System-Level Performance Evaluation of Three-Dimensional Integrated Circuits

by

Arifur Rahman

B.S., Polytechnic University (1994)
S.M., Massachusetts Institute of Technology (1996)

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 2001

Author ....
Department of Electrical Engineering and Computer Science
January 12, 2001

Certified by...
Rafael Reif
Associate Department Head and Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Professor of Electrical Engineering, Graduate Officer

System-Level Performance Evaluation of Three-Dimensional Integrated Circuits

by

Arifur Rahman

Submitted to the Department of Electrical Engineering and Computer Science
on January 12, 2001, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

As the critical dimensions in VLSI design continue to shrink, system performance of integrated circuits (ICs) will be increasingly dominated by interconnect delay [1]. For the technology generations approaching 50 nm and beyond, innovative system architectures and interconnect technologies will be required to meet the projected system performance [2]. Interconnect material solutions such as copper and low-k inter-level dielectric (ILD) offer only a limited improvement in system performance. Significant and scalable solutions to the interconnect delay problem will require fundamental changes in system design, architecture, and fabrication technologies.

Three-dimensional (3-D) ICs can alleviate interconnect delay problems by offering flexibility in system design, placement and routing. They (3-D ICs) can be formed by vertical integration of multiple device layers using wafer bonding, recrystallization, or selective epitaxial growth. The flexibility to place devices along the vertical dimension allows higher device density and smaller form factor in 3-D ICs. The critical signal path that may limit system performance can also be shortened to achieve faster clock speed. By 3-D integration, device layers fabricated with different front-end process technologies can be stacked along the 3rd dimension to form systems-on-a-chip [3]. In this thesis work, opportunities and challenges for 3-D integration of logic networks, microprocessors, and programmable logic have been explored based on system-level modeling and analysis. A stochastic wire-length distribution model has been derived to predict interconnection complexity in 3-D ICs. As more device layers are integrated, the 3-D wire-length distribution becomes narrower compared to that of 2-D ICs, resulting in a significant reduction in the number and length of semi-global and global wires. In 3-D ICs with 2-4 device layers, 30% – 50% reduction in wire-length can be achieved. Besides performance modeling, thermal analysis has also been performed to assess power dissipation and heat removal issues in 3-D ICs. The total capacitance associated with signal interconnects and clock networks can be reduced by 3-D integration, leading to lower power dissipation for system performance comparable to that of 2-D ICs. However, for higher system performance in 3-D ICs, power dissipation increases significantly, and it is likely that innovative cooling techniques will be needed for reliable operation of devices and interconnects.

Thesis Supervisor: Rafael Reif
Title: Associate Department Head and Professor of Electrical Engineering and Computer Science

Acknowledgments

I would like to thank my advisor, Prof. Rafael Reif, for providing support for this research project and creating a lively and stimulating atmosphere. His mentorship has been an inspiration to me. The brainstorming sessions with Dr. James Chung during the early stages of this project were integral to defining my thesis. I would also like to thank Prof. Anantha Chandrakasan and Dimitri Antoniadis for participating in my thesis committee and providing feedback at various stages of this research project. I would also like to thank Prof. Jeffry A. Davis at the Georgia Institute of Technology for helping me understand his wire-length distribution models as well as Andy Fan for keeping me up to date about his 3-D technology development work.

I have benefited both academically and socially by interacting with many people within MIT. The photography related experiments and discussions with Gabor Csanyi and Indranath Neogy, the fishing trips with Dr. Ilya Lyubomirsky, and the workout routines with Dr. Dennis Okumu Ouma were essential for relieving stress and enriching my experience at MIT. I would like to thank my colleagues in Reifgoup, Ritwik Chatterjee, Kuan-Neng Chen, Shamik Das, Andy Fan, Simon Karecki, Wendy Mao, Laura Pruette, and Minghao Qi for their company and making a pleasant work environment. Syed Alam and Andy Fan read some of the chapters of this thesis and provided useful feedback. Mathew Varghese helped me get started with ANSYS for thermal analysis. I would also like to thank Ihsan Djomehri, James Fiorenza, Dr. Abraham Kim, Dr. Huey Le, Vikas Mehrotra, Hasan Nayfeh, Debroah Hodges-Pabon, Tae Park, Tamba Tugbawa, and Dr. Andy Wei for making my stay at MTL enjoyable.

I am indebted to the support and encouragement I have received over the years from my wife, Oni, our parents, and family members. It would have been a tough journey without their love and affection.

This research project was funded by MARCO Focused Research Center for Interconnects under Task V, Materials and Processing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the last two decades, the historical progression of semiconductor industry can be attributed to cost per performance reduction due to transistor scaling and better manufacturing capabilities and the growing demand for integrated circuits (ICs) in consumer and industrial applications [4, 5]. High performance microprocessors, application specific integrated circuits (ASICs), and static or dynamic random access memories (SRAMs or DRAMs) continue to rely on transistor and interconnect scaling to meet density and performance targets. Though transistor scaling improves both speed and density, interconnect scaling improves density at the expense of degraded interconnect delay [1]. In older technology generations, interconnect delay represented a small fraction of the cycle time and did not have any significant impact on overall chip performance. However, in today's and future high-performance ICs, interconnect delay will have a significant impact on overall system performance [1].

To meet the wiring density requirement, interconnect's pitch is reduced and the number of interconnect levels is increased in successive technology generations. Interconnect's aspect ratio is often increased to reduce the wiring resistance. Though increasing the aspect ratio of interconnects leads to smaller resistance per unit length, it also results in higher coupling capacitance to neighboring interconnect lines on the same wiring level. In the current technology generation, 60% – 80% of total wiring capacitance is due to lateral coupling capacitance. If the lateral coupling capacitance is too large, it may cause serious signal integrity problems [6]. In an ideal situation, if the wire-length is scaled down by the same factor as the wiring cross-section and pitch, the RC delay of a scaled interconnect would remain the same. However, this is not the case. With higher number of transistors per chip, the wiring complexity of an IC is expected to increase, leading to an increase in wire-length and chip area [2, 7]. In successive technology generations, it is also necessary to reduce interconnect delay so that overall system performance can be enhanced.

Al and SiO_2 have been the materials of choice for multi-level metallization for many years. However, due to deteriorating performance of scaled interconnect lines in current and future technology generations, interconnect solutions based on low resistivity metal such as Cu and low-k inter-level dielectric (ILD) will be needed. Technology solutions based on Cu and low-k ILD are expected to extend interconnect scaling several technology generations. However, there is a limit to how low the dielectric constant of a realizable ILD material can be. In 35 nm technology node, it is expected that the effective dielectric constant of an ILD layer would be in the range of $1 - 1.5$ [2]. Also, there is no alternative to Cu, as an interconnect metal layer, in sight yet. In the recent international technology roadmaps for semiconductors (ITRS) it has been suggested that beyond Cu and low-k, interconnect solutions based on innovative technology and/or system architecture will be required to meet the projected system performance requirements [2, 7]. Some of these potential solutions are [2, 7, 8, 9]:

- Three-dimensional integration
- On-chip or off-chip optical/RF interconnects
- System architectures exploiting communication locality

A graphical illustration of these technology- and system architecture-based solutions are provided in Fig. 1-1. Three-dimensional integration is expected to provide several key advantages such as higher packing density, high-speed operation, and integration of mixed technologies [10, 11]. Three-dimensional IC has the potential to alleviate future interconnect delay problems by offering flexibility in system design, placement, and routing.

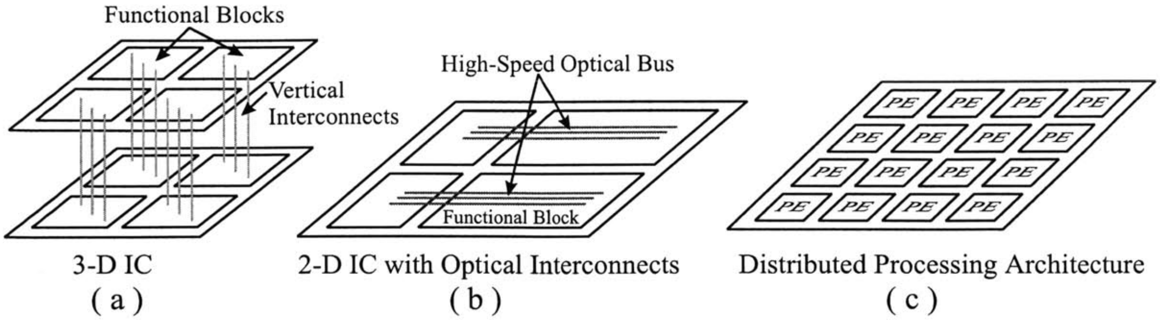


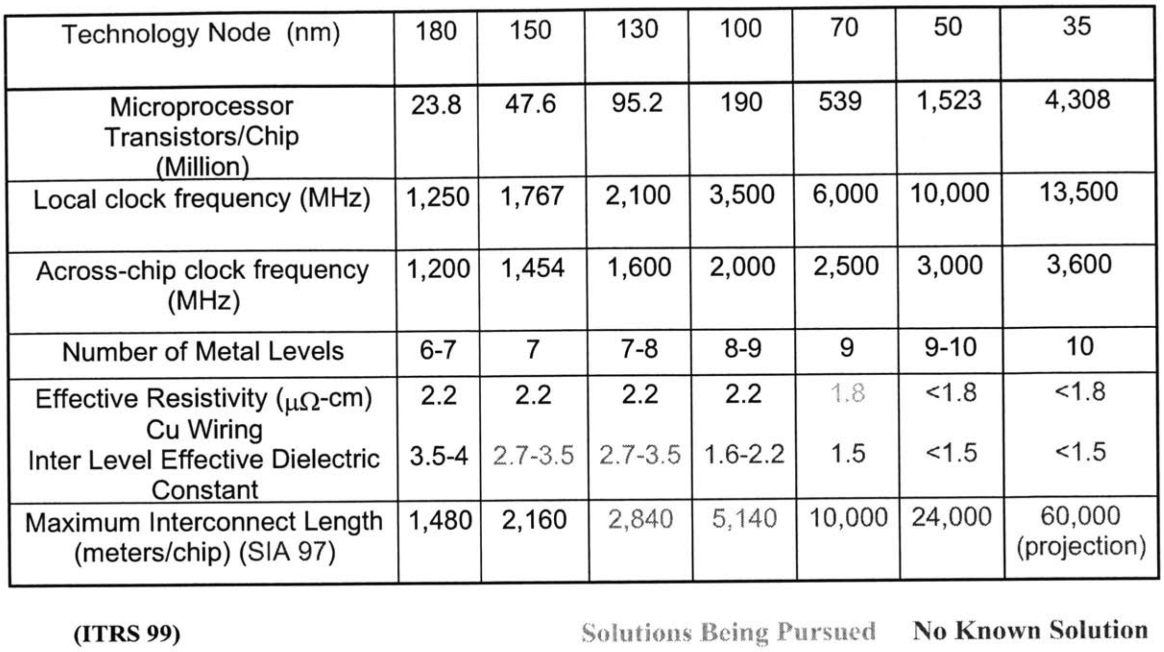
Figure 1-1: (a) Three-dimensional IC with short and high-density vertical interconnections between device layers. (b) Two-dimensional IC with high-speed on-chip optical buses for global signal distribution. (c) Clustered/tiled processing architecture with an array of processing elements (PEs).

Optical and RF interconnects are attractive for high-speed chip-to-chip communications [9, 12]. However, there is also interest in RF and optical interconnect technologies for on-chip clock and global signal distribution [9]. One of the benefits of optical interconnect is that it is essentially free from crosstalk, and power dissipation in global optical interconnect is much smaller compared to its electrical counterpart [9].

System architectures that exploit communication locality usually take advantage of tiled/clustered processing architecture [13]. Each cluster or processing element (PE) consists of an ALU and local registers. Computational tasks are mapped onto an array of distributed processing elements to minimize the need for frequent across-chip communication [13, 14]. When global communication is required, shared packet networks rather than dedicated wires can be used to make more efficient use of on-chip communication resource [13]. Though these competing solutions are quite different from each other, they are all targeted to minimize the impact of interconnect delay, cross-talk, etc. on overall system performance.

1.2 Issues Associated with Long Wires

Some of the commonly encountered interconnect related issues such as routability, delay, signal integrity, etc. are generally associated with semi-global and global wires [15, 16]. Though there are fewer long wires in an IC compared to short wires, they (long wires) take



Technology Node (nm)	180	150	130	100	70	50	35
Microprocessor Transistors/Chip (Million)	23.8	47.6	95.2	190	539	1,523	4,308
Local clock frequency (MHz)	1,250	1,767	2,100	3,500	6,000	10,000	13,500
Across-chip clock frequency (MHz)	1,200	1,454	1,600	2,000	2,500	3,000	3,600
Number of Metal Levels	6-7	7	7-8	8-9	9	9-10	10
Effective Resistivity ($\mu\Omega$-cm) Cu Wiring	2.2	2.2	2.2	2.2	1.8	<1.8	<1.8
Inter Level Effective Dielectric Constant	3.5-4	2.7-3.5	2.7-3.5	1.6-2.2	1.5	<1.5	<1.5
Maximum Interconnect Length (meters/chip) (SIA 97)	1,480	2,160	2,840	5,140	10,000	24,000	60,000 (projection)

(ITRS 99) Solutions Being Pursued **No Known Solution**

Table 1.1: Projected system performance and interconnect parameters for scaled CMOS technologies in high-performance microprocessors [2, 7].

up a significant fraction of the wiring tracks or routable area. Design methodologies for signal integrity, repeater insertion, etc. are also dictated by the performance requirements of long wires.

It is expected that total on-chip wire-length in microprocessors will increase from 820 m/chip in 250 nm technology node to 24000 m/chip in 50 nm technology node [2, 7] (see Table 1.2). This phenomenal increase in total wire-length is due to higher wiring requirements for a 1000× increase in the number of transistors per chip. Taking into account higher number of available interconnect levels in 50 nm technology node and the reduction in wiring pitch, complete wireability would require roughly a 4× increase in chip area. By examining the stochastic wire-length distribution of an IC, we find that typically the number of semi-global and global interconnects are much smaller than short or local interconnects [17]. However, the contribution of long wires to total wire-length is much higher than that of short wires. For example, as shown in Fig. 1-2, 20% of total number of wires contribute to 90% of total wiring need. Due to the significant contribution of long wires to total wire-length, the wiring-limited chip area, number of interconnect levels, etc. are

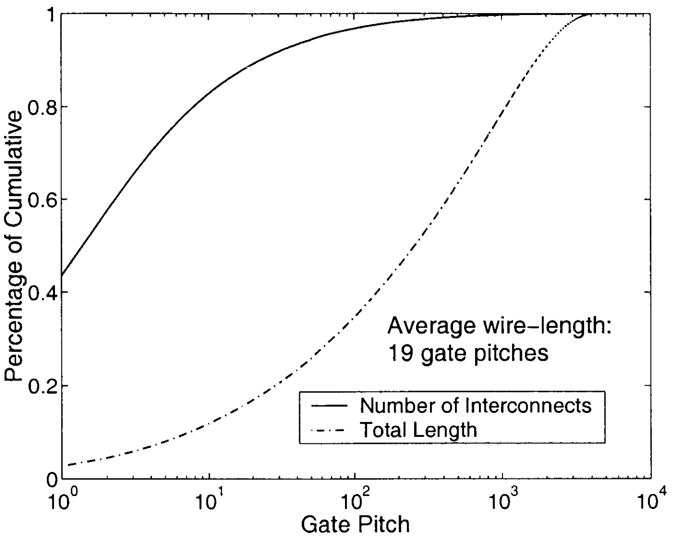


Figure 1-2: Percentage of total number of wires and the corresponding contribution to total wire-length in a 2-D IC with ten million logic gates. Gate pitch is the unit for average separation between adjacent logic gates. Rent's constant and exponents are 4 and 2/3, respectively. Rent's Rule is given by $T = kN^p$, where N is the number of logic gates, k is Rent's constant, and p $(0 \leq p \leq 1)$ is Rent's exponent.

Technology Node (nm)	500	180	100	70	50	35
Chip-Edge Wire-Length (mm) Width: 4F Pitch: 8F	12	18.4	20.2	21.6	23.1	24.7
Chip-Edge Length Wire Delay (ns)	.28	1.7	4.1	4.9	10.8	24.9
FO4 Inverter Delay (ps)	180	64	36	25.2	18	12.6
Chip-Edge Length Wire Delay (number of clock cycles)	.1	1.7	7.2	12.2	37.6	123

Table 1.2: Comparison of device and interconnect delay in scaled technology nodes. The parameters used to estimate delay are consistent with the projections in semiconductor technology roadmap [2, 7]. The FO4 inverter delay (in picosecond) is estimated using the heuristic, $\tau_{FO4} = 360 \cdot L_{drawn}$, where L_{drawn} is the drawn channel length in μm [18]. F is the minimum feature size. A clock cycle is approximated by 16FO4 delay [18].

generally determined by the long wiring need of an IC. It is expected that total on-chip wire-length in a microprocessor will be 60 Km/chip in 35 nm technology node, creating a tremendous pressure on interconnect design, routability, and physical implementation (see Table 1.2).

Besides routability constraints for intermediate and long wires, their RC delay is also expected to increase in scaled technology nodes. Solutions based on reversed scaling of interconnects such that long wires have "fat" cross-sectional dimensions may enhance their performance but at the expense of lower wiring density and larger chip area. In Table 1.2, RC delay of chip-edge length global wires and intrinsic device delay for microprocessors, implemented in scaled technology nodes, are shown. For a first order calculation, $16 - 20$ stages of FO4 (fan-out of 4) inverter delay can be used as a model for cycle time [18]. From Table 1.2, in scaled technology nodes, the number of clock cycles required to send signals across chip in non-repeated global lines will increase significantly which can have serious implications on overall system performance. To reduce long wire delay, repeaters have to be inserted every $1\ mm - 2\ mm$ in high-performance circuits [13]. However, this measure leads to an increase in chip area and power dissipation. Due to insertion of repeaters on global wires, the increase in chip area can be as much as 20% for microprocessors beyond 100 nm technology nodes [19]. Even with repeater insertion, the number of clock cycles required to send signals across-chip will increase in scaled technology generations, and interconnect delay of long wires will have a significant impact on overall system performance.

Signal integrity issues in interconnect lines due to coupled noise from neighboring lines have a strong dependency on interconnect's length, and generally, semi-global and global wires are more susceptible to coupling noise [20]. To reduce the coupling noise, separation between adjacent interconnect lines on the same wiring plane have to be increased and/or repeaters have to be inserted. However, both of these measures increase the wiring-limited chip area.

Some of the design issues associated with long wires can be resolved at the expense of larger chip area, higher power dissipation or by increasing the number of interconnect levels. An alternative approach to resolve the long wire problems is by taking advantage of 3-D integration technology which will allow higher packing density and shorter interconnect length compared to conventional 2-D integration [11]. As a result, significant reduction in interconnect delay and chip area can be achieved in wiring-limited ICs.

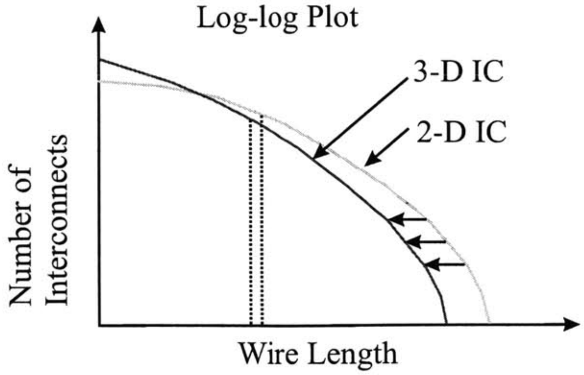


Figure 1-3: The wire-length distribution in 2-D and 3-D implementation of logic circuits. Three-dimensional integration results in a narrower wire-length distribution with fewer and shorter semi-global and global wires and higher number of local wires.

1.3 Opportunities for Three-Dimensional Integration

In the previous section, some of the commonly encountered design challenges associated with long wires have been discussed. Three-dimensional ICs are expected to reduce the semi-global and global wiring requirements significantly, therefore, allowing higher packing density in wiring-limited ICs and smaller interconnect delay [10, 11]. To illustrate the advantages of 3-D integration, in Fig. 1-3 the projected wire-length distribution of 2-D and 3-D logic circuits is shown. By mapping a 2-D IC in 3-D, the number and length of long wires can be reduced significantly at the expense of higher number of short wires. As a result of narrower wire-length distribution in 3-D integration, total and average wire-length will also become shorter, leading to smaller wiring-limited chip area and higher system performance. Three-dimensional integration can also result in a significant reduction in the number of repeaters for intermediate and long wires [21]. Based on system-level modeling and analysis work that will be presented in this thesis, we find that significant improvements in system performance and packing density are achievable by 3-D integration in applications such as high-performance logic, microprocessors, programmable devices, etc.

