of dimensional analysis, (3.15) can be expressed as a function of dimensionless quantities:

$$\frac{L_m}{\mu l_v} = f\left(\frac{r_v}{l_v}, \frac{d_v}{l_v}\right) \tag{3.16}$$

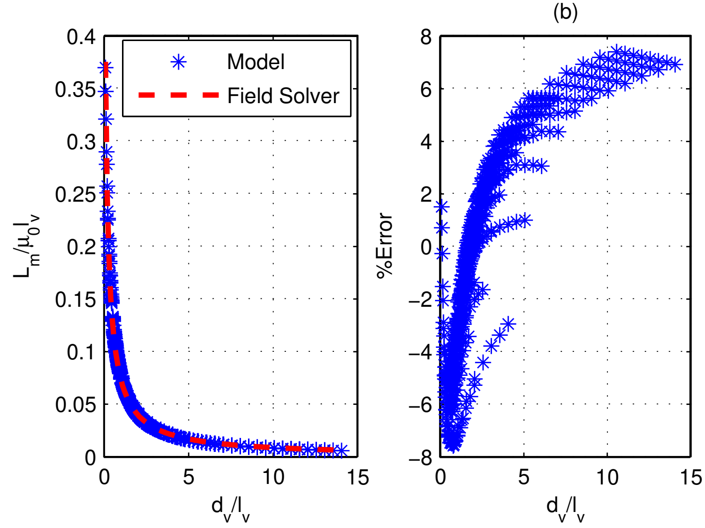


Figure 3.26: *Variation of mutual inductance (L_m) of Two TSVs with distance*

Figure 3.26 depicts the variation of mutual inductance between two parallel TSVs with $\frac{p_v}{l_v}$ for various $\frac{r_v}{l_v}$ combinations. The empirical formula modelling this behaviour is:

$$\frac{L_m}{\mu l_v} = 0.199 ln\left(1 + 0.438\frac{l_v}{d_v}\right) \tag{3.17}$$

The maximum error in this model is contained to within 8%.

A snapshot of some of the data points illustrating the accuracy of the models is shown in the following tables. The extracted inductance values for a 3×3 TSV bundle for $l_v = 70\ \mu m$, $p_v = 101\ \mu m$ and $r_v = 25\ \mu m$, are shown in Table 3.4. The values predicted from the proposed models and the error in comparison to the simulated values are shown in Table 3.5 and Table 3.6 respectively.

L	Middle	E	N	NE	NW	S	SE	SW	W
Middle	17.82	4.735	4.737	3.392	3.391	4.734	3.391	3.39	4.733
E	4.735	17.82	3.39	4.736	2.158	3.39	4.736	2.158	2.408
N	4.737	3.39	17.8	4.737	4.736	2.409	2.158	2.158	3.391
NE	3.392	4.736	4.737	17.77	2.41	2.158	2.41	1.71	2.158
NW	3.391	2.158	4.736	2.41	17.76	2.158	1.71	2.41	4.738
S	4.734	3.39	2.409	2.158	2.158	17.82	4.737	4.733	3.39
SE	3.391	4.736	2.158	2.41	1.71	4.737	17.77	2.41	2.158
SW	3.39	2.158	2.158	1.71	2.41	4.733	2.41	17.81	4.733
W	4.733	2.408	3.391	2.158	4.738	3.39	2.158	4.733	17.82

Table 3.4: *Extracted inductance values for a TSV bundle $l_v = 70\ \mu m$, $p_v = 101$ and $r_v = 25\mu m$.*

L	Middle	E	N	NE	NW	S	SE	SW	W
Middle	17.71	4.641	4.641	3.404	3.404	4.641	3.404	3.404	4.641
E	4.641	17.71	3.404	4.641	2.228	3.404	4.641	2.228	2.474
N	4.641	3.404	17.71	4.641	4.641	2.474	2.228	2.228	3.404
NE	3.404	4.641	4.641	17.71	2.474	2.228	2.474	1.785	2.228
NW	3.404	2.228	4.641	2.474	17.71	2.228	1.785	2.474	4.641
S	4.641	3.404	2.474	2.228	2.228	17.71	4.641	4.641	3.404
SE	3.404	4.641	2.228	2.474	1.785	4.641	17.71	2.474	2.228
SW	3.404	2.228	2.228	1.785	2.474	4.641	2.474	17.71	4.641
W	4.641	2.474	3.404	2.228	4.641	3.404	2.228	4.641	17.71

Table 3.5: *Predicted inductance values for a TSV bundle $l_v = 70\ \mu m$, $p_v = 101\ \mu m$ and $r_v = 25\ \mu m$ from (3.8) and (3.15).*

Error	Middle	E	N	NE	NW	S	SE	SW	W
Middle	0.61	2	2	-0.36	-0.37	2	-0.37	-0.4	2
E	2	0.61	-0.4	2	-3.3	-0.4	2	-3.3	-2.7
N	2	-0.4	0.51	2	2	-2.7	-3.3	-3.3	-0.39
NE	-0.36	2	2	0.31	-2.7	-3.2	-2.7	-4.4	-3.3
NW	-0.37	-3.3	2	-2.7	0.25	-3.2	-4.4	-2.7	2.1
S	2	-0.4	-2.7	-3.2	-3.2	0.61	2	2	-0.42
SE	-0.37	2	-3.3	-2.7	-4.4	2	0.33	-2.7	-3.3
SW	-0.4	-3.3	-3.3	-4.4	-2.7	2	-2.7	0.54	2
W	2	-2.7	-0.39	-3.3	2.1	-0.42	-3.3	2	0.61

Table 3.6: *Error between predicted and simulated inductance values for a TSV bundle $l_v = 70\ \mu m$, $p_v = 101\ \mu m$ and $r_v = 25\ \mu m$ from (3.8) and (3.15).*

3.5 Summary

This chapter outlined trends in TSV parasitic parameters for typical geometries and materials for a general three-dimensional integration technology. The variation of each parameter with respect to its physical dimensions was thoroughly investigated,

starting from an isolated TSV to a TSV bundle with 3×3 in parallel. The capacitive and inductive coupling in a bundle scenarios in a bundle has also been considered.

Simple yet accurate models for estimating delay and signal integrity of TSV a bundle are necessary for successful early analysis in 3-D integrated circuits. A detailed methodology for modelling of parasitic parameters using physical dimensions and material constants using dimensionless analysis was discussed, and compact closed-form equations for modelling resistive, inductive and capacitive parasitic parameters of through-silicon vias are proposed. Compact models are useful in system-conceptual level explorations of 3-D ICs. Specifically, they can be used for prediction of parasitic parameters in the estimation for comparison of performance and signal integrity related metrics.

4

Signalling Techniques for On-Chip Global Interconnects

Different signalling techniques are used for on-chip global interconnects for efficient transmission of data. This chapter looks into those techniques at different hierarchical levels extending from 2-D planar chips to vertically stacked chips with TSVs.

4.1 Introduction

Shrinking of the minimum feature size used in fabrication of ICs has resulted in exponential growth of performance and functionality over the past four decades. The integration of millions of devices on a single die however poses many difficult engineering challenges, most notably in power management and on-chip communication. As chip complexity and area grow, despite the best efforts to exploit locality with innovative architectural solutions, the average distance across which a bit has to be transferred has increased, and interconnection delay is a key bottleneck in modern digital design. Scaling of wires and tighter integration has also resulted in signal integrity problems which only add to the interconnection woes; cross-talk between signal lines results in signal corruption and variable delay, depending on the respective switching patterns.

A key technique in reducing propagation delay and signal degradation is repeater insertion. Although very effective and simple, this has an adverse effect on power consumption, and it has been estimated that over 50% of the power in a high performance microprocessor is dissipated by repeaters charging and discharging interconnects [*21*, *22*, *23*]. Furthermore, over 90% of this power is concentrated in only 10% of the interconnects; *i.e.* those which are classed as global and run for a significant fraction of the die length.

In this chapter discusses on-chip global signalling techniques proposed for two and three-dimensional integrated circuits in detail. The discussion starts with general design strategies for on-chip interconnects that can be used at different levels of hierarchy. Then, a smart repeater that consumes less energy, and is suitable for exactly these kinds of global interconnections is proposed [*112*]. Finally, signal transmission characteristics of TSV based vertical interconnects are discussed.

4.2 Design Methodologies for On-Chip Interconnects

Design methodologies for interconnects may be discussed under several categories: technological, layout and routing, circuit and signaling, and architectural levels. The technological level methodologies and their limitations in the nanometer era have been outlined in section 1.3, therefore this section restricts to discuss the other three levels only.

4.2.1 Layout and Routing Level

Layout techniques are usually integrated in placement and routing tools to reduce undesired induced effects in bus wire structures. A few such methods are: wire spacing, shielding, wire ordering, wire swizzling, wire sizing and shaping. Moreover, these techniques are limited by the available bus area, via blockage, nature of the coupling etc.

Wire spacing is the simplest technique that can be used to reduce crosstalk as the farther the aggressor, the lower the crosstalk noise. The added benefit of this method is the improved manufacturability of the design and the reduced power consumption. Furthermore, the increase in wire loop inductance caused by wire spacing will offset some benefits of the reduction in coupling capacitance. Placing a GND or VDD wire - *shield wire* - between two signal wires may effectively reduce the capacitive coupling, and hence dynamic delay. The effective capacitance of the interconnect is almost fixed and no longer depends upon the signal switching activity. With shielding, the normalized peak crosstalk noise can be reduced to less than 5% of Vdd for RC interconnects with lengths ranging up to 2 mm [*113*]. However there is always a trade-off between the maximum cross-talk allowed and the total bus wire area. More effectively, since the inductive coupling effect decays much slower than the capacitive effect, spacing is not so effective when the inductance is dominant. However, inserting a shield line between two wires reduces the loop inductance, since the current return path is adjacent. Due to the importance of the on-chip clock signal, the clock distribution network in a high speed circuit is generally shielded on both sides in the same layer [*114*]. The primary drawback of the shielding technique is the overhead of the metal resources. *Active shielding* is a variation of shielding methodology in which the shield lines on either side are not connected to GND or VDD, but connected to the signal itself [*115*]. Then, with the signal, shield lines too switch in-phase, and reduces the effective driver load, thereby increasing the performance by trading off additional area as well as power.

Additionally, inter-wire coupling can be reduced by asserting a maximum length over two sensitive wires can be routed next to each other, known as net ordering [*116*, *117*]. The net-ordering technique, however, is less efficient in reducing long range inductive coupling. In [*117*], the net-ordering and shield insertion techniques are simultaneously performed to minimize both capacitive and inductive coupling. In bus wire structures, wires can be swizzled; they are split into several segments, and the wire order in each segment are changed. In this technique the maximum run length for two particular wires is reduced [*118*] and both the capacitive and inductive coupling among the wires averages out for each wire, reducing both the worst case delay and the delay uncertainty [*119*, *118*]. [*118*] claims that it can provide up to 31.5% reduction in worst-case delay and 34% reduction in delay

uncertainty. This technique may require more vias in order to effectively swizzle wires in a bus.

Wire sizing is very efficient to optimize delay in non-coupled RC lines, because resistance of the wire is reduced with wider wires. For coupled lines, still the delay reduction by resistance outperforms that due to increased capacitance. For inductive wires, as is shown in [*37*], low-frequency inductance does not decrease by more than 10% by doubling the wire width, and the reduction at high frequencies is even smaller due to the skin and proximity effects.

Widening the interconnect decreases the resistance while increasing the capacitance, and there will be an optimum wire size for a given cost function. In many cases, propagation delay was the target cost function, but power dissipation and bandwidth is also introduced. By explicitly characterizing the relationship between the interconnect parameters and wire geometries, the trade-offs among the delay, bandwidth, and power of the global interconnect can be found [*120*, *121*]. As the inductance became important, new optimization algorithms were introduced [*120*]. There is detailed discussion given in [*122*] that the width of an inductive interconnect affects the power consumption in wires.

Wire shaping in RC or RLC dominated wires can improve their speed, and the optimum shaping which minimizes the delay is a decaying exponential function from the driver towards the load [*123*, *124*]. However exponential shaping is more difficult to implement than uniformly sized wires. The research described in [*124*] claims that wire tapering improves the speed by only 3.5% as compared to uniform wire sizing with optimum repeater insertion.

Widely used rules of thumb to optimize RLC nets are [*29*]: to provide as many as close return paths to a signal as possible, and to use larger than minimal wire spacing because the resultant reduction in coupling capacitance is greater than the increase in loop inductance.

4.2.2 Circuit Level

In this section, circuit level techniques that reduce crosstalk, propagation delay, and power consumption are reviewed. For several decades, buffer (or repeater) insertion has been the effective methodology used for improving interconnect performance, but during recent years there has been several other techniques proposed. This thesis proposes such a circuit level solution for global on-chip interconnects.

As the resistance and coupling capacitance significantly increases with wire length, the global wire delay increases exponentially. The most common method of reducing this delay over long interconnects is to insert repeaters at appropriate positions; it makes wire delay proportional to the length. Since the buffer insertion breaks a wire into several smaller sections, it efficiently reduces the inductive effects by shortening the current return path. When t_{rep} is the repeater delay, an approximation equation for the line delay (t_d) with repeater circuits is given by:

$$t_d = R_d c_w L + \frac{r_w c_w L^2}{2k} + (k-1)t_{rep}. \tag{4.1}$$

It can be seen from (4.1) that when k is equal to one, *i.e.* when there are no repeaters, the line delay has two terms; one is proportional to length L and one proportional to L^2. The loading of the driver is represented by the first term and the

second term is the RC line delay. As L becomes larger the square law dependence causes t_d to increase very rapidly. However, with the insertion of repeaters, an additional term to t_d is introduced; *i.e.* delays in the buffers itself. Therefore, buffer insertion is not useful for short-wires, but more effective for longer (global) wires. Repeater insertion is effective only when the wire time constant ($r_w c_w L^2$) is at least equal to seven times the time constant of a repeater ($R_d(C_d + C_g)$) [*6*]. The delay optimum repeater size and number of repeaters are given by

$$H_{opt} = \sqrt{\frac{R_d(C_L + cwL)}{r_w L C_g}}$$

$$k_{opt} = L\sqrt{\frac{0.4 r_w c_w}{0.69 r_w L C_g}}.$$

This method changes the delay dependence on the wire from quadratic to linear:

$$t_{rc} = 2.5L\sqrt{R_d C_d r_w c_w} \quad (4.2)$$

However, introducing repeaters increases the wire capacitance by HkC_g, which is equal to $0.73C_w$ - 73% higher than a interconnect without buffers [*14*].

For RLC interconnects there are several approaches proposed in the literature, which are simply extensions of Backoglu's method. Out of them two widely used approaches to insert repeaters are: Ismail's method [*89*] and Venkatesan's method [*125*]. For convenience the closed-form equations for optimum buffer parameters for a RLC interconnect presented in [*89*] are presented:

$$h_{opt} = \frac{1}{[1 + 0.16(T_{L/R})^3]^{0.24}}\sqrt{\frac{R_{drv} c_w}{r_w C_{g\,\min}}} \quad (4.3)$$

$$k_{opt} = \frac{1}{[1 + 0.18(T_{L/R})^3]^{0.3}}\sqrt{\frac{r_w c_w L^2}{2R_{drv} C_{drv}}} \quad (4.4)$$

where,

$$T_{L/R} = \sqrt{\frac{l_w r_w}{R_{drv} C_{drv}}}$$

As evident from the aforementioned two works, the inductance affects the optimal repeater number and repeater size, and also the RLC interconnect has a fewer number of repeaters and smaller sized repeaters than RC interconnect does.

In addition to that, there are several variations of repeater insertion techniques and optimization strategies proposed and currently being investigated to overcome some of the challenges in repeater insertion. A scheme proposed in [*126*] staggers the repeaters so that opposing transitions only persist for the length of the offset between repeaters, and become best-case patterns for the remainder, resulting in a delay reduction. Inserting latches instead of repeaters [*127*, *128*], elastic interconnects [*129*], and regenerative repeaters [*130*, *131*] are some other techniques to name a few.

Many innovative alternatives to the traditional repeater have also been proposed during the last two decades such as low-swing bus techniques [*132*], which reduce power consumption in interconnects; differential signalling [*133*], which minimizes

both inductive and capacitive crosstalk; current-mode signalling [*134*, *135*], which improves performance and reduces power consumption.

Moreover, the Transient Sensitive Accelerator (TSA) [*136*], Charge Recycling Technique (CRT) [*137*], Boosters [*138*], the TAGS receiver [*139*], the Aggressor-Aware Repeater [*140*], and the Capacitor Coupled Trigger and Accelerator combination [*141*] are some advanced circuit solutions which have been proposed as interconnect solutions that outperform traditional repeaters, in general.

The work done in [*142*] and [*143*] also seeks to reduce the delay by avoiding simultaneous switching similar to [*126*], but they accomplish this by introducing static delays in the repeaters rather than by physical offsets in the placement. They report an overall reduction in the delay for the worst-case pattern of up to 20%, but this scheme dissipates more power for transitions in the same direction, due to additional charging and discharging of the coupling capacitance. [*144*] and [*145*] report average energy savings of upto 25% by introducing a delay dependant on the relative transition pattern between two adjacent wires, but this additional delay introduces a timing penalty. The worst-case pattern for the delay is also the worst case pattern for the energy, and hence any energy saving is at the cost of an increase in the cycle time, which may not always be possible.

4.2.3 Architectural and System Level

As on-chip communication has became a challenge that is limited by physical constraints, interconnect planning, design and optimization has to be tackled at architectural and system level where significant optimization opportunities can be exploited. Methods such as interconnect-centric design flow, interconnect-centric architectures, signal encoding techniques, package-intermediate interconnects fall into this category.

A procedure that has already been used for critical path design for several technology generations is interconnect-centric design [*146*, *147*, *148*]. In this approach, interconnect design including interconnect planning, interconnect synthesis, and interconnect layout are optimized (often at the expense of other circuit features) at every level of the design process. This approach has the distinct advantage of using current technology to optimize performance in the design areas where interconnect is a bottleneck. It suffers from two specific disadvantages. First, appropriate interconnect design tools and design models are not available to implement this approach over all designs, so much of this work becomes custom. Second, to carry this approach to its fullest benefit often requires a major revision of standard design and layout practices, which are inconsistent with the advantages offered by scaling and technology changes that have been used in the past to follow Moores law.

Signal encoding is a powerful method to improve performance and/or reduce noise induced by crosstalk in interconnects. Encoding for example avoids worst-case patterns, and thereby, the transmitted data is modified in such a way to reduce delay, crosstalk, or power consumption. For instance, Error Control or Transition Coding Techniques [*149*, *150*, *151*, *152*, *153*, *154*, *155*] overcome the effects of inter-symbol interference. The relatively complex codec circuitry causes additional delay and consumes more power, rendering the coding ineffective in many cases [*156*]. Even otherwise, these schemes mostly address the problem of reducing transitions on a given wire, which is less important than reducing the relative

switching activity *between* lines, given that the aspect ratio of on-chip interconnect emphasises the coupling capacitance over the self capacitance.

Taking a signal off-chip and bringing a signal back on-chip is known as *package-intermediate-interconnect* [*157*, *158*], which entail chip-to-package parasitics that include the pad capacitance, and bond wire or ball-grid solder ball. Even taking into account the off-chip drivers and chip-to-package parasitics, off-chip wires are much faster than on-chip wires for transmitting a signal for the length of a die edge, for a relatively large die. This is because the fast off-chip traces more than make-up for the chip-to-package parasitics by outperforming the RC lines. The opportunity exists to take advantage of this phenomenon by running wires off-chip and bypassing long chip-edge to chip-edge length RC lines. As shown in [*157*] the actual saving will of course depend on the specific layout, for example, this technique of avoiding long on-chip wires by running them off-chip to realize Package-Intermediate Interconnects, is reported to yield a saving of up to 40%, even considering the chip-to-package parasitics.

4.3 SMART Driver Circuit

The smart repeater exploits the fact that in a parallel wire structure, the effective capacitance of a given wire is dynamic; *i.e.* it is a function of not only the physical geometry, but also the relative switching pattern described by the bits on the wire in question (the victim) and the adjacent wires (aggressors). With a traditional repeater, since the drive strength is static, the result is a spread of the propagation delay, with the repeater strength being essentially too much for every bit pattern *other* than the worst-case pattern. In the proposed repeater, the drive strength is dynamically altered depending on the relative bit pattern, by partitioning it into a *Main Driver* and *Assistant Driver*. For a higher effective load capacitance both drivers switch, while for a lower effective capacitance the assistant driver is quiet [*159*]. By disconnecting part of the repeater when it is not needed, the total load capacitance to the previous stage is reduced, resulting in reduced energy consumption for those instances. It is theoretically shown that the potential average saving in energy can be as much as 25% over a traditional repeater for typical global wire lengths in nanometre technologies.

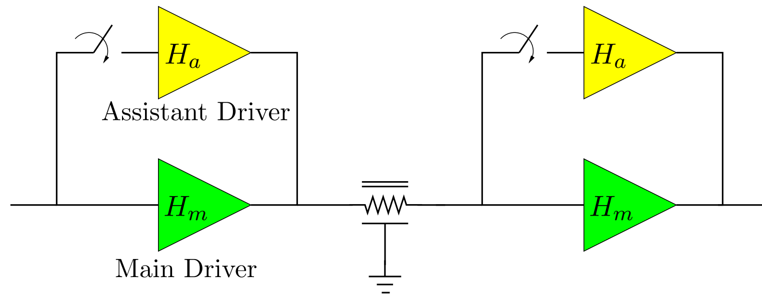


Figure 4.1: *Basic schematic of the proposed driver scheme.*

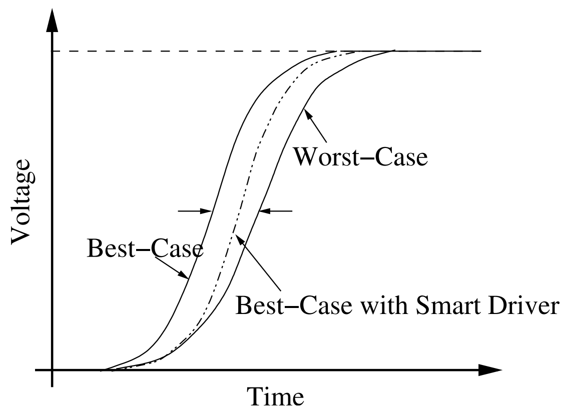


Figure 4.2: *Method of Jitter reduction using SMART driver.*

4.3.1 Limitations in Existing Driver circuits

In general, not only do these alternatives to traditional repeaters require much effort in circuit design similar to library cell design, but they also lack a clear high-level abstraction; in contrast, performance metrics such as delay and energy consumption can easily be quantified in terms of a few critical design parameters for the traditional inverting repeater [*6*], resulting in easy amalgamation in CAD flows at different levels of hierarchy from initial signal planning to detailed place and route [*160*].

A secondary advantage of the repeater circuit proposed here is that the relatively minor increase in circuit complexity required to obtain the energy saving and delay equalization described above can be completely abstracted in the performance analysis. A design methodology similar to that for traditional single-wire inverting repeaters is presented, including an RC equivalent circuit and closed-form expressions for the first-order approximation to the delay. Therefore this repeater can be very easily modeled in tasks such as delay calculation, signal integrity analysis and timing driven optimisation in any CAD flow for physical design.

4.3.2 The Concept

The effective interconnect capacitance varies with the transitions of neighboring lines and can be written as $C_s + \lambda C_c$, where C_s is the self capacitance of the wire, λ is the switch factor and C_c is the inter-wire capacitance. In this work, different switch factors are used for delay and power estimation (given in Table 4.1) based on the experimental validation in [*161*] which proposes power-based switch factors that are slightly different from the delay-based ones. The variation of the effective capacitance with the relative switching pattern introduces a spread in the arrival time at the far end of the wire. To demonstrate this a pair of coupled lines is used as a constituent unit for a bus. For two simultaneously switching lines, sixteen possible switching combinations can be identified. These can be categorized into five different groups according to the effective capacitance as follows. *Group 1:* Both switch in the same direction; *Group 2:* Both lines are quiet (at 0 or 1); *Group*

3: One line is switching while the other is quiet at 0; *Group 4:* One line is switching while the other is quiet at 1; *Group 5:* The lines switch in opposite directions.

To ensure error-free operation, timing constraints have to be satisfied for the switching pattern that causes the worst-case delay, which are the ↑↓ *and* ↓↑ combinations. Since the effective load is highest for these patterns, the size of the buffer designed statically for the worst-case delay is much larger than would be necessary for the same timing requirements for other patterns [*20*]. Now this worst-case condition occurs only twice out of 16 possible input switching patterns, with a probability of 1/8 for simultaneously switching lines if the transitions are equally distributed as in a random bit stream. For the 14 other cases, the wire is driven faster, which just translates to slack which typically cannot be used, consuming energy unnecessarily. The driver proposed here changes its drive strength depending on the neighbour's switching direction by using some simple logic. A basic schematic of the proposed *SMART* repeater is shown in Figure 4.1. If the switching pattern belongs to Groups 1, 3, or 4, a single inverter (the Main driver) drives the interconnect. When a switching pattern in Group 5 occurs, another inverter (the Assistant) also drives the line, increasing the total drive strength appropriately. By disconnecting the Assistant driver when it is not needed, part of the parasitic capacitance is disconnected for the majority of the switching patterns, leading to a saving in the average energy consumption.

The other useful feature in the smart driver is its ability to reduce jitter while saving energy. The SMART Driver achieves this energy saving by delaying the response for the best-case without affecting the worst-case, so that the variation in delay is as small as possible [*159*]. In other words, the concept is to make the response slower in the face of non worst-case input patterns, which reduces the effective load capacitance. This incurs no penalty, as the cycle delay has to be set to the worst-case delay anyway. In Figure 4.2 the curves with solid lines represent the output response of a conventional driver, for minimum effective capacitance (*Best-Case*) and maximum effective capacitance (*Worst-Case*).

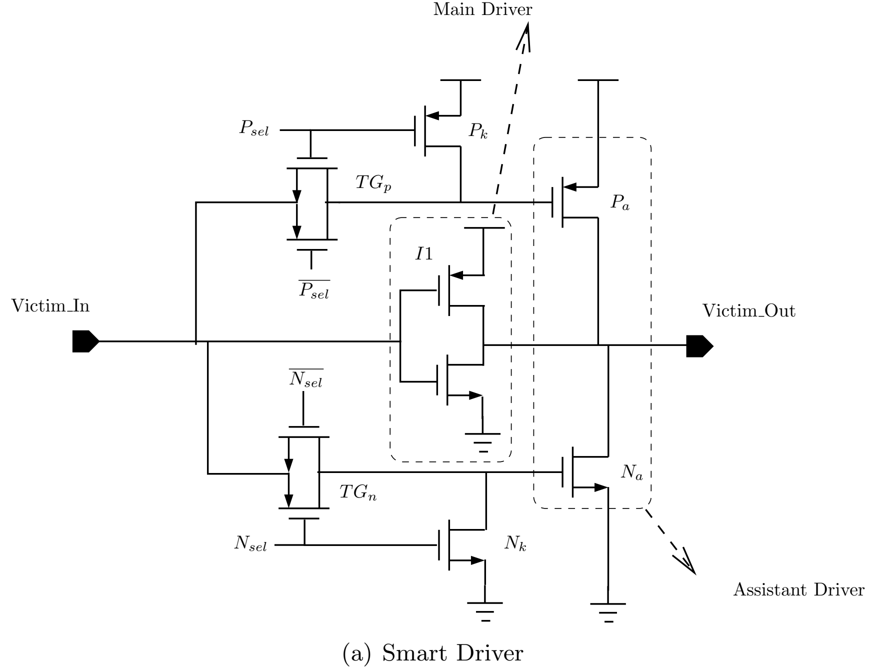
4.3.3 Circuit Realization

In the example 0.18 micron technology, it is difficult to change the state of the assistant before the input completes its transition due to the delay in logic elements. Hence in the implementation, a decision is made prior to the next transition about whether or not it constitutes a worst-case pattern. This decision is based on the relative logic values of the aggressor and the victim at the current time. Since the Assistant Driver needs to switch on for the worst-case patterns described in Group 5 in Table 4.1, anytime the current state has opposing logic values on the victim and aggressor, the assistant is turned on. This actually turns the assistant on for two other patterns which are not worst-case, namely patterns 10 and 12 in Table 4.1, which reduces the energy saving from the theoretical maximum, but allows a robust and fairly simple circuit implementation.

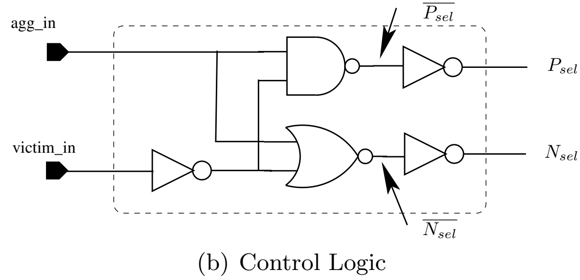
The simplified schematic of the Smart driver is shown in Figure 4.1, and the complete schematic in Figure 4.3. The transistors *Pa* and *Na* form the Assistant driver, whereas the Inverter *I1* is the Main Driver. Two transmission gates (*TGp* and *TGn* in Figure 4.3(a)), drive the pull-up and pull-down networks of the Assistant Driver. The weak transistors *Pk* and *Nk* act as keepers ensuring that the

Group	Case	Switching Event on		Switch Factor		Effective Wire Capacitance	
		wire i	wire j	Delay-Based(λ,μ)	Power-Based	Traditional driver	Smart driver
1	1	$\downarrow$	$\downarrow$	$0, 0$	0.25	$C_{w_trad} + 0.25C_c/k$	$C_{w_smrt} + 0.25C_c/k$
	2	$\uparrow$	$\uparrow$	$0, 0$	0.25	$C_{w_trad} + 0.25C_c/k$	$C_{w_smrt} + 0.25C_c/k$
2	3	0	0	n.a.	n.a.	0	0
	4	0	1	n.a.	n.a.	0	0
	5	1	0	n.a.	n.a.	0	0
	6	1	1	n.a.	n.a.	0	0
3	7	0	$\uparrow$	$0.57, 0.65$	1	0	0
	8	$\uparrow$	0	$0.57, 0.65$	1	$C_{w_trad} + C_c/k$	$C_{w_smrt} + C_c/k$
	9	0	$\downarrow$	$0.57, 0.65$	1	0	0
	10	$\downarrow$	0	$0.57, 0.65$	1	$C_{w_trad} + C_c/k$	$C_{w_trad} + C_c/k$
4	11	1	$\uparrow$	$0.57, 0.65$	0	0	0
	12	$\uparrow$	1	$0.57, 0.65$	0	C_{w_trad}	C_{w_trad}
	13	1	$\downarrow$	$0.57, 0.65$	0	0	0
	14	$\downarrow$	1	$0.57, 0.65$	0	C_{w_trad}	C_{w_smrt}
5	15	$\uparrow$	$\downarrow$	$1.51, 2.20$	1.75	$C_{w_trad} + 1.75C_c/k$	$C_{w_trad} + 1.75C_c/k$
	16	$\downarrow$	$\uparrow$	$1.51, 2.20$	1.75	$C_{w_trad} + 1.75C_c/k$	$C_{w_trad} + 1.75C_c/k$

Table 4.1: *Switching Activities on the lines and the variation of effective capacitance.*



(a) Smart Driver



(b) Control Logic

Figure 4.3: *Complete circuit schematic of the proposed smart driver (a), and its control logic.*

Assistant Driver is turned off properly when the corresponding transmission gate is disabled. The control signals of the transmission gates Psel and Nsel are determined as:

$$P_{sel} = \overline{Agg_In + \overline{Victim_In}} \tag{4.5}$$

$$N_{sel} = \overline{Agg_In.\overline{Victim_In}} \tag{4.6}$$

When the victim input is at logic 0 and the aggressor is at logic 1, the next victim stage would be logic 1, and this might be a worst-case pattern if the aggressor also changes its state. In this case Na is switched on and Pa switched off since Pa is not needed during this discharging period. This is achieved by setting $N_{sel} = 0$ and $P_{sel} = 0$ (Refer Equations (4.5) and (4.6)).