It is expected that future system-on-a-chip (SOC), as shown in Fig. 1-4, would require integration of digital, analog, and communication functionalities [22]. For portable applications, it will be desirable to have a small form factor for such systems. One of the challenges for implementing such SOC is the difficulty for monolithic integration of functional blocks

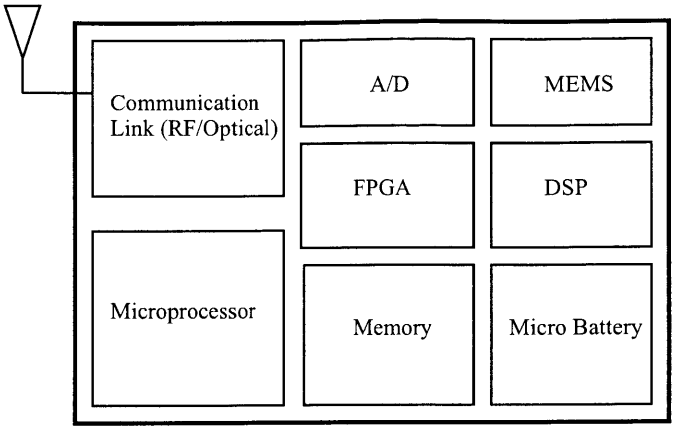


Figure 1-4: Illustration of various functional blocks in a future system-on-a-chip (SOC).

that would require different process technologies. For conventional (2-D) monolithic implementation, it would be difficult to optimize the CMOS process technology for all functional blocks. An alternative approach would be package-level integration of various components that are fabricated with optimized process technologies at the expense of higher cost and lower bandwidth (due to I/O limitation for off-chip communication) between functional blocks. Three-dimensional integration can alleviate these difficulties associated with both package-level and 2-D monolithic implementation of SOC. In 3-D integration a functional block or groups of functional blocks can be fabricated in different wafers using optimized process technologies and then integrated vertically to form SOC. To take advantage of the high bandwidth interconnection paths between device layers in 3-D integration, it may also be necessary to make innovative changes in the system architecture.

1.3.1 Approaches to Three-Dimensional Integration

The concept of 3-D integration was demonstrated as early as in 1979 [23]. Though there have been significant research efforts in developing 3-D integration technologies in late 80's [10, 24, 25, 26, 27], commercialization of these technologies never materialized. Technology solutions by scaling the feature size to reduce cost per performance were simply too attractive compared available 3-D integration technologies. Also, the performance of ICs was device-limited, and there was no urgent need for technologies targeted towards reducing interconnect's RC delay. However, with the growing importance of interconnect delay, it is now necessary to explore both near- and long-term interconnect solutions. It is speculated that beyond Cu and low-k, 3-D integration technologies can play a pivotal role in reducing interconnect delay and enabling monolithic integration of mixed technologies for SOC applications.

Approaches to 3-D integration can be classified by the following enabling technologies:

- Package-Level Integration
- Thin Film Technology (epitaxial growth, re-crystallization, etc.)
- Wafer Bonding

There are various package-level technologies for 3-D integration. In [28, 29] reviews of various packaging approaches to 3-D integration have been provided. In package-level 3-D integration, bare dies or packaged chips are stacked vertically, as shown in Fig. 1-5, and interconnections between them are formed at the chip periphery. Using package-level integration techniques, dense memory modules can be fabricated [30, 31, 32, 33]. Area interconnections formed by through-wafer vias or solder balls in the case of face-to-face bonding can also be used to provide vertical interconnections between stacked multi-chip modules (MCMs). Utilizing the area interconnection technology, vertical integration of ASICs and memory, signal processing circuits, and microprocessors have been implemented [34, 35, 36, 37, 38, 39]. The drawback in package-level 3-D integration technology is the lower density of vertical (inter-device layer) interconnects compared to monolithic 3-D integration. Some or the area interconnect technologies for wafer scale integration (WSI) also have very poor yield [35, 36].

Presently, there are several fabrication process technologies, based on selective epitaxial growth (SEG) or re-crystallization, that can be used to grow single crystal or recrystallized poly-Si, separated by dielectric material, on top of an existing substrate [10, 26, 40, 41, 42, 43]. In selective epitaxial growth, as shown in Fig. 1-6, seed windows are opened on an oxidized Si substrate. Epitaxial growth is initiated through the seed window in the presence of dichlorosilane which supplies the silicon [44, 10, 42, 40, 41]. The quality of devices fabricated on epitaxially grown layer can be as good as those in bulk silicon. However, the high processing temperature ($600\ ^0C - 1000\ ^0C$) in SEG can cause significant degradation

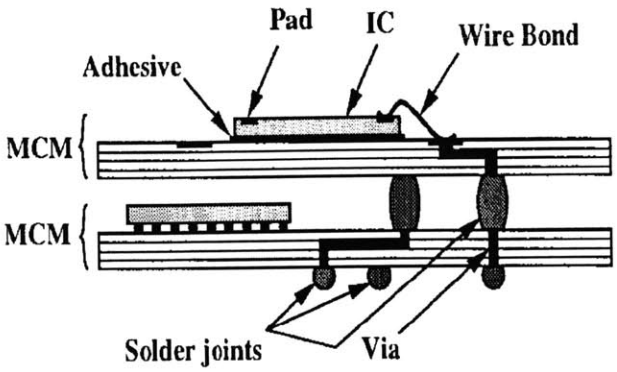


Figure 1-5: Package-level 3-D integration by vertical stacking of multi-chip modules [28].

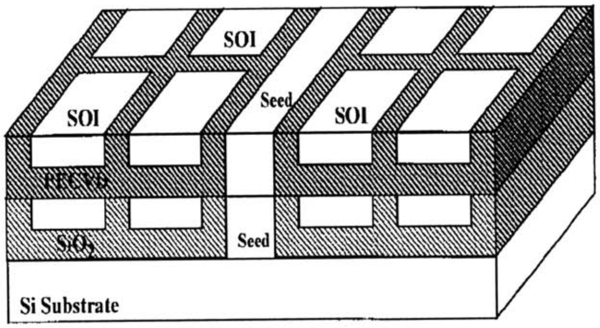


Figure 1-6: Selective epitaxial growth of Si islands through seed windows for monolithic 3-D integration [44].

in the quality of devices on lower device layers. Moreover, due to the low melting points of some interconnect metals and barrier layers, only poly-silicon interconnect levels are allowed in lower device layers. Another popular technique for fabricating second Si layer is to deposit polysilicon on a fully processed wafer, recrystallize the polysilicon film by intense laser beam or Ge-seeded lateral crystallization, and fabricate devices on the re-crystallized Si layer [10, 26, 43]. However, in these recrystallization techniques, it is difficult to control the grain size variations, and the quality of thin film transistors is not as good as bulk Si devices.

Low-temperature ($< 450\ ^0C$) copper-copper or polyemide based wafer bonding can be used to bond two fully processed wafers for fabricating 3-D ICs [45, 46, 47]. In 3-D IC technology based on wafer bonding, after the fabrication of individual wafer, top wafer is attached to a handle wafer, and it is thinned down from the back side. Then the backside of top wafer is bonded to the front side of the bottom wafer (i.e. back-to-front bonding); after the bonding process, the handle wafer is released. Through wafer vias for inter-device layer connections can be formed before or after the bonding step [46, 47]. Ideally, the processing steps required to bond two wafers can be repeated to bond as many wafers as desired. However, thermal issues, processing cost or complexity, etc. may dictate the number of wafers that can be profitably integrated. Using low-temperature wafer bonding technology, functional blocks that require different process technologies can be placed in different device layers to form 3-D SOC. A cross-sectional view of a 3-D IC formed by Cu-Cu wafer bonding is shown in Fig. 1-7. One of the limitations of wafer bonding technology is the lack of precision (± 3 μm) in aligning Cu or metallic bumps using IR alignment to form inter-device layer interconnections. However, even with ± 3 μm alignment tolerance, the number of available inter-device layer wiring tracks will be sufficient to meet the semi-global and global wiring need of a 3-D IC with a few device layers.

Compared to non-monolithic approaches to 3-D integration, the density of inter-device layer interconnections in monolithic approach is 30× to 50× higher. The relative cost for integrating various modules or components monolithically is generally smaller compared to packaging alternatives [48]. The potential improvements in form factor, power dissipation, or I/O bandwidth by monolithic 3-D integration may reduce the system cost compared to 2-D or 3-D package-level integration. Presently, Cu-Cu wafer bonding technique is being explored at MIT to demonstrate the feasibility of monolithic 3-D ICs. A detailed description of the process flow for fabricating 3-D ICs based on Cu-Cu wafer bonding is presented in Appendix A.

For clarity and to avoid confusion about the terminology for describing a 3-D IC, as shown in Fig. 1-8, following definitions will be used in this thesis work:

- *Interconnect level*: Combination of one metal layer of wires and one inter-level dielectric (ILD) with vias.
- *Interconnect layer*: A set of contiguous interconnect levels. E.g. multiple interconnect levels in a modern CMOS process.

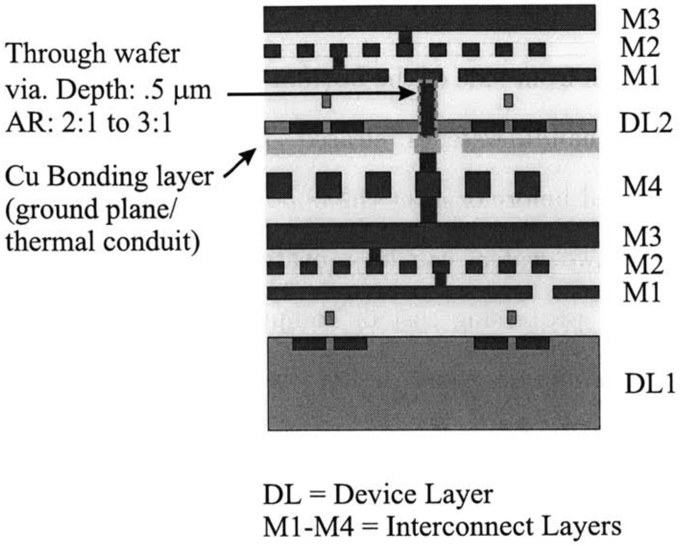


Figure 1-7: Cross section of a 3-D integrated circuit formed by face-to-back Cu-Cu wafer bonding.

- *Device layer*: A planar layer of silicon in which transistors are fabricated. E.g. silicon wafer in bulk-Si CMOS IC or the thin silicon layer in a SOI-CMOS IC.
- *Stacked device layer*: Two or more device layer without intervening metal wires. E.g. a system of silicon layers in which one of the layers is used as seed for crystalline growth by selective epitaxial growth (SEG) or by laser recrystallization.
- *Sandwich interconnect*: Two or more interconnect layers sandwiching a thin device layers from top and bottom.
- *Stratum*: A combination of one device layer with a single interconnect layer, or sandwich interconnect.
- *Multi-strata IC*: Integrated circuits formed by stacking or bonding multiple strata and interconnecting them.
- *Intra-stratum interconnect*: Metal wires and vias which form electrical connection between two terminals within the same stratum.
- *Inter-stratum interconnect*: Metal wires and vias which form electrical connection between two terminals on different strata in multi-strata IC.

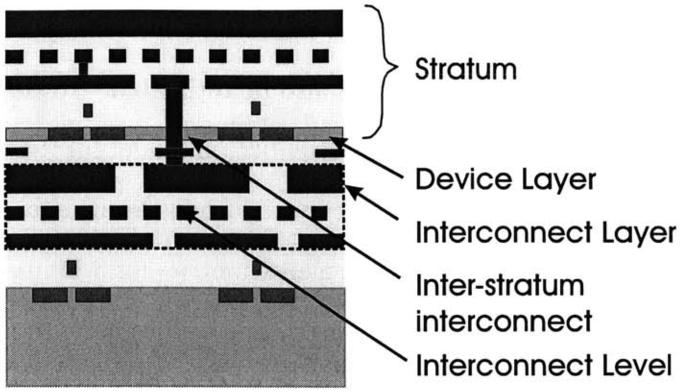


Figure 1-8: Key definitions and terminologies for describing a generic 3-D IC.

1.4 System-Level Performance Evaluation of Integrated Circuits and Contribution of Thesis Work

In system-level modeling and analysis, key performance metrics such as chip area, cycle time, interconnect parameters (number of interconnect levels, wiring pitch, etc.), etc. are estimated based on models representing system architecture, interconnection complexity, and device and interconnect technologies [49, 50, 51, 52]. Over the last several years, system-level modeling has been used quite successfully to assess the impact of scaling device and interconnect parameters and modifying system architecture on overall system performance [49, 50, 51, 52]. Circuit and system designers can use system-level modeling to estimate chip performance and wiring requirements at an early stage of a design cycle before routing and placement. Based on the feedback from system-level modeling, adjustments in chip design as well as wiring parameters can be made such that the final design can fit within the available Si area and required performance targets can be met.

Though some of the benefits of 3-D integration can be easily speculated, they have to be quantified so that key decisions for focusing future direction of research in this area can be made. It is also essential to examine the benefits and limitations in various approaches to 3-D integration so that cohesive research efforts can be directed to the most viable 3-D technology. In this thesis work, system-level modeling and analysis framework for 2-D ICs is extended to 3-D ICs to estimate key performance metrics and also to perform trade-off analysis for various approaches to 3-D integration. Based on system-level modeling, key interconnect parameters such as wiring pitch, number of interconnect levels per stratum, inter-stratum interconnect's density, etc. can also be estimated. To examine the interconnection complexity, a stochastic wire-length distribution model for 3-D ICs is derived and integrated in the system-level modeling framework. Various limiting cases of inter-stratum connectivity are also be examined. Key performance metrics for 3-D implementation of various VLSI circuits such as high-performance logic, microprocessors, reconfigurable circuits, etc. are estimated. For high performance logic circuits, it is found that significant reduction in critical path delay and chip area can be achieved by 3-D integration. Generally, system performance can be improved by increasing the number of strata. However, due to thermal issues, process complexity, yield, etc. it may not be profitable to integrate more than 4-5 strata.

Besides performance modeling, system-level analysis work is extended to examine thermal issues in 3-D ICs. Power dissipation in 3-D ICs is modeled by taking into account power dissipation due to driving device, interconnect, and clock wiring capacitance. Based on our analysis, due to the significant reduction in wiring capacitance in 3-D circuits, it will be feasible to reduce their total power dissipation compared to 2-D ICs for comparable system performance. However, due to the higher thermal resistance in 3-D ICs, for comparable amount of total power dissipation, junction temperature (in 3-D ICs) will be slightly higher than that of 2-D ICs. In this thesis work, detailed thermal analysis will be conducted to address both power dissipation and heat removal issues. Device- and package-level thermal analysis will be conducted based on numerical analysis and analytical models. Some recommendations will also be made to reduce the die temperature in 3-D ICs.

The thesis write-up is organized as follows: in Chapter 2, background on system-level modeling and analysis is provided. It is followed by the derivation of stochastic wire-length distribution of 3-D ICs and some case studies in Chapter 3. In Chapters 4 and 5, simulation results of system performance for various applications of 3-D ICs are presented. Thermal issues in 3-D ICs are discussed in Chapter 6, followed by Conclusions and Summary in Chapter 7.

Chapter 2

Background: System-Level Modeling

In this chapter the methodology and modeling framework for estimating key performance metrics such as clock frequency, chip area, power dissipation, etc., of two- and three-dimensional integrated circuits are presented. An accurate estimation of these metrics can be made using computer aided design (CAD) tools at the end of a design cycle, after layout and verification. However, sometimes it is necessary to estimate chip performance and technology requirements at a very early stage of the design cycle to assess the impact of integrating new technologies or system architectures and also to make sure a design, when implemented in silicon, will fit within the available area. In order to make such projections, it is desirable to have a modeling framework that can be implemented easily and requires simple and consistent models representing system architecture, wiring complexity, and technology requirements. A set of analytical equations and empirical models, representing the dependencies of system performance on both architectural and technology dependent parameters, are often used for a priori system-level modeling of key performance metrics [49, 50, 51, 52]. This type of modeling framework has been applied quite successfully to assess the impact of integrating new technologies such as Cu and low-k or modifying system architecture on system performance of 2-D ICs [49, 50, 51, 52].

Over the years, several simulators have been developed based on system-level modeling that include SUSPENS (Stanford University System Performance Simulator), RIPE (Rensselaer Interconnect Performance Estimator), GENESYS (Generic System Simulator), and BACPAC (Berkeley Advanced Chip Performance Calculator) [49, 52, 50, 51]. In general, these system-level simulators require descriptions about the device and interconnect technologies and system architecture. Then equations governing the dependencies between device and interconnect parameters and chip performance are solved to estimate chip area, clock frequency, wiring pitch, etc. The beauty of system-level modeling is that the dependencies between key technology parameters of interest are represented in a consistent way. For example, increasing the wiring pitch in an IC may initially lead to smaller interconnect delay; however, it may also increase the chip area and wire length, and at some point it may not be justifiable to increase the wiring pitch. In system-level modeling, such intricate dependencies between various technology dependent parameters are represented in a consistent way and also exploited to obtain solutions for minimum chip area at the expense of lower clock frequency or maximum clock frequency at the expense of non-optimum chip area [53].

Though there have been significant research efforts on system-level modeling for conventional 2-D ICs, applications of system-level modeling to assess the impact of integrating new technologies such as 3-D or optical interconnects on system performance have not been explored. In this thesis work, system-level modeling framework will be extended to assess the impact of 3-D integration on system performance. In this Chapter, before discussing the opportunities for 3-D ICs, some of the commonly used models for system-level analysis of 2-D ICs will be presented. An extension of the modeling framework to 3-D ICs will be presented in Chapter 3.

2.1 Methodology and the Modeling Framework

The first researcher to propose a system-level modeling framework that incorporated information about architecture, technology, and packaging was Bakoglu [49, 54]. His system-level simulator, SUSPENS, predicts clock frequency, chip size, and power dissipation based on a set of analytical equations as shown in Table 2.1 [49]. In his modeling framework, chip area, D_c, is assumed to be wiring-limited, determined by the on-chip wiring need of an IC and the wiring pitch. To estimate the total wire-length and chip area, an IC is represented by an array of logic gates, and the stochastic wiring requirements between these logic gates are estimated using Rent's Rule. The chip area, number of interconnect levels, and wiring pitch are determined such that the total wiring between logic gates can be accommodated within the available wiring area [49]. The canonical logic gate delay, T_g, is represented by its equivalent resistance, R_{tr} and capacitance, C_{tr}. A typical logic gate is loaded by a similar gate and connected by an average-length wire of length $\overline{R}$. The cycle time is represented by the sum of several stages of logic gate delay, specified by the logic depth, f_{ld}, and interconnect delay of a chip-edge length wire. This cycle time model is also the basis for all other system-level simulator such as SUSPENS, RIPE, GENESYS, and BACPAC.

Step		Calculation
1	$\overline{R}$	$= \frac{2}{9}\left(7\frac{N_g{}^{p-0.5}-1}{4^{p-0.5}-1} - \frac{1-N_g{}^{p-1.5}}{1-4^{p-1.5}}\right)\frac{1-4^{p-1}}{1-N_g{}^{p-1}}$
2	d_g	$= \frac{f_g\overline{R}p_w}{e_w n_w}$
3	D_c	$= \sqrt{N_g}d_g$
4	l_{av}	$= \overline{R}d_g$
5	R_{gout}	$= f_g\frac{R_{tr}}{k}$
6	C_{gin}	$= 3kC_{tr}$
7	T_g	$= f_g R_{gout} l_{av} C_{int} + f_g R_{gout} C_{gin} + R_{int} C_{int}\frac{l_{av}^2}{2} + R_{int} l_{av} C_{gin}$
8	f_c	$= \left(f_{ld}T_g + R_{int}C_{int}\frac{D_c^2}{2} + \frac{D_c}{v_c}\right)^{-1}$
9	C_{TOT}	$= \frac{D_c^2 n_w e_w C_{int}}{p_w} + 3C_{tr}kN_g f_g$
10	N_p	$= K_p N_g^{\beta}$
11	P_c	$= \frac{1}{2}f_c f_d C_{TOT} V_{DD}^2 + \frac{1}{3}\frac{1}{2}N_p f_c f_d C_{OUT} V_{DD}^2$

Table 2.1: A complete list of equations used for estimating the key performance metrics of CMOS microprocessor in SUSPENS [49].

The system-level modeling approach in SUSPENS has been quite useful in predicting system performance of ICs. However, there are some limitations in the modeling framework and methodology. SUSPENS does not consider memory elements, and it assumes the same wiring pitch for all interconnect levels. In todays high-performance ICs, inverse scaled interconnect scheme is used, where longer wires have larger wiring pitch to minimize interconnect delay [15]. Assuming the same wiring pitch in all interconnect levels in system-level modeling may lead to spurious estimate of chip area and cycle time. Recent system-level simulators such as RIPE, GENESYS, and BACPAC have incorporated many refined models to remove some of the limitations in SUSPENS. For example, in BACPAC, chip area models based on a cluster of functional blocks with 50K-100K logic gates have been proposed to reflect commonly used design practices in ASIC [51]. In GENESYS, very detailed and elaborate models for device characteristics, interconnect architecture, and memory hierarchy are incorporated [50]. Unlike other simulators, GENESYS also incorporates complete wiring requirements based on stochastic wire-length distribution. Also, no knowledge about the semi-global or global wiring pitch is required, and the wiring pitch is determined to optimize chip area or system performance. The methodology used in this thesis work to estimate the key performance metrics of 3-D ICs is very similar to the methodology used in SUSPENS and GENESYS. A very generic description of the methodology which is the basis for system-level simulation is summarized in Fig. 2-1.

2.2 Models for System-Level Simulation

2.2.1 Gate Delay

To estimate intrinsic gate delay as well as delay associated with logic gates driving capacitive and/or resistive loads, information about the device technology is needed. There are many published models to estimate the delay of CMOS inverters and logic gates [49, 55]. Most of these models require many input parameters including mobility, doping profile, etc. In our modeling framework, we use a simpler model to estimate gate delay based on equivalent resistance and load capacitance of logic gates [49, 51]. Using a lumped element model,

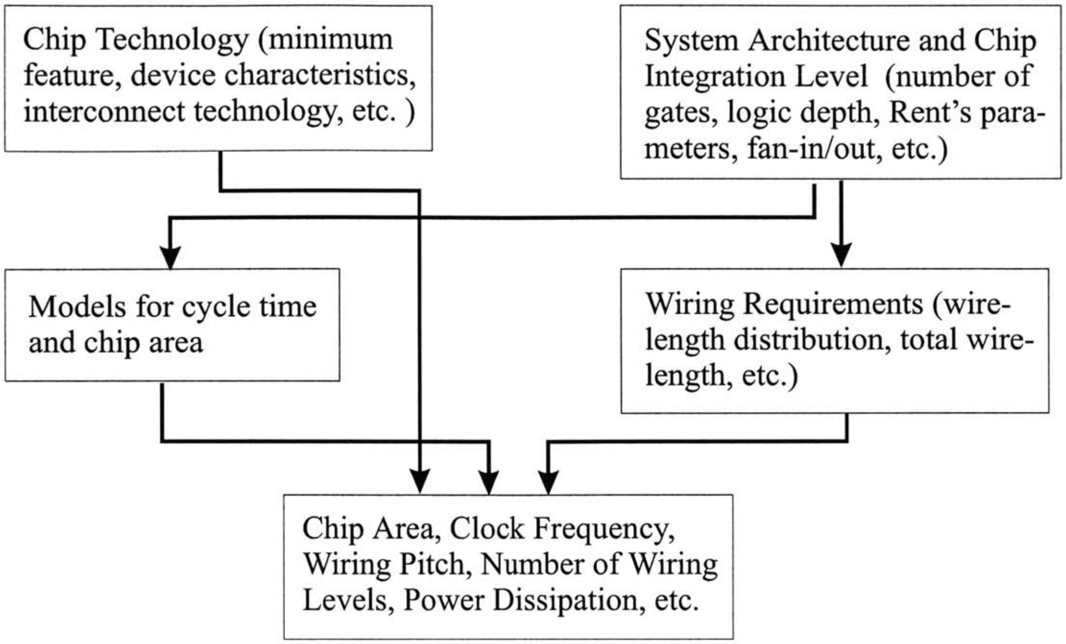

Chip Technology (minimum feature, device characteristics, interconnect technology, etc.)
System Architecture and Chip Integration Level (number of gates, logic depth, Rent's parameters, fan-in/out, etc.)
Models for cycle time and chip area
Wiring Requirements (wire-length distribution, total wire-length, etc.)
Chip Area, Clock Frequency, Wiring Pitch, Number of Wiring Levels, Power Dissipation, etc.

Figure 2-1: A graphical illustration of the methodology which is the basis for system-level simulation.

50% delay of a CMOS inverter, as shown in Fig. 2-2, is approximately equal to $0.69 R_{tr} C_L$, where R_{tr} is the equivalent resistance of the nMOS or pMOS transistor and C_L is the load capacitance [49]. An alternative equation for 50% gate delay in terms of drain saturation current, I_{dsat}, is $C_L V_{dd}/(2I_{dsat})$. By equating these two models of gate delay and taking into account the fact that average charging current is smaller than I_{dsat}, R_{tr} can be estimated. It is given by [51]

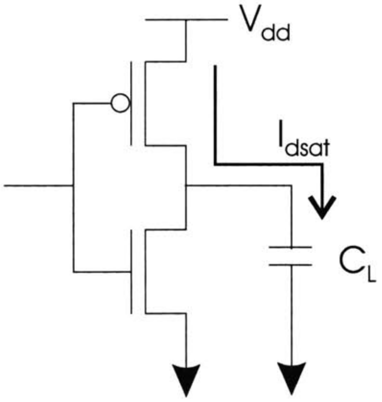


Figure 2-2: Schematic of a CMOS inverter driving a capacitive load of capacitance, C_L

$$R_{tr} = \alpha_{tr} \frac{V_{dd}}{I_{dsat}}, \tag{2.1}$$

where V_{dd} is the supply voltage, I_{dsat} is the drain saturation current and α is a constant. Based on HSPICE simulation as well as published results, it has been found that a typical value of α_{tr} is 0.805 [51]. Though various models exist for predicting I_{dsat}, we use a fixed value of $I_{dsat}/W = 600\ \mu A/\mu m$, as described in the SIA and International Technology Roadmap for Semiconductors (ITRS), to estimate R_{tr}. It is assumed that by properly scaling device technology parameters such as doping concentration, channel length, gate oxide thickness, etc., $I_{dsat}/W = 600\ \mu A/\mu m$ can be achieved in scaled technologies.

To estimate load capacitance, C_L, we assume the output terminal of an inverter is connected to the input terminal of a similar inverter; C_L is estimated by summing the dominant components of C_L: gate capacitance, C_g, and drain junction capacitance, C_{db}. As discussed in [56], gate capacitance includes gate to channel and overlap capacitance, and it can be approximated by $C_{ox}WL$, where C_{ox} is the gate oxide capacitance per unit area, and W and L are drawn channel width and length, respectively. The drain junction capacitance is contributed by the reversed biased source-bulk and drain-bulk pn junctions. It consists of bottom plate and side wall capacitance between the source/drain and the substrate [56].

In system-level modeling, an IC is often represented by an array of equivalent logic gates such as NAND gates with specific fan-in and fan-out. The 50% delay associated with a logic gate is represented by a lumped element model, and it is given by $0.69R_gC_L$, where R_g is the equivalent gate resistance, represented in terms of R_{tr}, and C_L is the load capacitance. As discussed earlier, the dominant contribution to C_L are C_g and C_{db}, and their values can be estimated knowing the device parameters (doping, gate oxide thickness, junction depth, etc.) and gate layout. Typical layout of a 3-input NAND gate is shown in Fig. 2-3. By taking into account design rules, we find the area of a 3-input NAND gate is roughly $16F \times (14F + 3W_n)$, where F is the minimum feature size and W_n is the width of an nMOS transistor [56].

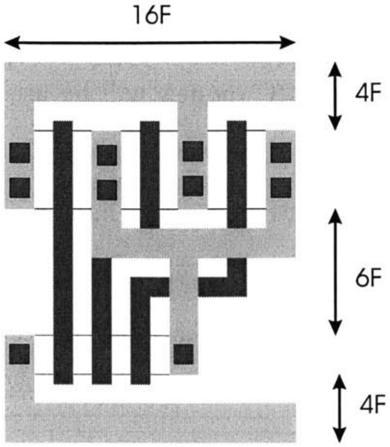


Figure 2-3: Layout of a 3-input NAND gate gate. F is the minimum feature size.

2.2.2 Interconnect Delay

Models for Estimating Interconnect Delay

In Chapter 1, the importance of interconnect delay on system performance of VLSI circuits has been discussed. In order to assess the impact of integrating new interconnect technology, accurate models for estimating interconnect delay are necessary. Historically, interconnects have been modeled as lumped capacitors to estimate gate or interconnect delay. If the wiring capacitance is comparable or larger than gate capacitance, the canonical gate delay, as described in the previous section, should be estimated by including the wiring capacitance in C_L. With the scaling of CMOS technology, the cross-sectional area of interconnects has been scaled down while the interconnect length has increased. The resistance of interconnects has therefore increased significantly, requiring the use accurate RC models [57]. Currently, inclusion of interconnect's inductance in delay models is also becoming necessary with faster on-chip rise times and longer wire lengths [58, 59]. Wide wires that are generally used in clock distribution networks and signal buses can exhibit significant inductive effect. When inductive effects are important, distributed RLC models must be used to estimate interconnect delay. However, based on our observation from system-level modeling and simulation and also from reviewing published work, we speculate the use of RLC models for estimating interconnect delay and cross-talk to only very specific wiring nets [16, 58, 59]. The choice of these wiring nets depends upon a set of design guidelines[1] governed by the wiring resistance, impedance, capacitance, and driver's equivalent resistance [58]. For simplicity, RC models will be used throughout our system-level modeling.

The RC circuit model of a logic gate driving an interconnect is shown in Fig. 2-4. R_g is the equivalent gate capacitance, R_{int} and C_{int} are interconnect's resistance and capacitance, and C_L is the load capacitance. The 50% delay due to step input for the equivalent RC circuit shown in Fig. 2-4 is given by [49]

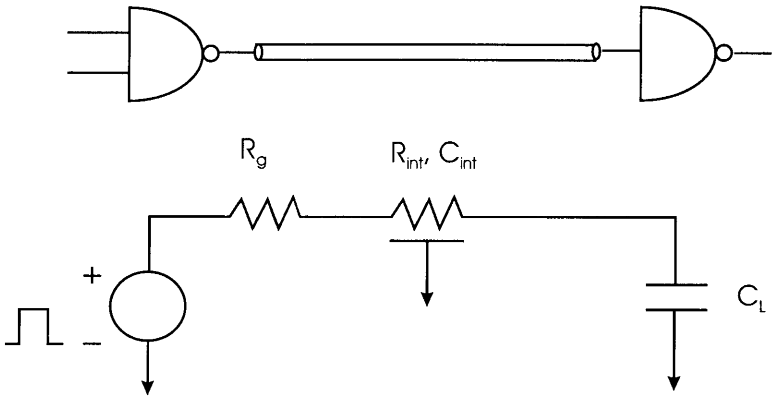


Figure 2-4: RC circuit model of an interconnect driven by a logic gate. R_g is the equivalent gate capacitance, R_{int} and C_{int} are interconnect's resistance and capacitance, and C_L is the load capacitance.

$$\begin{aligned} T_{50\%} &= 0.4R_{int}C_{int} + 0.7(R_gC_{int} + R_gC_L + R_{int}C_L) \\ &\approx 0.4R_{int}C_{int} + 0.7R_gC_{int} \quad for\ C_L \ll C_{int}. \end{aligned} \tag{2.2}$$

If R_g is comparable to R_{int},

$$T_{50\%} \approx 1.1R_{int}C_{int} = 1.1r_{int}c_{int}l^2, \tag{2.3}$$

where r_{int} and c_{int} are interconnect's resistance and capacitance per unit length and l is interconnect's length. The RC lumped element model in Eq. 2.2 is reported to have less than 4% error over the entire range of parameters [49, 57]. When there is a fan-out, $f.o.$,

[1] A design rule for including inductive effects in interconnect modeling can be $\frac{T_{rise}}{2\sqrt{LC}} < l_{wire} < \frac{2}{R}\sqrt{\frac{L}{C}}$, where T_{rise} is the signal rise time, l_{wire} is the wire-length, and R, L and C are equivalent resistance, inductance, and capacitance of the interconnect line [16].

the 50% delay is given by

$$T_{50\%} = 0.4R_{int}C_{int} + 0.7(f.o.R_gC_{int} + f.o.R_gC_L + R_{int}C_L). \tag{2.4}$$

Eq. 2.4 is useful for estimating delay through a critical path which consists of a set of logic gates with specific fan-out and connected by average length wires.

The wiring resistance depends on the effective resistivity, ρ_{eff}, of the interconnect metal, its cross-sectional area, and length. Typically, an interconnect line is encapsulated by a barrier material such as Ti or TaN which may have a higher resistivity, ρ_b, than the resistivity of the core interconnect metal, ρ_{core}. The value of effective resistivity is estimated by treating the interconnect line as two resistors in parallel with different resistivity and cross-sectional area [50]. For different technology nodes, the targeted/desired effective resistivity of interconnect metal can be found from the technology roadmap (ITRS) [2], and it is related to the resistivity of the barrier and the core material and wiring dimensions by the following equation [50]:

$$\rho_{eff} = \frac{\rho_b\rho_{core}}{\rho_{core} + (\rho_b - \rho_{core})(1 - \frac{2t_b}{W_\rho})(1 - \frac{2t_b}{H_\rho})}, \tag{2.5}$$

where W_ρ and H_ρ are the width and height of the interconnect line and t_b is the thickness of the barrier layer.

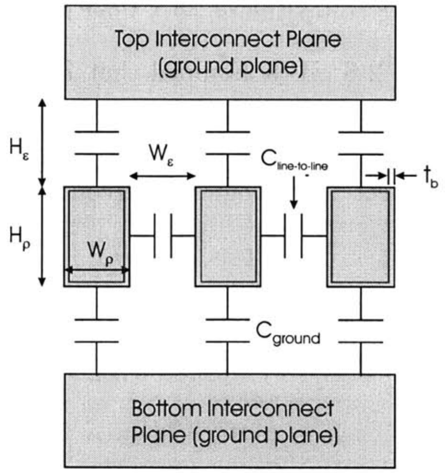

Top Interconnect Plane (ground plane)
H_ε
W_ε
C_line-to-line
t_b
H_ρ
W_ρ
C_ground
Bottom Interconnect Plane (ground plane)

Figure 2-5: Cross-section of interconnect lines and models for representing their equivalent capacitance. H_ρ and W_ρ are the height and width of interconnect lines, respectively; H_ϵ is the thickness of the ILD layer and W_ϵ is the spacing between adjacent interconnect lines. t_b is the thickness of the barrier material.

The capacitance of an interconnect line can be estimated by using a quasi-analytical model which treats the neighboring wiring planes in the multi-level wiring network as ground planes [60]. The cross-section of interconnect lines and models for estimating wiring capacitance are shown in Fig. 2-5. The total interconnect capacitance per unit length is given by

$$c_{int} = 2c_{ground} + 2c_{line-to-line}, \tag{2.6}$$

where c_{ground} is the line-to-ground capacitance and $c_{line-to-line}$ is the line-to-line or coupling capacitance. The values of c_{ground} and $c_{line-to-line}$ including the fringing effects are given by [60]

$$\begin{aligned}
\frac{c_{ground}}{\epsilon} &= \frac{W}{H_\epsilon} + 1.086(1 + 0.685e^{-(H_\rho/1.343S)} - 0.9964e^{-(S/1.421H_\epsilon)}) \\
&\quad \cdot(\frac{S}{S + 2H_\epsilon})^{0.0476}(\frac{H_\rho}{H_\epsilon})^{0.337} \\
\frac{c_{line-to-line}}{\epsilon} &= (\frac{H_\rho}{S})(1 - 1.897e^{(-H_\epsilon/0.31S)-(-H_\rho/2.474S)} + 1.302e^{-H_\epsilon/0.082S} \\
&\quad -0.1292e^{-H_\rho/1.326S}) + 1.722(1 - 0.6548e^{-W/0.3477H_\epsilon}) \\
&\quad \cdot e^{-S/0.651H_{epsilon}}.
\end{aligned} \tag{2.7}$$

The geometrical variables in Eq. 2.7 are defined in Fig. 2-5. Total interconnect capacitance per unit length and its components, as a function of interconnect's aspect ratio, $\gamma = H_\rho/W_\rho$, are plotted in Fig. 2-6. It is assumed that $H_\rho = H_\epsilon$ and $W_\rho = W_\rho$. We find that when $\gamma > 1$, interconnect's capacitance increases linearly with aspect ratio. In today's and future multi-level interconnect technologies, interconnect's aspect ratio is expected to be more than one [2]. Using Eq. 2.7, when $\gamma > 1$, $H_\rho = H_\epsilon$, and $W_\rho = W_\rho$, a simplified formula can be found by curve fitting for total wiring capacitance per unit length, and it is given by

$$c_{int} = (4.07 + 2\gamma)\epsilon_o\epsilon_r. \tag{2.8}$$

Based on Eq. 2.8, coupling capacitance becomes the dominant component of wiring capacitance when the aspect ratio $\gamma > 2$. Unless otherwise specified, Eq. 2.8 is used to estimate wiring capacitance and interconnect delay in this thesis work.

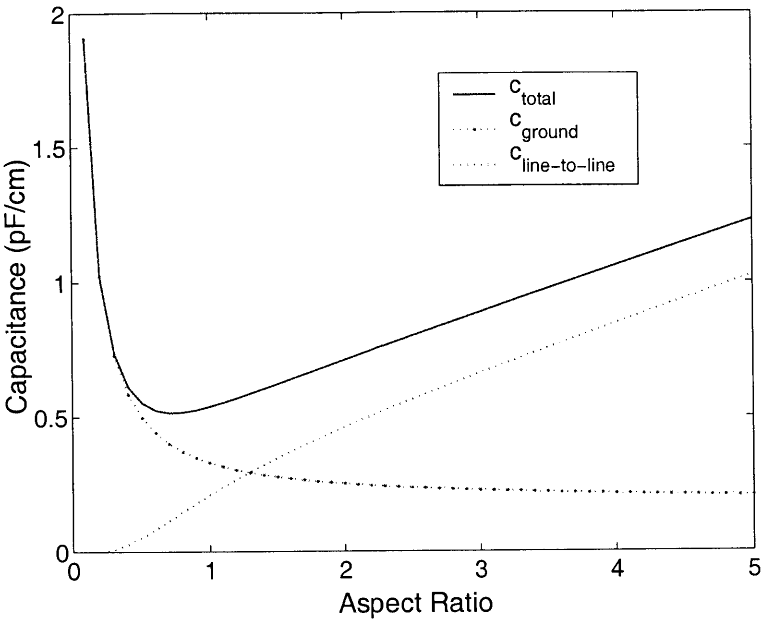


Figure 2-6: Wiring capacitance as a function of interconnect's aspect ratio, $\gamma = H_\rho/W_\rho$. It is assumed that $H_\rho = H_\epsilon$, $W_\rho = W_\epsilon$, and $\epsilon = \epsilon_o\epsilon_r = \epsilon_o$. When the aspect ratio is more than one, total wiring capacitance increases linearly with aspect ratio.

Generally, for estimating the wiring capacitance, neighboring interconnects are treated as ground planes. However, if neighboring interconnects on the same wiring plane switch in the opposite direction compared to the interconnect of interest, the coupling (line-to-line) capacitance is effectively doubled due to Miller effect. This is the worst case scenario and very unlikely for all wiring nets.

Models for Estimating Number and Size of Repeaters

When the resistance of an interconnect line is comparable to or larger than the on-resistance of the driver, delay of a logic gate driving an interconnect load is proportional to the square of interconnect's length, as described in Eq. 2.3. If the interconnect is divided into subsections and repeaters are inserted in each subsection, overall delay becomes linear with interconnect's length [49]. In today's high-speed digital designs, repeaters are necessary to reduce interconnect delay of long wires and also to improve signal integrity. It is expected that in future CMOS technologies, repeaters have to be inserted every 1 mm – 2 mm [13].

The number and size of repeaters can be estimated for various design requirements.

For example, the number and size of repeaters that minimize interconnect delay can be estimated by differentiating the equation governing interconnect delay of a repeated line with respect to repeater's size and number and setting it to zero [49]. The number of repeaters, n_r, and their size, s_r, that minimize interconnect delay are given by

$$\begin{aligned} n_r &= \sqrt{\frac{0.4 R_{int} C_{int}}{0.7 R_o C_o}} \\ s_r &= \sqrt{\frac{R_o C_{int}}{R_{int} C_o}}, \end{aligned} \tag{2.9}$$

where R_o and C_o are the equivalent on-resistance and capacitance of a minimum size inverter; R_{int} and C_{int} are interconnect's resistance and capacitance, respectively. It should be noted that repeater's size, s_r, is represented in units of minimum size inverter. Though it is desirable to insert optimum number and size of repeaters to reduce interconnect delay of long wires, it may increase the chip area and power dissipation. Based on our case studies for microprocessors, we find that the increase in chip area can be as high as $15\% - 20\%$ due to insertion of repeaters in 100 nm technology node. This conclusion is also consistent with other published results [19].

Considering the increase in power dissipation due to insertion of repeaters, in some cases, it may be more appropriate to estimate the number of repeaters and their size based on a different design guideline. For example, using a procedure similar to the one used for estimating number and size of repeaters for minimum interconnect delay, J. Eble proposed a methodology to estimate the number and size of repeaters that minimize energy-delay product [50]. The number of repeaters, n_r', and their size, s_r', for energy efficient design are given by [50]

$$\begin{aligned} n_r' &= \frac{1}{\sqrt{3}} \sqrt{\frac{0.4 R_{int} C_{int}}{0.7 R_o C_o}} \\ s_r' &= \frac{0.6}{\sqrt{3 (0.4)(0.7)}} \sqrt{\frac{R_o C_{int}}{0.7 R_{int} C_o}}. \end{aligned} \tag{2.10}$$

It should be noted that the number and size of repeaters for the energy efficient design differ from the minimum delay design by only a constant.