The decision as to whether the assistant should be on or off should be taken at the point the output voltage of the driver (at node $Victim_Out$) reaches the

Victim	Agg	Pa_{sel}	Na_{sel}
0	0	0	1
0	1	0	0
1	0	1	1
1	1	0	1

Table 4.2: *Selection Logic output for the present Victim and Aggressor States*

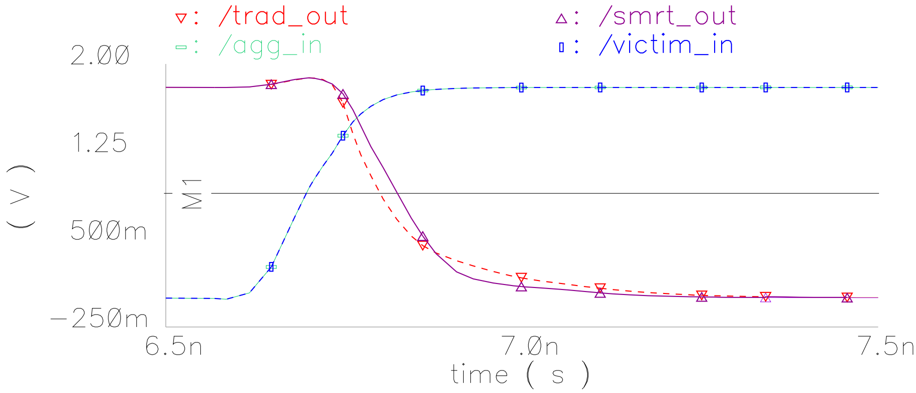


Figure 4.4: *Waveforms when the aggressor and victim switch in the same direction, with only the Main Driver being active.*

threshold value, as this gives sufficient time for the selection logic to drive the control signals to the appropriate logic value for the next transition. Hence the propagation delay of the selection logic, T_{logic}, should conform to the following inequality:

$$T_{driver} \leq T_{logic} < T_{clk} \tag{4.7}$$

where T_{driver} is the maximum propagation delay of the driver (From Node $Victim_In$ to $Victim_Out$) and T_{clk} is the clock period. The output of the selection logic should be available just after the interconnect is driven, and this is the maximum clock rate that can be achieved with this proposed scheme. The lower bound of the inequality ensures that the assistant driver is kept on until the output has crossed the threshold voltage ($V_{DD}/2$). The transmission gates are sized so as to reduce the path resistance, and are driven by cascaded buffers, to minimize the propagation delay.

Figures 4.4 and 4.5 show the simulations results at the far end and near end of a 2.5mm long wire driven by a smart repeater and a traditional repeater (inverter). When the aggressor and victim switch in opposite fashion, there is very little difference between the waveforms produced by the SMART and traditional repeaters. This is to be expected, and shows that the selection logic functions appropriately. When the aggressor and victim switch in the same direction, the Assistant Driver is off, and hence the output of the SMART driver is slower, at both near and far ends. This too is as expected, since the SMART Driver deliberately slows down

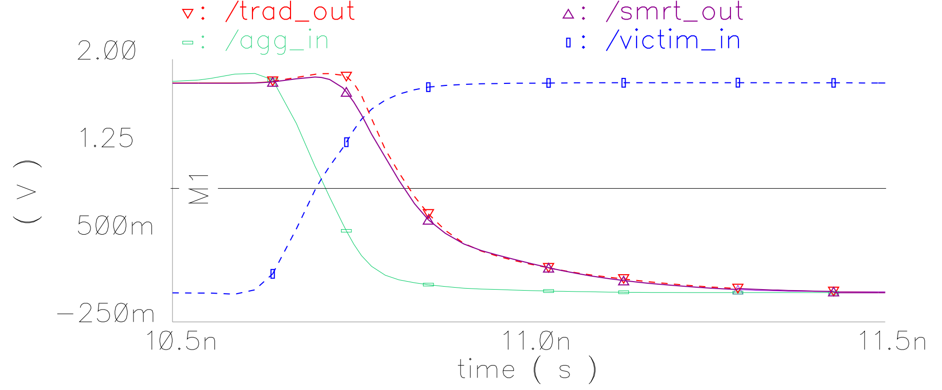


Figure 4.5: *Waveforms when the aggressor and victim switch in opposite directions, with both drivers active.*

the waveform by reducing the drive strength.

4.3.4 Noise Resiliency of the SMART Driver

Along with the delay performance, the noise resilience of the proposed Smart driver is of paramount importance. In our implementation, skewed inverters are not being used while using complementary logic with a switching threshold of $\frac{V_{DD}}{2}$ throughout the control circuitry. The only exception is the transmission gate switch pair, which are protected by the keepers, MP_k and MN_k.

When the present state has similar logic values on the victim and aggressor, $Psel$ and $Nsel$ are set to logic zero and logic one respectively. With this set up, transmission gates are switched off, and the keepers are switched on. Thus, gate inputs of MP_a and MN_a are disconnected from the node $Victim_in$ and are connected with V_{DD} and GND respectively, ensuring that both transistors work in the cut-off region.

When the present state has opposing logic values on the victim and aggressor, the next state would also be having opposing logic values on them, creating a worst-case switching pattern which requires the support of the assistant driver. In this situation, the transmission gates must be switched on and keeper transistors must be in cut-off, thereby connecting the $Victim_In$ node with the gate input of either MP_a and MN_a. When the present logic of the victim is zero, next possible logic value on the victim is one, and switching on the transmission gate TG_n and switching MN_k is sufficient. While the pull-down path is active, pull-up path is deactivated. In the event the present logic value on the victim is one and that of the aggressor is zero, TG_p will be turned off.

4.3.5 Energy Saving of the SMART Driver

(A) Energy Modelling

This section describes estimating energy components in detail with a particular emphasis on comparisons between traditional repeaters and the SMART driver. The energy consumed by the traditional driver is denoted E_{trad}, while the energy consumed by the SMART repeater comprises the energy consumed in the driving transistors, E_{smrt}, and also the energy consumed in the selection logic, E_{sel}.

Dynamic Energy Considering a single section of a buffered wire, the effective capacitance when two wires are coupled together is given in Table 4.1 for all possible switching patterns. In accordance with common terminology, the size of a traditional inverting repeater is defined in terms of multiples of a minimum sized repeater as H_t. Since the driving portions of the SMART driver are two inverters, they can be characterized in a similar fashion as H_m and H_a, which denote the sizes of the Main and Assistant drivers respectively. The total static capacitive load of the traditional driver, C_{w_trad}, can be defined as $\frac{C_s}{k} + H_t(C_{dmin} + C_{gmin})$; *i.e.* the sum of its own parasitic drain capacitance, the self capacitance of the wire, and the gate capacitance of the target load (a repeater for the purpose of this analysis) at the end of the wire. Here C_{gmin} and C_{dmin} are the gate capacitance and the drain diffusion capacitance of a minimum sized inverter, while k is the number of repeaters. Similarly, C_{w_smrt} can be described as $\frac{C_s}{k} + H_tC_{dmin} + H_mC_{gmin}$.

The energy dissipation per cycle depends on whether or not switching transitions occur, and on the relative switching pattern as given in Table 4.1. If all switching events are random uniformly distributed events with no correlation between neighbouring lines, the average energy dissipation per transition for wire i can be obtained by averaging out the dynamic energy consumption for each pattern. It can be shown to be:

$$E_{avg}^{dyn} = \frac{1}{16}\sum_{k=1}^{16}\frac{V_{DD}^2}{2}C_{eff}^k \tag{4.8}$$

Using equation 4.8, the dynamic energy dissipation for a wire buffered with k traditional repeaters is:

$$E_{trad}^{dyn} = \frac{V_{DD}^2}{32}\left(8C_{w_trad} + 6\frac{C_c}{k}\right) \tag{4.9}$$

and that for a wire buffered with k SMART repeaters is:

$$E_{smrt}^{dyn} = \frac{V_{DD}^2}{32}\left(4C_{w_trad} + 4C_{w_smrt} + 6\frac{C_c}{k}\right) \tag{4.10}$$

The dynamic energy consumption of the selection logic is:

$$\frac{1}{2}a_iC_{logic}V_{dd}^2 \tag{4.11}$$

where C_{logic} is the total effective load capacitance including parasitic capacitances, and a_i is the activity factor.

Short Circuit Energy Since the effective wire capacitance C_w changes dynamically as discussed at the beginning of this section, t_{sc} was calculated by averaging the value over 16 different switching patterns in the case of the traditional repeater. In the case of the SMART driver, in addition to the changing wire load, the load presented by the downstream repeater changes between $H_m C_{gmin}$ and $H_t C_{gmin}$, while the driver resistance changes between R_{dmin}/H_m and R_{dmin}/H_t. This effect was taken into account when calculating t_{sc} for the SMART driver.

In modeling the short circuit power consumed in the selector logic, the series connected PMOS/NMOS combination is represented by am equivalent single PMOS/NMOS device for the purpose of computing the driving resistance. This resistance is multiplied by the load capacitance to obtain t_{sc}, which is:

$$t_{sc_gate} \approx R_{gout}(C_{dout} + C_{gin}) \tag{4.12}$$

where R_{gout} is the equivalent output resistance of the gate, C_{dout} is the output capacitance , and C_{gin} is the input capacitance.

Leakage Energy The leakage energy dissipation of traditional driver and smart driver is estimated as described in section 2.5.3.

(B) Dynamic Energy Saving

An approximate expression for the saving in energy can be obtained by considering only the dynamic energy component. Subtracting Equation (4.10) from (4.9) gives:

$$\Delta E_{avg} = \frac{3}{16}(H_t - H_m) C_{gmin} V_{dd}^2 \tag{4.13}$$

This expression also neglects the energy consumed in the selection logic, and is derived to show the qualitative relationship between the relative sizing of the Main and Assistant Drivers and the energy saving. Since $H_t = H_m + H_a$, (4.13) reveals that the larger the Assistant driver, the greater the energy saving.

4.3.6 Design Methodology

A high-level design methodology for sizing the main driver and assistant driver is discussed in this section. The design methodology is highly dependent on the characterization of devices and the accuracy of the interconnect timing model. However, it is always better to make an estimation at a higher level in the design flow in order to be able to model the performance early in the design cycle. The methodology for smart repeater insertion is the same as traditional repeater insertion that is based on propagation delay.

(A) Smart Repeater Delay Modelling

The delay analysis for repeater insertion uses the characterization of a minimum-sized repeater in terms of an output resistance R_{dmin}, input gate capacitance C_{gmin} and output drain-diffusion capacitance C_{dmin} already introduced in Section 2.5.2. This simplification is justified for initial global signal planning and incremental physical optimization when detailed parasitics are not available to justify a more expensive analysis.

Delay Analysis with both Drivers Switching With the linearisation of the driver, the equivalent circuit for one repeater segment can be shown to be the circuit in Figure 4.6. Hence the 50% delay for the wire can be expressed as

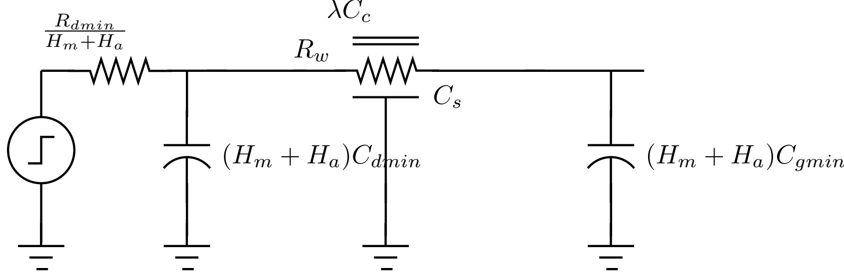


Figure 4.6: *Equivalent Circuit for the case when both drivers are switching.*

$$\begin{aligned} T_{MA} &= k\left\{0.7R_d\left(C_d + \frac{C_w}{k} + C_g\right) + 0.7\frac{R_wC_g}{k}\right. \\ &+ \left. 0.4\frac{R_w}{k}\frac{C_s}{k} + \lambda_i\frac{R_w}{k}\frac{C_c}{k}\right\} \end{aligned} \quad (4.14)$$

where k is the number of repeaters, and the parasitics are $R_d = \left(\frac{R_{dmin}}{H_m} \parallel \frac{R_{dmin}}{H_a}\right) = \frac{R_{dmin}}{H_m+H_a}$, $C_g = C_{gmin}(H_m + H_a)$, $C_d = C_{dmin}(H_m + H_a)$ and $C_w = C_s + \mu_i C_c$. Here H_m and H_a are the sizes of the Main and Assistant drivers respectively, and λ, μ is the switching factor. Since the Assistant driver switches only when adjacent lines switch in opposite directions, i=3. To simplify the delay equation, the following time constants are defined: $t_{Dout} = R_{dmin}C_{dmin}, t_{DWs} = R_{dmin}C_s, t_{DWc} = R_{dmin}C_c, t_{Din} = R_{dmin}C_{gmin}, t_{WD} = R_wC_{gmin}, t_{Ws} = R_wC_s$ and $t_{Wc} = R_wC_c$. This results in:

$$\begin{aligned} T_{MA} &= 0.7k(t_{Dout} + t_{Din}) + \frac{0.7(t_{DWs} + \mu_3 t_{DWc})}{(H_m + H_a)} \\ &+ 0.7t_{WD}(H_m + H_a) + 0.4\frac{t_{Ws}}{k} + \frac{\lambda_3 t_{Wc}}{k} \end{aligned} \quad (4.15)$$

Delay Analysis with the Assistant Quiet When the Assistant driver is quiet while the Main driver is switching, the gate capacitance of the Assistant will not add to the load, as it is disconnected by a switch in which the input capacitance is negligible compared to the Assistant driver's input capacitance (see Figure 4.7). However the parasitic drain-diffusion capacitance will always add to the load.

The delay expression is now

$$\begin{aligned} T_M &= k\left\{0.7\frac{R_{dmin}}{H_m}\left[C_d + \frac{C_s + \lambda C_c}{k} + H_mC_{gmin}\right]\right. \\ &+ \left. 0.7\frac{R_w}{k}H_mC_{gmin} + 0.4\frac{R_w(C_s + \lambda C_c)}{k}\right\} \end{aligned}$$

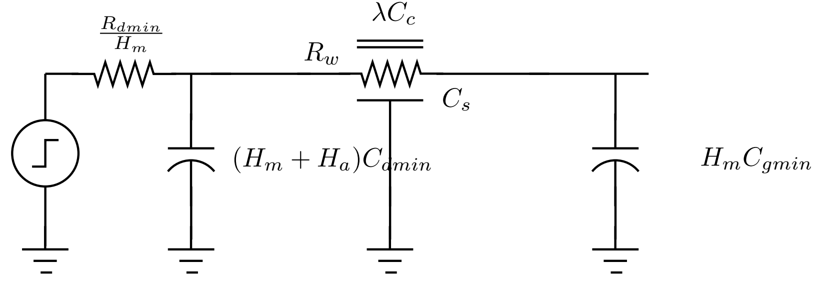


Figure 4.7: *Equivalent Circuit for the case when the Main driver is switching.*

Replacing the time constants defined earlier, this is reduced to:

$$\begin{aligned} T_M &= 0.7k\left[t_{Dout}\left(1+\frac{H_a}{H_m}\right)+t_{Din}\right]+0.7H_m t_{WD} \\ &+ \frac{0.7(t_{DWs}+\mu_i t_{DWc})}{H_m}+0.4\frac{t_{Ws}}{k}+\frac{\lambda_i t_{Wc}}{k} \end{aligned} \tag{4.16}$$

where $i = 0, 1$
Equations (4.15) and (4.16) are the two principal delay equations of the SMART driver for its two states of Main and Assistant drivers switching, and Main driver switching while the Assistant driver is quiet.

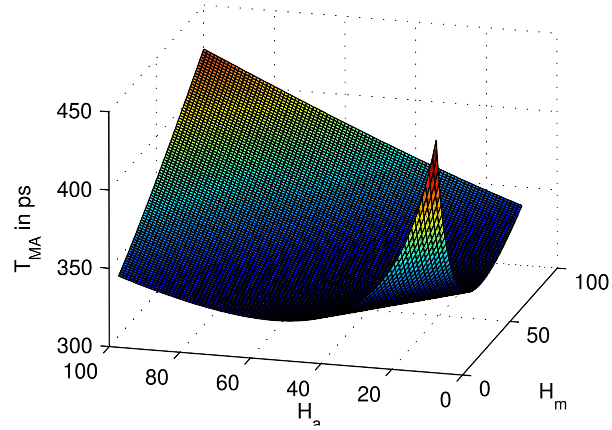


Figure 4.8: *The Variation of T_{MA} with H_m and H_a.*

(B) Delay Balancing with the SMART Driver

As explained earlier, the SMART driver saves energy by reducing the capacitive load for certain switching combinations, which in turn is achieved by switching off

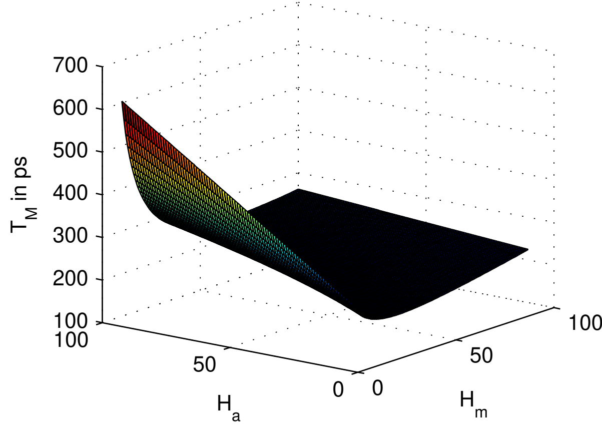


Figure 4.9: *The Variation of T_M with H_m and H_a.*

part of the driver. This means that the driver is essentially slower for the switching combinations that give rise to a lower capacitive load, and hence reduces jitter - the variation between the best-case and worst-case delays. This is a secondary benefit of the driver, and here we present a design methodology for sizing the SMART driver to minimize jitter.

The delay equations (4.15) and (4.16) predict a global minimum for the delay for *optimal* k,H_m and H_a values. The variation of T_{MA} and T_M with H_m and H_a is shown in Figures 4.8 and 4.9. T_{MA} is a convex function of H_m and H_a, and hence also of $(H_m + H_a)$. T_M is a convex function of H_m, while it has a linear dependence on H_a for a given value of H_m. This is a consequence of the fact that the assistant driver contributes a parasitic capacitance to the load while not contributing any drive strength for the switching combinations represented by T_M.

Since the Assistant driver switches only for the worst-case switching pattern defined by Group 5 in Table 4.1, the size of the Assistant driver, H_a, can be used to tune the delays for the other switching combinations defined by Groups 1 and 3-4. The expressions in (4.16) and (4.15) represent the delay for all these switching combinations. For clarity of explanation, say T_1, T_2 and T_3 are the wire delays for Groups 1, 3-4, and 5 respectively. Hence $T_1 = T_M|_{\lambda=0}$, $T_2 = T_M|_{\lambda=1}$ and $T_3 = T_{MA}$. Now increasing H_a increases T_M (see Figure 4.9), and hence H_a can be sized so that either $T_1 = T_3$ or $T_2 = T_3$ ($T_1 = T_2 = T_3$ is not possible because the relative delay variation between T_1 and T_2 is not a function of H_a).

The delay variation can be quantified as

$$\Delta T = T_{MA} - T_M.$$

By setting $\Delta T = 0$, delay balancing can be achieved. Substituting for T_{MA} and T_M from (4.16) and (4.15) and using the relation $H_{mDB} = H_t - H_{aDB}$ the following quadratic for H_{aDB} can be obtained.

$$AH_{aDB}^2 + BH_{aDB} + C = 0 \tag{4.17}$$

where

$$
\begin{aligned}
A &= 0.7t_{WD} \\
B &= 0.7\left[kt_{Dout} - t_{WD}H_t\right. \\
&+ \left.\frac{t_{DWs} + 2t_{DWc}}{H_t}\right] + \frac{(\lambda_2 - \lambda_1)t_{Wc}}{k} \\
C &= 0.7(\lambda_3 - \lambda_2)t_{DWc} + \frac{0.4(\mu_3 - \mu_2)t_{WC}}{k}
\end{aligned}
$$

Now sizing H_{aDB} to equalise T_1 and T_3 results in T_2 being larger than T_3, which may not always be possible due to constraints on T_3, the worst-case delay. However equalising T_2 and T_3 does not result in any such adverse effect. Here H_t and k can be calculated according to the strategy adopted for optimal repeater insertion, H_t can be calculated from (4.18) and k from (4.19). Where $H_t = H_m + H_a$

$$\frac{\partial T_{MA}}{\partial H_t} = 0 \Rightarrow H_t = \sqrt{\frac{t_{DWs} + mu_3 t_{DWc}}{t_{WD}}} \tag{4.18}$$

$$\frac{\partial T_{MA}}{\partial k} = 0 \Rightarrow k = \sqrt{\frac{0.4t_{Ws} + \lambda_3 t_{Wc}}{0.7(t_{Dout} + t_{Din})}} \tag{4.19}$$

Note also that H_a is not a function of the wire length, but is solely dependent on the crosstalk capacitance.

Feature size (nm)	180	130	90	65	45	32
L_{eff} (nm)	120	49	35	24.5	17.5	12.6
V_{dd} (V)	1.8	1.3	1.2	1.1	1.0	0.9
I_{dsat} ($\mu A/\mu m$)	554	1000	1100	1150	1200	1250
t_{ox} (nm)	4.2	1.6	1.4	1.2	1.1	1
V_{th} (V)	0.53	0.288	0.284	0.289	0.292	0.295
$I_{off}(nA/\mu m)$	20	30	50	70	100	150
$freq$ (GHz)	1.0	1.6	2.0	2.5	3.0	3.5
R_{dmin} ($k\Omega$)	8.27	14.15	16.62	20.82	25.40	30.48
C_{gmin} (fF)	2.31	0.43	0.25	0.14	0.077	0.043
C_{dmin} (fF)	2.00	0.49	0.33	0.22	0.15	0.10
Width(w)(nm)	525	335	205	145	102	70
Aspect Ratio (AR)	2.1	2.1	2.1	2.2	2.3	2.4
k_{ILD}	3.5	3.3	2.8	2.5	2.1	1.9
$r_w(\Omega/mm)$	38	93	249	475	919	1870
$c_s(fF/mm)$	36	34	29	26	22	20
$c_c(fF/mm)$	101	96	81	75	65	61

Table 4.3: *Buffer and Wire Parameters for Various Future Technologies based on ITRS [162] projections and [163].*

4.3.7 Energy and Delay Model Validation

The implementation has been carried out in a UMC 0.18 μm CMOS technology, with a V_{DD} of 1.8 V. All simulations are carried out using Cadence Spectre.. A

typical global metal layer is used for routing the bus, with a minimum pitch of 1050 nm. The wire and buffer parameters are given in Table 4.3. Each interconnect section is modelled as a distributed line with ten π segments, including capacitive coupling to one adjacent wire. It is trivial to extend the coupling to two neighbouring wires. This two-wire representative unit of a bus structure was fed with two uniformly distributed pseudo random bit sequences (PRBS) with a cycle time of 1ns and rise/fall times of 350 ps. In this particular design the selection logic propagation delay T_{sel} is 567 ps, which defines the maximum cycle time that can be achieved using this scheme.

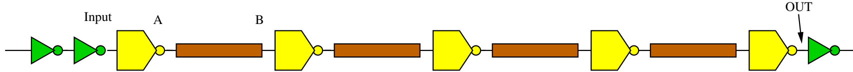


Figure 4.10: *Interconnect Link used for the Simulation - Total length is* 10mm, *and one segment is 2.5 mm.*

For the parameters described in Table 4.3, the driver parameters necessary to minimise delay have been derived using the delay balancing methodology outlined earlier. These results are presented in Table 4.4.

	Driver	Energy Dissipation (fJ)			E_{avg}(fJ)
		Group 1	Group 3/4	Group 5	
Model	Trad.	1507	1888	2195	939
	Smart	934	1274	2195	789
	Selector	77	77	77	77
	ΔE	33%	28%	-4%	7.8%
Simu	Trad.	1530	1735	1997	893
	Smart	994	1248	2065	753
	Selector	57	102	215	71
	ΔE	31%	22%	-14%	8%

Table 4.4: *Energy Dissipation for each switching group* ($H_a = 104$)

The ability of the proposed equations to calculate the true optimal buffer sizes and numbers to minimise delay was verified by simulating with different sizes and numbers around the predicted point. The deviation was found to be insignificant. As shown in Table 4.4, there is a good match for the calculated and simulated energy values, but the delay values deviate by approximately 25%. This is due to the inaccuracy in the Elmore delay, which however has the well known attribute of fidelity; *i.e.* the results of the design optimisation are similar to the results obtained using a more expensive model as the delay metric.

When Ha is within approximately 20% of H_t, the jitter prevalent in the SMART driver is similar to that of the traditional driver. However as the relative size of H_t increases, the jitter of the SMART driver is drastically reduced, as is evident in Figure 4.12.

The near-end and far-end crosstalk are respectively 0.113 V and 0.213 V with the smart driver whereas they are 0.069 V and 0.206 V with a traditional driver. There is a slight increase in the peak crosstalk voltage with a smart driver compared

to that of traditional but the peak crosstalk at the far-end is about 12% of V_{dd}, which is in the normal acceptable range of 20% of V_{dd}.

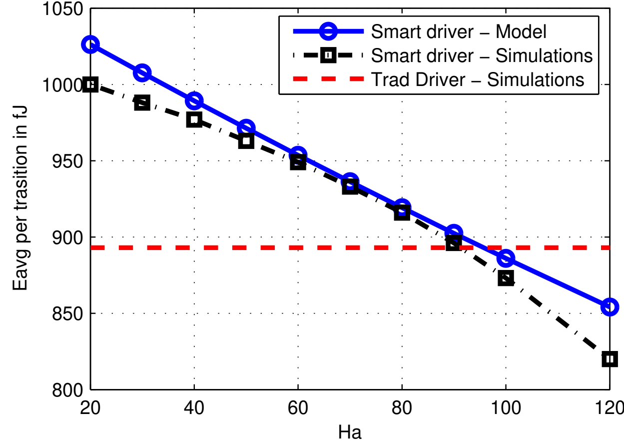


Figure 4.11: *Variation of Energy per transition with H_a.*

The simulations show that the energy models derived for the traditional and smart repeaters are accurate to within 95% of their simulated values as evidenced in Figure 4.11 and summarised in Table 4.4. As predicted by the model, increasing the size of the Assistant driver will increase the energy saving, although at the cost of increased delay, if the size is increased beyond the optimal (refer Figure 4.12).

It is evident from Table 4.4 that the energy loss introduced by the extra selection logic for switching patterns in Group 5, where both the Assistant and Main drivers switch, is more than offset by the energy saving for those patterns in Groups 1, 3 and 4 where the Assistant does not switch. On average, assuming equally likely occurrences of all patterns, the total energy saving is around 10%.

4.3.8 Impact of Technology Scaling

In this section the potential of the Smart Driver to save energy in future technology nodes is investigated. As the feature size decreases, the short circuit energy increases fairly sharply, which adversely affects the energy saving due to the fact that the Smart driver has a few transistors in the selector logic. However this is offset to some degree due to the relative decrease in area and the associated dynamic energy consumption of the selection logic in comparison to the driving inverters. Since global wires are scaled selectively, the wire parasitics remain approximately the same, or are worse, and the driving transistors see no reduction in size. In contrast, the selection logic can be implemented with minimum sized transistors, and the dynamic energy consumed becomes truly negligible. An analysis was carried out using ITRS predictions to derive the relevant technology parameters, as summarised in columns two through six in Table 4.3. The predicted total average energy saving in driving global length wires is shown in Table 4.5, highlighting the usefulness of the smart driver right up to the 32 nm node.

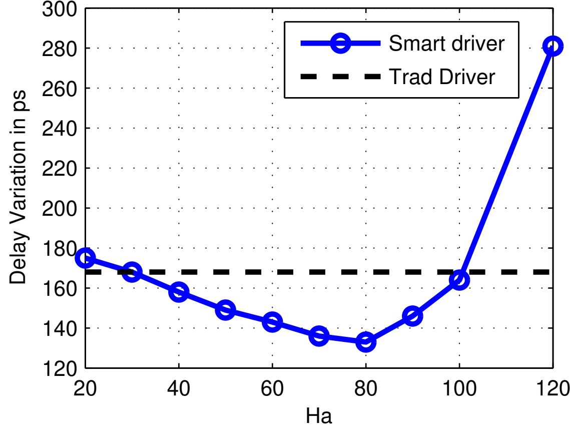


Figure 4.12: *Variation of Delay variation with H_a.*

Tech. node (nm)	130	90	65	45	32
k	14	24	36	54	84
H_t	325	268	277	278	282
H_a	202	162	163	158	154
E_{smrt}/E_{trad}	0.74	0.75	0.77	0.80	0.83

Table 4.5: *Energy Saving for Future Generations.*

4.4 Vertical Signal Transmission Methodologies

A wide variety of interconnect technologies suitable for layer-to-layer communication in 3-D ICs have been developed recently. Some promising techniques are: wire-bonding, micro-bumps, through-silicon-vias, and coupling type or contact-less interconnects. Individual dies are stacked and can be wire-bonded [*7*]; the connections between chips are made through the board or chip-carrier and back to other chips in the stack. In most of the contemporary handheld devices this approach is used. This approach is limited by the resolution of wire-bonding equipments and a large number of I/Os in the IC stack limit the applicability of the wire-bonding technique. To protect the pad from tearing off due to mechanical stresses during bond process, all metal layers are required.

In micro-bump technology, connections are made using solder or gold bumps on the surface of the die. The micro-bump pitch is typically around 50-500 μm [*164*]. An epoxy routing tier has micro bumps bonded to it and this brings the signals to the edges of the cube, where the different tiers are then stacked together. Since assembly related mechanical stresses are less, the pads require a maximum of two layers. Here, dies are assembled into a cube. Compared to wire bond technology, micro bump technology provides greater vertical interconnect density.

Through-Via interconnects are a promising technology for wafer-level three-dimensional integration [*99*, *100*, *101*, *102*]. Several world leading companies including IBM, IMEC, Intel, Samsung, NEC, Elpida, and Tezzaron are developing

TSV methodologies optimized for their applications. The key enabling technologies for wafer-level-stacking include: electrically isolated Through-silicon vias, thinning of wafers (usually less than 50 μm), precision alignment of wafer-to-wafer or die-to-wafer, and bonding.

In wafer-level stacking, wafers can be attached to one another with organic glues, oxide bonding, or metal bonding using either a face-down or face-up approach, and electrical connections between wafers are provided by vias [*165*]. The approach used to create through-wafer interconnects can be via-first or via-last. Via-last processes create the interconnect after the wafers are bonded, using a *drill and fill* sequence [*166*]. Via-first processes build the through-wafer via on each wafer prior to the bonding process. Both methods have the potential to offer the greatest interconnect density with the disadvantage being the greatest cost, but the via-first process is generally more efficient and cost-effective [*167*]. Each of the bonding processed mentioned can support via-last or via-first processes, but for the metal bonding process, via-first is preferred [*165*].

Non-wiring interconnect solutions such as AC coupling and RF wireless methods have also been been proposed in the literature [*168*, *169*, *170*, *171*]. This method eliminates the signal interconnect connection to the periphery of the IC as well as inter-tier routing, and the major benefit is it does not sacrifice active layer area. This particular technology is most suitable for AC signal transmission, but for DC signals such as power and ground connections, through silicon vias should still be used. The distance between the tiers, the rise/fall times of the signal, and the dielectric constant of the gap decide the density of these interconnects. However, the power supply between chips is provided by the help of bumps. The capacitive coupling type communication requires the tiers to be face-to-face and the plates should be placed as close as possible for better coupling. And, it limits the number of tiers to two. Either high-k dielectric or trench formation is used to achieve better capacitive coupling. Inductive coupling is more suitable wherein separation of the coupling elements is of the order of the lateral dimension of the coupling elements. However, in inductive coupling the feasibility of using free active area beneath inductors for circuitry needs to be investigated.

RF wireless interconnects (RF-WIs) have been researched for many years for board level communications and it can be extended to 3-D ICs [*172*]. This is simply a form of a WLAN inside a chip with transceivers, on-chip antennas and signal generation and detection circuits.

4.4.1 Signal Transmission Characteristics of TSV interconnects

As the use of TSVs is a fairly a recent concept, their effects of signalling within 3-D ICs are not well documented. The goal of this section is to provide an assessment of the effect a TSV has on signal integrity within a realistic context and providing recommendations on suitable model for TSV, sizing of drivers, repeater insertion to benefit the integrity and efficiency of the vertical bus system.

(A) Significance of TSV parasitics

First of all, a suitable model for a TSV should be investigated. A lone TSV, which does not include any coupling effect with other TSVs, driven by a 10X inverter

and loaded by a minimum sized inverter in a 0.35 μm technology is simulated in a SPICE environment to determine its latency. The driver was found to be the optimum size, given the TSV parasitics, by a series of sweeps and the minimum-sized inverter represents the pin load for each vertical interconnect. A 50 ps rise time was employed throughout all of the simulations. Simulations were performed for the entire range of resistance, inductance and capacitance values as determined by a field solver. The resistance was swept from $0 - 500\ m\Omega$, and the output waveforms were plotted to observe variations in delay. As can be seen in Figure 4.13, the TSV resistance within the considered range is so small that it has no observable effect on the output waveform. The inductance was then swept from 0 - 500 pH, the extracted range, for rise times down to 1 ps, revealing no significant contribution as seen in Figure 4.14. Finally, the capacitance was swept from 0 - 500 fF showing a significant effect on latency of the output waveform, as seen in Figure 4.15. These results appear to show that the electrical model for a cylindrical TSV can be reduced to a purely capacitive model. The resistive and inductive parasitics are small enough to be neglected in any delay simulations, which reduces the complexity of the electrical model significantly.

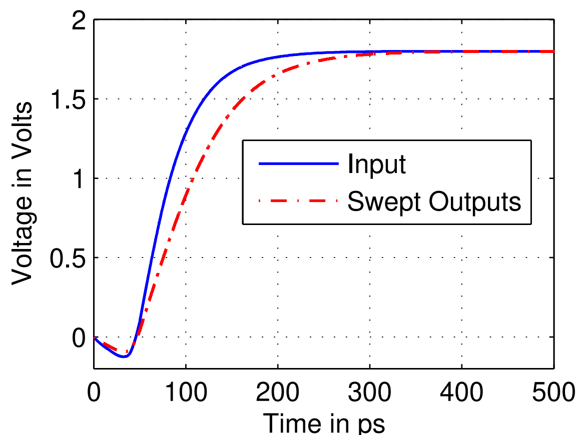


Figure 4.13: *TSV Resistance Sweep.*

In addition to parametric sweeps, simulations were conducted to determine if distributed models were necessary to attain accurate results. The model was segmented into 2, 5, and 10 sections and output waveforms examined to show no significant effect on the signal from increasing the number of segments within the parasitic range determined by the field solver. The relatively low resistive and inductive terms reduce the necessity for a distributed model for simulation of signals within the considered range.

(B) Crosstalk in a TSV Bundle

This section investigates the effects of crosstalk between TSVs organized in a 3x3 bundle. By employing electrical models derived from the field solver simulations, various switching patterns are simulated to analyze crosstalk effects between these

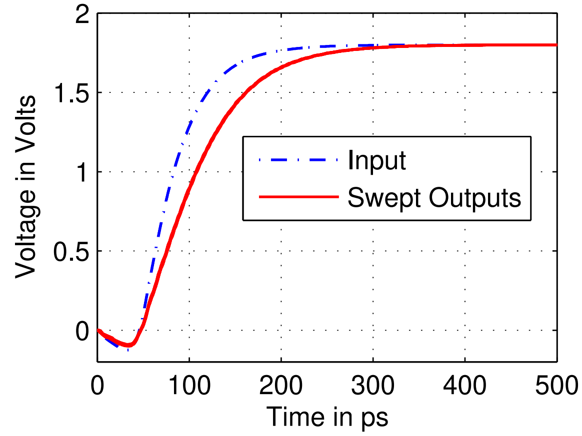


Figure 4.14: *TSV Inductance Sweep.*

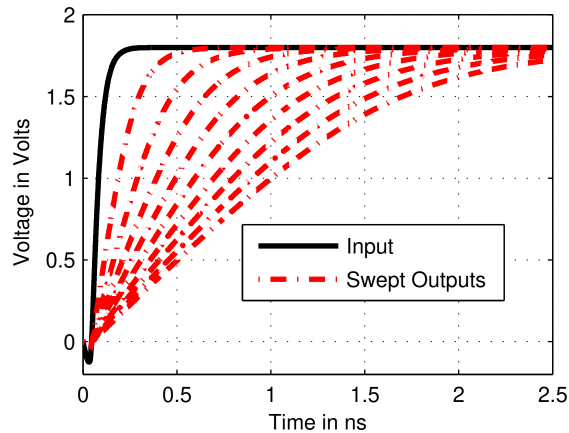


Figure 4.15: *TSV Capacitance Sweep.*

structures. The coupling capacitance between two TSVs is a function of radius, length and inter-via spacing, as well as dielectric barrier thickness and permit detrimental effects on bandwidth and signal integrity, the crosstalk between adjacent structures must be examined to determine the most efficient use of area and TSV sizing to maximize signal throughput and reliability.

The first simulations performed determine the individual contributions of mutual inductance and capacitance terms when 8 aggressors switch simultaneously on a silent victim net. For up to twice the maximum mutual inductance extracted by the field solver for the considered range, the coupling turns out to be insignificant for rise times down to $1ps$. This is born out in Figure 4.16 which shows minor oscillations near the aggressor transition points. The capacitive coupling on the other hand is significant with a coupled noise amplitude of up to 15% of V_{dd}. Figure 4.17

shows a subset of waveforms within the considered geometrical range, illustrating this.

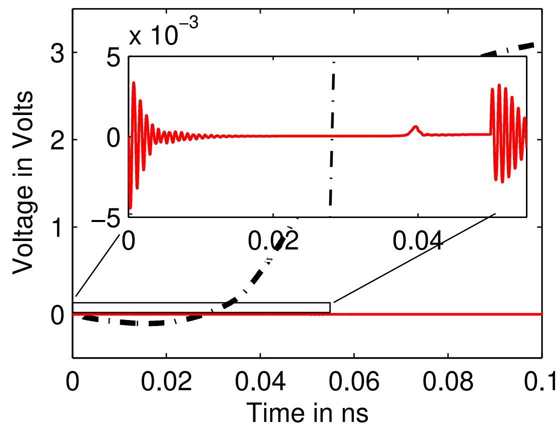


Figure 4.16: *Inductive coupling on a Silent victim.*

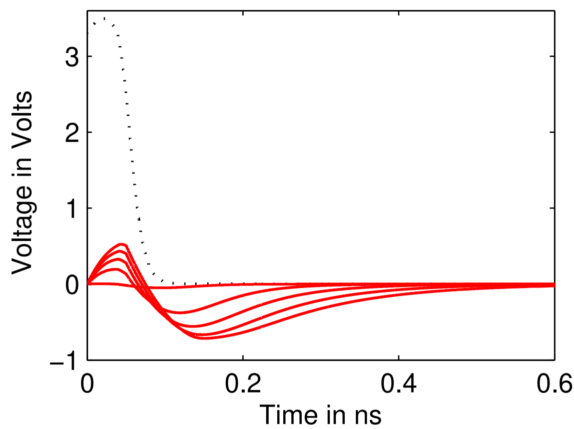


Figure 4.17: *Capacitive coupled noise on silent victim.*

The effect of inductive coupling on the victim net does not appear to be large enough to justify the modeling of parasitic mutual inductance in a 3×3 bundle for signals. It is however possible that the simultaneous switching of many different aggressors, including non-adjacent ones in a larger bundle, can produce a more significant effect. This is due to the inductive coupling having measurable effects over a long range. Capacitive coupling however needs to be considered at the outset. Simulations were performed for a variety of TSV geometries and pitches to provide a clearer picture of the capacitive crosstalk within a 3×3 TSV bundle. Simulations were performed to highlight the effect of crosstalk on delay in a bundle when the 8

surrounding TSVs remain quiet, switch in the same direction and in the opposite direction.

The simulation result demonstrates significant crosstalk on delay effects in spite of the relatively low interconnect density. For example, within a TSV bundle with a pitch of 100 μm and lengths and radii of 20 μm and 40 μm respectively, the 50% delay of the victim, 393 ps, is greater than that for an isolated TSV, 157 ps. For the selected geometrical configuration, the delay variation between best-case and worst-case switching patterns was 36 ps to 135 ps, a 4-fold difference over the minimum delay. As the interlayer connectivity in a 3-D IC has to be achieved by a high dimensional TSV bundle accounting for coupling effects will become paramount in designing high performance, reliable systems.

Given that the investigations reveal a lumped capacitive equivalent circuit as being sufficiently accurate, the switching pattern dependent delay within a bundle can be accurately captured by a first-order Elmore delay model. For the entire range of geometrical configurations considered, the delay can be accurately estimated by (4.20) where the empirically determined switch factors λ and μ for the various switching patters are defined in Table 4.6. Here C_s is the self capacitance, C_{lat} the lateral coupling capacitance, C_{diag} the diagonal coupling capacitance and R_d the driver resistance.

$$t_{d_tsv} = 0.69R_d \left(C_s + \lambda C_{lat} + \mu C_{diag}\right) \tag{4.20}$$

Switching Pattern			λ	μ
Victim	Lateral	Diagonal		
↑	↑	↑	0	0
↑	–	–	3.4	5.2
↑	↓	↓	9.0	10.6

Table 4.6: *Switch Factors for delay estimation*

The minimum accuracy of this equation over the entire range was greater that 92%, principally due to the negligible parasitic resistance inherent in the TSV. These switch-factorbased delay equations facilitate the integration of TSV interconnects into established on-chip static timing methodology.

(C) Signal Integrity Simulations

In order to fully capture the effect of crosstalk on delay and coupled noise amplitude under real-world conditions, simulations were carried out with pseudo-random bit streams (PRBS) at the victim and aggressor inputs in a 3×3 bundle to generate the eye diagrams at the output. All drivers were size 10 inverters while every TSV was loaded with a minimum sized inverter. The example geometry chosen was a bundle with radii, length and pitch of 15 μm, 20 μm, 50 μm respectively.

The eye diagram in Figure 4.18 shows the response of the victim line when the PRBS speed is 10 GBPS with signal rise and fall times of 10 ps. It is clear that the eye is very narrow and the variation in delay has widened to an unacceptable level. At this speed the crosstalk completely overpowers the signal on the victim.

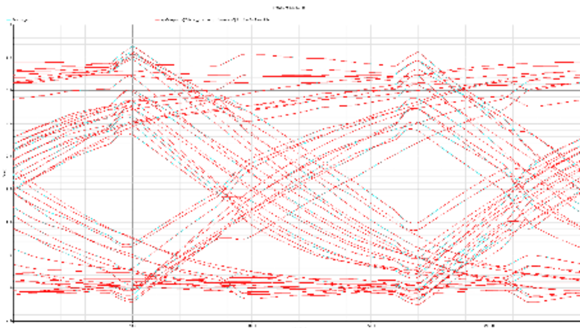


Figure 4.18: *Victim eye diagram 10 GBPS.*

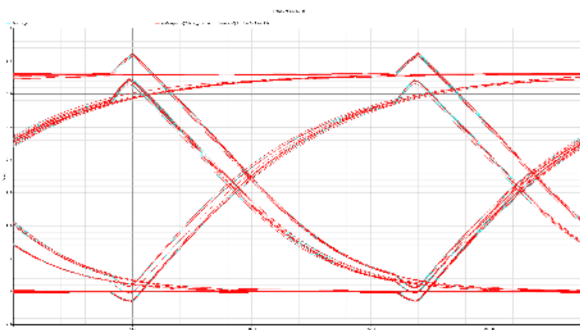


Figure 4.19: *Victim eye diagram 10 GBPS with lateral TSVs shielded.*

Since the lateral (N, E, S, and W) neighbours in the bundle contribute the majority of the capacitive coupling, an obvious strategy to counteract capacitive crosstalk at high signaling speeds is to use these lines as shields. As seen in Figure 4.19, this effectively eliminates the majority of the coupling and allows for higher bit rates through the interconnect.

It is clear that that judicious shielding has opened up the eye and reduced the delay variation significantly. The main drawback to this method is that although higher bit rates can be achieved, significant area loss occurs due to the unusable grounded lines. Investigation into optimal configurations for TSV sizing, spacing and shielding has to be performed to determine best configuration for the highest bandwidth achievable in a 3-D device.

4.4.2 Signalling Link Design for Layer-to-Layer Communication

Shown in Figure 4.20 is a general TSV based layer-to-layer interconnect link for 3-D ICs. It consists of two on-chip repeater inserted global wires in layer i and $(i+1)$, and their interconnecting TSV. Given the fact that a TSV can be modelled as a lumped capacitor rather than a RC wire segment [*173*], the TSV which connects the global wires in layers i and $(i+1)$ acts as capacitive load, because TSV resistance

is considerably smaller than that of a typical on-chip global wire resistance. It is important to mention that when the TSV dimensions are very small the overall capacitance is comparable to a minimum sized gate. For example, for an isolated TSV with r_v =5 μm, l_v = 20 μm, and d_b = 0.2 μm, the capacitance is more or less 4 fF, whereas with r_v = 40 μm, it is around 25 fF. A middle TSV in a closely packed bundle on the other hand can have a significant parasitic capacitance; field solver simulations revealed that with r_v = 5 μm and s_v = 20 μm M TSV has C_s = 0.5 fF, C_{lat} = 9.02 fF and C_{diag} = 3.00 fF for l_v = 140 μm, and C_s = 0.298 fF, C_{lat} = 1.31 fF and C_{diag} = 0.458 fF for l_v = 20 μm. The two resulting are effective capacitances are 113.48 fF and 16 fF respectively, when all the neighboruing TSVs switch in the opposite direction. The first case represents a TSV spanning several layers in which its capacitance is more or less similar to a 1 mm long global wire in nanometer technology. This is an example which amply demonstrates the need of early analysis of the overall vertical link to achieve desired signal integrity; rather than just connecting just two wires in different layers, TSVs require properly sized drivers such as cascaded buffers fabricated before connecting them vertically.

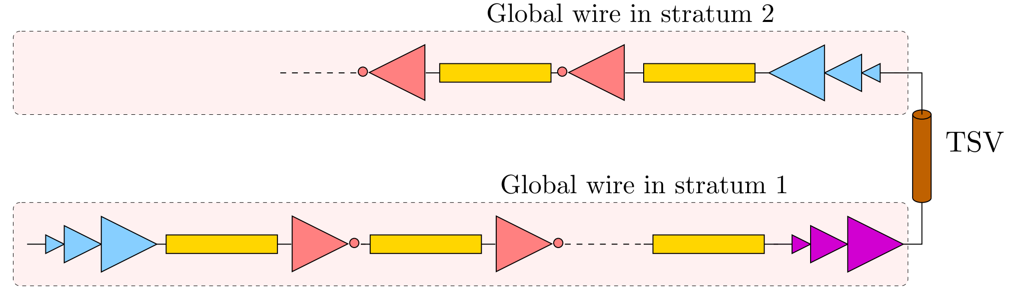


Figure 4.20: *Repeater inserted global interconnect in 3-D IC.*

When the technology scales down, C_{gmin} reduces (see Table 4.3), and therefore, a minimum sized inverter takes a significantly higher duration to charge a comparatively large capacitive load. In such a case the usual and intuitive method is to use a progressively increasing sized (say, u) chain of inverters, starting with a minimum sized inverter. Then the number of inverter stages required in the cascaded buffer is given by [*28*]:

$$N = \frac{ln(x)}{ln(u)}, \tag{4.21}$$

With x being the ratio of load capacitance (C_{tsv}) and input capacitance of the first inverter (C_{gmin}). Also, depending on the switching pattern C_{tsv} varies as:

$$C_{tsv} = C_s + \lambda C_{lat} + \mu C_{diag} \tag{4.22}$$

Sizing a TSV driver for the best case for smaller pitches will result in a smaller number of driver stages, because C_s is comparatively small. However, the usual and intuitive approach is to size the drivers for the worst-case coupling scenario, when all the neighbours switching in the opposite direction to that of victim. The

total delay of the cascaded buffer in this case is:

$$\tau_{drv} = N\left[0.69R_d(C_d + uC_g)\right] \tag{4.23}$$

$$= \frac{ln(x)}{ln(u)} \underbrace{0.69R_dC_d}_{=\tau_0} \left(1 + u\frac{C_g}{C_d}\right) \tag{4.24}$$

Equating the derivative of τ_{drv} with respect to u to zero produces the optimal scaling factor, and thereby the following implicit relation for u is found.

$$1 + \frac{C_d}{uC_g} = ln(u) \tag{4.25}$$

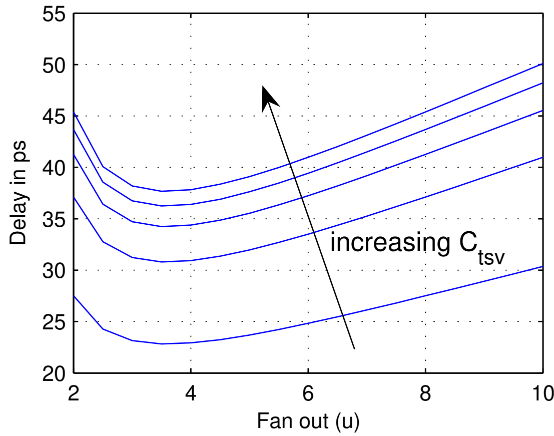


Figure 4.21: *The variation of delay of a cascaded buffer with u for various C_{tsv} values with $C_{gmin} = C_{dmin} = 0.1e\ fF$, and $R_{dmin} = 20\ k\Omega$.*

Increasing u results in increasing the minimum delay point. However, setting $C_d = C_g$, the approximate value for most current technologies, u evaluates to be 3.6 [28]. For real circuits with parasitics, a fanout of 4 (FO4) per stage or $u = 4$ is good, giving roughly minimum delay and also reducing the number of stages required. As Figure 4.21 depicts, for u to be anywhere in the range of 2-6, the delay is within few percent of optimal. Therefore, using a precise value of $u = 4$ is not very important; depending on the needs of the circuit a higher or lower value can be used.

4.5 Summary

In this chapter existing and proposed on-chip global signalling technologies were presented with an emphasis on their advantages and disadvantages. For a smart repeater circuit, a comprehensive delay and energy analysis, and a design methodology to obtain the optimal repeater configurations for minimising delay while also minimising jitter has been presented. The issue of reducing energy consumption is

addressed by exploiting the switching-pattern-dependent delay of repeater inserted global wires. The proposed smart repeater circuit was implemented in an UMC 0.18 μm CMOS technology and tested for proof of concept. The average energy saving was shown to be around 10%, and the jitter reduction to be 20% for a data rate of 1 GB/s. A comprehensive delay and energy analysis was presented, including a design methodology to obtain the optimal repeater configurations for minimizing delay while also minimizing jitter. Further, as processes scale, the selector latency shrinks, and higher data rates can be achieved. The total energy saving that can be achieved by the SMART driver in future nanometer technologies is found to be in the range of 20% - 25%.

Vertical signal transmission technologies have also been discussed in the chapter with an emphasis on signal transmission characteristics of TSVs in isolation as well as in a bundle. For the considered range, simulations show that resistance and inductance are mostly negligible for latency and SI considerations and therefore signal propagation through an isolated TSV as well as a TSV in a bundle can be analyzed by considering the capacitance alone. Tests were also carried out to determine if the via should be treated as a lumped or distributed model. The results show that no benefit is conferred by considering a distributed model due to the relatively low resistance and a single lumped section is sufficiently accurate.

Furthermore, crosstalk effects between TSV structures in a 3×3 bundle were examined. Capacitive crosstalk is far greater than inductive crosstalk, such that inductance can be ignored in most cases. Due to the reduced complexity of the TSV electrical model as proposed in this thesis, simplified delay formulae based on the Elmore delay and empirical switch factors were proposed to estimate delay in a TSV bundle with a maximum error contained to within 8% over the entire simulated range. These equations allow for preliminary assessment of delay for worst, nominal and best case switching scenarios in accordance with well-established timing analysis practice. Simulations were carried out using eye diagrams to further investigate SI issues, demonstrating the effect of capacitive coupling in a TSV bundle with random switching patterns. Shielding the lateral TSVs in a bundle was shown to increase signal reliability and allow for faster speeds through the structures. It is expected that this study will provide the basis for further explorations through the recommendation of the equivalent circuits as well as the investigations on the relative importance of the various parasitic terms, providing insight into signaling schemes over TSV interconnects.

5

IC Cost Modelling at the System Conceptual Level

Several interdependent models are required to adequately capture essential cost and performance characterization of Si chips and package implementations at an early stage of the design cycle. This chapter introduces a system-level methodology to estimate the cost of a typical integrated circuit which can be used in early trade-off analyses.

5.1 Introduction

It is becoming increasingly common for electronics to be used in conjunction with other technologies in the operation of an overall system, to provide complex control functions more compactly and at lower cost than could be provided otherwise. Therefore, the design of electronic systems is placing increasing demands on designers with respect to factors such as performance, reliability, cost, and design time. This is leading to a move away from the more traditional and ad hoc design methods toward the adoption of a more structured approach to the design process.

The silicon cost was the dominant factor in all cost calculations and estimating chip cost was a simple matter of determining die size. Silicon remains a major variable in the equation, but it has become necessary to think outside the die. As IC designs become increasingly complex, factors other than pure die size can have a huge impact on the final chip cost. This chapter outlines a basic approach which enables to estimation of the final cost of the IC or module based on the available gate count, processing technology and choice of package.

5.2 Cost Analysis and Modelling

The objectives of cost modeling in the early stage of the design cycle, sometimes known as the *system conceptual and planning stage*, is to predict the final cost of a product, to analyze its cost structure and to identify a cost effective optimum production solution for the electronic system. The last objective is of great interest to industry, being a critical criterion of product success. The main parameters

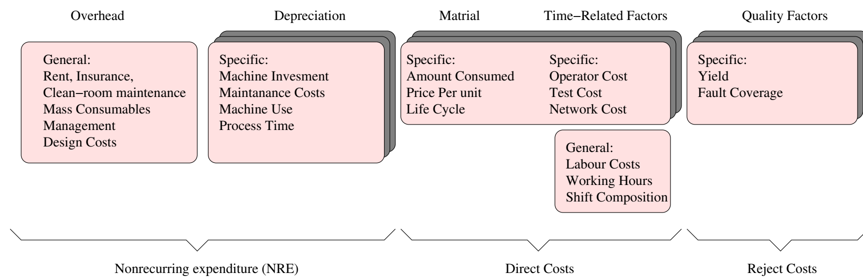


Figure 5.1: *Cost Model Parameters (reproduced from [174]).*

required to predict the product cost structure and final cost are: material and component costs, labour costs including design, manufacturing and test generation, equipment cost and amortization, and yield and test costs.

- Design Effort: These costs are very significant in low-volume products. However, for mass-produced products, these costs are less important.
- Silicon Area: Based on the technology and production line used for manufacturing, a price per square meter of processed silicon can be estimated. It is obvious that the larger the area the higher the cost.
- Production Yield: Since the production of defective devices reflects on the price of fault-free ICs, the yield of the manufacturing line is an important parameter in determining the price of mass-produced ICs.
- Package Costs: The cost of the package itself and of packaging a device are relatively high. Minimizing the number of defective ICs being packaged and optimizing the packaging yield will reduce the total costs.
- Test Costs: Increasing complexity of ICs, means functional tests are more complicated, and take increasing amounts of time on expensive general-purpose Automatic Test Equipment (ATE). Currently, test-related costs may constitute a significant part of the total costs of an IC (even up to 50%, especially for complex mixed-signal ICs).

To completely assess the cost of a potential product implementation all the above elements and related input parameters need to be considered. Table 5.1 summarizes a simplified IC fabrication and a module assembly process flow and key input parameters for each step.

5.2.1 Rent's Rule

E.F. Rent of IBM found an interesting relationship between the number of signal pins on a circuit and the number of logic components in it [13]. On a log-log plot, these data points describe a straight line, and the relation the form:

$$N_p = k_p \cdot N_g^p \tag{5.1}$$

Step	Process	Key Parameters
1	Wafer Fabrication	Die Area, Number of IO Defects per unit Area, Process cost, Number of Mask Layers
2	Wafer-Level test	Number of Gates, Test Coverage Defectivity of the incoming die
3	Dicing	Number Up
4	Burn-In	Burn-in Time, Burn-in Costs
5	Die Test	Number of Gates, Test coverage Defectivity of the incoming die
6	Assembly	Cost and Quality of Incoming materials, Number of dies, Assembly Cost, Assembly yield
7	Test module	Number of dies, Number of gates, Defectivity of the incoming module Test Coverage
8	Rework	Cost of Rework, Assembly Yield, Rework Yield

Table 5.1: *IC Fabrication and Module Assembly Flow and key Inputs for Cost Prediction[175]*

System/Chip Type	ρ	K_g
Gate array	0.5	1.9
Microprocessor	0.45	0.82
Static Memory	0.12	6
High-speed Computer (Chip and Module level)	0.63	1.4
High-speed Computer (Board and System level)	0.25	82
Functionally complete logic chip	0.21	7
Functionally incomplete logic chip	0.434	3.2

Table 5.2: *Rent's Constants for various systems or chip types*

where N_p is the number of pins, ρ Rent's exponent$(0 < \rho < 1)$, K Rent's coefficient, and N_g is the number of logic gates or logic partition on the chip. This relationship was first revealed by Landman and Russo in [*176*] and also by Chiba [*177*].

Usually, ρ is constrained to values in the range of 0 and 1, inclusive, the two extreme values interpreted as representing completely serial and completely parallel circuits, respectively. It is one way of measuring how many logic gates in the circuit block communicate to the outside world. Many of the gates in a circuit will only need to interact within the circuit block and some of them will need to interact to the outside in order to achieve the desired functionality. For example, microprocessors have few external connections as the chip is designed such that interaction is mostly within gate's but in gate arrays, since the functionality is typically defined after the chip is floor-planned, there are a lot more external pins.

Backoglu, in his well known book [*6*], has examined the parameters for a variety

of systems such as gate arrays, microprocessors, memory, random logic, and boards and systems. That examination is valid only for homogeneous systems, but not for more complex systems, where several different architectures are integrated to form a SoC; heterogeneous systems, for example, contain a memory, datapath, and some random logic. A new heterogeneous Rent's rule described in [*178*] argues that the same power-law expression as 5.1 is valid for heterogeneous systems with equivalent K and ρ parameters:

$$K_{eq} = \sqrt[N_{g_eq}]{\left(\prod_{i=1}^{n} K_i^{N_{g_i}}\right)} \tag{5.2}$$

$$\rho_{eq} = \frac{\sum_{i=1}^{n} \rho_i N_{g_i}}{N_{g_eq}}$$

where K_i and ρ_i are the usual Rent's rule parameters, N_{g_i} is the number of gates in i^{th} the block and $N_{g_eq} = \sum_{i=1}^{n} N_{gi}$.

5.2.2 Wire-Length Modelling

Early estimation of achievable wiring length is important as signal delay and power consumption depend on the interconnect length.

Rent's rule can be applied recursively to smaller and smaller logic blocks, and by comparing the external communication requirements of different size blocks, the average wire length can be determined. The most simple method to find the average interconnection length is by assuming that exactly half of the wires go to a module's nearest neighbours (F_p) and exactly half go to the next next nearest neighbour ($2F_p$), where F_p is the gate pitch. Hence, the average wire length $\overline{R_m}$ in gate pitches is equal to 1.5. However, there are both theoretical and empirical treatments for wire length prediction based on Rent's rule, for instance Donath's [*179*], Gamal's [*180*] Feuer's [*181*], Mikhail's [*182*] and Davis's [*183*] models.

In Donath's Model, a probabilistic, hierarchical scheme is used and a closed-form result for average wire length obtained. The average wire-length derived from Donath's model is an upper bound and hence gives a conservative performance values. By partitioning the design into hierarchical divisions and calculating the number of connections between the partitions via Rent's Rule, the average interconnection length, which is given by gate pitch, can be determined[*179*].

$$\overline{R_m} = \begin{cases} \frac{2}{9}\frac{1-4^{(\rho-1)}}{1-N_g^{(\rho-1)}}\left(7\frac{N_g^{(\rho-0.5)}-1}{4^{(\rho-0.5)}-1} - \frac{1-N_g^{(\rho-1.5)}}{1-4^{(\rho-1.5)}}\right) & \text{for } \rho \neq 0.5 \\ \frac{2}{9}\frac{1-4^{\rho-1}}{1-Ng^{\rho-1}}\left(7log_4 N_g - \frac{1-N_g^{\rho-1.5}}{1-4^{\rho-1.5}}\right) & \text{for } \rho = 0.5, \end{cases} \tag{5.3}$$

Davis'theory is widely used in many works in order to find both average interconnect length and interconnect length distribution. Davis's expression of inter-

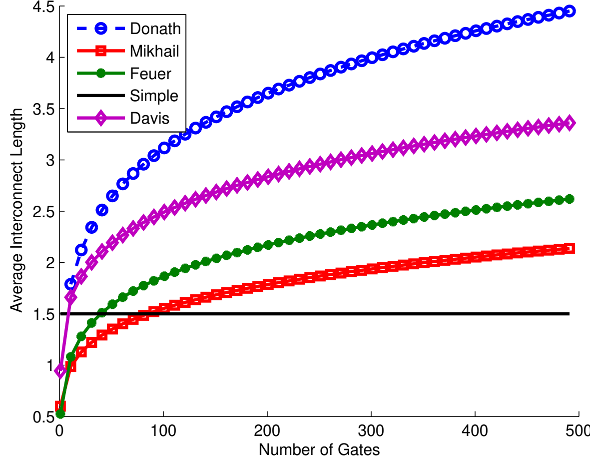


Figure 5.2: *Comparison of Average Interconnect Estimation Methods*

connect length distribution is

$$I(l) = \begin{cases} \frac{k\Gamma f_g}{(f_g+1)}\left(\frac{m^3}{3} - 2\sqrt{N_g}m^2 + 2N_g m\right) m^{(2p-4)/2} & \text{when } 1 < m \leq \sqrt{N_g} \\ \frac{k\Gamma f_g}{(f_g+1)}\left(2\sqrt{N_g} - m\right)^3 m^{(2p-4)/6} & \text{when } \sqrt{N_g} < m < 2\sqrt{N_g} \end{cases} \tag{5.4}$$

where,

$$\Gamma = \frac{2N_g(1 - N_g^{p-1})}{\frac{(2^{2p-1}-2p-1)N_g^p}{p(2p-1)(p-1)(2p-3)} - \frac{1}{6p} + \frac{2\sqrt{N_g}}{2p-1} - \frac{N_g}{p-1}}$$

Then the average interconnect length is:

$$\overline{R}_m = \frac{\left[\frac{1}{p} - \frac{\sqrt{N_g}}{p-0.5} - \frac{1}{6\sqrt{N_g}(p+0.5)} + N_g^p\left(\frac{-p-1+4^{p-0.5}}{2(p+0.5)(p-0.5)p(p-1)}\right)\right]}{\left[N_g^{p-0.5}\frac{-2p-1+2^{2p-1}}{p(2p-1)(p-1)(2p-3)} - \frac{1}{6p\sqrt{N_g}} + \frac{1}{p-0.5} - \frac{\sqrt{N_g}}{p-1}\right]} \tag{5.5}$$

In actual dimensions, when d_g is gate dimension in, for example, *nanometers*, the average interconnection length is $l_{av} = \bar{R}_m d_g$.

5.3 Bare Die/Packaged Chip Cost Analysis

The cost modeling methodology for a bare-die or a packaged die is shown in Figure 5.3. If not provided by the IP vendor the area of a digital module implemented in some target technology can be estimated in a straightforward manner, using gate information and technology scaling. In this process the key input parameter is the gate count in a chip, which allows one to calculate the number of I/Os required for external communications, core area and hence, the required die area. The very

basic cost formula for a bare die is just the cost of a wafer divided by the product of number of dice per wafer and die area. However, due to manufacturing defects not all the dice obtained from a wafer function as desired. Hence, the number of dies which can be obtained from a wafer and effective yield need be incorporated to the model to the model along with chip design cost, fabrication cost, packaging, and testing cost for to estimate the final cost.

However, the area of an analog chip depends not only on the number of transistors and their sizes (in practice, minimum sized transistors are not used in analog circuits), but also on the circuit architecture. For example, in a Voltage-Controlled Oscillator (VCO), the area of the on-chip inductor may be hundreds of times larger than that of a transistor. In an Analogue-to-Digital-Converter (ADC) or Digital-to-Analogue-Converter (DAC), on-chip resistors and capacitors may occupy a large fraction of the total area. Full custom design experiences are necessary to estimate the size of an analog chip. The models for core area described below are for digital implementations; it is assumed that area information for analog blocks is available. However, all models following on from the core area are valid for mixed-mode systems. Instances where variations with respect to pure digital systems occur are identified and supported by appropriate models. The following sections describe the methodology for area, yield, and testing cost analysis for a bare or packaged die.

A detailed description of the chip-level calculation of parameters such as average interconnect length, gate packing density and then chip area are given in this section.