2.2.3 Cycle Time

The cycle time of an IC such as microprocessor is generally determined by the signal delay though any of the following paths: integer ALU, cache access, instruction decoder, and register access [61]. A generic representation of the minimum cycle time in a flip-flop based system, is given by [49]

$$T_c = t_{latch} + t_{logic} + t_{setup} + t_{skew} \tag{2.11}$$

An equivalent representation of the cycle time is based on the concept of logic depth. Logic depth is defined as the number of logic stages in the critical path between two clocked latches, including the latch delay and setup time (see Fig. 2-7). Typically, to estimate cycle time, logic gates in the critical path are approximated by NAND gates with average fan-in and fan-out of 2-3 [15]. Logic depth in high-performance microprocessor such as DEC's Alpha can range from 10 for Alpha 21364 to 16 for Alpha 21064 [51]. However, processors in the PowerPC family and ASIC based designs have much higher logic depth, and they can range from 15-30 [51]. Logic depth in a microprocessor can be reduced by exploiting the micro-architecture. Instruction level parallelism and pipelining are two of the commonly used techniques for reducing the logic depth in high performance designs. Reduced instruction set (RISC) microprocessors generally have a smaller logic depth than complex instruction set (CISC) microprocessors, and it is due to simpler instruction decoding units and additional

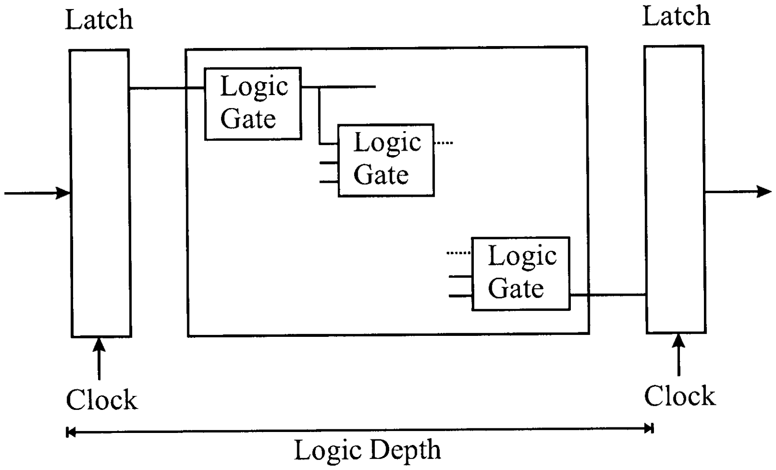


Figure 2-7: The concept of logic depth in integrated circuits.

pipelining and parallelism in RISC processors [49]. Using the concept of logic depth, a generic model for cycle time is given by

$$\begin{aligned} T_c &= f_{ld}T_g + T_{long\ wire} + T_{time-of-flight} \\ &= f_{ld} \cdot [0.4 r_{int} c_{int} l_{ave}^2 + 0.7(f.o.R_g c_{int} l_{ave} + f.o.R_g C_g + r_{int} C_g l_{ave})] \\ &\quad +1.1 r_{int} c_{int} l_{chip-edge}^2 + \frac{l_{chip-edge}\sqrt{\epsilon_r}}{c_o} \end{aligned} \tag{2.12}$$

where f_{ld} is the logic depth, r_{int} and c_{int} are interconnect's resistance and capacitance per unit length, R_g is the gate output resistance, C_g is the gate input capacitance, l_{ave} is the average wire-length, $l_{chip-edge}$ is the length of a chip-edge length wire, and $f.o.$ is the average fan-out. The first term in cycle time equation (Eq. 2.12) takes into account delay through f_{ld} stages of logic gates. The second term in Eq. 2.12 is the delay of a chip-edge length global wire, and the third term is its time-of-flight delay. In a full-custom design, designers try to avoid long wire delay in the critical path by careful and customized routing and placement of key macro cells; however, in a non-custom design, implemented in ASICs or FPGAs, the critical path may include long wires. The delay due to a long wire in the critical path can be reduced by inserting optimum even number of repeaters based any of the repeater insertion criteria discussed earlier.

2.2.4 Stochastic Wire-Length Distribution of 2-D Integrated Circuits

The estimation of wiring requirements (total and average wire-length) is the key element in system-level modeling. The chip area, clock frequency, and power consumption of an IC are largely determined by the wiring requirements [49, 15]. The number of interconnect levels, wiring pitch, etc. in logic circuits are determined such that the total wiring need between logic gates can be accommodated within the available wiring tracks. On the other hand, wiring need in memory elements such as SRAM and DRAM is much smaller compared to that of logic circuits, and their chip area is device-limited.

In this sub-section, the derivation of wire-length distribution of 2-D ICs, based on J. Davis' methodology [17], will be provided. The wire-length distribution along with interconnect delay criteria can be used to estimate the chip area, wiring pitch, etc. of ICs accurately [53]. Though there are many published models for estimating average and total wire-length in 2-D ICs [15, 62, 63], knowing the complete wiring histogram or wire-length distribution provides a better insight into selecting interconnect parameters such as width, spacing, and aspect ratio. Also, the wire-length distribution can be used to optimize interconnect parameters to achieve minimum chip area or maximum clock frequency [53]. The number and size of repeaters for long interconnects can also be estimated using the wire-length distribution of semi-global and global interconnects.

Rent's Rule

The analytical modeling of wire-length in ICs has been a very popular topic in the physical design community for many years [62, 64, 65, 66]. An earlier work on stochastic wire-length prediction showed that wire density emanating from logic blocks in 2-D gate arrays followed Poisson distribution [64]. Wire-length prediction based on fractal analysis of interconnection complexity has also been proposed [66]. An alternative approach for wire-length prediction that has gained considerable amount of popularity is based on a well established empirical relationship, commonly known as Rent's rule. By applying Rent's rule iteratively to a 2-D array of random logic network, models for average wire-length and wire-length distribution can be derived [17, 62, 65].

Rent's Rule describes the interconnection complexity as a function of the module or system size in well-partitioned designs. It is modeled by a relationship between the number of logic gates, N, within a module and the number of interconnection terminals, T, of the module. It is given by [67]

$$T = kN^p, \tag{2.13}$$

where k is the average number of terminals per logic gate, and Rent's exponent, p ($0 \leq p \leq 1$), is a constant for a given logic graph. Rent's exponent is a measure of interconnection complexity of a design, and reported values of p are in the range of 0.12 to 0.8 [49, 67, 68]. In Fig. 2-8, interconnection complexity in various circuits, exhibiting lower and upper bounds of Rent's exponents and a typical value, is illustrated. The first circuit performs a serial operation and the number of I/O terminals is independent of the number of logic gates in the module (i.e. $T = kN^0 = k$). The interconnection complexity in the second circuit can be represented by Rent's exponent $0 < p < 1$. The third circuit performs operations in parallel, and the number of I/O terminals, $T = kN^1 = kN$.

Generally, memory circuits (SRAMs or DRAMs) are associated with smaller values of Rent's exponent, and logic circuits are associated with higher values of Rent's exponent [49]. Various approaches for implementing a digital using full-custom, semi-custom, FPGA, etc. may also result in different Rent's exponents [49]. Typically, Rent's exponent is smaller for full-custom designs and larger for FPGA based designs.

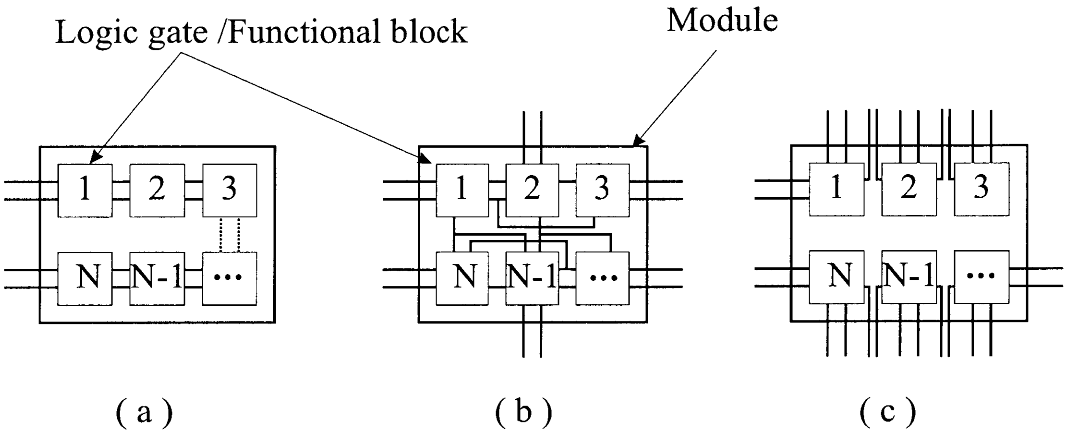


Figure 2-8: Interconnection complexity in three circuits exhibiting Rent's exponent $p = 0$ in (a), $0 < p < 1$ in (b), and $p = 1$ in (c).

Landman and Russo provided the early compelling evidence of Rent's Rule by partitioning logic graphs in scientific computers into modules and observing a power law relationship between the number of pins (I/O terminal) and number of gates per module [67]. Bakoglu also also examined a variety of circuits including microprocessor, ASIC, memory, and gate array chips, and found similar power law relationship, as shown in Fig. 2-9. Though Rent's rule was originally established as an empirical relationship [67], later it was suggested to be a consequence of logic design practices. Using principal of self-similarity and fractal analysis [69], it has been suggested that the existence of Rent's rule is a direct consequence of hierarchical approaches to logic design [70, 62]. Using fractal analysis and associating fractal dimension to a design, theoretical values of Rent's exponent have also been sug-

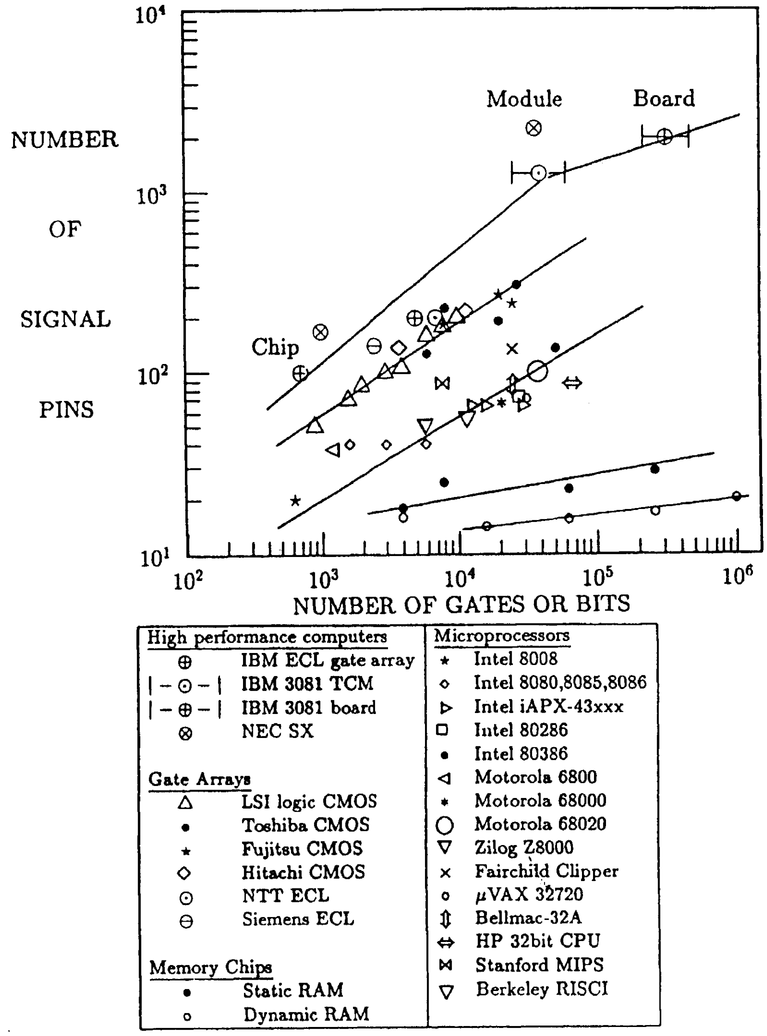


Figure 2-9: Correlation between the number of signal pins (I/O terminals) and the number of logic gates for various digital systems [49].

gested [70, 62, 66, 71].

Derivation

In J. Davis' methodology, wire-length distribution of a 2-D IC is derived by applying Rent's rule iteratively throughout an entire monolithic system of N logic gates [17]. It is assumed that the same Rent's parameters can be applied within a subset of N logic gates. Although an IC may be composed of sub-circuits with different Rent's parameters, equivalent Rent's parameters can be found to describe the interconnection complexity of the system [72], and J. Davis' methodology can be used to estimate the wire-length distribution.

To derive the point-to-point wire-length distribution, $f_{2D}(l)$, of a 2-D IC with N_t transistors, the integrated circuit is partitioned into N logic gates, where $N = N_t/\phi$; ϕ is a function of the average fan-in and fan-out in the system [49]. The average separation between adjacent logic gates is called gate pitch, and it is equal to $\sqrt{A_c/N}$, where A_c is the chip area.

The complete wire-length distribution is determined by superimposing the wire-length distribution of all logic gates without double counting. To illustrate the procedure for estimating wiring requirements of a single logic gate, consider the corner element of a square array of logic gates as shown in Fig. 2-10. The gates in Fig. 2-10 are grouped into three distinct but adjacent blocks, A, B, and C. A single closed path can encircle one, two, or all three of these logic blocks. Number of logic gates in A, B, and C are given by N_a, N_b, and N_c, respectively. The number of interconnects between logic gates in A and C can be found by conserving all I/O terminals for logic gates in A, B, and C.

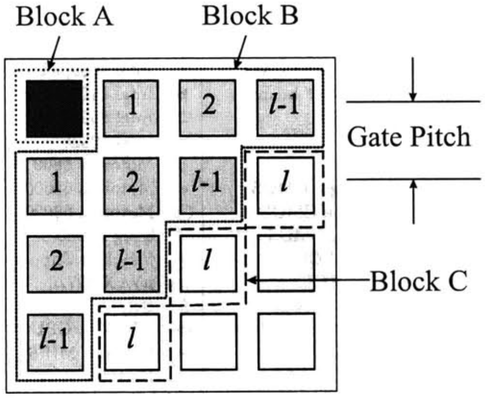


Figure 2-10: The procedure for estimating the wire-length distribution in 2-D ICs.

For example, by applying the principle of conservation of I/O terminals to the system shown in Fig. 2-10 we find,

$$T_A + T_B + T_C = T_{A-to-C} + T_{A-to-B} + T_{B-to-C} + T_{ABC}, \qquad (2.14)$$

where the variables are defined in Table 2.2. By applying the conservation of I/O terminals

Variable	Definition
T_A	# of I/O's of block A
T_B	# of I/O's of block B
T_C	# of I/O's of block C
T_{A-to-B}	# of I/O's connecting block A to B
T_{A-to-C}	# of I/O's connecting block A to C
T_{B-to-C}	# of I/O's connecting block B to C
T_{AB}	# of I/O's of block $A+B$
T_{AC}	# of I/O's of block $A+C$
T_{BC}	# of I/O's of block $B+C$
T_{ABC}	# of I/O's of block $A+B+C$

Table 2.2: Definition of variables for conservation of I/O terminals

between adjacent blocks, we also find

$$\begin{aligned} T_{A-to-B} &= T_A + T_B - T_{AB} \\ T_{B-to-C} &= T_B + T_C - T_{BC}. \end{aligned} \tag{2.15}$$

Substituting Eq. 2.15 in Eq. 2.14, we can estimate the number of I/O terminals for making interconnections between logic gates in A and C [17],

$$\begin{aligned} T_{A-to-C} &= T_{AB} - T_B + T_{BC} - T_{ABC} \\ &= k[(N_a + N_b)^p - (N_b)^p + (N_b + N_c)^p - (N_a + N_b + N_c)^p]. \end{aligned} \tag{2.16}$$

To calculate the number of interconnections between logic gates in A and C from Eq. 2.16, we define a variable α_f ($1/2 \leq \alpha_f < 1$). α_f is defined in terms of average fan-out in the system, f.o., as [62]

$$\alpha_f = \frac{f.o.}{f.o. + 1}. \tag{2.17}$$

By multiplying the number of I/O terminals for interconnections between logic gates in A and C by α_f, the total number of point-to-point interconnections can be found, and it is given by

$$I_{A-to-C} = \alpha_f k[(N_a + N_b)^p - (N_b)^p + (N_b + N_c)^p - (N_a + N_b + N_c)^p]. \tag{2.18}$$

In Fig. 2-10, the complete stochastic wire-length distribution for the corner element A can be found by tabulating Eq. 2.18 for all values of l ($1 \leq l \leq 2\sqrt{N} - 2$), in Fig. 2-10. This process is repeated for all logic gates, avoiding multiple counting. By superimposing the wire-length distribution for individual gates, the wire-length distribution of the entire system can be found.

An analytical expression for stochastic wire-length distribution can be derived by making several approximations. If we assume average partitioning strategies are similar to partial Manhattan circle as shown in Fig. 2-11, we can find simplified expressions for N_b and N_c. Using the partial Manhattan circle approximation in Fig. 2-11, $N_a = 1$, $N_b = l(l-1)$, and $N_c = 2l$. Equation 2.18 along with the simplified expressions for N_a, N_b, and N_c can be used to calculate the expected average number of interconnects between a gate pair separated by length l in a partial Manhattan circle, and it is given by

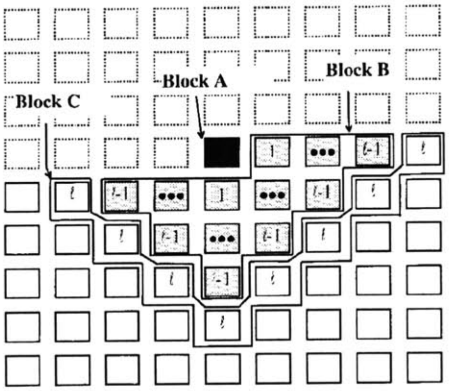


Figure 2-11: A partial Manhattan circle partitioning strategy for estimating N_a, N_b, and N_c [17].

$$\begin{aligned} I_{exp-2D}(l) &= \frac{\alpha_f k}{2l}[(1+l(l-1))^p - (l(l-1))^p + (l(l+1))^p - (1+l(l+1))^p] \\ &\simeq \alpha_f k \frac{p}{2}(2-2p)^{(2p-4)}. \end{aligned} \quad (2.19)$$

To complete the derivation of stochastic wire-length distribution, number of gate pairs $M_{2D}(l)$ separated by length l must be determined. In a 2-D square array of N gates,

$M(l)_{2D}$ is given by [17]

$$\begin{aligned} M_{2D}(l) &= \frac{l^3}{3} - 2l^2\sqrt{N} + 2Nl \quad & 1 \leq l < \sqrt{N} \\ & \frac{1}{3}(2\sqrt{N} - l)^3 & \sqrt{N} \leq l < 2\sqrt{N} - 2 \end{aligned} \tag{2.20}$$

The complete wire-length distribution function is given by

$$f_{2D}(l) = \Gamma_{2D} M_{2D}(l) I_{exp-2D}(l), \tag{2.21}$$

where Γ_{2D} is a normalization constant. $f_{2D}(l)$ is normalized such that

$$\Gamma_{2D} = \frac{I_{total}}{\sum_{l=1}^{2\sqrt{N}-2} M_{2D}(l) I_{exp-2D}(l)}, \tag{2.22}$$

where I_{total} is the total number of interconnects in a system. I_{total} can be found by partitioning a system or IC into many progressively smaller hierarchical levels and estimating the number of interconnects at each hierarchical level using Rent's Rule, and it is given by [62]

$$I_{total} = \alpha_f k N (1 - N^{p-1}). \tag{2.23}$$

The total and average point-to-point wire-length, L_{total} and l_{ave}, can be determined directly from the stochastic wire-length distribution function, and they are given by

$$\begin{aligned} L_{total} &= \sum_{l=1}^{2\sqrt{N}-2} l f_{2D}(l) \\ l_{ave} &= \frac{L_{total}}{\sum_{l=1}^{2\sqrt{N}-2} f_{2D}(l)}. \end{aligned} \tag{2.24}$$

To make a more realistic wire-length prediction, point-to-point wire-length must be converted to an equivalent net length. This is achieved by multiplying L_{total} and l_{ave} by a correction factor, χ. The value of χ depends on net models and average fan-out [17]. For example, for the linear net model as shown in Fig. 2-12, total point-to-point wire-length is $(2l + 3l + \cdots + (f.o. + 1)l)$ and the equivalent net-length is $2l(f.o.)$, resulting in $\chi = \frac{4}{f.o.+3}$.

The derivation of stochastic wire-length distribution has been verified by J. Davis using

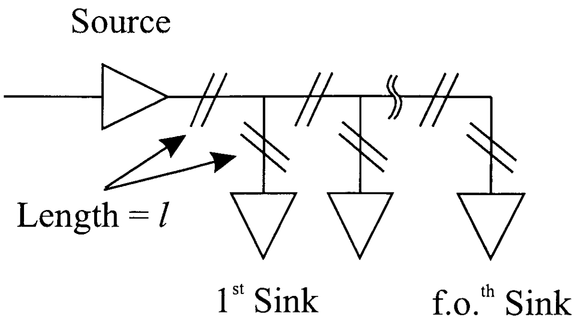


Figure 2-12: Linear net model for converting point-to-point wire-length to net length.

wire-length data for various microprocessors [17]. The comparison between actual data and estimated wire-length distribution is shown in Fig. 2-13. sylvester:Globwire

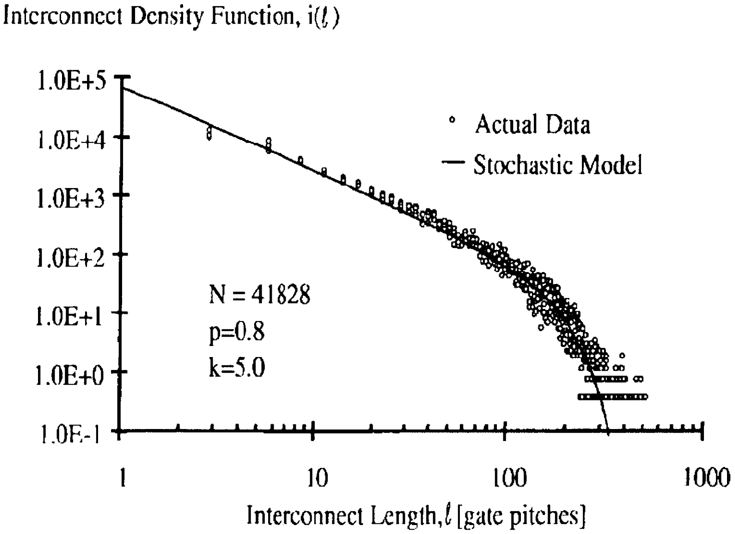


Figure 2-13: Comparison of stochastic model and actual data of wire-length distribution for a microprocessor [17].