5.3.1 Packing Density and Area

The packing density of a chip is basically limited by transistors when all the transistors/blocks can be placed right next to each other. However, there is a situation where transistors cannot be placed right next to each other when it is not possible to wire them without consuming additional space. The area occupied by the transistors and their interconnects is termed the core area (A_{core}) of the chip. This area can either be interconnect-capacity limited or transistor-area limited depending on the logic type and the available resources such as number of metal layers [*6*]. For example, memories usually have a very regular structure and do not require many interconnection levels, resulting in a very high packing density. However, digital logic circuits are less regular and require more connectivity, resulting in the area being either interconnect-limited or gate area limited.

(A) Packing Density Limited Core Area

Packing density limited core area is estimated from:

$$A_{die} = N_g A_g \tag{5.6}$$

where A_g is the area of an average logic gate. For static CMOS, the average logic gate is considered to be a 2-input NAND gate with a fanout of 3 identical NAND gates; for dynamic logic, [*184*] proposes a 2-input NOR with fan-out of an inverter as the representative gate. The reasons for these choices are that the NAND gate is

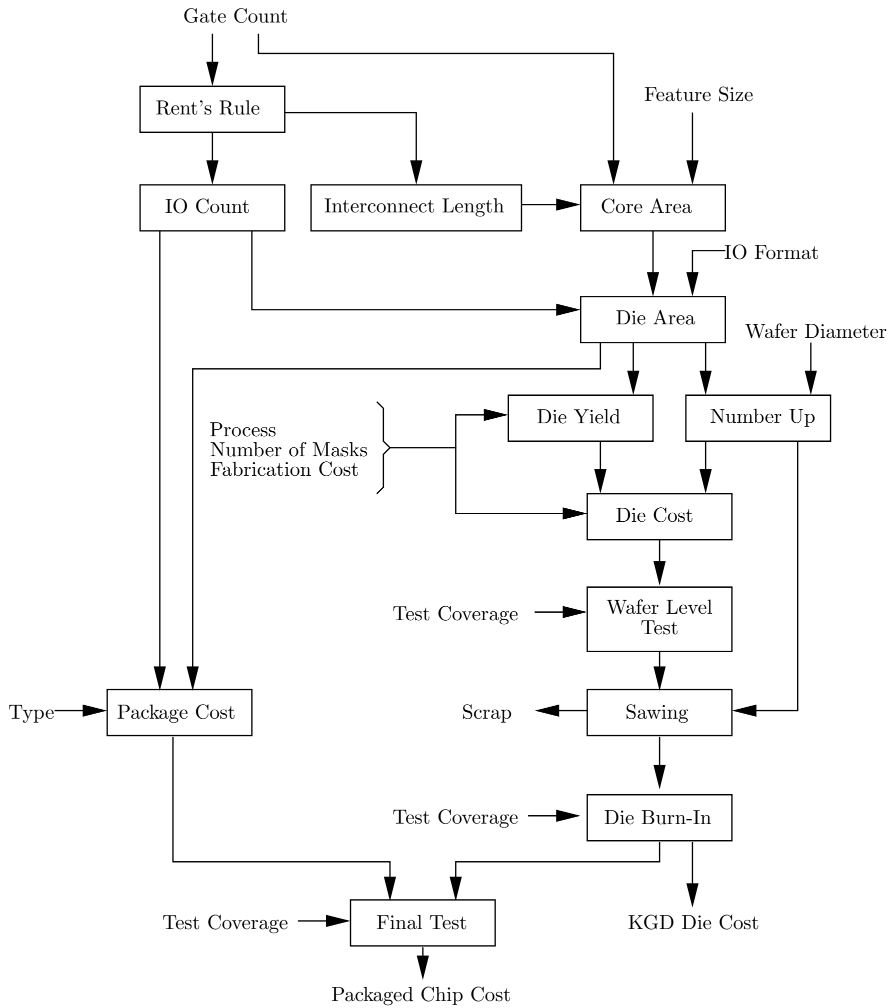


Figure 5.3: *The Overall Cost Modelling Flow for a Chip*

one of the commonest gates in random logic and is widely used in density metrics; NOR gates are widely used in high-speed dynamic logic.

The estimation of A_g in (5.12) can be carried out based on the height and width of the cell layout, determined by the contacted metal pitches of the local metal layers. As per [*184*], the size of a 2-input NAND gate for a standard drive strength is 4 metal pitches across by 16 metal pitches, i.e. $4p_{wL} \times 16p_{wL}$.

(B) Interconnection Area Limited

In many contemporary logic-intensive chips, the chip area is limited by wiring capacity. The average gate dimension can be calculated by equating the demand

for interconnections and the available wiring resources [*6*].

$$\text{interconnection available per gate} = e_w \frac{d_g^2}{p_w} n_w \tag{5.7}$$

$$\text{interconnection required per gate} = f_g l_{avg}, \tag{5.8}$$

where p_w is wiring pitch, e_w is wiring efficiency, n_w is the number of wiring levels, f_g is the fan-out of a typical gate, and d_g is the gate dimension:

$$d_g = \frac{f_g \bar{R}_m P_w}{e_w n_w} \tag{5.9}$$

where f_g refers to the gate fanout, p_w to interconnection pitch, n_w to the number of interconnection layers, e_w to the utilization efficiency of interconnects, and $\overline{R_m}$ to the average interconnect length. This model was further validated and in [*21*], and also used in [*185*]. However in modern technologies, the number of available wiring levels is much higher and the variation in wire pitch between the lowest and highest levels is significant; for example the pitch of global wires is typically several times that of local wires. Additionally, the higher the number of levels, the greater the congestion introduced by the presence of vias and studs needed for the interconnection of adjacent wiring levels. Therefore, (5.9) requires a refinement in order to be used with multi-level interconnect structures. One proposal, in [*21*], is to use an average value for p_w, while another, in [*186*], is to estimate $\frac{p_w}{n_w}$ considering only local and global wires. Also, due to unequal usage of power and clock lines on the different metal layers and via blockage, the wiring efficiency for signal wires vary from one level to another. In this work, the change in wiring pitch and different wiring efficiency factors for each layer, as well as the effects of via blockage are considered by modifying (5.9) to get:

$$d_g = \frac{1+f_g}{2} \frac{\overline{R_m}}{\sum_{i=1}^{n_w} \frac{e_{w,i} k_{p,i}}{p_{w,i}}}, \tag{5.10}$$

where $k_{p,i}$ and $e_{w,i}$ are the wiring utilisation factor (modelling the effect of via blockage) and wiring efficiency factor (modelling the routing efficiency) respectively, for the ith layer. Such an approach is suggested in [*184*] and [*187*]. The term $\frac{(1+f_g)}{2}$ is a correction to take into account the fact that a logic gate fans out to several gates. The term $k_{p,i}$ is the fraction of metal layer i occupied by wires, while e_{w_i} can be expressed as the product of three factors as

$$e_{w,i} = e^i_{rout}(1 - b^i_{PGC})(1 - b^i_{via}) \tag{5.11}$$

where e^i_{rout} is the efficiency of the routing tool for the ith layer (approximately constant over all layers), b^i_{PGC} the blockage due to power/ground/clock nets, and b^i_{via} the blockage due to vias [*187*]. The work [*88*] estimates that in the case of two layers of identical wire pitch, the top layer blocks around 12%-15% of the wiring capacity of the lower layer, and further recommends that the blocking percentage between levels of varying pitch be scaled proportionally with pitch. Hence b^i_{via} on the first wiring layer can be between $10\% - 50\%$, and much smaller on higher metal

levels [*188*]. By contrast, b_{PGC} for the topmost level is around 30%-40%, and less than 3% for the lower levels [*184*]. It is possible however to assume a constant e_w for all wiring layers by a process of averaging the different values in a first-order approximation.

5.3.2 Chip Size

The chip core size is estimated taking the maximum of packing-density limited area and interconnect capacity limited area. That is:

$$A_{core} = \max\left\{N_g d_g^2, N_g A_g\right\} \tag{5.12}$$

The methodology to estimate A_g and d_g have been presented in the previous section.

Parameter	Alpha 21164	Pentium
Die Area (mm^2)	299	163
Memory Area (mm^2)	102	44
CPU Logic Area (mm^2)	180	111.8
I/O Pad Area (mm^2)	17	7.2
Transistor Count (M)	9.3	3.1
Memory Transistors (M)	6.71	0.971
Technology	0.5 μm CMOS	0.6 μm BiCMOS
n_w	4	4
Wiring Pitch Values ($p_{w,i}$) (μm)	1.125,1.125,3,3	1.4,1.7,1.7,3.5
Rent's Constant p	0.35	0.35
A_g (μm^2)	81	125
e_w	0.26	0.32
N_g (M)	0.648	0.532
f_g	3	3
Gate-Area-Limited Area (mm^2)	52	89
Interconnect-Limited Area (mm^2)	172	108

Table 5.3: *Validation of Area Models for Two Microprocessors*

The integration mixed signal systems in a single die is a merging of several technologies, such as logic, memory, analog/RF, and this results in increased process complexity and a area change. For example merging logic circuits together with memory results in lower circuit density and hence larger circuit area, than their logic only or memory only counter parts. For example, in a UMC 0.18 μm technology a 6T-SRAM cell is about 4 μm^2 for logic and SRAM intensive product but it is 5.6 μm^2 when merged. In this case there is a 1.4 times larger cell area, when combined process area is used. The increased process steps for merging different technologies are mentioned in Table 5.4 [*189*]. If memory, logic and RF systems are merged into one chip, the total areas are [*190*, *191*]:

$$A_{mem \cup logic} = \alpha A_{mem} + \beta A_{logic} \tag{5.13}$$

$$A_{SoC} = \alpha A_{mem} + \beta A_{logic} + \gamma A_{RF} \tag{5.14}$$

The total number of mask layers after merging is:

$$N_{mem \cup logic} = N_{mem} + N_{logic} - N_{mem \cap logic} \quad (5.15)$$

$$\begin{aligned} N_{mem \cup logic \cup RF} &= N_{mem} + N_{logic} + N_{RF} - N_{mem \cap logic} \\ &- N_{mem \cap RF} - N_{logic \cap RF} \\ &+ N_{mem \cap logic \cap RF} \end{aligned} \quad (5.16)$$

In this analysis it is assumed that $N_{mem \cap logic \cap RF} = 0$.

When it comes to packaging the core, the number of I/Os to be connected to the outside must be arranged around the periphery and may require a larger perimeter in order to place them according to the minimum peripheral pitch. A simple and yet reasonably accurate model for the die size has been introduced in [*191*], which is:

$$A_{die} = \max\left\{(\sqrt{A_{core}} + 2P_p)^2, \left(\frac{N_p P_p}{4} + 2P_p\right)^2\right\} \quad (5.17)$$

$$A_{pkged_chip} = (\sqrt{A_{die}} + 2L_{bnd})^2 \quad (5.18)$$

However, in addition to the die area model described above, there are some other models proposed in the literature such as one in Sandborn *et.al.* in [*192*]. It describes a relationship for a peripherally bonded die with a single row of bond pads on all sides; the die area is given by the maximum of two limits given by:

$$A^1_{peripheral} = \left(2l_{pd} + P_p \left\lfloor \frac{N_p}{4} \right\rfloor\right)^2, \quad (5.19)$$

$$A^2_{peripheral} = N_p w_{pd} l_{pd} + (1 + kN_p)^2 \quad (5.20)$$

where N_p is the total number of pads, l_{pd} and w_{pd} are the length and width of a pad respectively, P_p is the pad pitch, and k is the fractional increase in the core die area necessary to accommodate redistribution of IO to the periphery of the die, which is approximately a constant in the range of $0.00074 - 0.00079$ [*192*]. The peripheral I/O limited area is given by (5.19) and the peripheral redistribution limited area is given by (5.20).

Assuming that the active circuitry cannot be placed under the bond pads, for an area array bonded die, the die area is given by the maximum of I/O limited chip area and the bond pad limited area as:

$$A^1_{areaarray} = (P_p * \lfloor N_p \rfloor)^2, \quad (5.21)$$

$$\text{and,} \quad A^2_{areaarray} = N_p w_{pd} l_{pd} + A_{core}, \quad (5.22)$$

respectively.

5.3.3 Die Yield Analysis

Ideally in a properly processed silicon wafer containing circuits, all of the circuits on the wafer should be functional circuits. The number of functional circuits can range very close to 100% to one or few circuits per wafer. The defects which contribute

Added Process	Logic	SRAM	Flash	DRAM	CMOS RF	FPGA	MEMS	FRAM	Chem. Sensors	Electro Optical
Logic	0									
SRAM	1-2	0								
Flash	4	3-4	0							
DRAM	4-5	3-4	7-9	0						
CMOS RF	3-5	5-9	6-9	6-10	0					
FPGA	2	2-4	4-6	3-7	5-7	0				
MEMS	2-10	3-12	6-14	6-15	5-15	4-12	0			
FRAM	4-5	3-4	7-9	2-3	7-10	6-7	9-15	0		
Chem. Sensors	2-6	3-7	6-10	6-11	5-11	4-8	4-16	6-11	0	
Electro Optical	5-8	6-9	9-12	9-13	8-12	7-10	7-18	9-13	7-14	0

Table 5.4: *Added Process Complexity (number of mask level) for SoC Technologies, based on CMOS logic*

to chip-yield loss fall into three basic categories: random point defects, systematic defects, and parametric defects in the circuit [*193*, *194*].

Random defects are due to mechanisms that are not specifically tied to a particular wafer-process step. Some examples of random defects include the following: a foreign particle on the wafer that may cause a short between two interconnect lines; a short caused by a metal bridge between two lines; failure of a contact to open; a break in an interconnect line; and a pinhole in a transistors gate oxide. Such defects can cause a chip to either fail outright or not meet a performance specification.

Design, processing, or test operations can all add systematic defects to a chip. After all, these defects can result from any mechanisms that create spatial- or time-based variations on the chip. Systematic yield loss are usually corrected with tighter controls during chip processing because there exists a tie between design and systematic yield at process geometries. A common way to reduce systematic yield loss in a chip processed at fine geometries is by applying corrective operations which will adjust geometries so that the features printed on the chip more closely resemble the ones drawn during the chip layout design.

Parametric yield loss is caused by various factors, which represent the process and environmental variations of a chip from targeted, nominal values as well as the design implementation. Examples of this type of yield loss include statistical process variations; temperature and operating-voltage spreads; and geometry variations on a chip resulting in differences from nominal values in parameter values.

(A) Modeling Yield Loss

IC yield modeling in terms of fundamental parameters that are independent of the particular IC and processing process parameters is very important in several aspects, but is highly complex. With an IC yield model one can not only estimate cost but also compare the processing quality of different process lines and discover where improvements in the process facilities are required. Exhaustive simulations of parameter variations and sensitivity analysis are performed to determine the effects of the various parameter variations on critical design targets. A complete yield model should account for all sources of yield loss; it should ideally give insight into the possible uses of yield loss, and should quantify the yield losses resulting from design, process and random defects. Generally, IC yield is expressed as a function of D_o, the average number of defects per unit area, and A_c, the critical chip area [*195*]:

$$Y = f(D_o, A_c). \tag{5.23}$$

The first model used to predict IC yield was derived from the Poisson probability distribution function. The problem of estimating the yield of good chips is analogous to the statistical problem of placing k number of balls in N cells, and calculating the probability that a given cell contains k balls. The probability that a die has k number of defects is

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots \tag{5.24}$$

where $\lambda = D_oA$, the average number of defects per die. Then, for a die with no defects ($k = 0$),

$$Y_1 = e^{-\lambda} = e^{-D_oA}, \tag{5.25}$$

which is the classical Poisson model. This model requires that defects are perfect points and are uniformly distributed and spatially uncorrelated across a wafer. Many experts believe that the Poisson Model is too pessimistic, since defects are often not randomly distributed, but rather clustered in certain areas. Defect clustering allows less defects over large areas of the wafer than if the defects are randomly and uniformly distributed. As it explained in [*195*], the Poisson yield model was accurate enough when chip areas had been below 0.25 cm^2. As chip sizes increased, the Poisson model became inaccurate and tended to underestimate the yield and was rarely used in practice. Even early as 1964, B.T. Murphy argued that since the defect density varies widely from chip to chip, and from wafer to wafer, and even from run to run, the Poisson model tends to underestimate the yield. Hence, Murphy reasoned that the defect density needs to be summed over all chips and wafers using a normalized probability distribution function of defect densities. The yield of chips on a wafer, where defect density is nonuniform across the wafer, can be expressed as

$$Y = \int_0^\infty e^{-DA} f(D) dD \tag{5.26}$$

where $f(D)$ is the normalized distribution of defect density (pdf), with $\int_0^\infty f(D)dD \equiv 1$. Assuming a delta function distribution and a triangular distribution for D_o, (5.26) reduced to Poisson and Murphy [*196*] yield models respectively. With the passage of time, several other models were introduced by using different distribution functions. Some of these are Seed's model (assuming an exponential distribution of defects, $f(D) = \frac{e^{-D/D_o}}{D_o}$) [*197*], Dingwall's model [*198*], Moore's model [*199*], and Price's model [*200*]. Price's model is an extension of Seed's model using Bose-Einstein statistics where all the defects are indistinguishable.

However, the Gamma distribution can also be used to approximate the defect distribution of ICs [*201*]. The probability distribution function for Gamma distribution is:

$$f(D) = \frac{1}{\Gamma(\alpha)\beta^\alpha} D^{\alpha-1} e^{\frac{-D}{\beta}} \tag{5.27}$$

where α and β are the two distribution parameters and $\Gamma(\alpha)$ is the Gamma function. The average density distribution is given by $D_o = \alpha\beta$, the variance of D is given by $var(D) = \alpha\beta^2$, and the coefficient of variation is given by

$$\frac{\sqrt{var(D)}}{D_o} = \frac{1}{\sqrt{\alpha}}.$$

In the early 70's, two papers [*202*] and [*201*] described a yield function by applying Gamma distribution function to (5.26) which is presently known as the negative-binomial model. The elegance of this model is that through a relatively simple statistical analysis of defect density distribution data, a more accurate yield model can be derived.

Stapper's derivation in [*201*] describes that the most general way to describe the probability of a chip having k defects is by the compound probability distribution.

Model Name	Formula	Distribution of defects and other remarks
Poisson	$y_1 = e^{-ADo}$	Defect Distribution $f(D) = \delta(D - D_o)$
Murphy	$y_2 = \left[\frac{1-e^{-ADo}}{ADo}\right]^2$	Triangular Approximation of Gaussian
Seeds	$y_3 = \frac{1}{1+ADo}$	Exponential $f(D) = \frac{1}{D_o e^{-D/D_o}}$
Dingwall	$y_5 = \left(1 + \frac{AD_o}{3}\right)^{-3}$	$f(D)$ is not defined
Moore	$y_4 = e^{-\sqrt{ADo}}$	$f(D)$ is not defined
Price	$y_6 = \prod_{i=1}^{n} \frac{1}{(1+AD_i)}$	$f(D)$ is not defined. Derived from Bose-Einstien statistics where defects are "indistinguishable"
Negative Binomial	$y_7 = \left(1 + \frac{AD_o}{\alpha}\right)^{-\alpha}$	Gamma Distribution

Table 5.5: *Yield Models, where A is die area, and D_o is defect density per unit area. For Price model, n is the number defect producing mechanism or process step as the wafers pass through the line and D_i is the defect density of each defect producing mechanism or the defect density for each process step as the wafers pass through the line. α is the negative-binomial model is usually referred to as the cluster parameter. A comparison of yield function with D_oA is shown in Figure 5.4.*

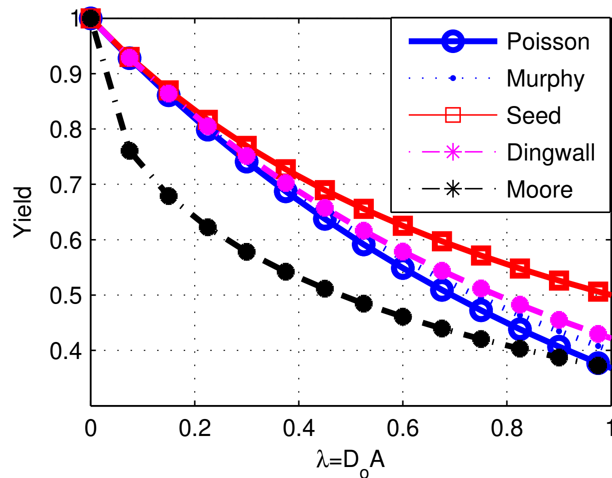


Figure 5.4: *Comparison of Several Yield Models. For all models, lowering defect density or using a smaller chip will result in higher die yield.*

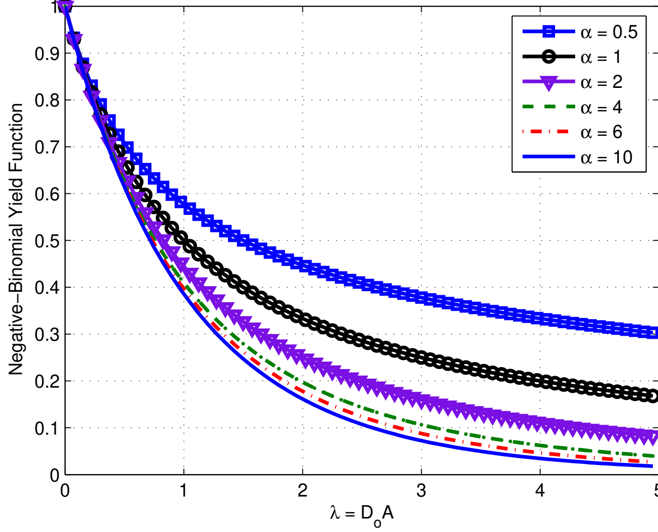


Figure 5.5: *Negative Binomial Yield Function against λ for different α values*

Then Murphy's definition for chip yield (5.26) takes the form:

$$P_k = \int_0^\infty e^m \frac{m^k}{k!} f(D)dD \tag{5.28}$$

$$= \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)} \frac{(A\beta)^k}{(A\beta+1)^{k+\alpha}} \tag{5.29}$$

Then the probability of having no defects on a chip is

$$Y_4 = \frac{1}{(A\beta+1)^\alpha} = \left(1 + \frac{AD_o}{\alpha}\right)^{-\alpha} \tag{5.30}$$

This is the well known negative-binomial yield function. α is usually referred to as a "cluster" parameter and increases with decreasing variance in the distribution of defects. This negative-Binomial yield function can also be expressed in terms of $1/\alpha$, i.e:

$$Y = \frac{1}{(1+SD_oA)^{\frac{1}{S}}} \tag{5.31}$$

where S is the shape parameter of the distribution of D equal to $\frac{var(D)}{D_o^2} = \frac{1}{\alpha}$.

The Gamma distribution is in general a skewed distribution stretching from zero to infinity. In the limiting case when $S \to 0$, the Gamma distribution reduces to a delta function and the yield (5.31) reduces to Poisson yield model. By selecting different values for α (or S), various yield models can be approximated [*195*]. To be more specific selecting α about 10 to ∞, 4.2,3 and 1, Poisson, Murphy, Dingwall and Seeds models can be emulated respectively.

Additionally to that the Gamma yield function is capable of representing quite a large variation in the shape of an experimental yield versus area curve. Therefore, Gamma yield function is known to be the most common function for representing IC yield. The shape parameter of the Gamma distribution varies considerably among different types of products manufactured in different processing flows. Also, many different types of defects affect the yield, and the parameters D_o and S vary considerably for different types of defects. Each defect mechanism can be characterized by its mean defect density D_{on}, the shape factor of the distribution of defects S_n, and the portion of the total chip area A_n that is susceptible to that particular defect. Then, for each type of defect, the yield is:

$$Y_n = \frac{1}{(1 + S_n D_{on} A_n)^{\frac{1}{S_n}}} \tag{5.32}$$

The overall yield is then the product of the yield for each known type of defect, that is

$$Y = \prod_{1}^{N} Y_n = \prod_{1}^{N} \frac{1}{(1 + S_n D_{on} A)^{\frac{1}{S_n}}} \tag{5.33}$$

Extending this model for different layers by assuming different defect density distributions for each mask layer, a yield function can be obtained. However, for the sake of simplicity, one can assume that the distribution of electrical defects are independent of the masking step, and with this assumption, a new yield function is derived [*203*].

$$Y_d = \frac{1}{(1 + SD_0 A)^{\frac{N}{S}}} \tag{5.34}$$

where N is the number of mask layers, and A is the chip area.

Oftentimes several yield models are implemented in different product developers and made refinements for already existing models to suit their process flow, for example Murphy's model for memory, Seed's model for gate arrays etc. What is common in all the models is the yield of a bare silicon die, Y_d, depends on electrical defects and each model assumes a particular defect density distribution: random for the Poisson model, triangular in the Murphy model, exponential in the Seeds model, and gamma for Price and the negative binomial model. The merit of each model can only be judged by how it approximates the actual yields, and there is no universal model. Among the seven models listed in table 5.5, negative-binomial model is probably the more powerful yield function which is widely used and can easily be fitted with experimental data.

5.3.4 Chip Cost Model

The wafer yield of the die, discussed in the above section, represents the actual yield. In the integrated circuit fabrication process, after all the process steps related to circuit fabrication, before the wafers cut into the dies, the whole wafer is tested. This is known as wafer level testing. They then go for a die level test and a burn-in process. This comes with an added cost. It is well-known that wafer-level or Die-level testing leads to early defect screening, thereby reducing packaging and

production cost. The wafer level yield of a die discussed in the previous section represents the actual die yield, and after the testing process some defective dies will be sent to packaging process due to the fact that all defects cannot be identified at the wafer or die level test and burn-in process. Also, in a testing process only a certain number of faults can be tested and then a question arises whether all those faults tested good or not? This problem has qualitatively discussed in [*204*].

(A) Fault Coverage, Defect Level and Yield

A chip/die testing process is characterized by the fault coverage level, which is defined as the fraction of defects that are identified in the test. If most incoming parts have a higher incoming yield, then even a relatively poor test that screens out only some of bad parts can still give low defect levels. On the other hand, if most incoming parts are bad, then the test must have a higher coverage to ensure that the number of bad parts passed are not a significant fraction in comparison to the few good parts. The chip yield after wafer-level or die-level test is computed from the fault coverage, and the actual yield of the die on wafer.

Assume that a given chip has exactly n_c number of defects/faults, the probability of a fault occurring is independent, and all defects/faults are equally likely with probability p (in this case, stuck-at-faults on signal lines are assumed to be the only type of faults that can exist in the circuit [*204*]). Also assume that $m(\leq n_c)$ is the number of faults which can be tested in the test process. This gives rise to a uniform distribution of faults. By definition the chip yield is a chip without any defects, and in other words it can be defined as the probability of each possible fault being absent. Then the chip yield can be expressed as:

$$Y_{in} \quad = \quad (1-p)^{n_c} \tag{5.35}$$

By definition the fault coverage F_c is the number of faults that can be tested divided by the total number of faults: $F_c = \frac{m}{n_c}$. A defective die could have at least one defect (defective die could have more than one defect), and identification of any defect is enough to reason out that the die is defective and should be scrapped. Using the Williams and Brown model described in [*204*], probability of accepting a chip with one or more faults in the testing process is

$$P_a = (1-p)^m - (1-p)^n_c.$$

Dies that pass the test might have good chips with probability $(1-p)^{n_c}$, and bad-chips with probability P_a. Hence, the fraction of chips which pass the test, the pass fraction (PF) is

$$PF \quad = \quad P_a + (1-p)^n = (1-p)^m = (1-p)^{n_c \frac{m}{n_c}} = Y_{in}^{F_c} \tag{5.36}$$

The defect level, which is also called the field reject ratio is equal to the probability that a bad chip is accepted, divided by the probability of accepting a bad chip plus the probability of a good chip:

$$DL \quad = \quad \frac{\text{Number of bad chips passed by the test}}{\text{Total number of chips passed by the test}} \tag{5.37}$$

$$\quad = \quad \frac{P_a}{P_a + (1-p)^{n_c}} = 1 - Y_{in}^{1-F_c} \tag{5.38}$$

The die yield after a testing process is the ratio of good chips passed by the test and total chips passed by the test, and can be expressed as:

$$Y_{out} \quad = \quad 1 - DL = Y_{in}^{1-F_c} \tag{5.39}$$

In [205] it is shown that fault clustering (multiple logical faults caused by a single fault, and multiple faults located on the same unit) decrease the test transparency probability, and a model for defect level is presented :

$$DL = \frac{(1-Y_d)(1-F_c)e^{-(n_0-1)F_c}}{Y_d + (1-Y_d)(1-F_c)e^{-(n_0-1)F_c}} \tag{5.40}$$

Here, n_0 is average number of faults on a defective chip and determined by exhaustive failure analysis and curve fitting.

(B) Chip cost model

The model computes the cumulative cost per die at the end of each process step as follows:

$$C_{1,i} = \frac{C_{1,i-1} + C_i}{PF} \tag{5.41}$$

where $C_{1,i-1}$ is the accumulated cost of all the steps up to but not including the present step, C_i, is the cost of the present step, and PF is the percent of the dies which pass the current step. For every process, which is followed by a testing process, the testing cost should also be included. The higher the fault coverage of the testing process, the higher the cost; the extra testing time results in extra cost including labour, and equipment usage costs. Assuming that a higher fault coverage

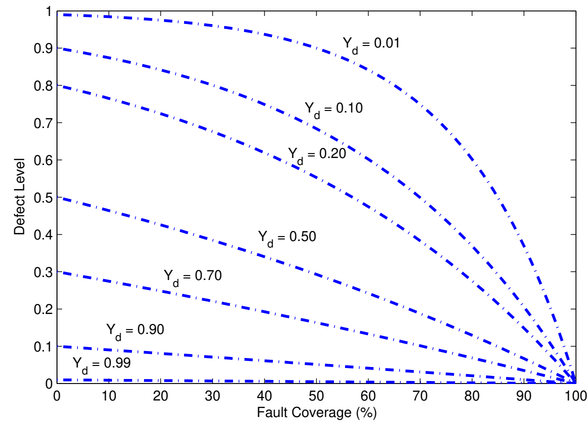


Figure 5.6: *Variation of defect level ($DL = 1 - Y_d^{1-F_c}$) with fault coverage for different yield values. Achieving defect level of few tens of parts per million requires near complete fault coverage for large circuits. It is a challenge to achieve such a higher fault coverage with the presence of many type of faults.*

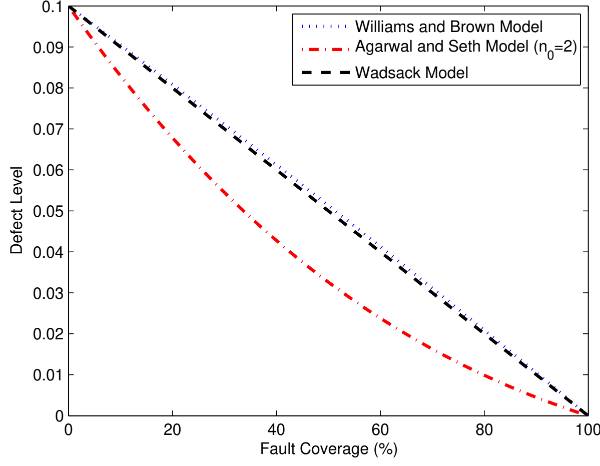


Figure 5.7: *Comparison of various defect level models*

level requires significantly increased testing time, the following exponential model correlating test coverage level with testing time, t_{test}, is proposed in [*206*]:

$$F_c = 1 - e^{-kt_{test}} \tag{5.42}$$

Here, k is an empirical constant that defines the steepness of the exponential function. It is assumed that 60 seconds is enough to achieve 99.99% coverage, in order to estimate the value of k [*206*]. Thus, k is calculated to be 0.1. Also, it is assumed that wafer-level testing achieves 80% fault-coverage, and testing after burn-in achieves a fault coverage level of 99%. Then, the testing cost can be assumed to be linearly propotional to the testing time [*206*]:

$$C_{test} = C_t t_{test} \tag{5.43}$$

Since the size of the die is already estimated in the previous section, the number of dies which can be fabricated on a wafer N_{die} is estimated by:

$$N_{die} = \frac{\pi d^2}{4d_c^2} - \frac{\pi d}{\sqrt{2}d_c} - 4 \tag{5.44}$$

After step one, wafer fabrication, die cost is estimated as follows:

$$C_1 = \frac{C_{wafer}(raw, process, mask)}{N_{die}Y_{die}} \tag{5.45}$$

It is hard to predict the cost for testing a die, but it depends on the number of pins that have to be tested. In this case we assume that dies are attached to a reusable carrier for a burn-in test so that the cost for the carrier is negligible when considering mass production. In this analysis, a constant value for chip test cost per module per lead ($ per lead) is assumed and it is multiplied by the number of leads to be tested to get the total chip test cost.