2.2.5 Models for Estimating Chip Area

The stochastic wire-length distribution can be used to estimate the wiring-limited chip area of logic circuits accurately [50]. In wiring-limited ICs, chip area, A_c, depends on the wiring pitch, p_w, number of interconnect levels, m, wiring efficiency (i.e. the utilization efficiency of wiring tracks), e_w, and the total wire-length, L_{total}. A first order approximation of chip area in terms of p_w, m, e_w, and L_{total} is given by [49]

$$A_c \simeq \frac{p_w L_{total}}{e_w m} \tag{2.25}$$

If the wiring pitch is increased, there are fewer wiring tracks available per interconnect level per unit area. As a result, m and/or A_c have to be increased for complete wireability. On the other hand, if the number of interconnect levels is increased with everything else being constant, there are more wiring tracts available per unit area. As a result, complete wireability can be achieved with a smaller chip area. In general, wiring-limited chip area can be reduced by making the wiring pitch smaller or increasing the number of interconnect levels until the chip area is not wiring-limited any more, and it is determined by the device-limited area.

Though Eq. 2.25 serves as a good starting point for estimating the chip area, it does not reflect some of the commonly used design practices in modern VLSI design. For example, multi-level interconnect structures consist of different wiring pitches: smaller wiring pitch for short wires and larger wiring pitch for intermediate and long wires [15]. Also, the wiring efficiency in different interconnect levels may not be the same [15]. In the following sections, a refined chip area model will be presented.

Wiring Efficiency

Wiring efficiency, e_w, represents the effective utilization of wiring tracks in each interconnect level. In a multi-level interconnect structure, vias are used to establish connections between metal wires on different interconnect levels and to contact transistor terminals. The blockage of wiring tracks due to vias, as shown in Fig. 2-14, can reduce the wiring utilization. A fraction of the layout area is also dedicated to route power, ground, and clock signals [15] and not available for routing signal wires.

The simplest model to represent the effective utilization of wiring tracks is to assume a constant wiring efficiency in all interconnect levels [49, 53]. Though this assumption may be reasonable for estimating chip area of ICs with a few interconnect levels, it may

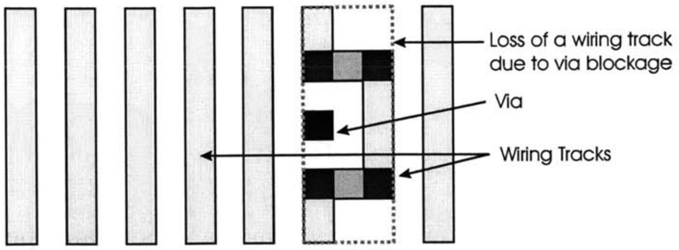


Figure 2-14: Illustration of reduction in wiring efficiency due to via blockage.

not be suitable for ICs with more than 6-7 interconnect levels where the value of e_w can vary significantly on different interconnect levels. As the number of interconnect levels is increased, the bottom most interconnect levels (local and semi-global wiring levels) are effected more than the top interconnect levels due to via-blockage.

G. S. Halasz, developed a model to estimate the wiring efficiency of interconnect levels based on his observation of design practices in IBM microprocessors [15]. In his model, it is assumed that power and ground signals take up 20% area or wiring tracks in each interconnect level. If all wiring pitches are identical, vias between 1^{st} and n^{th} interconnect levels reduce the wiring efficiency of $2^{\mathrm{nd}}, 3^{\mathrm{rd}}, \ldots., (\mathrm{n}-1)^{\mathrm{th}}$ levels by $12\% - 15\%$. If wiring pitches are not identical in adjacent interconnect levels, wiring efficiency in the lower interconnect level is generally higher compared to the scenario where wiring pitches are identical; for such cases, we assume the via blockage is scaled by the ratio of wiring pitches. We also assume the routing tools utilize 75% of the available routing area [73]. Based on these assumptions, average global, semi-global, and local interconnect level wiring efficiencies, e_{wg}, e_{wsg}, and e_l, in a three-tier interconnect architecture are given by

$$\begin{aligned} e_{wg} &= 0.8 \times 0.75[1 + (1-e_v) + \cdots + (1-e_v)^{m_g}]/m_g \\ e_{wsg} &= 0.8 \times 0.75[(1 - \frac{p_{sg}}{p_g} e_v)^{m_g} + (1-e_v)^{m_g+1} + \cdots + (1-e_v)^{m_g+(m_{sg}-1)}]/m_{sg} \\ e_l &= 0.8 \times 0.75[(1 - \frac{p_l}{p_{sg}} e_v)^{m_g+m_{sg}} + (1-e_v)^{m_g+m_{sg}+1} + \\ & \quad \cdots + (1-e_v)^{m_g+m_{sg}+(m_l-1)}]/m_l, \end{aligned} \tag{2.26}$$

where e_v is the via blockage factor, and m_g, m_{sg}, and m_l are the number of global, semi-global, and local interconnect levels and p_g, p_{sg}, and p_l are their wiring pitches, respectively.

Based on Eq. 2.26, wiring efficiencies in six and nine level interconnect structures with equal number of local, semi-global, and global wiring levels are shown in Fig. 2-15. It is assumed the wiring pitches are $2F$, $4F$, and $8F$, respectively, where F is the minimum feature size, and $e_v = 0.15$. As illustrated in Fig. 2-15, local interconnect levels are effected the most due to via blockage because vias from all semi-global and global interconnect levels to device layer create blockage in local interconnect levels.

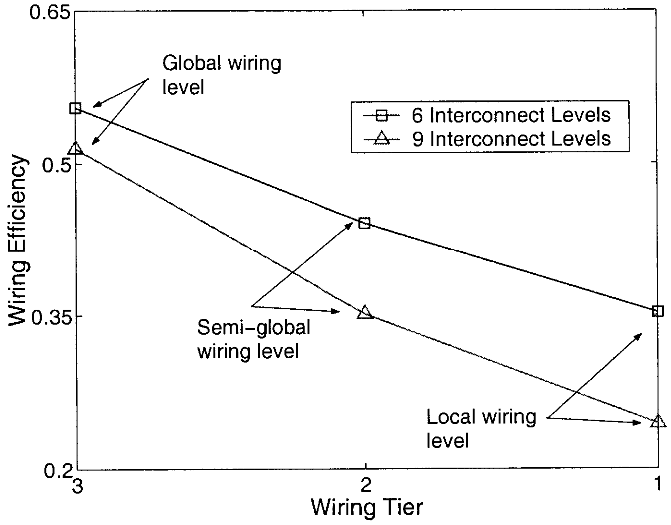


Figure 2-15: The average wiring efficiencies in global, semi-global, and local interconnect tiers, represented by tier 3, 2, and 1 respectively.

Conservation of Chip Area

Once the wiring efficiency, number of interconnect levels, and total wire-length are known, chip area and wiring pitches can be found such that the available layout area for local, semi-global, and global interconnects are larger than or equal to their required layout area. The required layout area for local, semi-global, and global interconnects, A^l_{req}, A^{sg}_{req}, and A^g_{req}, and their respective available area, A^l_{avail}, A^{sg}_{avail}, and A^g_{avail} are given by

$$A^l_{req} = \chi\sqrt{\frac{A_c}{N}} \cdot p_l \sum_{1}^{l_{lmax}} lf(l), \qquad A^l_{avail} = A_c \cdot m_l e_{wl}$$

$$A^{sg}_{req} = \chi\sqrt{\frac{A_c}{N}} \cdot p_{sg} \sum_{l_{lmax+1}}^{l_{sgmax}} lf(l), \qquad A^{sg}_{avail} = A_c \cdot m_{sg} e_{wsg}$$

$$A^g_{req} = \chi\sqrt{\frac{A_c}{N}} \cdot p_g \sum_{l_{sgmax}+1}^{l_{gmax}} lf(l), \qquad A^g_{avail} = A_c \cdot m_g e_{wg} \tag{2.27}$$

where χ is the point-to-point to net-length conversion factor. e_{wl}, e_{wsg}, and e_{wg} are the average wiring efficiencies of local, semi-global, and global interconnect levels; m_l, m_{sg}, and m_g are the number of local, semi-global and global interconnect levels and p_l, p_{sg} and p_g are their respective wiring pitches. l_{lmax}, l_{sgmax}, and l_{gmax} are the length of the longest local, semi-global, and global interconnects and they can be found using Eq. 2.28.

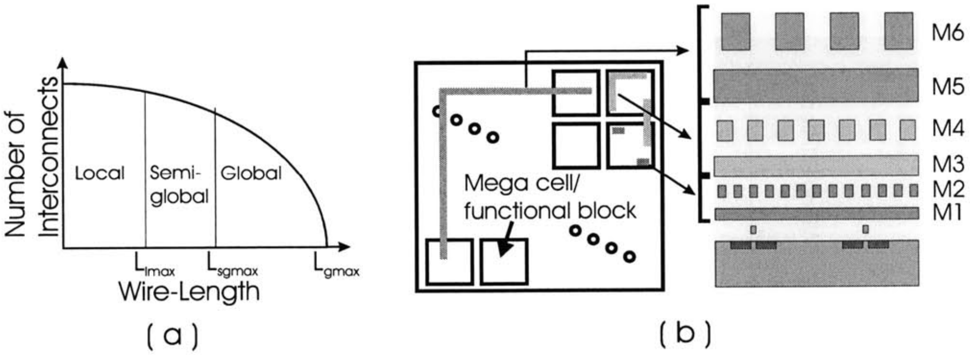


Figure 2-16: (a) Partitioning of the wire-length distribution into local, semi-global, and global regions. (b) Illustration of assigning interconnect levels based on wire length.

The length of the longest local, semi-global, and global wires, l_{lmax}, l_{sgmax}, and l_{gmax} to partition the wire-length distribution (see Fig. 2-16(a)) can be found using interconnect delay criteria. Typically, local wiring levels are used to route wires across several logic gates. In a design, synthesized with 30K- to 100K-gate mega cells, semi-global wiring levels can be used to route wires across their (mega cell's) semi-perimeter [18]. Global wiring levels are generally used to route signal wires between mega cells. Considering a design in 0.18 μm technology generation and synthesized with 100K-gate mega cells, as shown in Fig. 2-16(b), we estimate the ratio of interconnect delay of the longest semi-global wire and the clock period $\beta_{sg} \sim 15\%$. For simplicity, we also assume a similar delay constraint, $\beta_l \sim 15\%$, for the longest local wire. Without more detailed knowledge of a system, it is difficult to provide a better model to partition the wire-length distribution.

The wiring pitch and the maximum wire-length in any tier (local, semi-global, or global) can be found by solving the interconnect delay equation (Eq. 2.2) and assuming $R_{int} \sim R_g$.

The wiring pitch in any tier, p_w, and length of the longest wire in that tier, l_{max} (in gate pitches), are related to the interconnect parameters by the following equations:

$$\begin{aligned} p_w &= 2l_{max}\sqrt{\frac{A_c}{N}}\sqrt{\frac{1.1[(4.07+2\gamma)/\gamma]\epsilon_r\epsilon_o\rho}{\frac{\beta}{f_c}-L_{max}\sqrt{\frac{A_c}{N}}\frac{\sqrt{\epsilon_r}}{c_o}}} \\ l_{max} &= \frac{p_w^2}{2[4.4(4.07+2\gamma)/\gamma]\epsilon_r\epsilon_o\rho}\sqrt{\frac{N}{A_c}} \\ &\cdot\left[-\frac{\sqrt{\epsilon_r}}{c_o}+\sqrt{\frac{\epsilon_r}{c_o^2}+\frac{4\beta}{f_c}\cdot[4.4(4.07+2\gamma)/\gamma]\epsilon_r\epsilon_o\rho}\right], \end{aligned} \tag{2.28}$$

where β is ratio of interconnect delay of l_{max} gate pitch long wire and clock period; f_c is the clock frequency, ρ is the effective resistivity of the metal, γ is the aspect ratio of interconnects, c_o is the speed of light in free space, and ϵ_r is the relative dielectric constant of the ILD. It should be noted that in Eq. 2.28, it has been assumed that $\gamma > 1$ and wiring dimensions: $H_\rho = H_\epsilon$ and $W_\rho = W_\epsilon$.

In our analysis local wiring pitch, p_l, is chosen to be twice the minimum feature size. Semi-global wiring pitch, p_{sg}, is estimated to optimizing the chip area or clock frequency [53]. The global wiring pitch can be found such that the interconnect delay of a chip-edge length global wire is a fraction β_g of clock period/cycle time. In logic networks implemented in 0.18 μm technology generation, we find that chip-edge length wire delay corresponds to 30% – 40% of the critical path delay when repeaters are not used and $\sim$ 15% with optimum number of repeaters. In 100 nm technology node, interconnect delay of a long wire deteriorates significantly and based on Eq. 2.12, we find that long wire delay (without repeater insertion) can account for 80% – 90% of the cycle time.

2.3 Examples: Chip Area and Clock Frequency Estimation

Based on the methodology presented in [53], by solving Eq. 2.4, 2.27, and 2.28, semi-global and global wiring pitches can be found that optimize the chip area or clock frequency. Generally, the chip area can be reduced by making the semi-global and global wiring pitch smaller. However, reduction in p_{semi} shifts the partition between semi-global and global interconnects, l_{sgmax}, to the left and increases the total global wire-length that has to be routed in global wiring levels. Initially this measure may reduce the chip area; however, at some point the chip area begins to increase to accommodate the global wiring need. Generally, there exists a semi-global wiring pitch that minimizes the chip area in a three-tier wiring architecture [53].

Based on the models used in system-level analysis, there also exists an optimum value of clock frequency [53]. It (clock frequency) can be improved by increasing the wiring pitch. However, an increase in wiring pitch also results in an increase in chip area and wire-length. At some point, the improvement due to widening the wiring pitch is offset by an increase in delay due to longer wire-length. As a result, there is an optimum value of clock frequency. The concept of minimum chip area and optimum clock frequency has been investigated by J. Davis, and it is illustrated in Fig. 2-17.

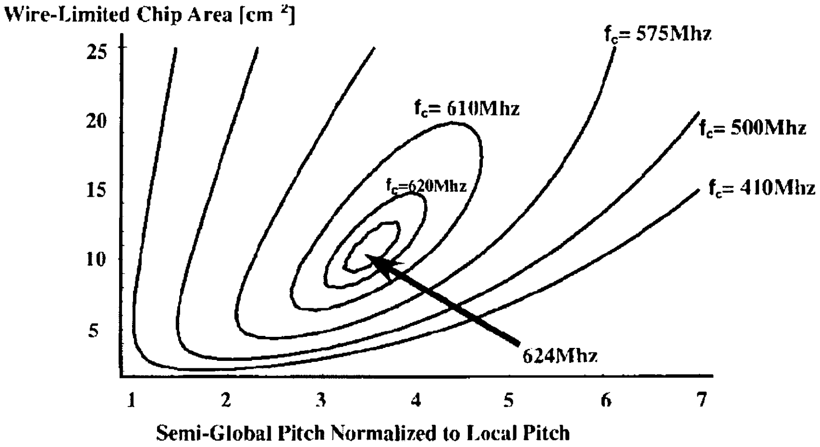


Figure 2-17: Clock frequency, f_c, versus chip area and semi-global wiring pitch of a wiring-limited random logic network implemented in 0.18 μm technology node [53]. Semi-global wiring pitch is varied to optimize the chip area or clock frequency. The input parameters are consistent with device and interconnect technologies listed in the SIA and ITRS roadmap [2, 7].

2.4 Extension of System-Level Modeling to 3-D ICs

The methodology described in this chapter, to estimate key performance metrics of 2-D ICs, can be extended easily to assess the benefits of 3-D integration. However, new models and refinement of existing models for wiring requirements, wiring efficiency, etc. will be needed. Some additional assumptions will have to be made regarding the cost/complexity for vertically integrating two wafers. The impact of additional (routing or delay) cost associated with inter-stratum connection on wire-length will also have to be included.

In the next chapter, extension of system-level modeling to 3-D ICs will be presented. A model for stochastic 3-D wire-length distribution will be derived, and wiring requirements in 2-D and 3-D integration will be compared. By decomposing the wire-length distribution into inter- and intra-stratum components, inter-stratum via density will be estimated. Various trade-off analyses to minimize the wiring-limited chip area or improve the clock frequency will be presented.

Chapter 3

System-Level Performance Modeling of Three-Dimensional (3-D) Integrated Circuits (ICs)

Over the years, there have been significant research efforts on system-level performance modeling of 2-D ICs [49, 50, 51, 52]. However, applications of system-level models to assess the impact of integrating new technologies such as 3-D or optical interconnects on system performance have not been explored. In some of the earlier works, asymptotic dependencies of average wire-length and wire-length distribution on system size were examined [70, 71, 74]. However, the analytical models based on those studies were not in suitable forms for integration with system-level modeling framework. For example, in [74] A. L. Rosenberg derives asymptotic upper and lower bounds of area and volume for 2-D and 3-D systems as a function of the number of blocks or modules in the system. However, his models don't take into account the impact of scaling interconnect parameters or integration of additional interconnect levels on system's size. In [70], D. Stroobandt extends Donath's methodology [62] to estimate wire-length distribution of 3-D ICs. However, only 3-D ICs with a cubic symmetry ($N_z = N^{1/3}$) are considered, where a realistic 3-D integrated circuit is likely to have a few strata with non cubic symmetry. Though the models for average wire-length in [63], based on hierarchical partitioning, can be integrated in system-level modeling framework, it is known that these models predict an upper bound of wire-length which can be as much as a factor of two higher than the measured wire-length [63]. Besides development of analytical models for wire-length prediction, there have been also some research efforts in developing routing algorithms and interconnection networks for 3-D ICs [75, 76]. However, implications of these physical design innovations on overall system performance and technology requirements were not discussed. In order to assess the impact of 3-D technologies, it is desirable to predict the wire-length more accurately and also be able to integrate wire-length estimation models easily in the system-level modeling framework. Based on extensive modeling and analysis, we find that to compare key performance metrics such as interconnect delay or chip area of 2-D and 3-D ICs, accurate estimations of both wire-length and wiring pitches are necessary. This can be achieved only by system-level modeling and analysis where the dependencies between chip area, device and interconnect parameters, delay, etc. are taken into account in a consistent way.

In this chapter, J. Davis' methodology for estimating the wire-length distribution is extended to 3-D ICs. The 3-D stochastic wire-length distribution model along with models for 3-D wiring efficiency, cost function, etc. are integrated in the system-level modeling framework to evaluate key performance metrics of 3-D ICs and to perform trade-off analysis. Various scenarios of inter-stratum connectivities are also examined by comparing the average and total wire-length. By resolving the 3-D wire-length distribution into intra- and inter-stratum components, the density of inter-stratum interconnects is projected and methodologies for reducing the inter-stratum via density are also discussed.

3.1 Stochastic Wire-Length Distribution of Three-Dimensional Integrated Circuits

3.1.1 Approaches to Wire-Length Prediction in 3-D ICs

The estimation of wiring requirements is a key component in system-level modeling and analysis. Most of the analytical models for average wire-length or wire-length distribution are based on either hierarchical or non-hierarchical partitioning [17, 70, 77, 62, 63]. In *hierarchical partitioning*, an IC is partitioned into progressively smaller hierarchical levels until the lowest level of the hierarchy consists of a single gate or a logic block. The wiring requirement at each hierarchical level is superimposed to estimate the total wire-length or wire-length distribution [70, 62]. The hierarchical partitioning for wire-length prediction represents the hierarchical approaches for partitioning and placement of a digital design, where groups of logic gates are interconnected to form a functional block or megacell, and many megacells are interconnected to form an integrated system. The interconnections within a megacell and between megacells are considered to be in different hierarchical levels.

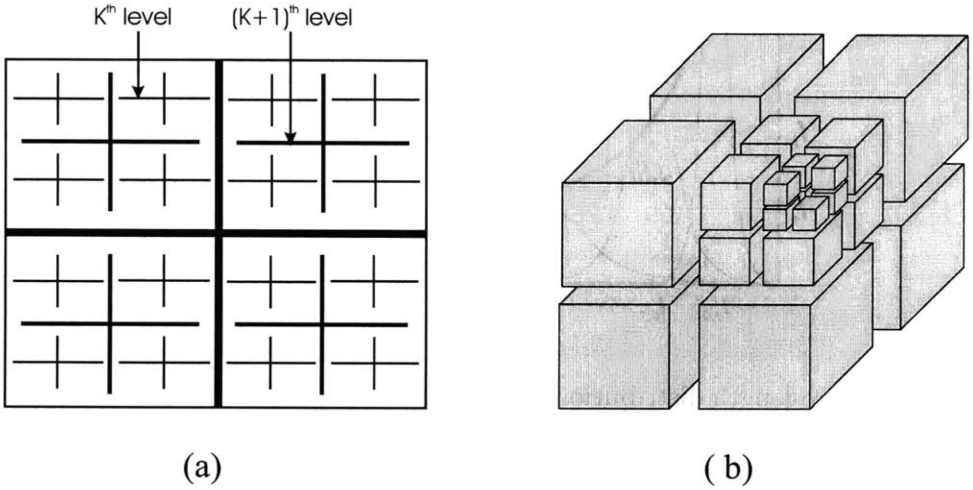


Figure 3-1: Hierarchical partitioning on the placement locations for 2-D and 3-D ICs.