The package cost is calculated using a price vs pin count assumption as in [*207*]. For a peripheral I/O single chip plastic package the cost is:

$$C_{pkg} = 0.01e^{1.16log(NIO)-2.09} \tag{5.46}$$

5.4 Board- or Package-Level Model

The cost of a module is summation of the cost for each chip/die, substrate cost, and interconnection cost. In the previous section, the methodology for determining a packaged chip or bare-die cost was discussed, and in this section looks how the module level cost is estimated.

The size of a module substrate depends on several parameters ranging from type of technology being used, for example MCM-L,C or D, to the number of dice on it, and even the thermal conductivity. Though several wiring limited module area estimations have been derived, the size of the substrate may not limited by the wire-ability. But it is dependent of total number of I/Os on the MCM, number of vias in the substrate, number of dice, and the thermal dissipation of the substrate.

5.4.1 Number of Pins per Chip

To determine the number of pins per chip Rent's rule can be applied.

$$N_p = K_p N_g^{\beta}. \tag{5.47}$$

Here, β is Rent's constant for the number of pins, and K_p a multiplicative constant. The constants for Rent's rule are different from those used for on-chip wire length calculations. Table 5.2 presents Rent's constants for a wide variety of products, including those for on-chip wire length calculations as well as board/module level pin count estimation. Depending on the way the electronic system is partitioned into modules/chips and the data transferring methodology (for example, serial, multiplexed, bidirectional) Rent's constants may differ.

5.4.2 Module Level Average Interconnection Length

Module level interconnection length can also be estimated as described in Section 5.2.2 which is used to calculate the average interconnection length for chips:

$$\overline{R_M} = \begin{cases} \frac{2}{9}\frac{1-4^{(\eta-1)}}{1-N_c^{(\eta-1)}}\left(7\frac{N_c^{(\eta-0.5)}-1}{4^{(\eta-0.5)}-1} - \frac{1-N_c^{(\eta-1.5)}}{1-4^{(\eta-1.5)}}\right) & \text{for } \eta \neq 0.5 \\ \frac{2}{9}\frac{1-4^{\eta-1}}{1-Ng^{\eta-1}}\left(7log_4 N_c - \frac{1-N_c^{\eta-1.5}}{1-4^{\eta-1.5}}\right) & \text{for } \eta = 0.5, \end{cases} \tag{5.48}$$

where η is the Rent's constant. In this case the number of gates is replaced by the number of chips, and the gate pitch is replaced by chip footprint dimension [*6*]. In addition to that Rent's constants to be used in that equation may be larger than that of ρ because laying out a board minimizing wire lengths is rather easier than a chip [*6*]. Then, the average interconnection length in actual units is

$$L_{AVG} = \overline{R_M} F_P, \tag{5.49}$$

where F_P is the chip footprint dimension.

5.4.3 Chip Footprint

The MCM substrate area A_{sub} can easily be estimated by the method outlined by Backoglu in [*6*]. If F_p, the chip footprint dimension is known, SoP area is:

$$A_{sub} = N_c F_p^2 \tag{5.50}$$

It is understood that if there is only one layer of chip carriers is available on the module, the footprint size cannot be smaller than chip carrier size, and footprint will be limited by the interconnection capacity of the module. As illustrated in [*6*], the interconnect-capacity limited substrate area is found by estimating the average interconnect length at the module level using $\overline{R_M}$. Assuming a fan-out of F_c (typically 1.5) for the chip pins, the number of interconnections are:

$$\frac{F_c}{Fc+1} N_c N_{p_mcm}$$

where N_c is the chip count, and N_{mcm} total number of chip I/Os and I/Os to and from the MCM. Then, the total interconnection length is the multiplication of number of interconnections and the average interconnection length as given by:

$$\frac{F_c}{Fc+1} N_c N_{p_mcm} \overline{R_M} F_P.$$

By equating supply and demand of interconnects, the interconnect-limited chip footprint F_p is estimated as:

$$F_p = \frac{F_c}{F_c+1} \frac{\overline{R_M} N_{p_mcm} P_{w_mcm}}{N_c e_w n_w}, \tag{5.51}$$

where n_w and P_{w_mcm} number and pitch of module wiring levels. Footprint can be limited by the dimension of the chip D_c or the dimension of a chip carrier P_c too, therefore the footprint dimension is given by the most limiting constraint:

$$F_p = \text{MAX} \left\{ \frac{F_c}{F_c+1} \frac{\overline{R_M} N_M P_{w_mcm}}{N_c e_w n_w}, D_c, P_c \right\} \tag{5.52}$$

Alternative to the method we have discussed, there are two other formulations for the module area: Hannemann's approach [*208*] and Moresco's approach [*209*].

Nevertheless the simplest approach discussed above assumed that the components to be arranged in a MCM substrate are homogeneous, which is usually not the case for mixed-signal system integration. It is understood that this restriction appears at two critical points [*210*]: (1) in the derivation of the wiring capacity limited footprint, and (2) in how the module size is determined. This limitation can be fixed by recomputing an effective chip count and corresponding average interconnect length for each component.

$$\text{Effective } Nc_i = \frac{NIO_{mcm}}{NIO_{chip_i}} \tag{5.53}$$

where NIO_{mcm} is the total number of IO connections in the whole module, and NIO_{chip_i} is the number of IO connections that the ith component possesses. Then,

to find the area of the module multiplying F_p^2 by chip count is no longer valid and hence the following summation must be used:

$$A_{SOP} = \sum_{i=1}^{N_c} F_{p_}i^2 \tag{5.54}$$

5.4.4 Yield and Cost Analysis

In traditional IC packaging approaches the module or system yield is a function of the yield of individual components and the yield of the integration methodology used. This is basically the multiplication of the yields of all the dice, substrate, and the bonding process, that is:

$$Y_{module} = \prod_{i=1}^{N} Y_{d_i} Y_{ass} Y_{sub}. \tag{5.55}$$

Thus, overall yield of a system can be uneconomically low for complex systems, unless particular attention is paid to test coverage and delivered die yield for bare dice, mainly through KGD (Known-Good-Dies) methods (KGD are fully tested unpackaged ICs that often comes with a guarantee). Even with a 90% probability of KGD, the resulting yield of an assembled module is unacceptable for systems with more than a few chips.

IC manufacturers often offer several levels of KGD, where each successive level entails a more rigorous test plan. High KGD levels often come with a quality and reliability guarantee, such as guaranteeing them to function on delivery, or to last through a certain time period. Industry aims to provide KGD that have at least as much quality and reliability as they would have if they were packaged.

In general there are four ways to ensure that a bare silicon die is *known good* [*211*]: 1) through the process under which the manufacturer fabricates the IC, 2) through the IC design making ICs testable (called design-for-testability), 3) through bare IC testing including electrical, mechanical and environmental tests, and 4) through sample packaged IC testing.

Assembled module costs are very complex to estimate; it includes the total cost for each chip including testing cost, assembly cost, substrate cost, rework cost, and finally the module test cost and packaging cost. Therefore, detailed costing of chips, substrate and interconnects is essential in analyzing the module cost. To analyze the relationship among all the assembly, yield, test, and repair parameters, and how they impact the final cost and quality of a module, a simple assembly, test, rework simulation model can be used.

Such a model is explained in detail in [*210*]. Each of the boxes in Figure 5.8 represents a specific process. The assembly box takes one or more inputs corresponding to various types of components to be assembled and the substrate that they are going to assembled upon. Each input is characterized by cost, yield, and a count of how many times it is going to be used. Also, the assembly process itself has a cost, and yield parameters, as does the components and the module. Therefore, the output cost of assembly process contains sum of substrate cost, and the cost for all the components including assembly cost. Output yield of the assembly is estimated by (5.55). After this process, module is tested using a process that

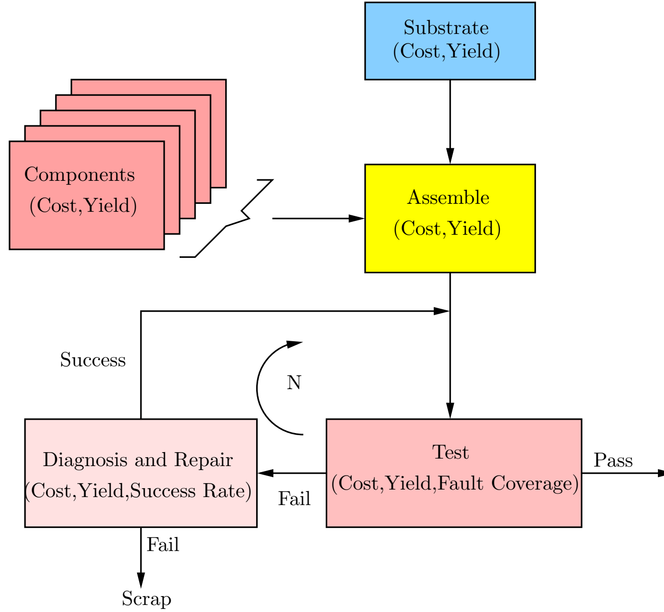


Figure 5.8: *Module Cost Estimation Flow. Adapted from [210]*

can be characterized by a testing cost and fault coverage. The modules that pass the test are assumed good and are passed to the next activity which is possibly another assembly process. The fraction of modules which pass the test is estimated by (5.36) with module yield as the input. However, since the testing process cannot detect all possible faults because the test coverage is usually less than 100%, some defective modules can escape and reduce the quality of the output modules.

The yield of units which exit the test (assuming no repair) is [*204*]:

$$Yield_{out} = \frac{\text{Number of good modules passed by the test}}{\text{Total number modules passed by the test}}, \tag{5.56}$$

and, the output cost from the test step is given by,

$$Cost_{out} = \frac{Cost_{input} + Cost_{test}}{PF}. \tag{5.57}$$

Then the fraction of bad modules tested $(1 - PF)$ are diagnosed for possible repair. The probability of being able to repair a module is known as the repair success rate. Also, it is worthwhile to mention that repairing does not guarantee that the unit is functioning properly, because the unit may have been improperly diagnosed or some failures introduced during the reworking process. The repairable modules are again subjected to the testing process. The diagnosis and repair/rework process has a cost and also affect the yield through the reworking process and its defects, new component defects, misdiagnoses. If a module fails to pass the test after being repaired the maximum allowable number of times (N), then it is scrapped.

The model used in [*191*] assumed only one repair/rework cycle, and after the rework process the modules are assumed to pass the test, and that there are no

scrapped units. If the assembly yield is Y_a, and after the assembly process $1 - Y_a$ units are sent back for reworking, the overall yield of the assembly process becomes $Y_a(2 - Y_a)$.

5.5 Summary

80% of the final product cost is commited during the first 20% of the design cycle and hence, it is very important to analyze the performance and cost of a chip at an early stage in the system conceptual level. The cost of a chip/module is predicated on its area. Taking technology parameters as inputs, a simple methodology for chip/module area estimation based on interconnect or device size limitations was discussed in this chapter. Process yield, die area, number of dies that can occupy a wafer, and the cost to produce a wafer determines a rough cost of the die. The number of mask layers for a chip and the process technology determine the factors to consider the cost per wafer. Finally, the testing process and the quality of testing that a chip undergoes, and the package type determines the final chip cost.

Also, this chapter presented models to estimate module cost, which is determined by the substrate area, assembly technology and its yield, testing process and its quality, and the repairing and reworking process.

6

Heterogeneous System-on-Chip Integration: 2-D or 3-D ?

Trade-off analysis is identified as the process of comparing performance gains in one region of the design space with associated performance losses in another. In this chapter using the system level cost models proposed in chapter 5 a generic methodology for cost, performance and other technological trade-offs is discussed.

6.1 Introduction

High performance electronic processor systems in portable applications need to satisfy increasingly stringent requirements on energy efficiency under ever more severe performance, cost, weight and technological restrictions. The solutions explored by the semiconductor industry to meet these challenges are migrating towards 3-D integration options. A major driver behind this trend is the plethora of implementation problems facing gigascale 2-D integration, ranging from technological to architectural. From a fabrication point of view, integrating disparate technologies such as sensors, MEMS structures, and other heterogeneous elements demanded by many applications on a single die is more challenging than connecting separate dies by external interconnections. The 2-D architecture also results in numerous bottlenecks due to area and routing congestion, such as the memory bottleneck in multimedia SoCs [*212*]. Recent developments in fabrication technology have resulted in 3-D integration being a potentially viable option for gigascale integration [*162*][*213*].

However, even as designers are presented with an extra spatial dimension, the complexity of the layout and the architectural trade-offs also increase. To get a true improvement in performance, an accurate analysis using detailed models at different hierarchical levels is crucial. Even though several previous works have addressed this issue [*190*][*214*][*215*], they mostly concentrate on isolated model development, or target some specific type of system.

In this chapter presents a cohesive analysis of the technological, cost and performance trade-offs for digital and also crucially mixed-mode systems, outlining the choices available at different points in the design and their ramifications. A

generic methodology for performance and cost estimations of 3-D systems that can be modified for different applications is also presented as well as a comprehensive set of estimation models as building blocks [*191*]. Finally, in order to validate the proposed metrics and methodology, two ubiquitous electronic systems are analyzed under various implementation schemes and the performance trade-offs discussed. These case studies are used to highlight the importance of a cost and performance trade-off analysis early in the design flow.

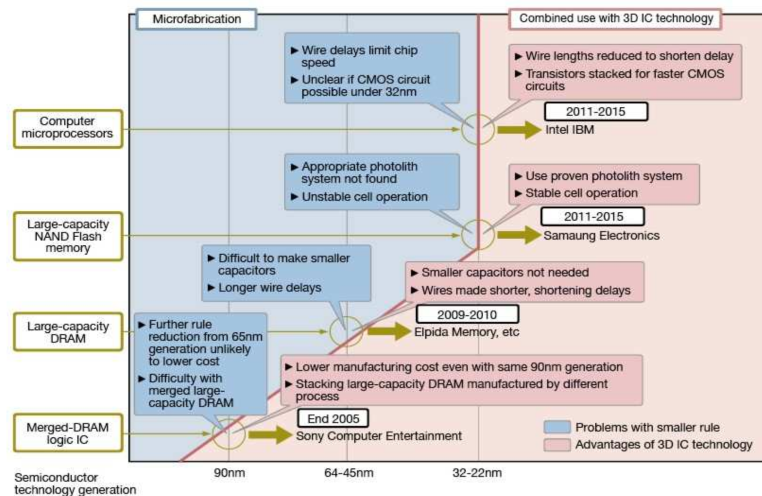


Figure 6.1: *The effect of smaller design rules is weakening in many types of ICs. In an effort to resolve the problem the semiconductor majors are investigating 3D IC technology, stacking chips, transistors and other elements vertically [24].*

6.2 Three-Dimensional Integration

As Akasaka has reported in [*216*], the idea of stacking multiple chips on top of each other dates back to the late 70s. In 1979, it was found that polysilicon deposited on an insulator can be melted and recrystallized by laser irradiation [*217*] and that the crystal perfection of the layer can be adequate to allow the fabrication of devices. The advent of SOI technology held out significant promise for 3-D integration. This fueled the research efforts for developing fabrication techniques and exploring the limits of 3-D integration [*39*, *216*, *218*, *219*].

6.2.1 Benefits of 3-D Integration over 2-D Planar

(A) Wire length Reduction and its implications

The principal benefit of 3-D integration is the reduction in the length of global interconnects and it is estimated in [*220*] that 3-D architectures reduce wiring length by a factor of the square root of the number of strata (layers) (m) used,

which is $\sqrt{m}$. For example, a 4-layer 3-D IC would have, on average, $\approx \sqrt{4} = 2$ reduction of interconnect length. 3-D integration also increases the global clock frequency by $m^{3/2}$. A second generic benefit of 3-D integration is a reduction in the total length of wiring required for a given system configuration. Along with this reduction comes a reduction in energy dissipation that varies roughly as the square root of the number of strata. The wire length reduction alone can reduce the interconnect energy and propagation delay by up to 51% and 54% respectively, at the 45 nm technology node [*221*]. However, the potential gain in performance is a strong function of the die-area [*191*]. The reduced parasitics for interconnects can significantly simplify the circuit and power distribution network design for high performance applications.

Davis et.al. presented a stochastic wire length distribution formula for 2-D planar ICs as discussed in Section 5.2.2. An extension to wire length distribution for 3-D homogeneous* ICs have been carried out in many works [*220*, *222*, *39*]. The work [*222*] gives rather simple closed-form equations which are suitable for back of the envelope calculations. 3-D Wire length distribution for a system with total of N gates in a square array architecture for clarity, distributed in m layers derived with neglecting the wires in length less than one gate pitch, because local wires have a insignificant impact on overall system performance. The 3-D circuit horizontal wire-length distribution function is as follows:

$$I(l) = \begin{cases} \Theta\left(\frac{l^3}{3} - 2l^2\sqrt{\frac{N}{m}} + 2l\frac{N}{m}\right) l^{2p-4} & \text{, for } 1 \leq l < \sqrt{\frac{N}{m}} \\ \frac{\Theta}{3}\left(2\sqrt{\frac{N}{m}} - l\right)^3 l^{2p-4} & \text{, for } \sqrt{\frac{N}{m}} \leq l \leq \sqrt{\frac{N}{m}} \end{cases}, \tag{6.1}$$

where

$$\Theta = \frac{\alpha k m^p \frac{N}{m}\left[1 - \left(\frac{N}{m}\right)^{p-1}\right]}{-\left(\frac{N}{m}\right)^p \frac{1+2p-2^{2p-1}}{p(2p-1)(p-1)(2p-3)} - \frac{1}{6p} + \frac{2\sqrt{\frac{N}{m}}}{2p-1} - \frac{\frac{N}{m}}{p-1}}.$$

It is important to note that when $m = 1$ is substituted in (6.1), it reduces to the Davis's 2-D wire-length distribution formula.

When d is the length of vertical wires in unit one device layer depth (distance between two neighbouring layers), the vertical wire-length distributions is derived:

$$V(d) = \frac{2\alpha k N(1 - N^{p-1} - m^{p-2} + m^{-1}N^{p-1})}{m(m-1)}(m-d) \tag{6.2}$$

where $d = 1, 2, \ldots, m-1$.

(B) Chip yield and cost

Another benefit in chip stacking comes from the economics and yield engineering, as cost and yield are the most important factor for chip manufacturing. From the yield models in Chapter 5, it is evident that the larger chip size gives a lower yield; chip yield decays exponentially with area. Therefore, chip yield increases when manufacturing a large number of smaller chips. Not only that, by partitioning a

*Homogeneous in this context means that all gate pairs are at a horizontal distance of l and are equally likely to connect to each other regardless of the layer they are in.

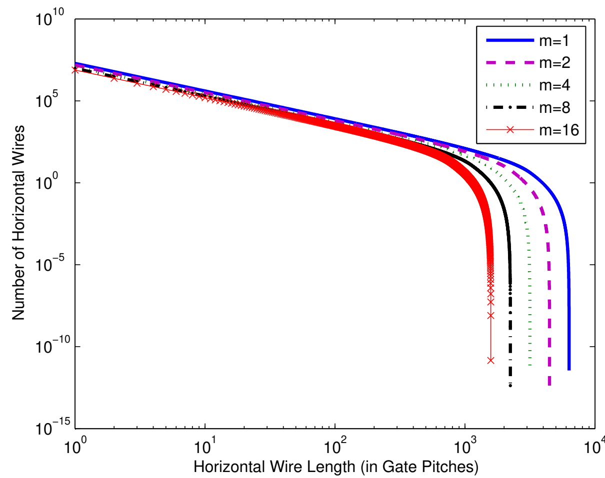


Figure 6.2: *Horizontal Wire-Length Distribution in a 3-D IC. Here, $m = 1$ corresponds to 2-D case. As the number of layers increase the number of global wires as well as local wires decrease.*

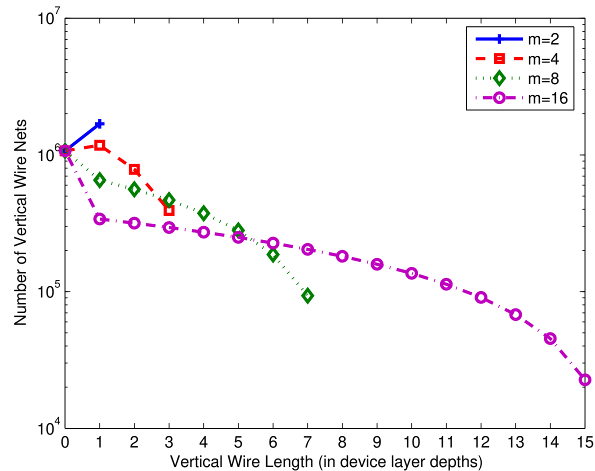


Figure 6.3: *Vertical Wire-Length Distribution in 3-D IC.*

large chip into several smaller chips, the otherwise unused extremities along the circumference of a wafer may be taken advantage of, thereby increasing dies per wafer.

The final footprint of the packaged system is also less for a 3-D implementation. Particularly portable systems such as cell phones and Personal Digital Assistants (PDAs) increased demand for signal processing, memory, sensors and wireless communication to be integrated in a single system that fits into one's hand.

(C) Heterogeneous Integration

With greater emphasis on increasing the functionality on a single die, different functional units such as digital, analog/RF, memory, opto-electronic, MEMS, image sensors, displays etc are being integrated on the same piece of silicon. These functional units inherently perform better in completely different process technologies such as silicon or non-silicon, and may not scale as fast as other technologies. For instance, a digital CMOS process's feature size scales at a faster rate than analog processes. Merging different technologies to make a bigger planar chip requires additional process steps and hence, increased mask count, which finally adds to the total product cost. Thus, with 3-D technology, different functional blocks can be manufactured separately and simply be integrated.

In mixed-signal systems, noise-sensitive analog/RF circuitry is prone to failure due to interference from their digital counterpart through the base silicon substrate [*39*]. 3-D integration aids in the solution for noise isolation as it separates the analog/RF and digital circuits into different substrates, with the metal or the dielectric bonding layer used in wafer-bonding technology providing an effective guard ring [*223*].

As Banerjee *et.al.* described in [*39*], a preliminary analysis shows a $30dB$ improvement in isolation by moving the RF portions of the circuit to a separate substrate. Moreover, since the second Si layer may not be continuous, good isolation between different analog and RF components such as the low-noise amplifier (LNA) and power amplifier can also be achieved.

6.2.2 Challenges for 3-D integration

Despite the great amount of research work carried out so far to realize 3-D integration, there are still numerous challenges to overcome to make 3-D integration a mainstream IC design paradigm.

(A) Thermal Integrity

Thermal integrity is a critical issue in all high-performance chips because system reliability is strongly dependent on the temperature.

The heat generated in a chip due to transistor switching is typically conducted through the silicon substrate to the package and then to the environment by a heat sink. With chip stack designs, ICs in the upper layers will also generate a significant fraction of the heat. These active layers are usually bonded or insulated from each other by layers of dielectrics (LTO, HSQ, polymide, etc.) which typically have much lower thermal conductivity than Si [*5*, *7*]. Therefore, it is difficult to remove the excess heat from chips or dies that have more than one degree of separation from the heat sink. The increased heat causes further leakage, which in turn increases the temperature, an undesirable cycle which can cause catastrophic breakdown. Other causes includes degradation in device performance, reduction in chip reliability due to increased junction leakage, electromigration failures [*224*].

Thermal management involves the control of temperatures of materials within a package. There are two basic cases for thermal management in 3-D IC systems: few layer and many layer systems [*7*]. In few layer systems, heat transport is vertically directed as in single ICs, since the component stack is mounted flat. In most cases,

the polymeric adhesive material may be a problem, but due to the total traversal distance being short, the high thermal resistive of polymer may not have an effect. However, two basic approaches can be considered in this case. First, the stack can be arranged in a way that places the most dissipative substrates closest to the heat removal surface. Second, it is possible to consider loading the polymeric material with insulators that conduct heat better.

In many layer systems, heat transport is more isotropic [*7*]: in this case two previously mentioned approaches apply, but the heat removal surface of choice may be different. Also, it is possible to include other heat removing structures such as periodic heat spreader layers or vertical columns or slugs of highly conductive materials such as Copper thermal vias (T-vias) that can can run the lateral extent of the entire ensemble of the 3-D assembly. However, thermal vias further increase the routing congestion [*225*, *226*]. Nevertheless, careful thermal-via placement in high performance systems have the potential to effectively control the temperature in 3-D ICs.

Some alternative methods that have been proposed, such as integrated micro-channel cooling [*227*, *228*] may also be a viable option. Moreover, it is shown in [*229*] that even though the increased temperature reduces the highest operating frequency, the overall system performance can still be comparatively better than in a 2-D implementation.

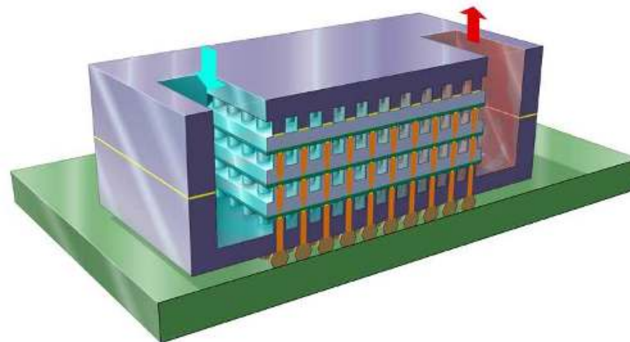


Figure 6.4: *IBMs Water Cooling Technique for 3-D Chips; tiny rivers as thin as a human hair (50 microns) of water flows between the individual chip layers in order to remove heat efficiently at the source [230].*

(B) Electromagnetic Interactions in 3-D ICs

Parasitic coupling effects among different layers in 3-D chips are expected to be present. For example, in a 3-D IC, additional coupling exists between the top layer metal of the ith active layer and the devices on the $(i+1)^{th}$ active layer. This needs to be addressed at the circuit design stage. However, for nanometer technologies, the aspect ratio of global tier interconnects is larger (usually ≥ 2) compared to local wires. Of these wires, the line-to-line (coupling) capacitance is the dominant portion of the overall interconnect capacitance, and therefore, the presence of an

additional silicon layer on top of a global metal line may not have an appreciable effect on the line capacitance per unit length. For technologies with global wires with a much smaller aspect ratio, the change in interconnect capacitance due to the presence of an additional silicon layer could be significant [*231*].

TSVs used for vertical communication between ICs in the stack may have signal integrity effects due to the substrate noise, and the devices in close proximity to TSVs may perform differently. Rousseau *et.al.* in [*232*] investigate the electrostatic impact that a TSV may have on a MOS transistor both in analogue and digital applications. They conclude that for digital mode of operation the electrostatic effect is negligible whereas in the analogue mode frequencies higher than 100 MHz involve parasitic coupling effects, and propose new design rules and layout methodology for 3-D ICs in order to ensure reliable electrical compatibility. Also, high frequency signal transients in one layer could simply couple with wires in neighbouring layers.

In 3-D ICs, the reduction of wire lengths will certainly help to reduce the inductance. When the substrate resistance is sufficiently low or the wafers are bonded through metal pads, the presence of another substrate close to the global wires will lower the loop inductance by providing shorter return paths [*39*].

(C) Reliability

Reliability of 3-D ICs due to electrothermal and thermomechanical effects between active layers and their interfaces is another concern that designers need to cope with [*39*, *232*]. Usually hot spots are inevitable due to the temperature distributions within a chip package, and as a consequence, the activation of failure modes in components and disruptions in physical structures inside the package due to differential expansion of dissimilar materials at their interfaces may occur [*233*].

Also the heterogeneous integration of technologies in 3-D architecture will increase the need to understand mechanical and thermal behavior of new material interfaces, thin-film-material thermal and mechanical properties, and barrier/glue layer integrity .

From a manufacturing point of view, yield issues might arise due to the mismatch between the individual die-yield of different active layers and the bonding technique to be used, which may affect the overall yield of 3-D chips. Yield issues would demand a careful trade-off between system performance, cost, and the 3-D manufacturing technology.

6.2.3 Three-Dimensional Integration Options

Vast numbers of research groups both in academia and the industry have vigorously studied 3-D integration technologies in recent years [*234*, *165*, *235*] using diverse processing steps - bare dies, packaged dies, MCM's and custom wafers stacked along the z-axis to form 3-D IC. The simplest way to distinguish among the various methods is by categorizing them s wafer-level and chip level methods. Several comprehensive descriptions of 3-D integration technologies and their comparisons are described in the literature [*165*, *235*] to mention a few.

Basically 3-D stacking was focused on chip stacking, and this method is widely known as 3-D SiP. Stacked die packages as such are very common in mobile products. This can be either chip-to-chip(C2C) or Package-on-Package (PoP) or MCM-

Figure 6.5: *System-In-Package (Die stacking using wire bonding)*

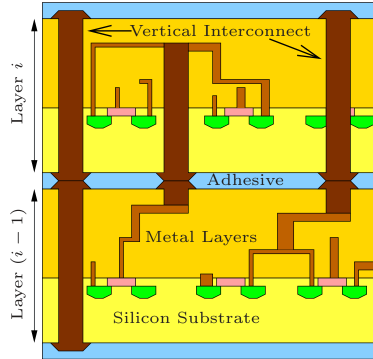


Figure 6.6: *Wafer Level Integration with vertical interconnects*

to-MCM bonding using epoxy or glues and creating electrical connections by wire-bonding techniques and this 3-D-C2C bonding makes it possible to stack KGDs in layers [*236*]. This method of integration is referred as 3D-SiP hereafter. Many of the 3D packaging systems that are manufacturing today are mostly memory modules, especially for *USB memory and CF, SD, XD memory*[*162*]. The most critical parameter is the thickness of each component. Up to eight assembled die packages are currently available in the market. This has been the approach for heterogeneous integration of mixed-signal systems, where different high-performance Intellectual Property (IP) blocks could also be integrated to achieve better performance, but the production throughput of this Chip-to-Chip bonding is very low compared to wafer level stacking.

Wafer-Level stacking can be performed in two ways; in the first case entire wafers are bonded into one single wafer and subsequently diced (3D-W2W), whereas in the later method, known-good-dies (KGDs) are bonded on top of a host wafer containing other KGD sites (3D-D2W) [*237*]. Furthermore, there are two primary wafer orientation schemes known as Face-To-Face and Face-To-Back. The former, provides the greatest layer-to-layer via density, and is suitable for two-layer organizations. Face-to-Back, on the other hand, provides uniform scalability to an arbitrary number of layers, except the reduced inter-layer via density. One of the keys to form vertical interconnections is to use through-hole vias, which is an integral part of the semiconductor process flow (discussed in Section 4.4). These vias can be formed after wafer fabrication (post-passivation) by etching a hole in Silicon, depositing a thin insulation layer (SiO_2) on the side walls and bottom,

filling the hole with a thick copper plug, and polishing away the excess copper plated on the wafer. However, due to the various limitations and drawbacks of these post-passivation through hole vias, there are new processes proposed recently to eliminate various short comings.

In general, there are three methods for three-dimensional integration: Wafer-to-Wafer (W2W), Die-to-Wafer (D2W), and SiP [*165*, *235*]. Chip-to-wafer 3-D integration technology can provide a high yield and a high flexibility in chip size, compared with wafer-to-wafer 3-D integration technology. In addition, a high fabrication throughput can be achieved in chip-to-wafer 3-D integration technology, compared with chip-to-chip 3-D integration technology. The fabrication throughput in the chip-to-wafer 3-D integration technology is lower than that in wafer-to-wafer 3-D integration technology because many chips used in the chip-to-wafer 3-D integration are mechanically aligned onto an LSI wafer chip by chip. To overcome this problem, another innovative approach to chip-to-wafer bonding with a highly precise alignment technique in batch process is under investigation. Thus, chip-to-wafer 3-D integration technology can simultaneously satisfy these requirements of production yield and fabrication throughput.

6.3 Early Estimation of Cost and Performance

The basic goal in electronic system design is to find a design methodology which balances performance with ease of manufacture, while minimizing cost. So, system implementation or manufacturing issues must be addressed early in the design cycle intelligently and quickly before major investments are committed or design work begins. It has been identified that decisions made within the first 20% of the total design cycle time will ultimately affect upto 80% of the final product cost [*210*]. Therefore, making appropriate design choices early in the design cycle is essential and will have a significant impact throughout the design and production lifecycles.