The hierarchical wire-length prediction methodology is illustrated in Fig. 3-1. Using this methodology a 2-D or 3-D IC with N logic gates is partitioned in L hierarchical levels, where $\mathrm{L} = \mathrm{logN/log4}$ or $\mathrm{logN/log8}$, respectively. In a 3-D cubic implementation, interconnections in k^{th} hierarchical level consists of connections between groups of size 8^k excluding the connections between groups of size 8^{k+1}. The wire-length distribution at each hierarchical level is superimposed to estimate the total wire-length distribution as shown in Fig. 3-2. Though in the original methodology, a symmetric (square or cubic) implementation of an IC is partitioned recursively into smaller symmetric modules, alternative portioning schemes such as row- and plane-wise partitioning can also be considered [63]. However, it has been found that row- or plane-wise partitioning results in longer average wire-length [63]. Though wire-length prediction techniques based on hierarchical partitioning are useful for predicting average and total wire-length, they tend to predict an upper bound of wire-length [62, 63]. This upper bound can vary by as much as a factor of 2× compared to measured data [62, 63]. Also, a realistic design may not be hierarchical down to gate level. As a result the number of hierarchical levels $L^{'}$ will be less than L which may effect the accuracy of wire-length prediction.

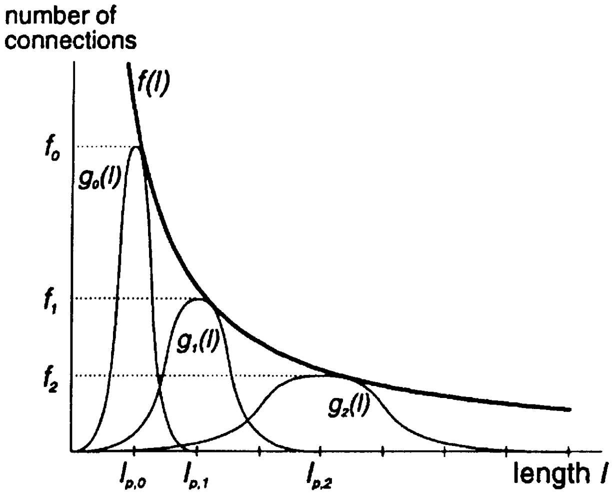


Figure 3-2: The wire-length distribution based on hierarchical partitioning [70]. $g_0(l)$, $g_1(l)$, and $g_2(l)$ are the wire-length distribution of 0^{th}, 1^{st}, and 2^{nd} hierarchical levels, respectively. The shorter wires generally belong to lower hierarchical levels, whereas the longer wires belong to upper hierarchical levels.

To avoid ambiguities in predicting the number of hierarchical levels and the granularity of a hierarchical level's size, *non-hierarchical* partitioning can be used for wire-length prediction [17, 77, 71]. In this approach, the partitioning and placement of a design in considered to be flat or non-hierarchical. An IC is modeled as an array of randomly placed homogeneous logic gates or modules, where the interconnection complexity for the entire system or sub-system is governed by the same Rent's parameters. The distribution of point-to-point connection lengths is a uniform and isotropic function which is determined by applying Rent's rule. The derivation also assumes an infinite extension of a 2-D array of logic gates neglecting the edge effects due to finite boundaries [77]. The procedure for deriving the wire-length distribution based on non-hierarchical partitioning is illustrated in Fig. 3-3.

J. Davis' methodology, presented in Chapter 2, is a refinement of existing non-hierarchical

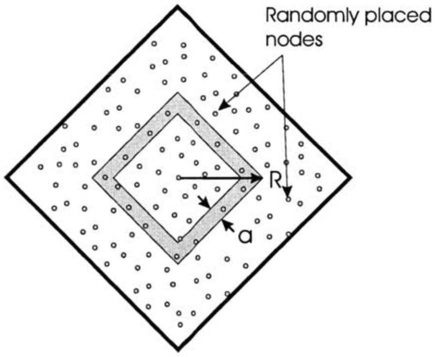


Figure 3-3: The procedure for estimating wire-length distribution based on non-hierarchical partitioning [77]. By applying Rent's rule, the number of interconnects between a node and a group on nodes within the band of radius R and area $\sim aR$ can be found. This process can be repeated for all possible values of R to derive the complete wire-length distribution.

approaches to wire-length prediction. As described in the Chapter 2, it incorporates both edge effects and conservation of I/O terminals to predict wire-length more accurately. Generally, both hierarchical and non-hierarchical approaches to wire-length prediction result in similar scaling behavior (i.e. the number and length of interconnects as a function of system's size). However, based on a survey of published results, non-hierarchical approaches tend to predict wire-length or wire-length distribution more accurately [17, 62, 63]. Extension of J. Davis' methodology to 3-D ICs for wire-length prediction [78, 79] has been a natural choice in our system level modeling and analysis work.

3.1.2 Derivation

In this section, J. Davis' methodology for wire-length prediction will be extended to 3-D ICs. We consider a 3-D IC with N_z strata and N logic gates. The vertical separation between adjacent strata is called stratal pitch, t_z, and it is represented in units of gate pitches. Rent's parameters, to a large extent, depend on the design itself and not on the implementation. If the placement of a design in 2-D and 3-D is optimal, differences between Rent's parameters in both implementations are likely to be very small. So, we assume same Rent's parameters are applicable to both 2-D and 3-D integration of an IC.

Similar to Eq. 2.21, we define the discrete wire-length distribution in 3-D ICs as

$$f_{3D}(l, t_z) \quad = \quad \Gamma' M_{3D}(l, t_z) I_{3D}(l, t_z) \qquad 1 \leq l \leq 2\sqrt{\frac{N}{N_z}} - 2 + (N_z - 1)t_z \tag{3.1}$$

where Γ' is a normalization constant, $M_{3D}(l, t_z)$ is the number of gate pairs in a 3-D IC (with t_z stratal pitch) at l Manhattan distance apart, and $I_{3D}(l, t_z)$ is the number of interconnections between these gate pairs. The derivation of wire-length distribution of 3-D ICs is illustrated in Fig. 3-4. Similar to 2-D ICs, $N_a = 1$, N_c is the number of logic gates l gate pitch apart from N_a, and N_b is the number of logic gates in between N_a and N_c. In 3-D ICs, N_b and N_c include logic gates located on multiple strata. The normalization constant Γ' is found such that the total number of interconnections, $I_{tot} = \sum_{l=1}^{l_{max}} f_{2D}(l) = \sum_{l=1}^{l_{max'}} f_{3D}(l, t_z)$, is conserved. In 3-D ICs,

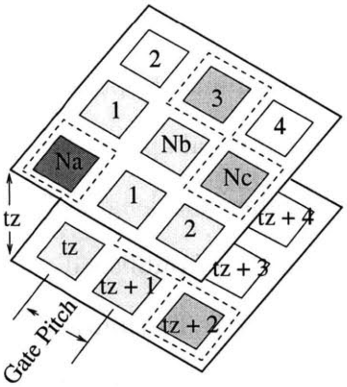


Figure 3-4: The derivation of 3-D wire-length distribution: $N_a = 1$ is the logic gate under investigation, N_c is the number of target logic gates at Manhattan distance l gate pitch, and N_b is the number of logic gates in between N_a and N_c.

$$M_{3D}(l, t_z) = M_{3D-intra}(l) + M_{3D-inter}(l, t_z), \tag{3.2}$$

where intra-stratum number of gate pairs

$$M_{3D-intra}(l) = N_z M_{2D}(l). \tag{3.3}$$

Inter-stratum number of gate pairs, $M_{3D-inter}(l, t_z)$, can be found by shifting $M_{2D}(l)$ by multiples of t_z and adding up the contribution for all possible gate pair combinations. It is given by

$$M_{3D-inter}(l,t_z) = \sum_{i=1}^{N_z-1} \beta_i M_{2D}(l-it_z)u(l-it_z), \tag{3.4}$$

where $u(l)$ is the unit step function; β_i is a constant that depends on the number of strata and the range of interconnects. In a 3-D integration scheme we define the range of interconnects, r, as the maximum allowable stratal pitch between the source and sink terminals of intra- and inter-stratum interconnects. If the range $r = N_z - 1$, $\beta_1 = 2(N_z - 1)$, $\beta_2 = 2(N_z - 2)$, $\cdots$, $\beta_{N_z-1} = 2$. Similarly, if $r = N_z - j$, $\beta_1 = 2(N_z - 1)$, $\beta_2 = 2(N_z - 2)$, $\cdots$, $\beta_{N_z-j+1} = 0$, $\cdots$, $\beta_{N_z-1} = 0$. For the example shown in Fig. 3-5, $r = N_z - 1$, $\beta_1 = 6$, $\beta_2 = 4$, $\beta_3 = 2$.

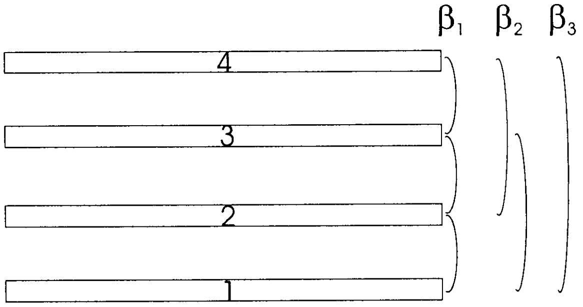


Figure 3-5: Illustration of inter-stratum gate pair combinations for a 3-D IC with four strata. For each combination there are two arrangements. As a result $\beta_1 = 6$, $\beta_2 = 4$, and $\beta_3 = 2$.

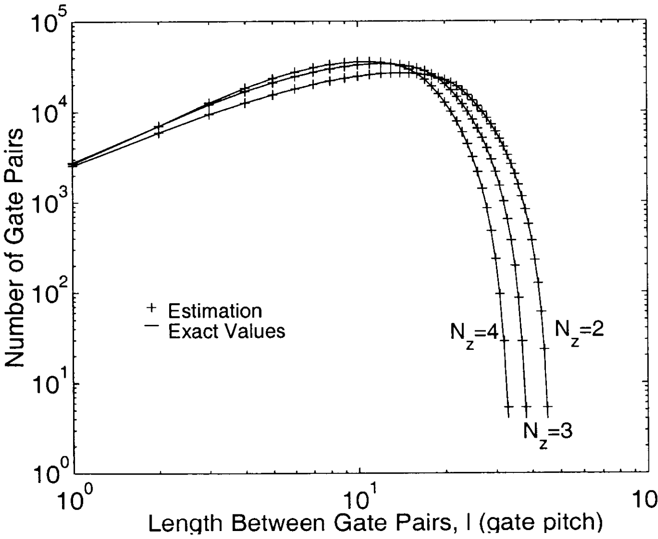


Figure 3-6: Number of gate pairs, $M_{3D}(l, t_z)$, vs. gate pair separation in 3-D ICs with 1000 logic gates and 2, 3 and 4 strata. It is assumed that $t_z = 1$.

In Fig. 3-6, analytically estimated values of $M_{3D}(l,t_z)$ and exact values of $M_{3D}(l,t_z)$, obtained by a computer enumeration, are shown. In estimating $M_{3D}(l,t_z)$, all possible 3-D gate pairs are included, and it is assumed that t_z is one gate pitch.

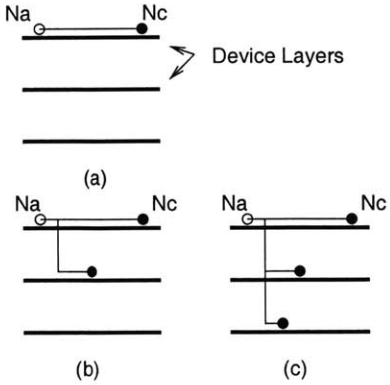


Figure 3-7: Procedure for estimating the average values of N_a, N_b, and N_c in 3-D ICs.

$I_{3D}(l,t_z)$ can be found by using Eq. 3.5.

$$I_{3D}(l,t_z) = \frac{\alpha_f k}{N_c}[(N_a+N_b)^p - (N_b)^p + (N_b+N_c)^p - (N_a+N_b+N_c)^p]. \tag{3.5}$$

Unlike the procedure used for 2-D ICs, values of N_b and N_c are calculated differently. As an example, the procedure for estimating N_a, N_b and N_c in a 3-D IC with three strata (device layers) is illustrated in Fig. 3-7. We observe that r can be restricted to 0, 1, and 2. So, average values of N_b and N_c for $r = 0, 1, \text{and } 2$ are used to estimate $I_{3D}(l,t_z)$. In general, in a 3-D IC with N_z strata, average values of N_b and N_c for $r = 0, 1, ..., N_z - 1$ are used to estimate $I_{3D}(l,t_z)$. This modeling approach also allows us to examine the effect of varying the upper bound of range, r_{upper}, on $I_{3D}(l,t_z)$ and $f_{3D}(l,t_z)$. The average values of N_a, N_b, and N_c, in a 3-D IC with N_z device layers are given in Eq. 3.6:

$$\begin{aligned}
N_a &= 1 \\
N_b(l,t_z) &\simeq (l(l-1) + 2(l-t_z)(l-t_z-1)u(l-t_z) + \cdots \\
&\quad +2(l-(N_z-1)t_z)(l-(N_z-1)t_z-1)u(l-(N_z-1)t_z) + l(l-1) \\
&\quad +2(l-t_z)(l-t_z-1)u(l-t_z) + \cdots + 2(l-(N_z-2)t_z)\cdot \\
&\quad (l-(N_z-2)t_z-1)u(l-(N_z-2)t_z) + \cdots + l(l-1))/N_z
\end{aligned}$$

$$\begin{aligned} N_c(l,t_z) &\simeq (2l + 4(l-t_z)u(l-t_z) + \cdots + 4(l-(N_z-1)t_z)u(l-(N_z-1)t_z) \\ &\quad +2l + 4(l-t_z)u(l-t_z) + \cdots + 4(l-(N_z-2)t_z)u(l-(N_z-2)t_z) \\ &\quad + \cdots + 2l)/N_z \end{aligned} \tag{3.6}$$

In cubic implementation of a 3-D IC, where $N_z = N^{1/3}$ and $t_z = 1$, $I_{3D}(l,t_z) \simeq \alpha kp(1-p)N_b(l,t_z)^{(p-2)} \simeq \alpha kp(1-p)l^{3(p-2)}$. When $N_z \ll N^{1/3}$, $I_{3D}(l,t_z)$ is estimated using the values of N_a, N_b, and N_c given in Eq. 3.6.

Using the methodology presented here, the wire-length distribution of a 3-D random logic network with 3.5 million logic gates is estimated. It is assumed that k = 4 and p = 2/3. These are typical Rent's parameters in a full- or semi-custom logic design in high-performance circuits such as microprocessors [15, 53].

Results

We consider two limiting cases of 3-D integration based on how efficiently the third dimension is utilized for inter-stratum interconnections. In the first case, we assume the system is partitioned and placed in multiple strata in a way that the number of inter-stratum interconnects is negligible compared to the number of intra-stratum interconnects. This interconnection scheme does not utilize the vertical dimension efficiently, and it can be represented by a 3-D interconnection scheme with $r_{upper} = r = 0$. The inter-stratum connections consist of a few global wires, and for estimating average or total wire-length their contribution can be ignored. The wire-length distribution in this case is approximated as $f_{3D}(l) \simeq N_z f_{intra}(l)$, where $f_{intra}(l)$ is the 2-D wire-length distribution of interconnects within each stratum. The wire-length distribution for this case is shown in Fig. 3-8.

In the other case $r_{upper} = N_z - 1$, and the system is partitioned in a way that there is comparable connectivity between logic gates on different and same strata. The inter-stratum connections consist of both gate- and module-level interconnects of all lengths. As a result there will be a significant number of inter-stratum interconnects. The wire-length distribution for $r_{upper} = N_z - 1$ can be estimated using the methodology described in Section 3.1.2, and it is shown in Fig. 3-9. The dotted region in Fig. 3-9 which may

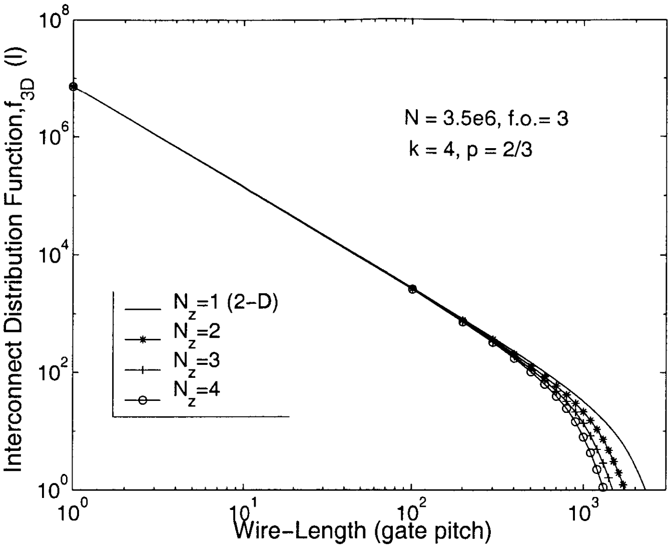


Figure 3-8: The wire-length distribution in 3-D ICs for negligible connectivity between logic gates on different strata ($r_{upper} = r = 0$).

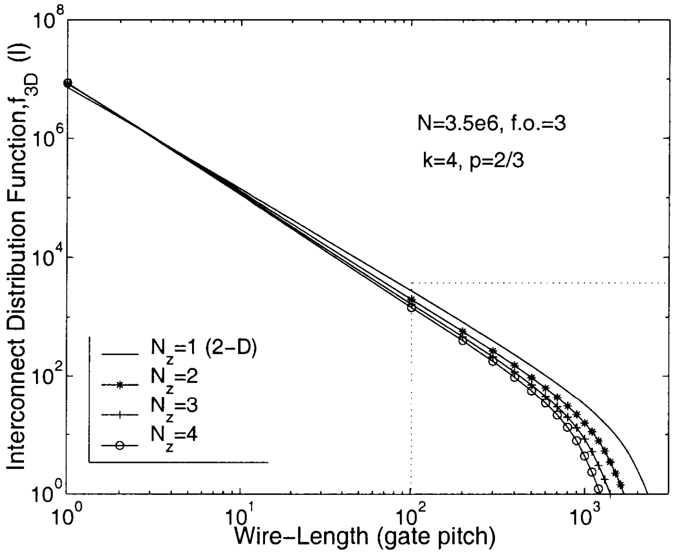


Figure 3-9: The wire-length distribution of 3-D IC for $r_{upper} = N_z - 1$.

correspond to the distribution of long or global wires is shown separately in Fig. 3-10.

For $r_{upper} = N_z - 1$, the wire-length distribution is narrower with higher number of short wires and fewer long wires than the case with $r = 0$. The total wire-length, $L_{tot} = \sum_{l=1}^{l_{max'}} l f_{3D}(l, t_z)$, and the average wire-length, $l_{ave} = \sum_{l=1}^{l_{max'}} l f_{3D}(l, t_z) / \sum_{l=1}^{l_{max'}} f_{3D}(l, t_z)$ for $r_{upper} = N_z - 1$ are also shorter compared to the case with $r = 0$. In Fig. 3-11 and Fig. 3-12, the total and average wire-length in 3-D ICs are plotted as a function of the number

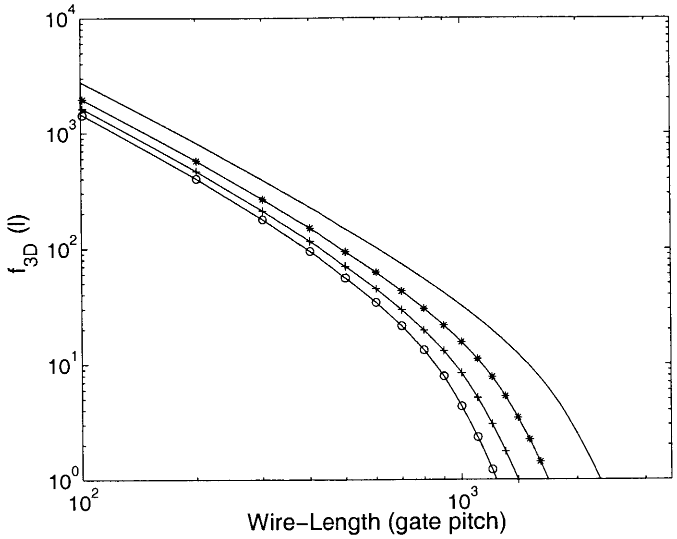


Figure 3-10: The dotted region of the wire-length distribution of Fig. 3-9.

of logic gates. The reduction in the average and total wire-length for $r = 0$ results from the physical shrinking of system's size. Whereas for $r_{upper} = N_z - 1$, both the reduction in system's physical size and the elimination of many global wires by shorter local or semi-global wires result in shorter average and total wire-length. Based on our simulation results, the reduction in average or total wire-length in 3-D ICs with $2 - 4$ strata can be as much as $30\% - 50\%$.

Our estimated average wire-length is roughly a factor of 2 smaller than the average wire-length prediction based on Masaki's methodology [63]. However, it is known that Masaki's methodology, based on hierarchical partitioning, computes an upper bound of average wire-length, and in 2-D ICs, estimated values of average wire-length are 1.5× to 2× higher compared to the experimentally obtained values [63, 80].