The possible implementations of mixed-signal Systems is shown in Figure 6.7. It ranges from traditional two-dimensional SoC implementation toward 2-D SoP or 3-D SiP or 3-D WLI. As designers face more dimensions for system integration, an optimal total solution should be pursued with an accurate trade-off analysis between different design metrics. First, a decision has to be made as to in which technology (SoC or SoP or SiP or WLI) a system should be implemented. The most important metrics for these decisions are probably the performance of the resulting systems and cost of implementation. So far, these kinds of trade-offs are made in a relatively crude and generic way, dependent on the designer's expertise. As system complexity further rises, more structured approaches are necessary.

The key challenges of early cost and performance estimation for system implementation are as follows[*190*]:

1. Lack of physical layout information
2. Lack of accurate models for performance estimation, mixed-signal isolation, IP module integration, and technology fusion
3. Lack of accurate and efficient computation algorithms since most of the systems are inherently complex.

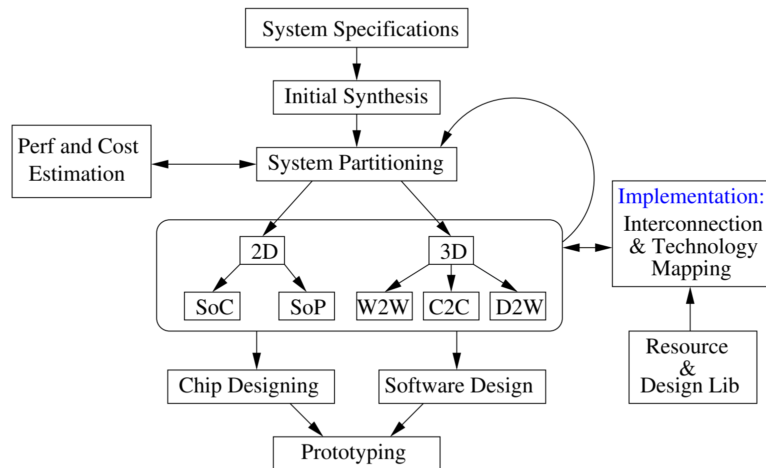


Figure 6.7: *Mixed-signal electronic system design flow and options.*

Furthermore, from a marketing perspective, a cost, performance and other technological trade-off analysis early in the design cycle provides more freedom in the process of decision making. Being the first with higher performance for a higher cost is as bad as with poor performance for a lower cost, because both will be rejected in the market.

6.3.1 Models for Trade-off Analysis

The trade-offs of 2-D integration of Mixed signal systems has qualitatively been analyzed in [*190*]. Shen *et.al.* analyse the mixed-signal isolation of complex electronic systems and how cost effective they are. This work extends those models toward 3-D integration with the emphasis on interconnect performance estimation, and include the costs related to KGD testing for all cost models, which was not discussed in [*190*].

The cost estimations for all implementation choices are based in area, and estimated from the model discussed in Chapter 5. In the performance estimations, the latency for the longest possible link is the characteristic metric used for comparison. For example in a planar system the latency between two diagonal corners is considered, while in a 3-D system, the delay from a corner on the bottom chip to a diagonally opposite corner on the top chip is considered. One of the challenges in 3-D integration is its increased vertical temperature profile and it is considered as a constraint for optimization assuming an allowable chip temperature for reliable operation of the entire circuit. In addition to that, the effect of temperature hinders the interconnect performance.

(A) Analytical Die Thermal Model for 2-D and 3-D Integration

Assuming the heat dissipates through the Silicon substrate, the average die temperature can be usually described using a one-dimensional heat equation when the

die size is much larger than its thickness (t) [*233*]:

$$T_{die} = T_{ambient} + \left(\frac{t}{kA}\right) P_{chip}, \tag{6.3}$$

where $T_{ambient}$ is the ambient temperature, P_{chip} is the chip power dissipation, A is the chip area, and k is the thermal conductivity of the material. The factor $\frac{t}{kA}$ in (6.3) is known as the effective thermal resistance (R) of the substrate layer and the package. In this analysis, the contribution to the chip temperature from interconnect joule heating is disregarded.

If the same assumption is made that the die size is much larger than its thickness, the maximum temperature in a 3D-IC occurs at the highest device layer. Then as described in [*233*] , the average die temperature of a 3-D IC with m layers is:

$$T_{3D} = T_{ambient} + \sum_{i=1}^{m} R_{(i-1),i} \sum_{j=i}^{m} P_j, \tag{6.4}$$

where $R_{(i-1),i}$ is the effective thermal resistance between the ith and $(i-1)^{th}$ layer including the glue layer where applicable, and P_j is the power dissipation in the kth active layer.

With the increasing power density of nanoscale chips, die temperatures are expected to rise substantially. The thermal problem is further aggravated by the fact that leakage power is exponentially dependent on temperature. Hence rising temperatures lead to larger leakage power dissipation and vice versa in a positive feedback relationship.

One effective way to alleviate the excessive heat generated in 3D-ICs is to incorporate dummy thermal vias (T-vias); the thermal conductivity of a die layer is significantly improved by the existence of thermal vias. When k_{thv} and k_{layer} are the thermal conductivity of a thermal via and the layer respectively and m is the fraction of area occupied by the thermal vias to the total area, the effective thermal conductivity of the layer is [*7*]:

$$k_{eff} = mk_{thv} + (1-m)k_{layer}. \tag{6.5}$$

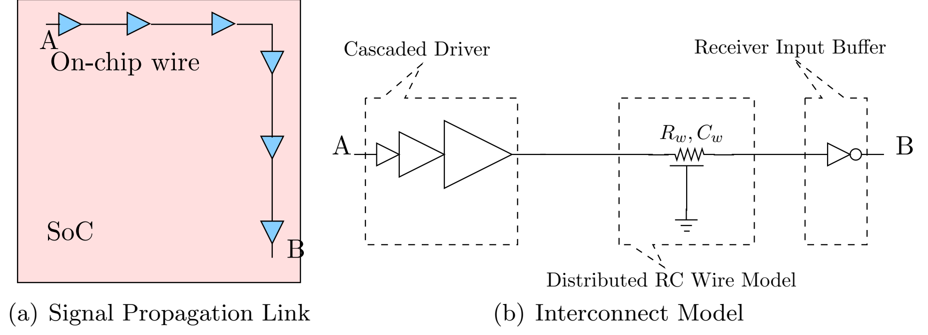
To estimate the thermal resistance, the effective thermal conductivity coefficient for each pair of layers, for example die and glue, should be found.

(B) Interconnect Performance Models

On-Chip Wire Delay Typically, on-chip global signal wires are highly resistive while the inductance is negligible. Hence signal transmission obeys the diffusion equation. The appropriate model therefore, is a distributed resistance-capacitance (RC) line [*6*, *20*]. A very good approximation to the delay over an RC dominated wire with capacitive load C_L connected at the far-end is given by the first order Elmore approximation is[*6*]:

$$t_{rc_d} = 0.693\left\{R_d(C_d + c_wL + C_L) + r_wLC_L\right\} + 0.377 r_w c_w L^2 \tag{6.6}$$

where R_d is the driving inverter's equivalent output impedance and C_d the self-loading drain diffusion capacitance, while c_w and r_w are the per-unit-length capacitance and resistance of the interconnect and L its length. The wire capacitance



(a) Signal Propagation Link (b) Interconnect Model

Figure 6.8: *Signal Propagation for Worst-Case Latency in a SoC.*

incorporates the coupling capacitance to adjacent wires, if necessary with an appropriate switching factor to allow for worst-case coupling [*112*], resulting in a combined total equivalent capacitance. It can be seen that the delay increases exponentially with length when the wire parasitics dominate. The most common method of reducing this delay over long interconnects is to insert repeaters at appropriate positions, which makes the wire delay linear with length. However, repeater insertion is effective only when wire time constant ($r_w c_w L^2$) is at least seven times the time constant of a repeater ($R_d(C_d + C_g)$) [*6*]. The 50% delay for a repeater inserted on-chip wire of length L is:

$$t_{rc} = k\left\{0.69\frac{R_d}{H}\left[H(C_d + C_g) + \frac{c_w L}{k}\right] + 0.69\frac{r_w L}{k}HC_g + 0.377\frac{r_w L}{k}\frac{C_w L}{k}\right\} + 0.69\frac{R_d}{H}(C_d + C_L) \quad (6.7)$$

where H and k are the delay optimal repeater sizes and numbers given by $H = \sqrt{\frac{R_d(C_L + cwL)}{r_w L C_g}}$ and $k = L\sqrt{\frac{0.4 r_w c_w}{0.69 r_w L C_g}}$, respectively.

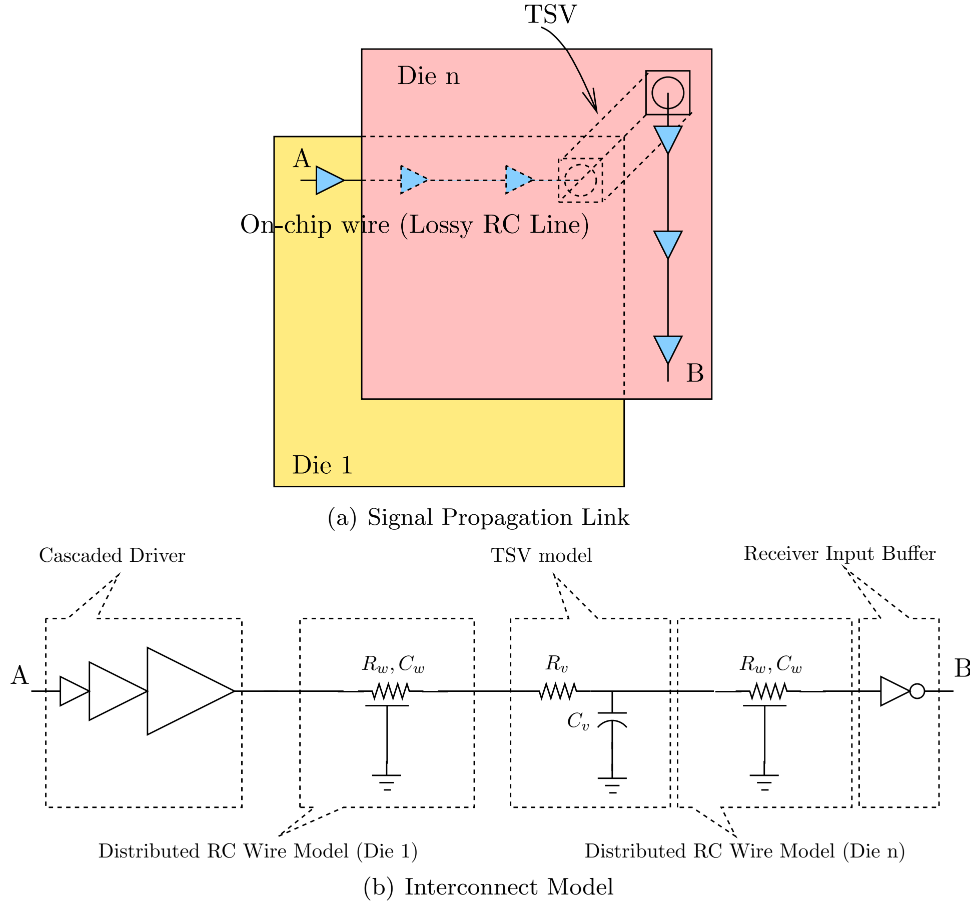
Finally the total propagation delay of the on-chip global wire, as shown in Figure 6.8(b), is the sum of the cascaded buffer delay (t_{drv}) at the near-end and the repeater-inserted delay of the RC wire:

$$t_{intra} = t_{drv} + t_{rc} \quad (6.8)$$

Off-Chip Wire Delay Inter-chip wires on a typical package substrate are characterized by conductors with low resistivity and a relatively large cross-section in a low-loss dielectric making losses due to shunt conductance negligible. Hence signal transmission exhibits transmission line behaviour. In a lossy transmission line, both RC and LC delays co-exist. For LC dominated wires, the signal propagation delay is equal to its time-of-flight.

$$t_{LC} = t_{tof} = L\sqrt{l_w c_w} \quad (6.9)$$

If a wire is a very resistive transmission line, the following empirical formula for adding the time-of-flight (t_{tof}) delay and conventional RC delay (t_{rc_tl}) was found



(a) Signal Propagation Link

(b) Interconnect Model

Figure 6.9: *Signal Propagation for Worst-Case Latency in 3-D Wafer-Level Chip Stack.*

in [*88*] to accurately predict the total wire delay:

$$t_{RLC} = (t_{tof}^{1.6} + t_{rc_tl}^{1.6})^{\frac{1}{1.6}} \tag{6.10}$$

For the inter-chip communication link shown in Figure 6.10(c), the following expressions can be derived:

$$t_{rc_tl} = 0.693 \left[Z_0(Cd + C_{pad} + C_{bnd} + 0.5C_L) + \frac{L_{bnd}}{Z_0} + r_w L(C_{pad} + C_{bnd} + C_L) \right] + 0.4 r_w c_w L^2 \tag{6.11}$$

Finally, the total delay for the inter-chip communication link is the summation of the cascaded driver delay (t_{drv}), the RLC-wire delay (t_{RLC}):

$$t_{inter} = t_{drv} + t_{RLC} \tag{6.12}$$

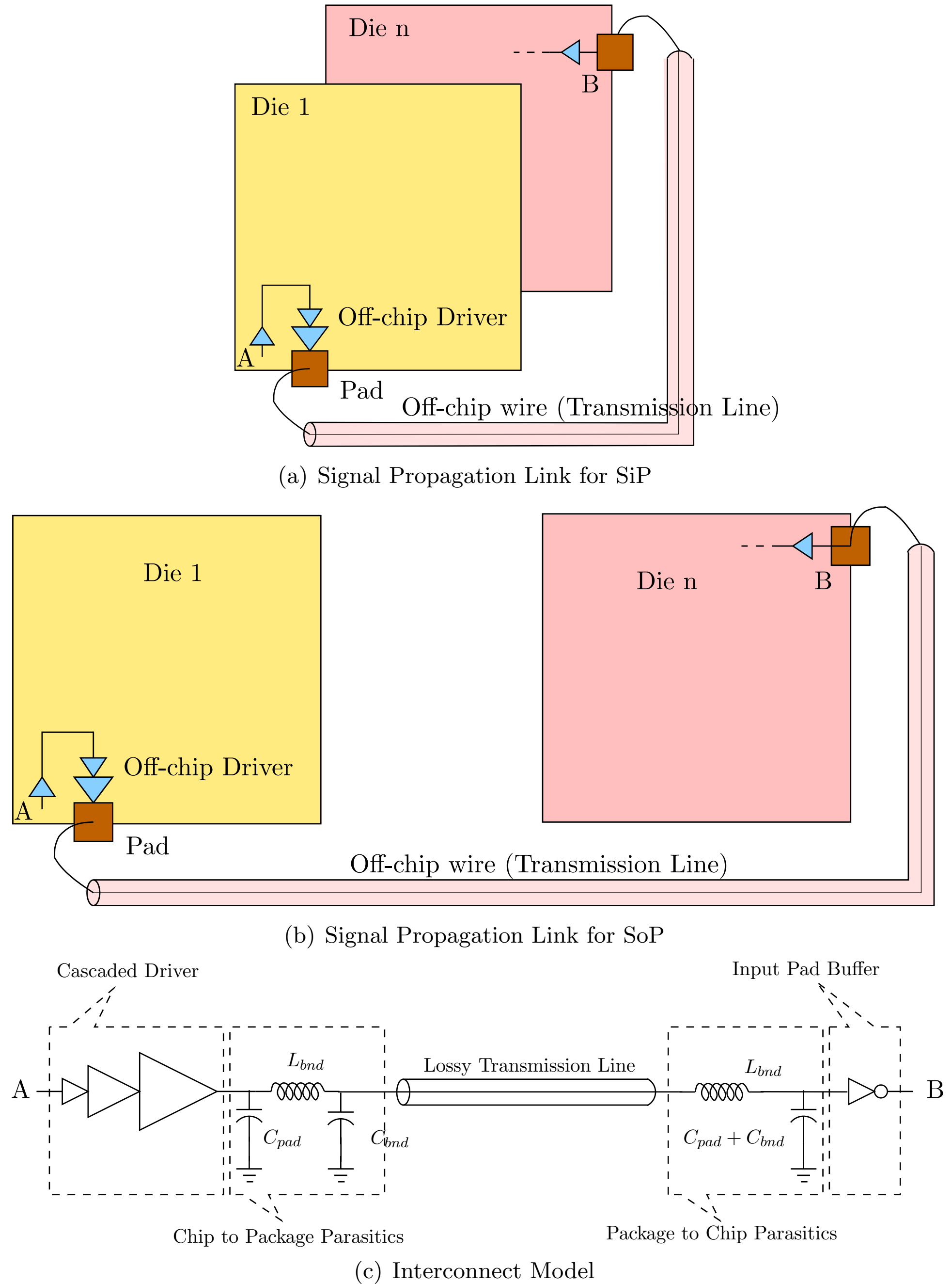


Figure 6.10: *Signal propagation for worst-case latency in 2D-SoP and 3D-SiP type arrangements.*

Cascaded Driver Delay The number stages, N, in the cascaded buffer can be calculated by pessimistically assuming that the first inverter in the cascaded buffer is a minimum sized inverter and is given by:

$$N = \frac{ln(x)}{ln(u)} \tag{6.13}$$

where x being the ratio of output resistance of the first inverter stage to characteristic impedance of the line, if an impedance matched line is necessary [238]. Otherwise, that means if the line is RC dominated, x is just the ration of the input capacitance of the first repeater in the RC wire or the wire capacitance itself and the first inverter in the cascaded buffer.

$$x = \begin{cases} \frac{R_d}{Z0} & \text{for LC wire} \\ \frac{C_{out}}{C_g} & \text{for RC wire} \end{cases} \tag{6.14}$$

Fan-out of 4 per stage ($u = 4$) is assumed and then total delay of the cascaded buffer is estimated from:

$$\tau_{drv} = 0.69NR_d(C_d + uC_g) \tag{6.15}$$

Thermal Effect on Interconnect Performance The driver resistance R_d and wire resistance r_w both increase with temperature. Usually R_d is expressed in terms of the saturation current of the device when the gate voltage is equal to the supply voltage:

$$R_d(T) = \frac{V_{dd}}{Kv_{sat}(T)W\left(V_{DD} - V_{th}(T)\right)^{\alpha}} \tag{6.16}$$

where K is a constant that is specific to a given technology, T is the temperature in Kelvin, V_{th} is the threshold voltage at temperate T, and v_{sat} is the saturation velocity. As validated in [239], when V_{DD} is sufficiently larger than V_{th}, the change in V_{th} with temperature is relatively insignificant compared to the change in v_{sat}. However, as V_{DD} is scaled down, V_{th} has a comparable and counter effect to the change in v_{sat}. Therefore for a 65nm CMOS technology the driver resistance can be taken as a constant with increasing temperature [239].

Wire resistance, R_w, increases linearly with temperature due to the change in the effective metal resistivity in relation to the barrier layer. In order to characterize the dependence of wire resistance with temperature, a linear relationship given by:

$$R_w(T) = \frac{\rho(T_0)l}{tw}\left[1 + t_{cr_bulk}(T - T_0)\right] \tag{6.17}$$

can be used [46]. In (6.17), $R(T)$ is the wire resistance at any given temperature T, $\rho(T_0)$ is wire resistivity at the reference temperature T_0, w and h are wire width and height, t_{cr_bulk} is the temperature coefficient of resistance (TCR) of the bulk material, which is around $t_{cr_bulk} = 0.0039^oC^{-1}$ [46].

6.4 Tradeoff Analysis for SoC, SoP and 3-D Implementations

The models and methodology proposed in this paper are demonstrated in a case study comprising a comparison of two mixed-signal systems, a wireless sensor and

a 3G mobile terminal. The *wireless sensor* contains a 2 Mb DRAM, an ASIC and Microprocessor with gate counts of 500k and 300k respectively, and an Analog/RF block occupying an area of 2 mm^2. It also contains a MEMS sensor with an area of 1 mm^2. The *3G mobile terminal* has a similar architecture, but with a larger memory of 128 Mb DRAM, and a CMOS image sensor with a pixel size of 1.75 $\mu m \times 1.75$ μm, and resolution of 8 $Megapixels$ instead of the MEMS sensor [*240*]. Further, in the analysis we consider the ASIC and Microprocessor together as a single logic block, treating our target system as comprising only four megacells: analog/RF, logic, memory, and a MEMS or CMOS image sensor. For all integration schemes, the underlying manufacturing process is a 65 nm, 11-metal, CMOS process with a wafer diameter of 300 mm and a lower-level wire pitch of 136 nm [*162*]. We also assume a peripheral in-line pad arrangement and wire bond packaging. All the other key parameters are listed in Table 6.1. The worst-case delay for 2-D systems is estimated diagonally from chip edge to chip edge, while it is estimated from one edge of the bottom chip to the diagonally opposite edge of the topmost chip for 3-D systems.

Notation	Parameter	Value
Do	Defect Density per m^2	250
S	Shape Factor	0.6
N_{dram}	DRAM Mask Layers	13
N_{logic}	Logic mask Layers	18
N_{RF}	CMOS RF Mask Layers	12
N_{MEMS}	MEMS Process Mask Layers	6
N_{CIS}	CMOS Image Sensor Process Mask Layers	10
D_{wafer}	Wafer Diameter	300 mm
C_{lgc}	Process cost per mask layer (logic)	700 $
C_{mixed}	Process cost per mask layer (mixed-signal)	1000 $
C_{mcm}	MCM-D cost per unit area per layer	1000 $
C_{asmb}	Cost of assembly per pin	0.01 $
C_{sub}	Cost of substrate	300 $
C_{3Dvia}	Cost of making a through hole via in WLP	0.01 $
C_{rewrk}	Cost of Rework	3 $
C_{SOI}	Cost of SOI substrate	2000 $
C_{wfr_tst}	Wafer Test Cost	0.1 $
FC_{wfr}	Wafer Test Coverage	80%
C_{burnin}	Die Burn-In and test Cost	0.2 $
FC_{die}	Die Test coverage	99%
C_{mod_tst}	Module/Chip test cost	0.3 $
FC_{mod}	Module/Chip test coverage	95%
Y_{MCMsub}	Yield of MCM substrate production	0.98
Y_{asmb}	Yield of assembly	0.97
Y_{3Dsub}	Yield of Wafer-Level 3-D stacking	0.98
α, β, γ	Area merging factors	2,1,1
K_p	Rent's Coeff. (ASIC,DRAM)	2, 4
	Rent's Coeff. (μP, module)	7, 1.4
ρ	Rent's Exp (ASIC,DRAM)	0.35, 0.12
	Rent's Exp (μP, module)	0.4, 0.63

continued overleaf

Notation	Parameter	Value
p_g	Global Metal Pitch	290 nm
p_i	Intermediate Wire Pitch	195 nm
p_l	Local Wire Pitch	152 nm
n_w	Number of interconnection layers (on-chip)	11
e_{rout}	Efficiency of routing tool	0.4
f_g	fanout of gates	3
P_p	Peripheral in-line pad pitch	60 μm
A_g	Gate Area	$1\mu m$
$A_{dramcell}$	DRAM Cell Area[*241*, *162*]	$0.05\mu m$
n_{w_mcm}	Number of interconnection layers (MCM-D)	8
P_{w_mcm}	Interconnect pitch (MCM-D)	20 μm
l_{bw}	Length of bondwire	1 mm
L_{bw}	Inductance of bondwire	2 nH
C_{bw}	Capacitance of bondwire	0.3 pF
R_d	Min. sized Buffer Output Resistance	20.8 $k\Omega$
C_g	Min. sized Buffer Input Capacitance	0.14 fF
C_d	Min. sized Buffer Output Capacitance	0.22 fF
R_v	Resistance of through-hole via[*165*]	0.35 Ω
C_v	Capacitance of through-hole via[*165*]	5 fF
C_{pad}	Capacitance of the bond pad	2 pF
t_{layer}	Total Thickness of a Die	20 μm
t_{glue}	Thickness of the glue layer in 3-D stack	2 μm
t_{Cu}	Thickness of Cu metalization layers per die	12 μm
k_{Cu}	Thermal Conductivity of Cu	385 $\frac{W}{mK}$
k_{ILD}	Thermal Conductivity of Dielectric	0.19 $\frac{W}{mK}$
k_{glue}	Thermal Conductivity of Glue layer	0.25 $\frac{W}{mK}$
k_{Si}	Thermal Conductivity of Si	148 $\frac{W}{mK}$
k_{pkg}	Thermal Conductivity of Package Material	0.35 $\frac{W}{mK}$
k_{board}	Thermal Conductivity of PCB	20 $\frac{W}{mK}$

Table 6.1: *Representative Values for a 65nm technology and summary of Notation for Major Parameters used in the analysis.*

Based on the manufacturers data, the power density for the constituent sub-modules in our case studies can be estimated. The power density for a DRAM is estimated to be 0.02 W/mm^2 [*242*], and for a logic block, 0.12 W/mm^2[*243*]. A CMOS Image sensor has an average power density of 0.016 W/mm^2. The power dissipation of the MEMS sensor is assumed to be 50 mW, while for the Analog/RF block it is assumed to be 500 mW. For the stacked arrangement, we assume that the logic block is closest to the heat sink and that the other blocks are in the following order from nearest the logic (and heat sink) to furthest: DRAM, Analog/RF block, and MEMS/CMOS Image sensor.

In contemporary IC design, a major design consideration is to maintain operating temperature at a level which is not detrimental to the desired performance, reliability, and durability. Usually in most of ICs, the circuits are often designed for the worst-case temperature of 125 ^oC [*244*]. However, DRAM data retention depends heavily operating on temperature, and should usually be maintained below

Parameter		On-Chip	Off-Chip
Physical	$W(nm)$	290	15
	$T(nm)$	319	5
	$H(nm)$	290	25
	$S(nm)$	145	50
	k_{ILD}	2.5	3.5
Electrical	$R_w(\Omega/mm)$	237	0.02
	$C_w(fF/mm)$	137	83
	$l_w(nH/mm)$	0.13	0.41
	$Z_0(\Omega)$	31	70

Table 6.2: *On-chip and off-chip wire parameters [69].*

approximately 85 ^oc. In this analysis, we assume that the ambient temperature is maintained at 35 ^oC without any loss of generality. The methodology allows for the viability of any operating temperature to be investigated.

6.4.1 Monolithic SoC

As mentioned earlier, the mixed-signal systems in consideration contain four different functional blocks and they may require different technologies for the implementation. In this circumstance the monolithic SoC area is estimated from the heterogeneous chip size estimation formula given in (5.14). Note that we assumed a MEMS-CMOS combined process for SoC implementation of the first system, the wireless sensor node. The total cost for an SoC implementation is given in (6.18).

$$C_{SoC} = \left[\left(\frac{C_{wafer}}{Y_{SoC}N_{die}} + C_{wafer_test}\right)\frac{1}{PF_w} + C_{burn_in}\right]\frac{1}{PF_b} + C_{pkg} \tag{6.18}$$

Multiplying the total power dissipation by the series combination of the substrate and package thermal resistances, we can estimate the average chip temperature.

6.4.2 2D-SoP

In the 2D-SoP implementation, we assume that four chips (DRAM, RF, Logic and MEMS/Image Sensor) are assembled as a multi chip module (MCM). Hence, the cost of implementing the MCM includes the total cost for each chip including testing cost, assembly cost, substrate cost, rework cost, and finally the MCM test and packaging costs.

The SoP can provide some reworking capability whereas SoC and wafer-level 3-D integration do not. If a single rework cycle is assumed for SoP, the yield in assembly is improved from Y_a to $(2 - Y_a)Y_a$. Then the cost for SoP is given by (6.20) and the overall yield as described in [*206*] is:

$$Y_{SoP} = Y_a \prod_{i=1}^{m} Y_{chip_i}^{(1-Fc)}, \tag{6.19}$$

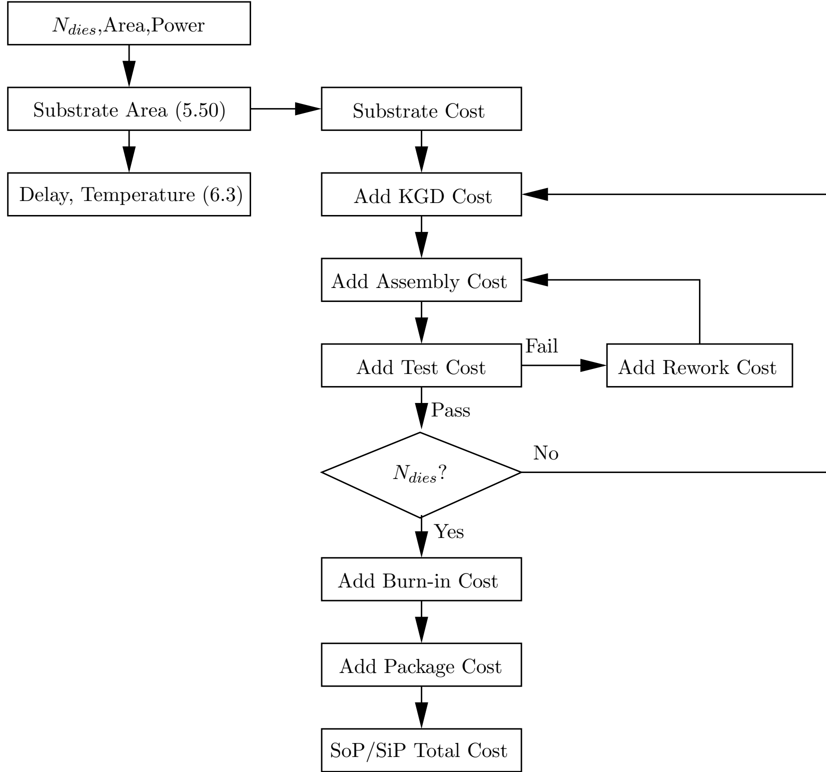


Figure 6.11: *SoP/SiP integration trade-off analysis flow*

where Y_{chip_i} is the yield for i th chip.

$$C_{SoP} = \left\{ \frac{\sum_{i=1}^{m} C_{kgd_i} + \frac{C_{substrate}}{Y_s} + C_{assembly} + C_{rework}}{Y_a} + C_{test} \right\} \frac{1}{PF_{SoP}} + C_{pkg} \tag{6.20}$$

The overall temperature is found by estimating the effective chip thermal resistance from $R_{eff_SoP} = \sum_{i=1}^{n} \frac{t_i}{k_i A_i}$ and then multiplying the total power dissipation of all chips by the series combination of thermal resistances R_{eff_SoP}, R_{pkg}(Package), and R_{subs}(substrate).

6.4.3 3D-SiP

A 3D-SiP implementation is similar to the SoP package integration, except that the SiP implementation integrates dies vertically. The cost formula is the same, but the MCM substrate area is reduced, compared to the 2D-SoP implementation. The thermal profile is also found in a similar manner, using (6.4).

$$C_{3D_SiP} = \left\{ \frac{\sum_{i=1}^{m} C_{kgd_i} + \frac{C_{substrate}}{Y_s} + C_{assembly} + C_{rework}}{Y_a} + C_{test} \right\} \frac{1}{PF_{3D_SiP}} + C_{pkg} \tag{6.21}$$

6.4.4 3D-WLI

The yield of each 3-D implementation method is the cumulative yield over all layers (m) and is given by:

$$Y_{3D} = Y_{2D} \prod_{i=1}^{m-1} Y_{2D_i} Y_a \tag{6.22}$$

where Y_{2D} is the fabrication yield of the 2D process, and Y_a is the yield loss due to the 3-D assembly process. The Y_a^{m-1} term in the equation takes into account the fact that integration of m layers of chips requires $m-1$ silicon growth or wafer bonding procedures. In the case of D2W stacking, die yield after KGD testing should be considered. Hence the overall yields for implementing our target system in 3D-W2W and 3D-D2W methods as described in [*206*, *234*] are as follows:

$$Y_{3D_w2w} = Y_{2D} \prod_{i=1}^{m-1} Y_{2D_i} Y_a \tag{6.23}$$

$$Y_{3D_d2w} = Y_{2D}^{(1-F_c)} \prod_{i=1}^{m-1} Y_{2D_i}^{(1-F_c)} Y_a \tag{6.24}$$

The total cost for 3-D Wafer-Level integration is given in (6.25) and (6.26).

$$C_{3D_W2W} = \left\{ \frac{\sum_{i=1}^{m} C_{die_i} + C_{bonding}}{Y_{a_3D_W2W}} + C_{test} \right\} \frac{1}{PF_{W2W}} + C_{pkg} \tag{6.25}$$

$$C_{3D_D2W} = \left\{ \frac{\sum_{i=1}^{m} C_{kgd_i} + C_{bonding}}{Y_{a_3D_D2W}} + C_{test} \right\} \frac{1}{PF_{D2W}} + C_{pkg} \tag{6.26}$$

Due to limitations in the wafer level processing, there is no possibility of reworking. In the case of D2W integration methodology, wafer level test and burning-in costs for each die as well as the final module test cost have been considered. However, in W2W technology, there is no die burn-in process to contribute to the cost.

It was assumed that standard test equipment can be used for testing of 3-D chips. If specialized equipment is to be used, their depreciation contribution to the cost has to be considered.

In a W2W integration methodology all dies must be of the same size in order to alleviate manufacturing difficulties, especially the precise alignment of wafers to

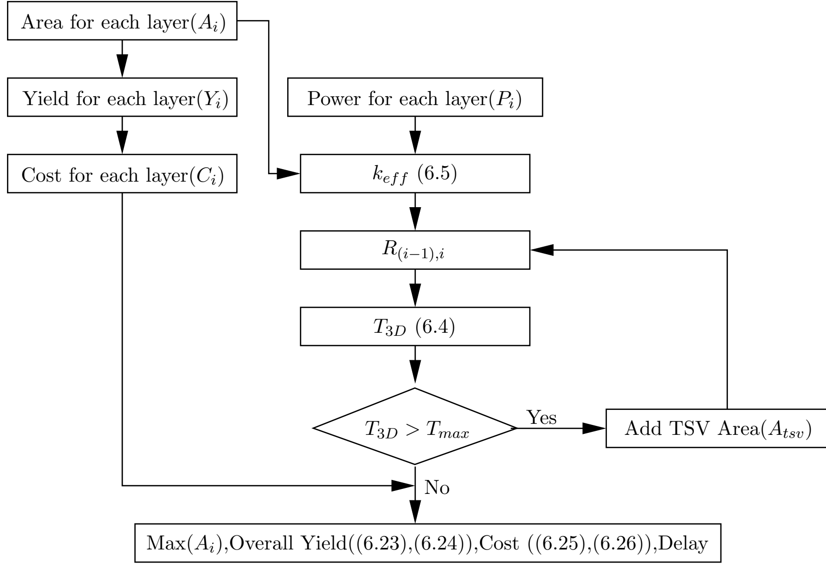


Figure 6.12: *3-D IC trade-off analysis flow.*

make the vertical interconnections and facilitate dicing, whereas for D2W integration the dies can be different in size. The thermal profile is calculated using (6.4), and when the topmost layer's temperature exceeds the allowable limit, T-via insertion is carried out. The area of state-of-the-art TSV is on the order of a few μm^2 [*165*], and inclusion of T-vias result in an area increase, and hence, yield reduction. Thus, the chip manufacturing cost increases.

6.5 Discussion

The results of the case studies are shown in Table 6.3. The following implementation options have been considered in the trade-off analyses: a single-chip planar SoC, and two-chip and four-chip arrangements of the different implementation options of 2D-SoP, 3D-SiP, 3D-W2W and 3D-D2W integration. In the two-chip arrangement, Logic and DRAM blocks have been merged to form one chip while the other two blocks have been merged to form the second chip. In the four-chip arrangement, each individual block constitutes a chip. Each case is discussed in detail in the remainder of this section.

For the mobile terminal, 3-D integration provides the most compact design compared to 2-D planar techniques. Irrespective of whether two or four layer stacking is carried out, the difference in the final footprint is approximately 5%, due to the area dominance of the image sensor. Since the area of the mobile terminal is relatively large, the yield of the SoC implementation is rather low, while all other implementations result in higher yields. As can be expected, 3D-W2W integration inherently results in a lower yield in comparison with other 3-D implementations, since untested dies are stacked together. In spite of this though, the final cost of

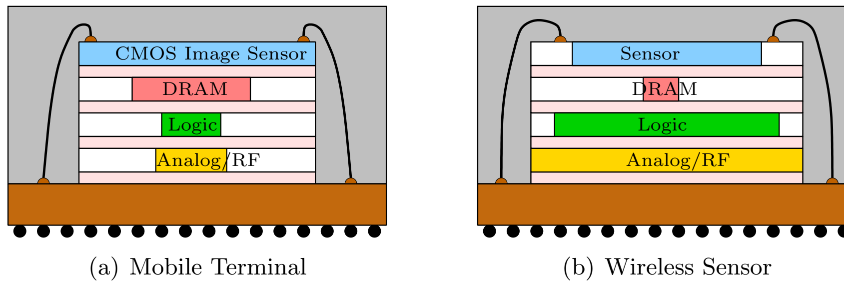
Mobile Terminal

	Single Chip (SoC)	Two Chips (Logic+DRAM,Analog/RF+IS)				Four Chips (Logic,DRAM,Analog/RF,IS)			
		2D-SoP	3D-SiP	3D-W2W	3D-D2W	2D-SoP	3D-SiP	3D-W2W	3D-D2W
Norm. Area	1.00	1.79	0.79	0.76	0.76	2.20	0.75	0.71	0.71
$Yield_{overall}$	0.56	0.98	0.98	0.88	0.98	0.98	0.98	0.84	0.94
Norm. Cost	1.00	0.54	0.66	0.39	0.47	0.71	0.74	0.54	0.76
Delay (ps)	311	203	171	277	277	213	170	271	271
$T_{top}(^oC)$	58	48	63	92	92	46	80	100	100

Wireless Sensor Node

	Single Chip (SoC)	Two Chips (Logic+DRAM,Analog/RF+Sensor)				Four Chips (Logic,DRAM,Analog/RF,Sensor)			
		2D-SoP	3D-SiP	3D-W2W	3D-D2W	2D-SoP	3D-SiP	3D-W2W	3D-D2W
Norm. Area	1.00	2.82	0.89	1.14	1.14	4.91	1.59	1.15	1.15
$Yield_{overall}$	0.95	0.98	0.98	0.96	0.98	0.98	0.98	0.92	0.94
Norm. Cost	1.00	2.21	3.01	1.30	2.52	4.60	4.48	1.25	4.01
Delay (ps)	132	170	151	155	155	187	158	156	156
$T_{top}(^oC)$	65	46	70	125	125	41	81	125	125

Table 6.3: *Results of cost and performance analysis for Wireless Sensor Node. For 3D-W2W and 3D-D2W integration, thermal-vias have to be inserted in order to limit the temperature inside the topmost chip.*



(a) Mobile Terminal (b) Wireless Sensor

Figure 6.13: *3-D arrangement for mobile terminal and wireless sensor. Die sizes are normalized to the largest die in each stack, and are approximately to scale.*

3D-W2W is the lowest among all options. This is due to the lower test cost in comparison to 3D-D2W, and lower assembley cost in comparison to 2D-SoP and 3D-SiP. The low yield of the SoC solution means that it is relatively expensive when compared to all of the other implementations. A 3D-SiP implementation is slightly more expensive than a 2D-SoP implementation due to the higher assembly cost for a 3-D stack. Overall, the 2-chip arrangement is a clear winner due to the lower assembly cost and higher yield in integrating two rather than four chips by stacking.

Interestingly, it appears that an SoC solution is the best choice for the wireless sensor node (Table 6.3), because all other implementations show a lower performance, higher cost, and for the most part, a larger area. A 3D-SiP implementation leads to the most compact system, although costing the most. The reason for the comparatively large area in 3D-W2W and 3D-D2W implementations is that the relatively small individual blocks result in a higher power density, and require the addition of a high number of T-vias for thermal management. The total area increases as a consequence, and the cost and worst-case delay increase accordingly. In this case T-vias occupy about 66% of the total area in the four-chip stack, and 49% in the two-chip stack. The wireless sensor node system is also quite small in comparison to the mobile terminal, and hence has a comparatively higher yield in all implementation choices. For all these reasons, a SoC solution may be the best option for such low area applications.

A comparatively elevated temperature can be seen in the block which is closest to the substrate (T_{top}) for 3-D implementations. As mentioned, this is the result of the increased power density caused by the relatively small area available for dissipation as opposed to the SoC implementation. This is the reason for the higher temperature for example in the 4-chip arrangement as opposed to the 2-chip arrangement in 3D-SiP technology. It should be borne in mind that the accuracy of the 1-D heat model for the particular implementation should be verified, and be replaced with a more accurate model wherever necessary.

One result that might seem counter-intuitive is that 3D-WLI technologies result in a higher worst-case delay in some cases, in spite of the reduction in the average interconnect length. For example the delay in 3D-WLI technologies is significantly higher than that in a 3D-SiP implementation for both case studies, and even than in

a 2D-SoP implementation for the mobile terminal. The reason for the increased wire delay in 3D-WLI is due to the use of *package-intermediate-interconnects* [*157*, *158*] in 2D-SoP and 3D-SiP implementations. For global signal transmission, three types of interconnects can be identified in general. These are on-chip wires on a top metal layer, off-chip Printed Circuit Board (PCB) type traces, and TSVs. The off-chip traces and TSVs exhibit fast transmission-line-like behavior whereas even the relatively wide global level on-chip lines are much more resistive, and exhibit diffusive (i.e. RC) behavior. Additionally, taking a signal off-chip and bringing a signal on-chip entail chip-to-package parasitics that include the pad capacitance, and bond wire or ball-grid solder ball. Finally the layer-to-layer TSV connection includes a pad capacitance in the signal path.

Even taking into account the off-chip drivers and chip-to-package parasitics, off-chip wires are much faster than on-chip wires for transmitting a signal for the length of a die edge, for a relatively large die. This is because the fast off-chip traces more than make-up for the chip-to-package parasitics by outperforming the RC lines. In 2D-SoP and 3D-SiP, the opportunity exists to take advantage of this phenomenon by running wires off-chip and bypassing long chip-edge to chip-edge length RC lines. This is demonstrated in Fig. 6.10(c). The actual saving will of course depend on the specific layout, but in [*157*] for example, this technique of avoiding long on-chip wires by running them off-chip to realize Package-Intermediate Interconnects, is reported to yield a saving of up to 40%, even considering the chip-to-package parasitics.

The layout and die sizes are a critical factor in determining the relative speed in different implementation technologies. If the layout permits communicating blocks to be placed vertically close to each other for example, vertical integration does provide an excellent opportunity to substantially reduce the communication delay. For the specific cases considered in the manuscript, the quantitative results based on accurate parasitics show that signal transmission from the corner of one chip to the diagonally opposed corner of another (A to B in Fig. 6.10) is faster in the 2-D SoP and 3-D SiP type implementations due to the outperformance of the long on-chip wires.

Another contributing factor is the increased temperatures in the higher-level layers, which has an adverse effect on device and interconnect performance, although this is of less significance.

6.6 Summary

Interconnect scaling happened to be the major bottleneck in high performance IC design and the growing need for heterogeneous integration of technologies in a single die pushed the traditional IC integration into the third dimension. 3-D integration provides an attractive chip architecture that can alleviate interconnect related challenges existing in 2-D chips such as delay and power dissipation, as well as making heterogeneous integration possible. However, there are many challenges that need to be overcome such as thermal, reliability and electromagnetic coupling effects.

Many research groups both in academia and industry have devised different 3-D integration options using diverse processing steps - bare dies, packaged dies, MCM's

or custom wafers are stacked along the z-axis and form the vertical interconnect for layer to layer communication. However, in general there are three major methods for 3-D integration: W2W, D2W, and SiP. Each method has its own pros and cons.

Among the various choices, finding an optimal solution for system implementation usually deals with cost, performance, power, thermal and technological trade-off analyses at the system conceptual level. Based on the quantitative area and yield models presented in Chapter 5, new yield and cost models, and simple yet useful thermal models, as well as performance metrics for evaluating 3-D integration options namely W2W, D2W and SiP were derived. These models have been combined in a cohesive cost and performance trade-off analysis methodology which is suitable for early analysis and design space explorations of future nanoscale electronic systems. In order to validate the proposed metrics and methodology, two ubiquitous electronic systems are analyzed under various implementation schemes and the performance trade-offs discussed. This case study is used to highlight the importance of a a-priory cost and performance trade-off analysis early in the design flow.

7

Conclusions

7.1 Summary

Advances in silicon processing technology and system integration has fueled growth of integrated circuits with paradigms such as system-on-chip (SoC), System-in-Package (SiP), System-on-Package (SoP), and Three-dimensional (3-D) integration. These have led to unprecedented design challenges due to complexity in the overall system and achieving performance and cost targets. The primary focus of this thesis is the design, modelling of system interconnection and their effects in massively integrated 3-D ICs, and the contributions are three-fold: (1) electrical modelling of through-silicon vias; (2) signaling techniques for global on-chip interconnects; (3) cost,performance and technological trade-offs for 2-D and 3-D heterogeneous ICs.

In the TSV modelling parasitic parameter extraction is carried out using a field solver to explore trends in typical technologies to gain an insight into the variation of resistive, capacitive and inductive parasitics including coupling effects. A detailed methodology for the generation of compact closed-form equations for modelling resistive, inductive and capacitive parasitic parameters using dimensional analysis is outlined, starting from an isolated TSV and proceeding to a bundle. Compact models are useful in system-conceptual level explorations of 3-D ICs. Specifically, they can be used for prediction of parasitic parameters in the estimation of performance and signal integrity related metrics without a need for an expensive field solver. Simulations show that error in circuit level metrics is within a few percentage.

Global on-chip interconnects throttle the performance gained through technology scaling, and several solutions for this interconnect bottleneck have been proposed in the literature at various design hierarchy levels. 3-D integration can potentially shorten the otherwise long global interconnects connecting critical blocks by placing them in vertical proximity. The inclusion of TSVs in global interconnect links has been analyzed; firstly through the development of a suitable model, and subsequently, the analysis of quality of signal transmission. Further, a novel smart repeater circuit suitable for on-chip global interconnects has been proposed along with a detailed delay and energy analysis as well as a design methodology to obtain the optimal repeater configurations for minimising delay and minimising jitter. The proposed abstraction means the smart repeater is easier to amalgamate

in CAD flows at different levels of hierarchy from initial signal planning to detailed place and route when compared to alternatives proposed in the literature. Further, as processes scale, the selector latency shrinks, and higher data rates can be achieved. The total energy saving that can be achieved by the SMART driver in future nanometer technologies is found to be in the range of 20% - 25%.

Signal integrity assessment of TSV bundles and signal transmission in TSV interconnects has also been carried out in this thesis. For a considered range of physical geometries, extensive circuit simulations were carried out to estimate delay and it was found that TSV resistance and inductance is negligible compared to driver resistance. However, the effect of TSV capacitance is significant for the considered range. By extensive simulations it was also found that the crosstalk effect of inductance is negligible. The results show that a lumped capacitive equivalent circuit is sufficiently accurate, and a switch-factor based delay model appropriate.

Finally the issues around challenges, opportunities and trade-offs in the different system integration options available have been studied. For the study area, yield, cost, thermal, and performance models have been reviewed and collated with modifications where necessary. These models have been combined in a cohesive cost and performance trade-off analysis methodology which is suitable for early analysis and design space explorations of future nanoscale electronic systems. Using example contemporary mixed-signal systems the use of the proposed methodology and models in analysing the impact of different implementations has been demonstrated. The case studies show that the implementation strategy must be carefully selected depending on the circuit complexity and architecture, as otherwise the move to 3-D may have a detrimental effect. Design choice early in the design cycle will have a significant impact throughout the design and production lifecycles, and it is expected that the models and methodology presented in this thesis will be an useful aid in this choice.

7.2 Future Work

The 3-D integration paradigm has already incited research activity at virtually all levels of design hierarchy stemming from technology to CAD tools, to packaging. Topics earmarked for future work are briefly explained below.

Design of efficient vertical interconnections in 3-D ICs is still immature; independent vertical interconnections to each layer from the package as well as from other layers obviously consume area, increase congestion, and create additional challenges such as yield loss. The design of vertical communication links under the physical constraints imposed by the vertical interconnects needs to be carefully studied. For example the relatively large pitch and footprint of TSVs restrict their parallelism, suggesting some sort of multiplexing scheme when routing a horizontal bus to a different layer. This is especially so as TSVs outperform on-chip wires and can support a much higher signal speed.

Parasitic coupling among different layers (vertical crosstalk) in a 3-D IC, discussed in Chapter 6, is expected to be present. The high frequency signal transients in one layer may couple with the wires in neighbouring layers. Also, coupling of substrate noise into vertical interconnections and its signal integrity degradation needs to be experimentally analyzed. Reduced order models to analyze this phe-

nomena at the system-conceptual level is desirable, which would otherwise need expensive 3-D field solvers.

With device scaling the current requirement per unit area of a chip is increasing 43% per year, and total chip current is increasing at about 61% per year (refer Table 1.1). Therefore, in order not to increase power supply noise, the power supply network impedance should be controlled to match the increase in device density and more wire resources need to be allocated to deliver higher current. This situation is further aggravated in 3-D ICs because TSVs contribute additional resistance and inductance to the supply network and the number of pins for power delivery is fundamentally limited by the the footprint of the chip. Some research topics in this regard are: (1) the impact of IR and $L\frac{di}{dt}$ supply noise due to TSVs in 3-D chips, (2) impedance characteristics of power delivery system of 3-D chips, and (3) derivation of suitable design guidelines such as architectural block placement and allocation of decoupling capacitances to efficiently deliver power to 3-D chip stacks.

Because of higher power density and increased thermal resistance between the tiers due to isolation, thermal management is at the forefront among design issues of 3-D ICs. The widely accepted design guideline is to place the most power dissipating tier very close to the cooling device in the primary heat flow path. Therefore, in applications where an IC's performance is limited by a single hotspot in a logic block, a 3-D implementation can actually help, using other layers of the IC as part of the heat sink for the hotspot and enabling higher performance. It is of paramount important to estimate 3-D IC temperatures accurately including hotspots in each layer to determine thermal-via insertion and placement of critical blocks. Therefore, a system-level thermal model for a 3-D IC for physical design implementations is crucial.

References

[1] J. S. Kilby, "Turning potentials into realities: The invention of the integrated circuit," *Nobel Lecture*, 2000. [Online]. Available: http://nobelprize.org/nobel_prizes/physics/laureates/2000/kilby-lecture%.pdf

[2] The International Technology Roadmap for Semiconductors(ITRS), 2005. [Online]. Available: http://www.itrs.net

[3] J. Meindl, Q. Chen, and J. Davis, "Limits on Silicon Nanoelectronics for Terascale Integration," *Science*, vol. 293, no. 5537, pp. 2044–2049, 2001.

[4] G. Moore, "Cramming More Components on Integrated Circuits," *Electronics (38)*, vol. 8, pp. 114–117, 1965.

[5] R. Tummala, *Fundamentals of Microsystems Packaging*. McGraw-Hill, 2001.

[6] H. B. Backoglu, *Circuits, Interconnections and Packaging for VLSI*. Addison-Wesley, 1990.

[7] R. K. Ulrich and W. D. Brown, Eds., *Advanced Electronic Packaging*, 2nd ed., ser. IEEE Press Series on Microelectronic Systems. Wiley-Interscience, September 2005.

[8] R. Fillion, C. Woychik, T. Zhang, and D. Bitting, "Embedded chip build-up using fine line interconnect," in *Proceeding of the 57th Electronics Components and Technology Conference*, 2007.

[9] R. R. Tummala, "Sop: what is it and why? a new microsystem-integration technology paradigm-moore's law for system integration of miniaturized convergent systems of the next decade," *IEEE Transactions on Advanced Packaging*, vol. 27, no. 2, pp. 241–249, 2004.

[10] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, pp. 23–29, July-August 1999.

[11] W. J. Dally and J. W. Poulton, *Digital Systems Engineering*. Cambridge University Press, 1998.

[12] V. Zhirnov, I. Cavin, R.K., J. Hutchby, and G. Bourianoff, "Limits to binary logic switch scaling - a gedanken model," *Proceedings of the IEEE*, vol. 91, no. 11, pp. 1934–1939, November 2003.

[13] E. F. Rent, "Microminiature PackagingLogic Block to Pin Ratio," *Memoranda*, vol. 28, 1960.

[14] T. Sakurai, "Superconnect Technology," *IEICE Transactions on Electronics*, vol. 84, no. 12, pp. 1709–1716, 2001.

[15] R. R. Tummala, "Moore's law meets its match (system-on-package)," *IEEE Spectrum*, vol. 43, no. 6, pp. 44–49, June 2006.

[16] J. D. Meindl, "Interconnect opportunities for gigascale integration," *IEEE Micro*, vol. 23, no. 3, pp. 28–35, May-June 2003.

[17] M. Bohr, "Interconnect scaling-the real limiter to high performance ulsi," in *International Electron Devices Meeting*, December 1995, pp. 241–244.

[18] D. Edelstein, J. Heidenreich, R. Goldblatt, W. Cote, C. Uzoh, N. Lustig, P. Roper, T. McDevitt, W. Motsiff, A. Simon, *et al.*, "Full copper wiring in a sub-0.25 μm cmos ulsi technology," in *International Electron Devices Meeting, Technical Digest.*, 1997, pp. 773–776.

[19] H. Bakoglu and J. Meindl, "Optimal interconnection circuits for vlsi," *IEEE Transactions on Electron Devices*, vol. 32, no. 5, pp. 903–909, 1985.

[20] D. Pamunuwa, L.-R. Zheng, and H. Tenhunen, "Maximizing throughput over parallel wire structures in the deep submicrometer regime," *IEEE transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 2, pp. 224–243, April 2003.

[21] D. Liu and C. Svensson, "Power consumption estimation in cmos vlsi chips," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 6, pp. 663–670, 1994.

[22] T. Sakurai, "Perspectives on power-aware electronics," in *International Solid-State Circuits Conference, Digest of Technical Papers*, vol. 1, 2003, pp. 20–26.

[23] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *Proceedings of the 2004 International Workshop on System Level Interconnect Prediction*, 2004, pp. 7–13.

[24] "Vertical stacking to redefine chip design," Nikkei Electronics Asia, April 2007. [Online]. Available: http://techon.nikkeibp.co.jp

[25] L.-R. Zheng, "Design, analysis and integration of mixed-signal systems for signal and power integrity," Ph.D. dissertation, The Royal Institute of Technology (KTH), Stockholm, Sweden, 2001.

[26] D. Pamunuwa, "Modelling and analysis of interconnects for deep submicron systems-on-chip," Ph.D. dissertation, The Royal Institute of Technology(KTH), Stockholm, Sweden, 2003.

[27] M. Shen, "Concurrent chip and package design for radio and mixed-signal systems," Ph.D. dissertation, The Royal Institute of Technology (KTH), Stockholm, Sweden, November 2005.

[28] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, 2nd ed. Prentice Hall Ltd, 2003.

[29] B. P. Wong, A. Mittal, Y. Cao, and G. Starr, *Nano-CMOS Circuit and Physical Design.* John Wiley and Sons, Inc., 2005.

[30] J. Poulton, "Signaling in high-performance memory systems," in *IEEE International Solid-State Circuits Conference (Tutorial)*, February 1999, presented at the.

[31] L. Lavagno, L. Scheffer, and G. Martin, Eds., *EDA for IC Implementation, Circuit Design, and Process Technology.* Boca Raton, FL, USA: CRC Press, Taylor and Francis Group, 2006, vol. 2.

[32] "Ansoft quick 3-d." [Online]. Available: http://www.ansoft.com/products/si/q3d_extractor/

[33] M. Kamon, M. J. Tsuk, and J. K. White, "Fasthenry: a multipole-accelerated 3-d inductance extraction program," *IEEE Transactions on Microwave Theory and Techniques*, vol. 42, no. 9, pp. 1750–1758, September 1994.

[34] K. Nabors and J. White, "Fastcap: a multipole accelerated 3-d capacitance extraction program," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 10, no. 11, pp. 1447–1459, Nov 1991.

[35] H. Wheeler, "Formulas for the Skin Effect," *Proceedings of the IRE*, vol. 30, no. 9, pp. 412–424, 1942.
[36] A. Deutsch, "Electrical characteristics of interconnections for high-performance systems," *Proceedings of the IEEE*, vol. 86, no. 2, pp. 315–357, February 1998.
[37] C. Cheng, J. Lillis, S. Lin, and N. Chang, *Interconnect Analysis and Synthesis.* John Wiley New York, 2000.
[38] S.-Q. Wang, "Barriers against copper diffusion into silicon and drift through silicon dioxide," *Materials Research Society Bulletin*, vol. 19, pp. 30–40, August 1994.
[39] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-d ics: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.
[40] K. Fuchs, "The conductivity of thin metallic films according to the electron theory of metals," in *Proceedings of Cambridge Philosophical Society*, 1938, pp. 100–108.
[41] E. H. Sondheimer, "The mean free path of electrons in metals," *Advanced Physics*, vol. 1, pp. 1–42, 1952.
[42] F. Chen and D. Gardner, "Influence of line dimensions on the resistance of cu interconnections," *IEEE Electron Device Letters*, vol. 19, no. 12, pp. 508–510, December 1998.
[43] A. F. Mayadas and M. Shatzkes, "Electrical-resistivity model for polycrystalline films: the case of arbitrary reflection at external surfaces," *Physics Review B*, vol. 1, no. 4, pp. 1382–1389, 1970.
[44] S. X. Shi and D. Z. Pan, "Wire sizing with scattering effect for nanoscale interconnection," in *Asia and South Pacific Conference on Design Automation*, 2006, pp. 503–508.
[45] Y. Travaly, M. Bamal, L. Carbonell, F. Iacopi, M. Stucchi, M. Van Hove, and G. Beyer, "A novel approach to resistivity and interconnect modeling," *Microelectronic Engineering*, vol. 83, no. 11-12, pp. 2417–2421, 2006.
[46] J. E. Sergent and A. Krum, Eds., *Thermal Management Handbook for Electronic Assemblies.* The McGraw Hill Companies, 1998.
[47] C. K. Hu and J. M. E. Harper, "Copper interconnect: Fabrication and reliability," in *Proceedings of the VLSI Technology, Systems, and Applications*, June 1997.
[48] N. Lu, M. Angyal, G. Matusiewicz, V. McGahay, and T. Standaert, "Characterization, modeling and extraction of cu wire resistance for 65 nm technology," in *IEEE Custom Integrated Circuits Conference*, September 2007, pp. 57–60.
[49] A. V. Mezhiba and E. G. Friedman, *Power Distribution Networks in High Speed Integrated Circuits.* Kluwer Academic Publishers, 2003.
[50] E. B. Rosa, "The self and mutual inductance of linear conductors," *Bulletin of the National Bureau of Standards*, vol. 4, pp. 301–344, 1908.
[51] A. Ruehli, "Inductance calculations in a complex integrated circuit environment," *IBM Journal of Research and Development*, vol. 16, no. 5, pp. 470–481, September 1972.
[52] L. Dworsky, *Modern transmission line theory and applications.* New York:

Wiley-Interscience, 1979.

[53] M. W. Beattie and L. T. Pileggi, "On-chip induction modeling: basics and advanced methods," *IEEE Transactions on Very-Large-Scale Integration (VLSI) Systems*, vol. 10, no. 6, pp. 712–729, December 2002.

[54] J. Choudhury, G. S. Seetharaman, and G. H. Massiha, "Accurate modelling of thin-film inductance for nano-chip," in *the Third IEEE Conference on Nanotechnology*, 2003, pp. 351–355.

[55] B. Young, *Digital Signal Integrity: Modeling and Simulation with Interconnects and Packages.* Prentice Hall PTR Upper Saddle River, NJ, USA, 2000.

[56] W. T. Weeks, L. L. Wu, M. F. McAllister, and A. Singh, "Resistive and inductive skin effect in rectangular conductors," *IBM Journal of Research and Development*, vol. 23, no. 6, pp. 652–660, November 1979.

[57] S. Mei and Y. Ismail, "Modeling skin and proximity effects with reduced realizable RL circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 4, pp. 437–447, 2004.

[58] B. Krauter and S. Mehrotra, "Layout Based Frequency Dependent Inductance and Resistance Extraction for On-Chip Interconnect Timing Analysis," in *Proceedings of the 35th annual conference on Design automation*, 1998, pp. 303–308.

[59] B. Mukherjee, L. Wang, and L. Pacelli, "A practical approach to modeling skin effect in on-chip interconnects," in *ACM Great Lakes Symposium on VLSI*, 2004, pp. 266–270.

[60] Y. Cao, X. Huang, D. Sylvester, T. King, and C. Hu, "Impact of on-chip interconnect frequency-dependent R (f) L (f) on digital and RF design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 1, pp. 158–162, 2005.

[61] H. Palmer, "Capacitance of a parallel-plate capacitor by the Schwartz-Christoffel transformation," *Transactions on American Institute of Electrical Engineers*, vol. 56, pp. 363–366, 1937.

[62] C. P. Yuan and T. N. Trick, "A simple formula for the estimation of the capacitance of two-dimensional interconnects in vlsi circuits," *IEEE Electron Device Letters*, vol. 3, no. 12, pp. 391–393, December 1982.

[63] E. Barke, "Line-to-ground capacitance calculation for vlsi: a comparison," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 7, no. 2, pp. 295–298, February 1988.

[64] T. Sakurai and K. Tamaru, "Simple formulas for two- and three-dimensional capacitances," *IEEE Transactions on Electron Devices*, vol. 30, no. 2, pp. 183–185, Feb 1983.

[65] N. P. van der Meijs and J. T. Fokkema, "VLSI circuit reconstruction from mask topology," *INTEGRATION, the VLSI journal*, vol. 2, pp. 85–119, 1984.

[66] J. Chern, J. Huang, L. Arledge, P. Li, P. Yang, T. Inc, and T. Dallas, "Multilevel metal capacitance models for CAD design synthesissystems," *IEEE Electron Device Letters*, vol. 13, no. 1, pp. 32–34, 1992.

[67] M. Lee, "A multilevel parasitic interconnect capacitance modeling andextraction for reliable VLSI on-chip clock delay evaluation," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 4, pp. 657–661, 1998.

[68] S. C. Wong, G. Y. Lee, and D. J. Ma, "Modeling of interconnect capaci-

tance, delay, and crosstalk in VLSI," *IEEE Transactions on Semiconductor Manufacturing*, vol. 13, no. 1, pp. 108–111, 2000.

[69] L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Accurate a priori signal integrity estimation using a multilevel dynamic interconnect model for deep submicron vlsi design," in *Proceedings of the 26th European Solid-State Circuits Conference*, 2000, pp. 352–355.

[70] C. S. Walker, *Capacitance, Inductance and Crosstalk Analysis Norwood*. Artech House, May 1990.

[71] R. Achar and M. Nakhla, "Simulation of high-speed interconnects," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 693–728, 2001.

[72] R. Chang, "Near speed-of-light on-chip electrical interconnects," Ph.D. dissertation, Stanford University, November 2002.

[73] T. Sakurai, "Closed-form expressions for interconnection delay, coupling, and crosstalk in vlsi's," *IEEE Transactions on Electron Devices*, vol. 40, no. 1, pp. 114–124, April 1993.

[74] A. Deutsch, P. Coteus, G. Kopcsay, H. Smith, C. Surovic, B. Krauter, D. Edelstein, and P. Restle, "On-chip wiring design challenges for gigahertz operation," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 529–555, April 2001.

[75] T. Dhaene and D. de Zutter, "Selection of lumped element models for coupled lossy transmissionlines," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 11, no. 7, pp. 805–815, July 1992.

[76] A. Deutsch, G. V. Kopcsay, P. J. Restle, H. H. Smith, G. Katopis, W. D. Becker, P. W. Coteus, C. W. Surovic, B. J. Rubin, R. P. Dunne Jr, *et al.*, "When are transmission-line effects important for on-chip interconnections?" *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 10, pp. 1836–1846, 1997.

[77] Y. Ismail, E. Friedman, and J. Neves, "Figures of merit to characterize the importance of on-chip inductance," *IEEE Transactions on Very Large Scale Integration(VLSI) Systems*, vol. 7, no. 4, pp. 442–449, 1999.

[78] A. Tsuchiya, M. Hashimoto, and H. Onodera, "Representative frequency for interconnect R (f) L (f) C extraction," in *Proceedings of the Conference on Asia South Pacific Design Automation*, 2004, pp. 691–696.

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

[80] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, no. 4, pp. 352–366, 1990.

[81] A. Odabasioglu, M. Celik, M. Celik, and L. Pileggi, "PRIMA: passive reduced-order interconnect macromodeling algorithm," in *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers*, 1997, pp. 58–65.

[82] P. Feldmann and R. W. Freund, "Efficient linear circuit analysis by pade approximation via the lanczos process," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 5, pp. 639–649, 1995.

[83] M. Silveria, M. Kamon, and J. White, "Efficient reduced-order modeling of frequency-dependent coupling inductances associated with 3-d interconnect

structures," *IEEE Transactions on Components, Packaging and Manufacturing Technology Part B: Advanced Packaging*, vol. 19, pp. 283–288, 1996.