We have also examined the effect of varying r_{upper} on average and total wire-length. Simulation results of average wire-length in a 3-D integrated circuit with four strata are shown in Fig. 3-13. For $r_{upper} = r = 0$, there is negligible or no connectivity between logic gates on different strata; when r_{upper} is one ($r = 0, 1$), only intra-stratum interconnections and inter-stratum interconnections between nearest stratum are allowed. As r_{upper} is increased, there is higher connectivity between logic gates on different strata and the average and total wire-length become shorter.

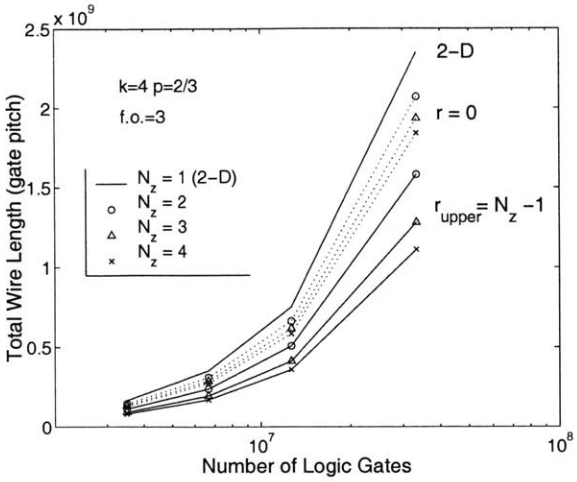


Figure 3-11: The total wire-length in 2-D and 3-D ICs.

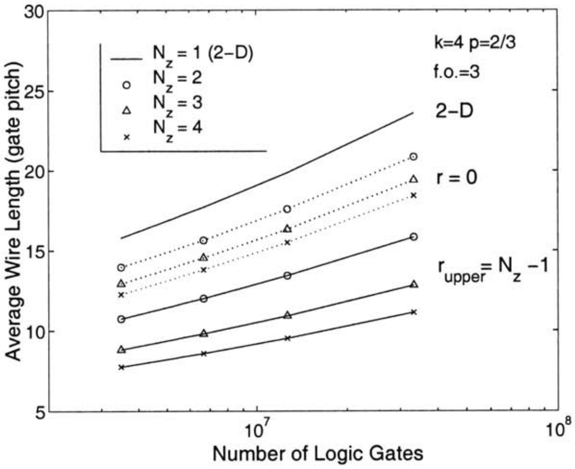


Figure 3-12: The average wire-length in 2-D and 3-D ICs.

3.1.3 Intra- and Inter-Stratum Components of Wire-Length Distribution

By decomposing the wire-length distribution into intra- and inter-stratum components, the number of intra- and inter-stratum interconnects and inter-stratum via density can be estimated. The dependency of inter-stratum interconnect's distribution on stratal pitch can also be examined. The wire-length distribution of intra-stratum interconnect is given by

$$f_{intra}(l, t_z) = \frac{f_{3D}(l, t_z)}{1 + \sum_{i=1}^{N_z - 1} \psi_i(l, t_z)}, \tag{3.7}$$

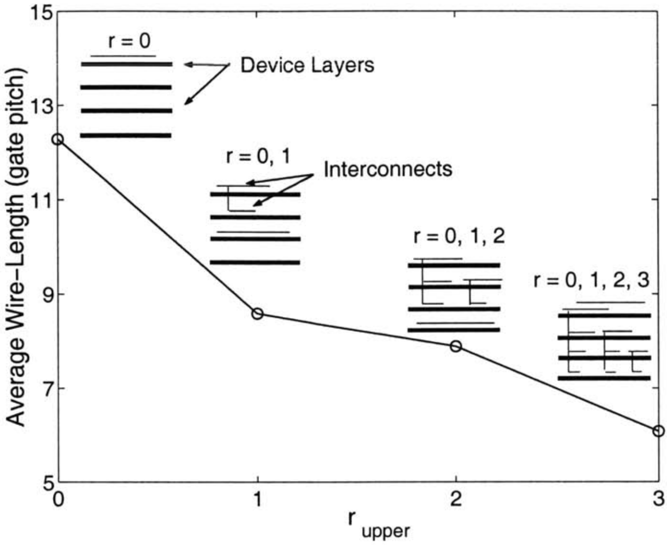


Figure 3-13: The average wire-length in a 3-D integrated circuit with four device layers for various upper bounds of r.

and the distribution of inter-stratum interconnects with it_z stratal pitch is given by

$$f^{i}_{inter}(l,t_z) = \frac{\psi_i(l,t_z)}{1+\sum_{i=1}^{N_z-1}\psi_i(l,t_z)} f_{3D}(l,t_z), \tag{3.8}$$

where

$$\psi_i(l,t_z) = \frac{f^{i}_{inter}(l,t_z)}{f_{intra}(l,t_z)} \simeq \frac{\beta_i M_{3D}(l,t_z)}{N_z M_{2D}(l)}. \tag{3.9}$$

In Fig. 3-14, the point-to-point wire-length distribution of intra- and inter-stratum components are presented. It is assumed that $N = 3.5$ million and $N_z = 2$. With two strata, the number of inter-stratum interconnects is roughly 17% of the total number of interconnects. As more strata are integrated, the number of inter-stratum interconnects increases. In the limiting case of 3-D integration with $N_z = N^{1/3}$ and $N \gg 1$, most of the interconnects are likely to be inter-stratum components.

There is a strong dependency of inter-stratum wire-length distribution on stratal pitch, t_z, as illustrated in Fig. 3-15. For larger values of t_z due to higher physical separation between adjacent strata, delay penalty or any other cost function, intra-stratum interconnections would be preferred over inter-stratum interconnects. As a result, the number of inter-stratum interconnects decreases as t_z is increased. The reduction in number of inter-

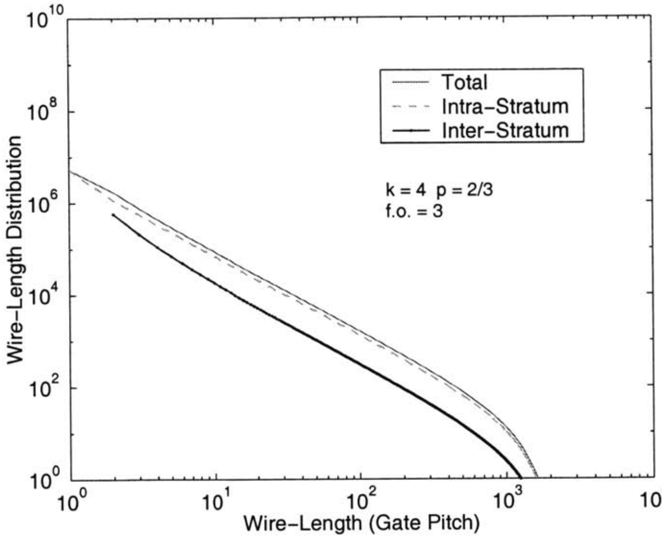


Figure 3-14: The inter-, intra-stratum, and total wire-length distribution of a 3-D IC with 3.5 million logic gates. Number of strata $N_z = 2$ and $t_z = 1$.

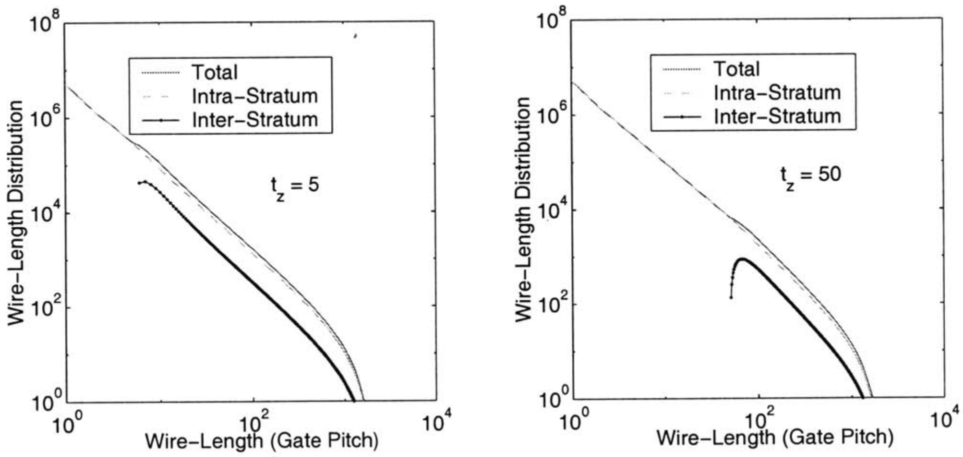


Figure 3-15: The inter-, intra-stratum, and total wire-length distribution of 3-D IC with 3.5 million logic gates and two strata for $t_z = 5$ and $t_z = 50$. Rent's parameters $k = 4$, $p = 2/3$ and $< f.o. >= 3$.

stratum interconnects will result in lower inter-stratum via density. However, the average and total wire-length will be higher. In Fig. 3-16, the percentage of interconnects due to inter-stratum components is shown for 3-D ICs with 3.5 million logic gates and Rent's parameters $k = 4$ and $p = 2/3$. As the number of strata increases, there are more inter-stratum interconnects; however, their number reduces significantly as t_z is increased. In Fig. 3-17, the dependency of total wire-length on stratal pitch is illustrated. The total wire-length begins to increase as stratal pitch, t_z, is increased. However, even with a stratal pitch of

100, total wire-length in 3-D integration is 10% – 20% smaller compared to that of 2-D integration.

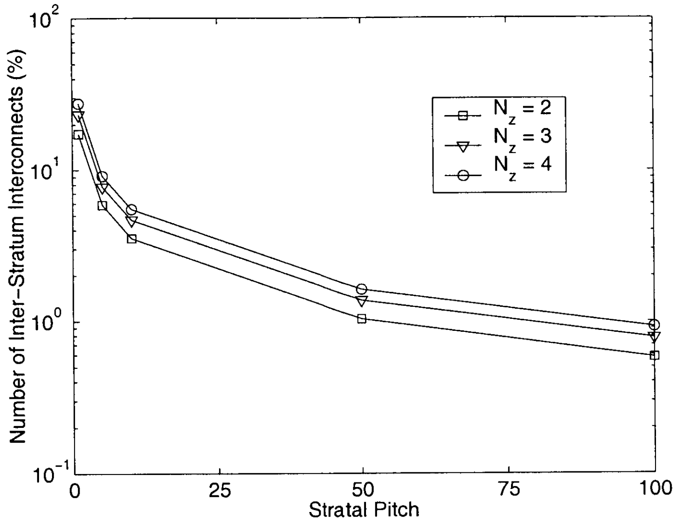


Figure 3-16: The percentage of total number of interconnects due to inter-stratum components vs. stratal pitch, t_z. Rent's parameters $k = 4$, $p = 2/3$, and $< f.o. >= 3$.

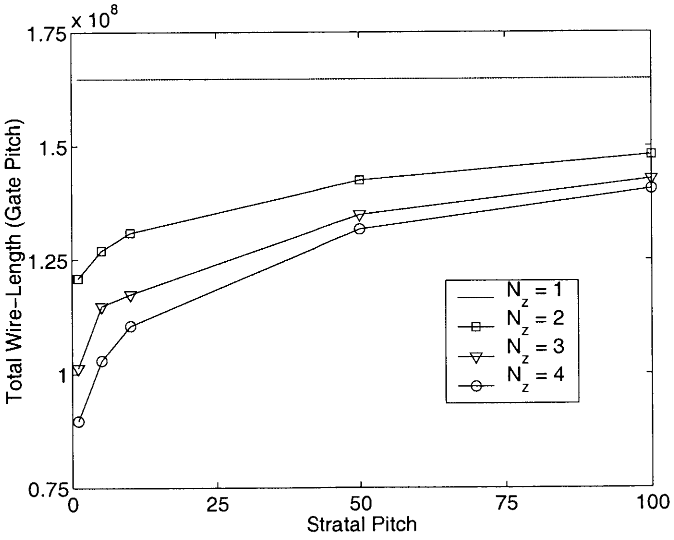


Figure 3-17: The total wire-length of 2-D and 3-D ICs as a function of stratal pitch, t_z. Rent's parameters $k = 4$, $p = 2/3$ and $< f.o. >= 3$.

In a 2-D IC, there are numerous short wires compared to long wires. However, the contribution of short wires to total wire-length or chip area is small compared to that of intermediate and long wires. By examining Fig. 3-16 and Fig. 3-17, we find that it is feasible to reduce the number of inter-stratum interconnects and their via density significantly at the expense of a small increase in total wire-length by eliminating short inter-stratum interconnect. A similar explanation based hierarchical partition can also be provided. Generally, shorter wires are associated with lower hierarchical partitioning levels. If the logic gates in lower hierarchical levels are placed in a single stratum rather than distributing them in multiple strata, there will be hardly any short inter-stratum interconnects. As a result, the number of inter-stratum interconnects can reduced significantly.

3.1.4 Wiring Efficiency

To estimate the wiring efficiency of multi-strata IC, we incorporate all previous assumptions for estimating 2-D wiring efficiency. In addition, via blockage due to inter-stratum connections are taken into account. For via blockage modeling, it is assumed that inter-stratum interconnections are between the top-most metal levels, and their via-blockage is also 15%. In a multi-strata IC, as more strata are incorporated, via-blockage due to inter-stratum interconnects may reduce the overall wiring efficiency, and it can have a significant impact on wireability.

3.1.5 Cost Function

To estimate the figures of merit of 2-D and 3-D ICs, it is important to make a fair comparison between different 2-D and 3-D technologies. The cost or cost per function of an IC depends on the equipment productivity, manufacturing yield, and the number of chips available per wafer [7]. The productivity and yield are tied strongly to the manufacturing process complexity. Our definition of a cost function is motivated by a scenario where one would like to fabricate the same number of 2-D and 3-D chips from a fixed number of wafers, and different strata (device layers) require similar front-end and back-end fabrication steps. We model the fabrication cost or complexity by a variable, cost function (c.f.), which is proportional to $(m + n_b)$, where m is the number of interconnect levels per stratum, and $n_b = N_z - 1$ is the number of inter-stratum bonding steps. By examining the processing steps necessary for our proposed wafer bonding scheme, we make a simplistic assumption that the complexity associated with the wafer bonding process is comparable to the complexity for integrating an additional interconnect level per stratum. The cost function also depends on chip area. In some of our analyses, the total chip area, $N_z A_c$, is kept constant. As a result, chip area per stratum is reduced by $1/N_z$, and N_z times more dies can be fabricated per wafer.

For example, consider a 2-D IC with chip size A_c, $c.f. \sim 6$, $m = 6$ and $n_b = 0$. For the same cost function in a 3-D IC with two strata, the chip size per stratum is $A_c/2$, $m = 5$, and $n_b = 1$. When comparing the system performance of 2-D and 3-D ICs, values of m, n_b, and the chip area are adjusted to keep the cost function constant.

3.2 Examples: Chip Area and Clock Frequency Prediction

By incorporating the models for wiring requirements, wiring efficiency, and cost function for 3-D ICs in the system-level modeling framework presented in Chapter 2, key performance metrics of 3-D ICs such as clock frequency, chip area, etc. can be estimated [11]. In this section simulation results of these performance metrics will be presented. For a fixed cost function, these results are estimated by solving equations governing the conservation of chip area (Eq. 2.27) and critical path delay (Eq. 2.12).

3.2.1 Clock Frequency

The critical path model described in Chapter 2 can be used to estimate the maximum clock frequency of 2-D and 3-D random logic circuits. Random logic networks can be considered as simplified representation of ASIC based designs. We begin by examining the impact of increasing the number of interconnect levels on the maximum clock frequency of a 2-D random logic network. Then the clock frequency is estimated for 3-D ICs with multiple strata.

We consider an IC with 3.5 million logic gates. The critical path consists of 3-input NAND gate, and it is assumed the average-length and chip-edge length interconnects in the critical path are routed in global interconnect levels. The logic depth in our critical path is 15, and the clock skew is 50 ps. The minimum feature size is 0.18 μm. The W/L (width/length) ratio of nMOS transistors in the critical path is 5, and Rent's parameters

Number of Interconnect Levels	6	7	8
m_l, m_{sg}, m_g	2, 2, 2	2, 2, 3	2, 3, 3
p_l, p_{sg}, p_g (μm)	0.36, 1.1, 1.84	0.36, 1.1, 2.33	0.36, 1.25, 2.55
Chip Area, A_c (cm^2)	3.5	4	3.6
Clock Frequency, f_c (MHz)	450	510	570
β_g	0.4	0.34	0.23

Table 3.1: Simulation results of system-level modeling and analysis of 2-D random networks. m_l, m_{sg}, and m_g are the number of local, semi-global, and global interconnect levels, respectively, and p_l, p_{sg}, and p_g are their wiring pitches. β_g is the fractional delay of the chip-edge length wire compared to clock delay.

are $k = 4$ and $p = 2/3$ which are typical for high-performance logic networks [15]. The interconnect materials are Cu and low-k ($\epsilon_r = 2.5$) ILD and interconnect's aspect ratio is 1.5. These values are typical for an IC in .18 μm technology generation [7].

In Table 3.1, the maximum clock frequency and interconnect parameters are listed for 2-D ICs with 6-8 total interconnect levels. As the number of interconnect levels is increased, more wiring area is available, and the wiring pitch can be increased. This measure results in smaller interconnect delay and higher clock frequency as illustrated in Table 3.1. When repeaters are inserted on the chip-edge length wire, critical path delay is mainly limited by gate and average-length wire delay, and the clock frequency for the cases listed in Table 3.1 is in the range of 620 $MHz - 640$ MHz.

Next, we estimate the maximum clock frequency of 3-D ICs with 2-4 strata by keeping the cost function, $c.f. \sim 6$, and total chip area, $A_c N_z = 3.5$ cm^2, constant. The simulation results for these case studies are shown in Fig. 3-18. Based on our critical path model, the improvement in clock frequency in 3-D ICs results from the reduction in interconnect delay of both average length and chip-edge length wires. Also, the total wire-length, L_{total}, in 3-D ICs is smaller than that of 2-D ICs. As a result, for comparable available wiring area which is also proportional to $\sim p_w \times L_{total}$, wiring pitch, p_w in 3-D ICs can be increased to reduce interconnect delay further. However, as more device layers are integrated, due to the constant cost function constraint, less wiring area is available in 3-D ICs. The wiring area is also reduced due to the via blockage of inter-stratum interconnects. To accommodate the required wiring need within the available wiring area, interconnect's pitch has to be reduced,

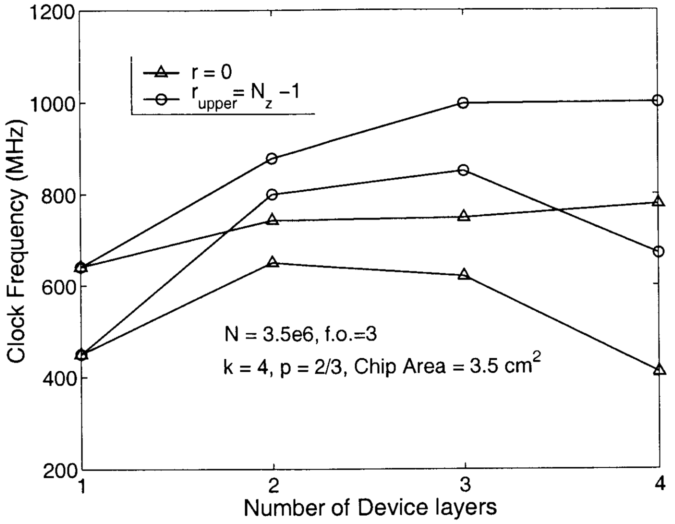


Figure 3-18: The clock frequency of 2-D and 3-D ICs with 3.5 million logic gates and minimum feature size of 0.18 μm.

Number of Strata	2	3	4
$r = r_{upper} = 0$			
m_l, m_{sg}, m_g	2, 1, 2	2, 1, 1	2, 0, 1
p_l, p_{sg}, p_g (μm)	0.36, 0.92, 2	0.36, 0.95, 1.65	0.36, X, 0.85
$r_{upper} = N_z - 1$			
m_l, m_{sg}, m_g	2, 1, 2	2, 1, 1	2, 0, 1
p_l, p_{sg}, p_g (μm)	0.36, 1.05, 2.64	0.36, 1.13, 2.4	0.36, X, 1.14

Table 3.2: Simulation results of system-level modeling and analysis of 3-D ICs for two limiting cases of connectivity governed by the maximum value of range, r. m_l, m_{sg}, and m_g are the number of local, semi-global, and global interconnect levels, respectively, and p_l, p_{sg}, and p_g are their wiring pitches.

and the improvement in clock frequency begins to slow down or diminish (see Fig. 3-18). This trend is observed clearly in the values of global wiring pitch listed in Table 3.2.

3.2.2 Chip Area

Similar to the earlier analysis, the impact of 3-D integration on wiring-limited chip area can evaluated by keeping the clock frequency and cost function constant and computing the total chip area that meets the wiring requirements. The wiring-limited total chip area, $A_c N_z$, of a 2-D IC with 3.5 million logic gates, 450 MHz clock frequency, and $c.f. \sim 6$ is 3.5 cm^2. The logic gate area of a 3-input NAND gate is roughly $300F^2$ and $460F^2$ for

$W/L = 1.5$ and $W/L = 5$, respectively. These area models lead to device-limited chip area of 0.33 cm^2 and 0.52 cm^2, respectively, which are much smaller than the wiring-limited chip area.