[84] M. Celik, N. Pileggi, and A. Odabasioglu, *IC interconnect analysis*. Kluwer Academic Publishers, 1996.

[85] F. Dartu, N. Menezes, J. Qian, and L. T. Pillage, "A gate-delay model for high-speed CMOS circuits," in *Proceedings of the 31st annual conference on Design automation*, 1994, pp. 576–580.

[86] F. Dartu, N. Menezes, and L. T. Pileggi, "Performance computation for precharacterized CMOS gates with RCloads," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, no. 5, pp. 544–553, May 1996.

[87] D. Sylvester and K. Keutzer, "System level performance modelling with bacpac - berkeley advanced chip performance calculator," in Workshop Notes, International Workshop on System Level Interconnect Prediction, 1999. [Online]. Available: http://www.eecs.umich.edu/~dennis/bacpac/

[88] G. Sai-Halasz, "Performance trends in high-end processors," *Proceedings of the IEEE*, vol. 83, no. 1, pp. 20–36, 1995.

[89] Y. Ismail and E. Friedman, "Effects of inductance on the propagation delay and repeaterinsertion in VLSI circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 2, pp. 195–206, 2000.

[90] S. Sapatnekar, *Timing*. Kluwer Academic Publishers, 2004.

[91] A. B. Kahng, S. Muddu, and E. Sarto, "On switch factor based analysis of coupled rc interconnects," in *Proceedings of the 37th Design Automation Conference*, 2000, pp. 79–84.

[92] M. Ghoneima and Y. Ismail, "Accurate decoupling of capacitively coupled buses," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2005, pp. 4146–4149.

[93] X. Huang, Y. Cao, D. Sylvester, S. Lin, T. King, and C. Hu, "Rlc signal integrity analysis of high-speed global interconnects," in *IEEE International Electron Devices Meeting, Technical Digest*, 2000, pp. 731–734.

[94] Y. Cao, X. Yang, X. Huang, and D. Sylvester, "Switch-factor based loop rlc modeling for efficient timing analysis," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 9, pp. 1072–1078, 2005.

[95] T. Sakurai and A. Newton, "Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, April 1990.

[96] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif, "Full chip leakage estimation considering power supply and temperature variations," in *Proceedings of the international symposium on Low Power Electronics and Design*, 2003, pp. 78–83.

[97] V. De and S. Borkar, "Technology and design challenges for low power and high performance [microprocessors]," in *Proceedings of International Symposium on Low Power Electronics and Design*, 1999, pp. 163–168.

[98] K. S. Khouri, N. K. Jha, M. Inc, and T. X. Austin, "Leakage power analysis and reduction during behavioral synthesis," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 10, no. 6, pp. 876–885, 2002.

[99] L. W. Schaper, S. L. Burkett, S. Spiesshoefer, G. V. Vangara, Z. Rahman, and S. Polamreddy, "Architectural implications and process development of

3-d vlsi z-axis interconnects using through silicon vias," *IEEE Transactions on Advanced Packaging*, vol. 28, no. 3, pp. 356– 366, 2005.

[100] P. R. Morrow, C. M. Park, S. Ramanathan, M. J. Kobrinsky, and M. Harmes, "Three-dimensional wafer stacking via Cu-Cu bonding integrated with 65-nm strained-Si/low-k CMOS technology," *IEEE Electron Device Letters*, vol. 27, no. 5, pp. 335–337, 2006.

[101] B. Swinnen, W. Ruythooren, P. De Moor, L. Bogaerts, L. Carbonell, K. De Munck, B. Eyckens, S. Stoukatch, D. Tezcan, Z. Tokei, *et al.*, "3D integration by Cu-Cu thermo-compression bonding of extremely thinned bulk-Si die containing 10 μm pitch through-Si vias," in *International Electron Devices Meeting*, 2006, pp. 1–4.

[102] D. M. Jang, C. Ryu, K. Y. Lee, B. H. Cho, J. Kim, T. S. Oh, W. J. Lee, and J. Yu, "Development and Evaluation of 3-D SiP with Vertically Interconnected Through Silicon Vias (TSV)," in *Proceedings of the 57th Electronic Components and Technology Conference*, 2007, pp. 847–852.

[103] S. Alam, R. Jones, S. Rauf, and R. Chatterjee, "Inter-strata connection characteristics and signal transmission in three-dimensional (3d) integration technology," *8th International Symposium on Quality Electronic Design (ISQED)*, pp. 580–585, March 2007.

[104] T. E. Lawrence, S. M. Donovan, W. B. Knowlton, J. Rush-Byers, and A. J. Moll, "Electrical characterization of through-wafer interconnects," in *IEEE Workshop on Microelectronics and Electron Devices*, 2004, pp. 99–102.

[105] L. L. W. Leung and K. J. Chen, "Microwave Characterization and Modeling of High Aspect Ratio Through-Wafer Interconnect Vias in Silicon Substrates," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 8, pp. 2472–2480, 2005.

[106] D. Khalil, Y. Ismail, M. Khellah, T. Karnik, and V. De, "Analytical model for the propagation delay of through silicon vias," in *Proceeding of the 9th International Symposium on Quality Electronic Design*, March 2008, pp. 553–556.

[107] I. Savidis and E. G. Friedman, "Electrical modeling and characterization of 3-d vias," in *IEEE International Symposium on Circuits and Systems*, May 2008, pp. 784–787.

[108] J. H. Wu, J. Scholvin, and J. A. del Alamo, "A through-wafer interconnect in silicon for RFICs," *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1765–1771, 2004.

[109] G. Box and N. Draper, *Empirical model-building and response surfaces.* John Wiley and Sons Inc., 1991.

[110] H. L. Langhaar, *Dimensional Analysis and Theory of Models.* John Wiley and Sons Inc., 1951.

[111] E. d. S. Q. Isaacson and M. d. S. Q. Isaacson, *Dimensional Methods in Engineering and Physics*, 1st ed. Edward Arnold (Publishers) Ltd, 1975.

[112] R. Weerasekera, D. Pamunuwa, L. Zheng, and H. Tenhunen, "Minimal-Power, Delay-Balanced Smart Repeaters for Global Interconnects in the Nanometer Regime," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 16, no. 5, pp. 589–593, May 2008.

[113] J. Zhang and E. G. Friedman, "Effect of Shield Insertion on Reducing Crosstalk Noise between Coupled Interconnects," *IEEE International Sym-*

posium on Circuits and Systems (ISCAS 2004), pp. 529–532, 2004.

[114] R. Escovar and R. Suaya, "Optimal Design of Clock Trees for Multigigahertz Applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, no. 3, pp. 329–345, 2004.

[115] H. Kaul, D. Sylvester, and D. Blaauw, "Active shields: a new approach to shielding global wires," in *Proceedings of the 12th ACM Great Lakes Symposium on VLSI*, 2002, pp. 112–117.

[116] T. Gao and C. Liu, "Minimum crosstalk channel routing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 15, no. 5, pp. 465–474, 1996.

[117] K. M. Lepak, M. Xu, J. Chen, and L. He, "Simultaneous shield insertion and net ordering for capacitive and inductive coupling minimization," *ACM Transactions on Design Automation of Electronic Systems*, vol. 9, no. 3, pp. 290–309, 2004.

[118] P. Gupta and A. Kahng, "Wire swizzling to reduce delay uncertainty due to capacitive coupling," in *Proceedings of the 17th International Conference on VLSI Design*, 2004, pp. 431–436.

[119] B. Soudan, "Controlling on-chip inductive coupling of signal busses through swizzling," in *the 14th International Conference on Microelectronics*, 2002, pp. 181–184.

[120] M. L. Mui, K. Banerjee, and A. Mehrotra, "A global interconnect optimization scheme for nanometer scale VLSI with implications for latency, bandwidth, and power dissipation," *IEEE Transactions on Electron Devices*, vol. 51, no. 2, pp. 195–203, 2004.

[121] X. C. Li, J. F. Mao, H. F. Huang, and Y. Liu, "Global interconnect width and spacing optimization for latency, bandwidth and power dissipation," *IEEE Transactions on Electron Devices*, vol. 52, no. 10, pp. 2272–2279, 2005.

[122] M. El-Moursy and E. Friedman, "Optimizing inductive interconnect for low power," *Canadian Journal of Electrical and Computer Engineering*, vol. 27, no. 4, pp. 183–188, April 2002.

[123] J. P. Fishburn and C. A. Schevon, "Shaping a distributed-RC line to minimize Elmore delay," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 42, no. 12, pp. 1020–1022, 1995.

[124] M. El-Moursy and E. Friedman, "Optimum wire tapering for minimum power dissipation in RLC interconnects," in *Proceedings of IEEE International Symposium on Circuits and Systems*, 2006.

[125] R. Venkatesan, J. A. Davis, and J. D. Meindl, "Compact distributed rlc interconnect models - part iv: unified models for time delay, crosstalk, and repeater insertion," *IEEE Transactions on Electron Devices*, vol. 50, no. 4, pp. 1094–1102, 2003.

[126] A. B. Kahng, S. Muddu, E. Sarto, and R. Sharma, "Interconnect tuning strategies for high-performance ics," in *Proceedings of the conference on Design, Automation and Test in Europe*, 1998, pp. 471–478.

[127] D. Audet, Y. Savaria, and N. Arel, "Pipelining communications in large vlsi/ulsi systems," *IEEE Transactions on Very-Large-Scale Integration (VLSI) Systems*, vol. 2, no. 1, pp. 1–10, March 1994.

[128] D. Audet and Y. Savaria, "An architectural approach for increasing clock frequency and communication speed in monolithic wsi systems," *IEEE Trans-*

actions Component and Packaging Manufacturing Technology, vol. 17, no. 3, pp. 362–368, August 1994.

[129] M. Mizuno, W. J. Dally, and H. Onishi, "Elastic interconnects: repeater-inserted long wiring capable of compressing and decompressing data," in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, 2001, pp. 346–347.

[130] M. Nekili and Y. Savaria, "Parallel regeneration of interconnections in vlsi and ulsi circuits," in *IEEE Symposium of Circuits and Systems*, 1993, pp. 2023–2026.

[131] I. Dobbelaere, H. Horowitz, and A. El-Gamal, "Regenerative feedback repeaters for programmable interconnection," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 11, pp. 1246–1253, November 1995.

[132] H. Zhang, V. George, and J. M. Rabaey, "Low-swing on-chip signaling techniques: Effectiveness and robustness," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 3, pp. 264–272, June 2000.

[133] Y. Massoud, J. Kawa, D. MacMillen, and J. White, "Modeling and analysis of differential signaling for minimizing inductive crosstalk," in *Proceedings of the 38th Design automation Conference*, 2001, pp. 804–809.

[134] A. Maheshwari and W. Burleson, "Current sensing techniques for global interconnects in very deep submicron (vdsm) cmos," in *IEEE Computer Society Workshop in VLSI*, 2001, pp. 66–70.

[135] R. Bashirullah, W. Liu, R. Cavin, and D. Edwards, "A 16gb/s adapative bandwidth on-chip bus based on hybrid current/voltage mode signaling," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 2, pp. 461–473, February 2006.

[136] T. Iima, M. Mizuno, T. Horiuchi, and M. Yamashina, "Capacitance coupling immune, transient sensitive accelerator for resistive interconnect signals of subquarter micron ulsi," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 4, pp. 531–536, April 1996.

[137] P. Sotiriadis, T. Konstantakopoulos, and A. Chandrakasan, "Analysis and implementation of charge recycling for deep sub-micron buses," in *Proceedings of the 2001 International Symposium on Low Power Electronics and Design*, 2001, pp. 364–369.

[138] A. Nalamalpu, S. Sirinivasan, and W. P. Burleson, "Boosters for driving long on chip interconnects-design issues, interconnect synthesis, and comparison with repeaters," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 1, pp. 50–62, January 2002.

[139] H. Kaul and D. Sylvester, "Low-power on-chip communication based on transition-aware global signaling(tags)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 5, pp. 464–476, May 2004.

[140] A. Katoch, S. Jain, and M. Meijer, "Aggressor aware repeater circuits for improving on-chip bus performance and robustness," in *Proceedings of the 29th European Solid-State Circuits Conference*, September 2003, pp. 261–264.

[141] H.-Y. Huang and S.-L. Chen, "Interconnect accelerating techniques for sub-100-nm gigascale systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 11, pp. 1192–1200, November 2004.

[142] K. Hirose and H. Yasuura, "A bus delay reduction technique considering crosstalk," in *Proceedings of the conference on Design, Automation and Test*

in Europe, 2000, pp. 441–445.

[143] K. Nose and T. Sakurai, “Two schemes to reduce interconnect delay in bi-directional and uni-directional buses,” in *Symposium on VLSI Circuits, Digest of Technical Papers*, 2001, pp. 193–194.

[144] M. Ghoneima and Y. I. Ismail, “Utilizing the effect of relative delay on energy dissipation in low-power on-chip buses,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 12, pp. 1348–1359, December 2004.

[145] A. K. Nieuwland, A. Katoch, and M. Meijer, “Reducing cross-talk induced power consumption and delay,” in *Lecture Notes in Computer Science, PATMOS 2004 Proceedings*, vol. LNCS 3254. Springer Berlin / Heidelberg, September 2004, pp. 179–188.

[146] J. Cong, “An interconnect-centric design flow for nanometer technologies,” *Proceedings of the IEEE*, vol. 89, no. 4, pp. 505–528, 2001.

[147] J. Nurmi, H. Tenhunen, J. Isoaho, and A. Jantsch, Eds., *Interconnect-Centric Design for Advanced SoC and NoC*. Kluwer Academic Publishers, 2004.

[148] I. Hatirnaz, S. Badel, N. Pazos, Y. Leblebici, S. Murali, D. Atienza, and G. De-Micheli, “Early wire characterization for predictable network-on-chip global interconnects,” in *Proceedings of the International Workshop on System Level Interconnect Prediction*. ACM Press New York, NY, USA, 2007, pp. 57–64.

[149] M. R. Stan and W. P. Burleson, “Bus-invert coding for low-power i/o,” *IEEE Transactions on Very Large Scale Integration Systems*, vol. 3, no. 1, pp. 49–58, 1995.

[150] Y. Shin, S.-I. Chae, and K. Choi, “Partial bus-invert coding for power optimization of system level bus,” in *Proceedings of the International Symposium on Low Power Electronics and Design*, 1998, pp. 127–129.

[151] P. P. Sotiriadis and A. Chandrakasan, “Bus energy minimization by transition pattern coding (tpc) in deep sub-micron technologies,” in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 2000, pp. 322–328.

[152] K.-W. Kim, K.-H. Baek, N. Shanbhag, C. L. Liu, and S.-M. Kang, “Coupling-driven signal encoding scheme for low-power interface-design,” in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, November 2000, pp. 318–321.

[153] Y. Shin, S.-I. Chae, and K. Choi, “Partial bus-invert coding for power optimization of application-specific systems,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 2, pp. 377–383, April 2001.

[154] M. Anders, N. Rai, R. Krishnamurthy, and S. Borkar, “A transition-encoded dynamic bus technique for high-performance interconnects,” *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 709–714, 2003.

[155] K. N. Patel and I. L. Markov, “Error-correction and crosstalk avoidance in dsm busses,” in *Proceedings of the International Workshop on System-Level Interconnect Prediction*, 2003, pp. 09–014.

[156] C. Kretzschmer, *é* K. Nieuwland, and D. Mullër, “”why transition coding for power minimization of on-chip buses does not work”,” in *Proceedings of the Conference on Design, Automation and Test in Europe*, February 2004, pp. 512–517.

[157] S. Afonso, L. Schaper, J. Parkerson, W. Brown, S. Ang, and H. Naseem, "Modeling and electrical analysis of seamless high off-chip connectivity (shocc) interconnects," *IEEE Transactions on Advanced Packaging*, vol. 22, no. 3, pp. 309–320, 1999.

[158] P. Mehrotra, V. Rao, T. M. Conte, and P. D. Franzon, "Optimal chip-package codesign for high-performance dsp," *IEEE Transactions on Advanced Packaging*, vol. 28, pp. 288–297, 2005.

[159] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Switching sensitive interconnect driver to combat dynamic delay in on-chip buses," in *PATMOS 2005 Proceedings, Lecture Notes in Computer Science*, vol. LNCS 3728, September 2005, pp. 277–285.

[160] D. Pamunuwa, S. Elassaad, and H. Tenhunen, "Modeling delay and noise in arbitrarily coupled rc trees," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 11, pp. 1725–1739, November 2005.

[161] P. Gupta and A. B. Kahng, "Quantifying error in dynamic power estimation of cmos circuits," in *Proceedings of the 4th International Symposium on Quality Electronic Design*. Washington, DC, USA: IEEE Computer Society, 2003, p. 273.

[162] The International Technology Roadmap for Semiconductors(ITRS), 2005. [Online]. Available: http://www.itrs.net

[163] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration," in *Proceedings of the 7th International Symposium on Quality Electronic Design*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 585–590.

[164] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P. Franzon, "Demystifying 3d ics: the pros and cons of going vertical," *Design & Test of Computers, IEEE*, vol. 22, no. 6, pp. 498–510, Nov.-Dec. 2005.

[165] R. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214–1224, 2006.

[166] T. S. Cale, J.-Q. Lu, and R. J. Gutmann, "Three-Dimensional Integration in Microelectronics: Motivation, Processing, and Thermomechanical Modeling," *Chemical Engineering Communications*, vol. 195, no. 8, pp. 847–888, 2008.

[167] K. Snoeckx, E. Beyne, and B. Swinnen, "Copper-nail TSV technology for 3D-stacked IC integration," *Solid State Technology*, vol. 50, no. 5, p. 53, 2007.

[168] R. J. Drost, R. D. Hopkins, R. Ho, and I. E. Sutherland, "Proximity communication," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1529–1535, September 2004.

[169] D. Mizoguchi, N. Miura, T. Sakurai, and T. Kuroda, "A 1.2 gbps non-contact 3d-stacked inter-chip data communications technology." *IEICE Transactions*, vol. 89-C, no. 3, pp. 320–326, 2006.

[170] A. Fazzi, L. Magagni, M. Mirandola, R. Canegallo, S. Schmitz, and R. Guerrieri, "A 0.14mw/gbps high-density capacitive interface for 3d system integration," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, 2005, pp. 101–104.

[171] A. Iwata, M. Sasaki, T. Kikkawa, S. Kameda, H. Ando, K. Kimoto, D. Arizono, and H. Sunami, "A 3d integration scheme utilizing wireless interconnec-

tions for implementing hyper brains," in *the IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, 2005, pp. 262–597 Vol. 1.

[172] M. F. Chang, V. P. Roychowdhury, L. Zhang, H. Shin, and Y. Qian, "RF/Wireless Interconnect for Inter-and Intra-Chip Communications," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 456–466, 2001.

[173] M. Grange, R. Weerasekera, D. Pamunuwa, and H. Tenhunen, "Exploration of through silicon via interconnect parasitics for 3-dimensional integrated circuits," in *IEEE International Symposium on Circuits and Systems*, 2009, submitted.

[174] M. Scheffler, D. Ammann, A. Thiel, C. Habiger, and G. Troster, "Modeling and optimizing the costs of electronic systems," *IEEE Design & Test of Computers*, vol. 15, no. 3, pp. 20–26, July-September 1998.

[175] C. F. Murphy, M. S. Abadir, and P. A. Sandborn, "Economic analysis of test process flows for multichip modules using known good die," *Journal of Electronic Testing: Theory and Applications, Kluwer Academic Publishers*, vol. 10, pp. 151–166, 1997.

[176] B. S. Landman and R. L. Russo, "On a pin versus block relationship for partitions of logic graphs," *IEEE Transactions on Computers*, vol. 20, no. 12, pp. 1469–1479, 1971.

[177] T. Chiba, "Impact of the LSI on High-Speed Computer Packaging," *IEEE Transactions on Computers*, vol. 100, no. 27, pp. 319–325, 1978.

[178] P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "Prediction of net-length distribution for global interconnects in a heterogeneous system-on-a-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 6, pp. 649–659, 2000.

[179] W. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Transactions on Circuits and Systems*, vol. 26, no. 4, pp. 272–277, 1979.

[180] A. E. Gamal, "Two-dimensional stochastic model for interconnections in master slice integrated circuits," *IEEE Transactions on Circuits and Systems*, vol. 28, no. 2, pp. 127–138, 1981.

[181] M. Feuer, "Connectivity of Random Logic." *IEEE Transactions of Computers*, vol. 31, no. 1, pp. 29–33, 1982.

[182] W. R. Heller, C. G. Hsi, and W. F. Mikhaill, "Wirability-designing wiring space for chips and chip packages," *IEEE Design and Test Magazine*, vol. 1, no. 3, pp. 43–51, 1984.

[183] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI). I. Derivation and validation," *IEEE Transactions on Electron Devices*, vol. 45, no. 3, pp. 580–589, 1998.

[184] B. Geuskens and K. Rose, *Modeling Microprocessor Performance.* Kluwer Academic Publishers, 1998.

[185] V. Garg, D. Stogner, C. Ulmer, D. Schimmel, C. Dislis, S. Yalamanchili, and D. Wills, "Early analysis of cost/performance trade-offs in mcm systems," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, vol. 20, no. 3, pp. 308–319, 1997.

[186] S. Takahashi, M. Edahiro, and Y. Hayashi, "Interconnect design strategy: structures, repeaters and materials with strategic system performance analysis (S^2PAL) model," *IEEE Transactions on Electron Devices*, vol. 48, no. 2,

pp. 239–251, 2001.
[187] R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal n-tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 6, pp. 899–912, 2001.
[188] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. Meindl, "A compact physical via blockage model," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, no. 6, pp. 689–692, 2000.
[189] The International Technology Roadmap for Semiconductors(ITRS), 1999. [Online]. Available: http://www.itrs.net
[190] M. Shen, L.-R. Zheng, and H. Tenhunen, "Cost and performance analysis for mixed-signal system implementation: system-on-chip or system-on-package?" *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 25, no. 4, pp. 262–272, 2002.
[191] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2007)*, 2007, pp. 212–219.
[192] P. Sandborn, M. Abadir, and C. Murphy, "The tradeoff between peripheral and area array bonding of components in multichip modules," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, vol. 17, no. 2, pp. 249–256, 1994.
[193] J. de Gyvez and D. Pradhan, Eds., *Integrated Circuit Manufacturability: The Art of Process and Design Integration*. IEEE Press, 1999.
[194] M. Sydow, "Compare logic-array to asic-chip cost per good die," *Chip Design Magazine*, February/March 2006.
[195] J. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, no. 2, pp. 60–71, May 1990.
[196] B. T. Murphy, "Cost-size optima of monolithic integrated circuits," *Proceedings of the IEEE*, vol. 52, no. 12, pp. 1537–1545, Dec. 1964.
[197] R. B. Seeds, "Yield and cost analysis of bipolar lsi," *IEEE Transactions on Electron Devices*, vol. 15, no. 6, pp. 409–409, Jun 1968.
[198] A. Dingwall, "High-yield-processed bipolar lsi arrays," *IEEE Transactions on Electron Devices*, vol. 16, no. 2, pp. 246–247, February 1969.
[199] G. E. Moore, "What level of LSI is best for you?" *Electronics*, vol. 43, pp. 126–130, 1970.
[200] J. E. Price, "A new look at yield of integrated circuits," *Proceedings of the IEEE*, vol. 58, no. 8, pp. 1290–1291, 1970.
[201] C. H. Stapper, "Defect density distribution for LSI yield calculations," *IEEE Transactions on Electron Devices*, vol. 20, no. 7, pp. 655–657, 1973.
[202] T. Okabe, M. Nagata, and S. Shimada, "Analysis of yield of integrated circuits and a new expression for the yield," *Electrical Engineering in Japan*, vol. 92, no. 12, pp. 135–141, 1972.
[203] A. George, J. Krusius, and R. Granitz, "Packaging alternatives to large silicon chips: tiled silicon on mcm and pwb substrates," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging*, vol. 19, no. 4, pp. 699–708, 1996.

[204] T. Williams and N. Brown, “Defect Level as a Function of Fault Coverage,” *IEEE Transactions on Computers*, vol. 100, no. 30, pp. 987–988, 1981.

[205] V. D. Agrawal, S. C. Seth, and P. Agrawal, “Fault coverage requirement in production testing of LSI circuits,” *IEEE Journal of Solid-State Circuits*, vol. 17, no. 1, pp. 57–61, 1982.

[206] Y. Deng and W. P. Maly, “2.5-dimensional vlsi system integration,” *IEEE Transactions on very large scale integration (VLSI) systems*, vol. 13, no. 6, pp. 668–677, June 2005.

[207] D. Ragan, P. Sandborn, and P. Stoaks, “A detailed cost model for concurrent use with hardware/software co-design,” in *Proceedings of the 39th IEEE/ACM Design Automation Conference*, 2002, pp. 269–274.

[208] R. Hannemann, “Physical Technology for VLSI Systems,” in *Proceedings of the International Conference on Computer Design*. Order from IEEE Computer Society, 1986, pp. 48–53.

[209] L. L. Moresco, “Electronic system packaging: the search for manufacturing theoptimum in a sea of constraints,” *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, vol. 13, no. 3, pp. 494–508, 1990.

[210] P. A. Sandborn and H. Moreno, *Conceptual Design of Multichip Modules and Systems*. Kluwer Academic Publishers, 1994.

[211] “Known good die testing,” NASA Electronic Parts and Packaging Program (NEPP), (Accessed: 2007, November 12). [Online]. Available: http://nepp.nasa.gov/index_nasa.cfm/1058/

[212] F. Catthoor, N. D. Dutt, and C. E. Kozyrakis, “How to solve the current memory access and data transfer bottlenecks: at the processor architecture or at the compiler level,” in *Proceedings of the Conference on Design, automation and Test in Europe*, 2000, pp. 426–435.

[213] E. Beyne, “3d interconnection and packaging: impending reality or still a dream?” in *IEEE International Solid-State Circuits Conference, Digest of Technical Papers*, 2004, pp. 138–139 Vol.1.

[214] C. C. Liu, J.-H. Chen, R. Manohar, and S. Tiwari, “Mapping system-on-chip designs from 2-d to 3-d ics,” in *IEEE International Symposium on Circuits and Systems*, 2005, pp. 2939–2942.

[215] P. Mercier, S. R. Singh, K. Iniewski, B. Moore, and P. O’Shea, “Yield and cost modeling for 3d chip stack technologies,” in *IEEE Custom Integrated Circuits Conference*, September 2006, pp. 357–360.

[216] Y. Akasaka, “Three-dimensional IC trends,” *Proceedings of the IEEE*, vol. 74, no. 12, pp. 1703–1714, 1986.

[217] M. W. Geis, D. C. Flanders, D. A. Antoniadis, and H. I. Smith, “Crystalline silicon on insulators by graphoepitaxy,” in *International Electron Devices Meeting*, vol. 25, 1979.

[218] R. Mukai, N. Sasaki, T. Iwai, S. Kawamura, and M. Nakano, “Indirect laser annealing of PolySilicon for three-dimensional IC’s,” in *International Electron Devices Meeting*, vol. 29, 1983.

[219] T. Kunio, K. Oyama, Y. Hayashi, and M. Morimoto, “Three dimensional ICs, having four stacked active device layers,” in *International Electron Devices Meeting, Technical Digest*, 1989, pp. 837–840.

[220] J. W. Joyner, P. Zarkesh-Ha, and J. Meindl, “A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3d-soc),” in *14th An-*

nual IEEE International ASIC/SoC conference, September 2001.

[221] M. Bamal, S. List, M. Stucchi, A. Verhulst, M. Van Hove, R. Cartuyvels, G. Beyer, and K. Maex, “Performance comparison of interconnect technology and architecture options for deep submicron technology nodes,” in *International Interconnect Technology Conference,*, 2006, pp. 202–204.

[222] R. Zhang, K. Roy, C.-K. Koh, and D. B. Janes, “Stochastic interconnect modeling, power trends, and performance characterization of 3-d circuits,” *IEEE Transactions on Electron Devices*, vol. 48, no. 4, pp. 638–652, April 2001.

[223] S. Kim, C. Liu, L. Xue, and S. Tiwari, “Crosstalk reduction in mixed-signal 3-d integrated circuits with interdevice layer ground planes,” *IEEE Transactions on Electron Devices*, vol. 52, no. 7, pp. 1459–1467, 2005.

[224] J. Tao, N. W. Cheung, C. Hu, H.-K. Kang, and S. S. Wong, “Electromigration performance of electroless plated copper/pd-silicide metalization,” *IEEE Eletron Device Letters*, vol. 13, no. 8, pp. 433–435, August 1992.

[225] B. Goplen and S. S. Sapatnekar, “Placement of thermal vias in 3-d ics using various thermal objectives,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 4, pp. 692–709, 2006.

[226] Z. Li, X. Hong, Q. Zhou, S. Zeng, J. Bian, W. Yu, H. H. Yang, V. Pitchumani, and C.-K. Cheng, “Efficient thermal via planning approach and its application in 3-d floorplanning,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,*, vol. 26, no. 4, pp. 645–658, 2007.

[227] J. M. Koo, S. Im, L. Jiang, and K. E. Goodson, “Integrated microchannel cooling for three-dimensional circuit architectures,” *ASME Journal of Heat Transfer*, vol. 127, pp. 49–58, 2005.

[228] B. Dang, M. Bakir, and J. Meindl, “Integrated thermal-fluidic i/o interconnects for an on-chip microchannel heat sink,” *IEEE Electron Device Letters*, vol. 27, no. 2, pp. 117–119, February 2006.

[229] G. Loi, B. Agrawal, N. Srivastava, S. Lin, T. Sherwood, and K. Banerjee, “A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy,” in *Proceedings of the 43rd Design Automation Conference*. ACM New York, NY, USA, 2006, pp. 991–996.

[230] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann, and H. Reichl, “Forced convective interlayer cooling in vertically integrated packages,” in *11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2008, pp. 1114–1125.

[231] S. A. Kuhn, M. B. Kleiner, P. Ramm, and W. Weber, “Interconnect capacitances, crosstalk, and signal delay invertically integrated circuits,” in *International Electron Devices Meeting*, 1995, pp. 249–252.

[232] M. Rousseau, O. Rozeau, G. Cibrario, G. Le Carval, M. Jaud, P. Leduc, A. Farcy, and A. Marty, “Through-silicon via based 3D IC technology: Electrostatic simulations for design methodology,” in *IMAPS Device Packaging Conference (2008)*, 2008.

[233] S. Im and K. Banerjee, “Full chip thermal analysis of planar (2-d) and vertically integrated (3-d) high performance ics,” in *International Electron Devices Meeting, Technical Digest*, 2000, pp. 727–730.

[234] E. Beyne, “The rise of the 3rd dimension for system intergration,” in *International Interconnect Technology Conference*, 2006, pp. 1–5.

[235] A. W. Topol, D. C. L. Tulipe, L. S. Jr., D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, no. 4/5, pp. 491–506, 2006.

[236] M. Dreiza, A. Yoshida, J. Micksch, and L. Smith, "Stacked package-on-package design guidelines," *ChipScale Review, Amkor Technology Inc.*, no. 7, July 2005.

[237] T. Fukushima, Y. Yamada, H. Kikuchi, and M. Koyanagi, "New three-dimensional integration technology using chip-to-wafer bonding to achieve ultimate super-chip integration," *Japanese Journal of Applied Physics*, vol. 45, no. 4, pp. 3030–3035, 2006.

[238] S. A. Kuhn, M. B. Kleiner, and W. Weber, "Performance modeling of the interconnect structure of a 3-dimensionally integrated risc-processor/cache-system," in *Proceedings of the 45th Electronic Components and Technology Conference*, 1995, pp. 592–599.

[239] J. C. Ku and Y. Ismail, "Thermal-aware methodology for repeater insertion in low-power VLSI circuits," in *Proceedings of the International Symposium on Low Power Electronics and Design*, 2007, pp. 86–91.

[240] Micron CMOS Image Sensor Part Catalog, March 2007. [Online]. Available: http://www.micron.com

[241] R. Waiser, Ed., *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices.* Wiley-VCH, September 2004.

[242] Micron 128MB SDRAM Part Catalog, 2007. [Online]. Available: http://www.micron.com

[243] ARM Cortex-A8 Processor Product Brief, March 2007. [Online]. Available: http://www.arm.com

[244] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, and W. Davis, "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits," in *Proceedings of the 43rd ACM/IEEE Design Automation Conference*, 2006, pp. 997–1002.