When multiple strata are integrated, significant reduction in total chip area, N_zA_c, can be achieved (see Fig. 3-19). There are two factors that lead to the reduction in chip area. The wiring-limited chip area is roughly proportional to the total wire-length, and the reduction in total wire-length will result in a smaller chip area. We also find that for fixed/comparable interconnect delay, which is also proportional to $\sim \frac{l^2}{p_w^2}$, a reduction in wire-length, l, can be accompanied by a reduction in wiring pitch, p_w. The reduction in wiring pitch will also lead to a smaller wiring-limited chip area. As more device layers are integrated, total chip area, N_zA_c, may become comparable to device-limited chip area, and further reduction in total chip area will not be feasible.

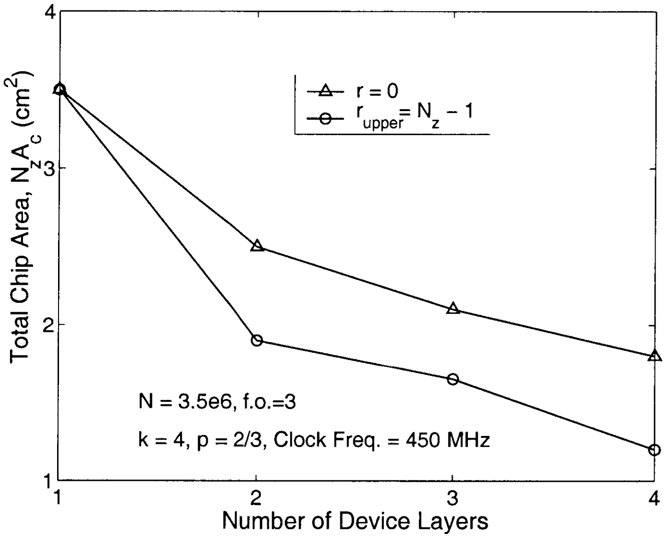


Figure 3-19: The total Chip area, N_zA_c, of 2-D and 3-D ICs for fixed clock frequency, $f_c = 450\ MHz$, and cost function, $c.f. \sim 6$. The device-limited chip area is .52 cm^2.

3.2.3 Sandwich Interconnects

In 2-D ICs with a few interconnect levels, wiring-limited chip can be reduced by integrating more interconnect levels. However, as the number of interconnect levels is increased, via blockage would impact severely the number of available wiring channels in lower interconnect levels. As result, it may not be feasible to reduce the chip size by increasing the number of interconnect levels [15]. However, it will be necessary to increase the number of interconnect levels to accommodate the higher wiring need in future ICs.

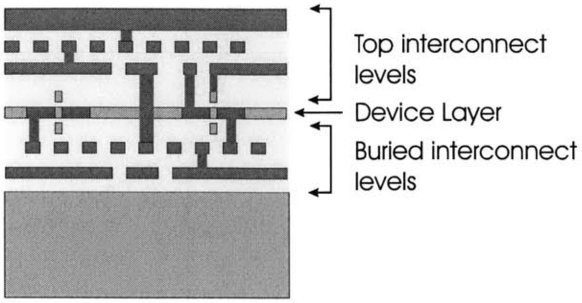


Figure 3-20: Cross section of a hypothetical sandwich interconnect technology.

To alleviate poor wiring efficiencies in lower interconnect levels, sandwich/buried interconnect technology may have tremendous potential. The cross section of a hypothetical sandwich interconnect technology is shown in Fig. 3-20. Typically, local interconnect levels have poor wiring efficiencies due to via blockage. By burying local interconnect levels, their wiring efficiencies can be improved. The fabrication technology of sandwich interconnects is most likely going to be very similar to that of double-gate SOI-based CMOS [81] or 3-D ICs. Short local wires with contacts to source, drian, and gate can be routed in the buried interconnect levels.

To evaluate the impact of sandwich interconnect technology on system performance, the maximum clock frequency and the associated chip area have been estimated as a function of number of total interconnect levels. We assume, in sandwich interconnect technology, only local interconnect levels are buried. The input parameters are the same as earlier case studies. The simulation results of maximum clock frequency and associated chip area are presented in Fig. 3-21, and the interconnect parameters are listed in Table 3.3. In sandwich interconnect technology, wiring efficiencies in all interconnect tiers are higher which lead to smaller chip area and shorter wire-length. As a result, the maximum clock frequency in sandwich interconnect technology is also higher compared to that of conventional 2-D ICs.

Though in our study, it has been assumed that only local interconnect levels are buried, an alternative solution by burying V_{DD} and ground interconnect levels can also be consid-

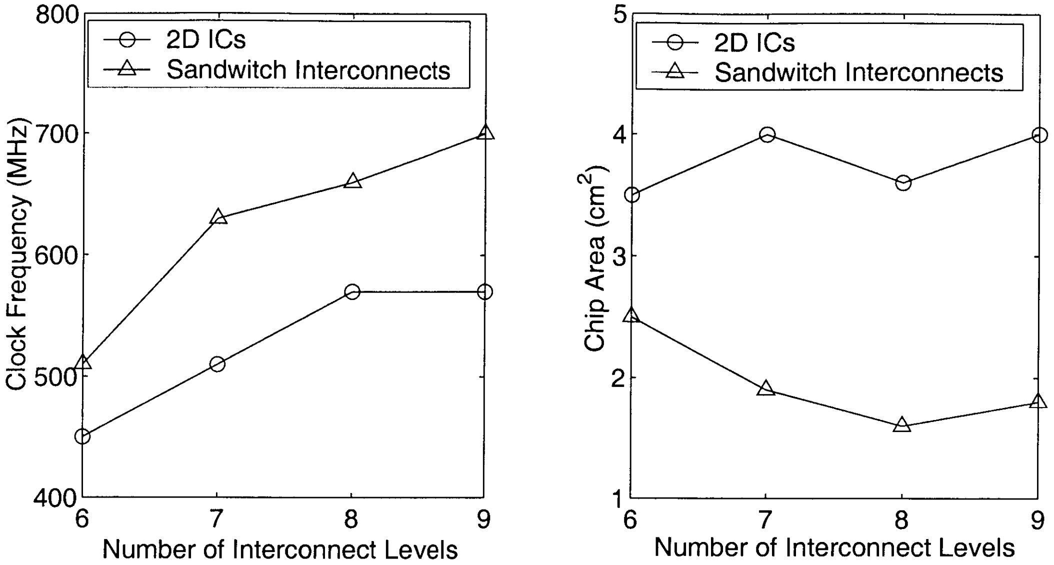


Figure 3-21: The maximum clock frequency and the associated chip area of conventional and buried interconnect technologies.

Number of Interconnect Levels	6	7	8	9
m_l, m_{sg}, m_g	2, 2, 2	2, 2, 3	2, 3, 3	2, 4, 4
p_l, p_{sg}, p_g (μm)	0.36, 1, 1.63	0.36, 0.9, 1.73	0.36, 1, 1.85	0.36, 1, 2.32

Table 3.3: Simulation results of interconnect parameters for system-level modeling and analysis of sandwich interconnect technology. m_l, m_{sg}, and m_g are the number of local, semi-global, and global interconnect levels, respectively, and p_l, p_{sg}, and p_g are their wiring pitches.

ered. However, based on system-level analysis and also considering the improvements in wiring efficiencies, we find that it is not worthwhile to dedicate two interconnect levels just for V_{DD} and ground plane.

3.3 Summary

In this chapter, the derivation of wire-length distribution of 3-D ICs has been presented. The dependencies of wire-length on stratal pitch, inter-stratum connectivity, etc. have been discussed. By integrating the wire-length models in system-level modeling framework, key performance metrics such as clock frequency, chip area, etc. have been estimated. We find that both 3-D integration and sandwich interconnect technology have the potential to reduce interconnect delay and increase gate density for future VLSI applications.

We have also examined the global wiring requirements in a 3-D system-on-a-chip (SOC) by extending the model for net-length distribution of heterogeneous logic network [82]. A detailed analysis of global net-length distribution between megacell or logic blocks in SOC can be found in Appendix B. We find that by 3-D integration, significant reduction in multi-terminal net-length can be achieved.

In this chapter, key performance metrics of random logic networks have been estimated. However, many VLSI applications require on-chip memory which have higher device density compared to logic circuits. In many cases, clock frequency alone is not always a good metric to assess system performance and cycles per instruction (CPI) should be used to make a fair comparison between various system architectures. In the next chapter, system-level modeling work will be extended by including on-chip memory and system performance will be estimated for microprocessors implemented in scaled technology nodes.

Chapter 4

Opportunities for Three-Dimensional Implementation of Microprocessors

In this chapter, system-level modeling and analysis will be performed to evaluate the opportunities for 3-D implementation of microprocessors. Various approaches for implementing 3-D microprocessors can be considered by partitioning logic and memory transistors in different strata or by distributing them in all strata [21, 83]. These approaches would result in higher device density and clock frequency and faster memory access time. To assess the technology requirements of 3-D microprocessors, it is necessary to develop chip area models that incorporate area requirements of both logic and memory transistors. Also, estimation of clock frequency alone may not be sufficient to assess the impact of 3-D integration on overall system performance, and cycles per instruction (CPI) or throughput should be used as system performance metric [50].

In this section, simplistic models are used to assess the technology requirements and system performance of microprocessors. We assume a microprocessor consists of an on-chip processor core and cache memory (L1) and additional off-chip cache memory (L2 and L3) and main memory. Generally, L1 cache is divided into instruction cache (I-cache) and data cache (D-cache), and often L2 cache memory is also integrated on-chip. As more transistors

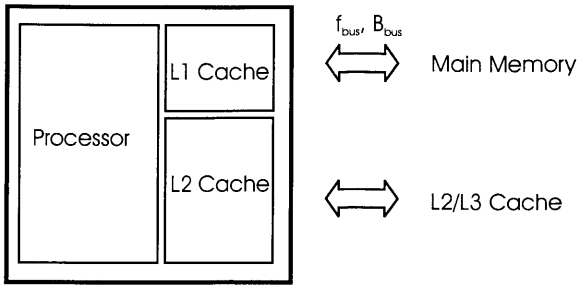


Figure 4-1: A simplistic model of a conventional microprocessor.

become available per chip, it is likely that L2 cache would be implemented on-chip in future technology nodes. Generally, on-chip and off-chip cache memory are implemented using SRAM to achieve fast access time, and (off-chip) main memory is implemented using DRAM to achieve high density. The phenomenal improvements in microprocessor's performance have created a significant demand for low latency and high-bandwidth operation of the memory system [84]. Also, the performance of the processor has been increasing at a faster pace than the performance of off-chip memory, creating a *processor-memory performance gap* [85]. To alleviate the performance gap, various system architecture-based solutions are currently being explored [85, 86]. In this chapter, rather than looking into the architectural solutions, we explore the impact of 3-D technology on system performance of traditional microprocessors using analytical models. The memory modules that are currently integrated off-chip can be combined with the processor core monolithically using 3-D technologies. This measure would reduce the memory access time, create unlimited bandwidth for memory access, and also reduce the form factor [83]. Moreover, by partitioning the logic and memory elements in multiple strata, significant reduction in interconnect delay, number of repeaters for global wires, and wiring-limited chip can be achieved [21].

In the following sections, simple models for estimating system performance and technology requirements of microprocessors will be presented. It will be followed by simulation results for 2-D and 3-D implementation of microprocessors in scaled technology nodes.

4.1 Microprocessor Model

We use the model of a microprocessor as shown in Fig. 4-1 in our analysis. The interface between the processor and memory is defined by the main memory access time, t_{mm}, bus frequency, f_{bus}, and the bus width, B_{bus}. The performance of the memory system is determined by memory access time, processor-memory bandwidth and memory organization (associativity, block size, etc.). L1 cache can often be in the critical path of a microprocessor. We assume the size of L1 cache is chosen such that its access time is small and does not effect cycle time. If there is a L1 cache miss (i.e. the information requested is not found in L1 cache), the requested block of data or instruction is fetched from L2 cache. However, if there is a L2 cache miss, main memory is accessed. The main memory access time can be 50-100 clock cycles, and the purpose of hierarchical cache memory is to minimize the delay penalty associated with accessing main memory.

4.1.1 Cycles Per Instruction (CPI) and Throughput

The number of cycles needed per instruction (CPI) in a microprocessor depends on the micro-architecture of the processor as well as the memory elements. Using a well-calibrated model for CPI, the logic and memory elements' contribution to overall system performance can be examined. Recently, an analytical model for CPI has been proposed based on empirical observation, and it has been also verified with existing/published data on microprocessor's performance. It is given by [50]

$$\begin{aligned} CPI &= CPI_{logic} + CPI_{memory} \\ &= E_c N_{logic}^{E_e} + m_{ref} m_{rate} m_{penalty} \end{aligned} \tag{4.1}$$

where the first term is due to the contribution of logic element (i.e. data path, functional units, etc.) and the second term is due to the contribution of memory element. E_c and E_e are empirical constants and N_{logic} is the number of logic gates; m_{ref} is the number of memory reference per instruction, m_{rate} is the memory miss rate per instruction, and $m_{pentalty}$ is the memory miss penalty (in number of cycles) per instruction. In Table 4.1, extracted

Architecture	E_c	E_e
Intel x86	51829	-.7725
DEC Alpha	4493.5	-.571
Power PC	1703.8	-.525
HP PA-RISC	184.41	-.3627
SPARC	6206.1	-.6115
MIPS	55.22	-.2488

Table 4.1: Extracted empirical parameters for projecting microprocessor's performance [50].

empirical values of E_c and E_e for various microprocessor architectures are presented [50]. These values are consistent with G. Sai Halasz's observation that typically 20% – 40% increase in circuit complexity (i.e. number of gate count) is needed to increase CPI by 10%. The dependency of CPI on the number of logic gates arises from the fact that methods for reducing CPI by pipelining, parallelism, addition of more datapaths and functional units, etc. would require higher logic count.

Typical values of m_{ref} is RISC and CISC architectures are .84 and 1, respectively [61]. Miss rate, m_{rate}, depends on the cache memory size and its organization [61]. Generally, doubling the cache size reduces the miss rate by 25% [87]. In [61] miss rate as a function of cache size and organization can be found. Miss penalty, $m_{penalty}$, depends on the bus frequency, memory access time, block size, etc. By combining these terms, the contribution of memory element to CPI in a 2-level cache configuration is given by [50, 61, 88]

$$\begin{aligned} CPI_{mem} &= m_{iref}m_{iratel}(t_{L2}f_c) + m_{dref}m_{dratel}(t_{L2}f_c) + \\ &\quad (m_{iref}m_{iratel} + m_{dref}m_{dratel})m_{rate2} \\ &\quad (t_{mm}f_c + \frac{8B_2}{B_{bus}}\frac{f_c}{f_{bus}}) \end{aligned} \qquad (4.2)$$

where m_{iref} and m_{dref} are L1 instruction and data cache reference per instruction and m_{iratel} and m_{dratel} are their respective miss rates. t_{L2} is the access time of L2 cache and B_2 is its block size. B_{bus} is the bus width, f_{bus} is the bus frequency, and f_c is the clock frequency.

Using Eq. 4.1 and cycle time, throughput for a system, instructions per second (IPS), can be found. It is given by

$$IPS = \frac{1}{T_c \cdot CPI} \tag{4.3}$$

Throughput is a more reliable metric for comparing system performance because it combines both architectural and technology dependent metrics into one equation.

4.1.2 Logic-Memory Partitioning

For system-level performance evaluation and technology requirements, it is essential to be able to determine the size of on-chip memory that results in optimum system performance. Generally, for a fixed number of total equivalent logic gates, N_{total}, and a specific system architecture, there is an optimum partitioning of total number of gates into logic and memory elements for minimizing CPI. If the memory size is too small, m_{irate1} and m_{drate1} would be very high and CPI_{mem} would increase. On the other hand, if there are not enough logic transistors, the processing power will not be sufficient and CPI_{logic} will be high. Assuming a power law relationship for miss rate, $m_{rate} \simeq \alpha_m C^{\beta_m}$ ($\alpha_m > 0$, $\beta_m < 0$), an optimum value of cache size C (in Bytes) can be found that minimizes CPI [50]:

$$CPI = E_c(N_{total} - 8C)^{E_e} + m_{ref}\alpha_m C^{\beta_m} m_{penalty} \tag{4.4}$$

Typically, devoting $20\% - 50\%$ of transistors to memory results in optimum system performance in conventional microprocessors [50].

For accurate system-level analysis, the logic gates in peripheral memory circuits and tag portion of the cache memory should also be accounted for. To reduce cache the access time in a large memory, multi-divided word line and bit line configurations are used which require additional peripheral and control logic. The "Tag" portion of the cache memory that provides a list of exactly what blocks are stored in the cache memory requires $\sim (C/B)$ storage bits, where B is the block size [50]. Although an exact model for the number of transistors devoted to peripheral and control logic and tag memory can be found, it requires a complete knowledge of the memory organization (i.e. associativity, number of word and bit line divisions, block size, etc.). Based on our observation and also reviewing published results, we find that peripheral logic, control logic, tag array, etc. account for only $2\% - 4\%$ of total number transistors devoted SRAM (data array).

4.1.3 Area Models

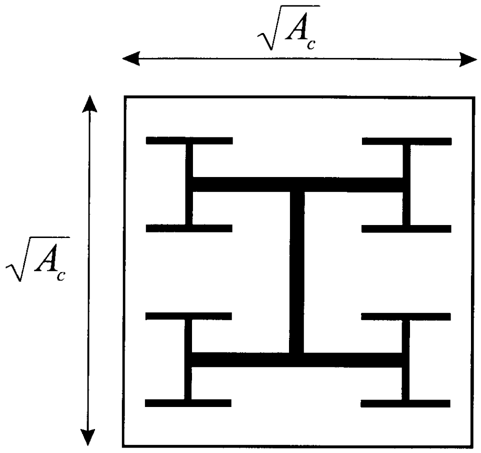


Figure 4-2: A H-tree clock distribution network.

To estimate the area of the logic element, the methodology presented in Chapter 2 can be used. In our area model, we also include the wiring area needed for an H-tree clock network. We assume interconnects on the lowest level of an H-tree network is pitch-matched to global wiring pitch. In each hierarchical level above the lowest level, wiring pitch is increase by 2× for impedance matching [49]. The wiring area needed for clock network, A_{clock}, is given by

$$A_{clock} = 1.5\sqrt{A_c}[\sum_{i=1}^{n_{level}} 2^{i-1}(p_g \times 2^{(n_{level}-i)})], \tag{4.5}$$

where $n_{level} = log_4 n_{latch}$; n_{latch} is the number of latches in the system, typically 2% of the total number of logic gates [50].

Unlike logic area, on-chip memory area is device-limited. In other words, by increasing the number of interconnect levels, the memory area cannot be reduced. By examining published results of SRAM cell size and chip area [89, 90, 91, 92], we find that on-chip memory area, including the peripheral and control logic area, of a N_{mem} bit SRAM can be represented by

$$A_{mem} = \beta_{mem} A_{cell} N_{mem}, \tag{4.6}$$

where β_{mem} is the ratio of actual chip area and cell size limited chip area, and it is determined empirically; A_{cell} is the memory cell size. Typically, $\beta_{mem} = 2.7 - 3$ and $A_{cell} = 130F^2 - 190F^2$ [89, 90, 91, 92].

4.2 Validation

We have validated our methodology for estimating chip area, wiring pitches, clock frequency, etc. by comparing our simulation results with available/published results. We assume in a full-custom design of high-performance microprocessors, it is unlikely to have a long wire in the critical path; so the long wire delay is omitted from our cycle time model. The simulation results of chip area, wiring pitches, etc. for two families of microprocessors are shown in Table 4.2 and Table 4.3. The composition of logic gates in the critical path is hard to determine and approximating them by 3-input NAND gates may contribute to some errors in clock frequency projection. Nevertheless, our estimated system performance metrics and interconnect parameters are in reasonable agreement with published results. To estimate on-chip memory area, $A_{cell} = 175F^2$ and $\beta_{mem} = 3$ have been assumed.

Parameters	Published Results	Estimated Values
Transistor Count	9.3 million	9.3 million
Memory	16KB L1, 96KB L2	16KB L1, 96KB L2
Chip Area	2.99 cm^2	2.7 cm^2
M1-M2 Pitch	1.125 μm	1 μm
M3-M4 Pitch	3.3 μm	3.8 μm
Clock Frequency (MHz)	300	345

Table 4.2: Comparison of published and projected system performance and interconnect parameters for Alpha 21164 (3.3V .5μm) microprocessor [93]. It is assumed that W/L ratio in the critical path is 30 and logic depth is 14, Rent's parameters k = 4, p = .6, average fan-out is 3, and I/O pad area is $\sim$ 10% of chip area.

Parameters	Published Results	Estimated Values
Transistor Count	3.3 million	3.3 million
Memory	16KB L1	16KB L1
Chip Area	.9 cm^2	.85 cm^2
M1 Pitch	.9 μm	.7 μm
M2-M3 Pitch	1.2 μm	1.4 μm
M4 Pitch	3 μm	3.38 μm
Clock Frequency (MHz)	200	155

Table 4.3: Comparison of published and projected system performance and interconnect parameters for Pentium ($3.3V$ $.35\mu m$) microprocessor [94, 95]. It is assumed that W/L ratio in the critical path is 30 and logic depth is 27, Rent's parameters k = 4, p = .6, average fan-out is 3, and I/O pad area is $\sim$ 10% of chip area.

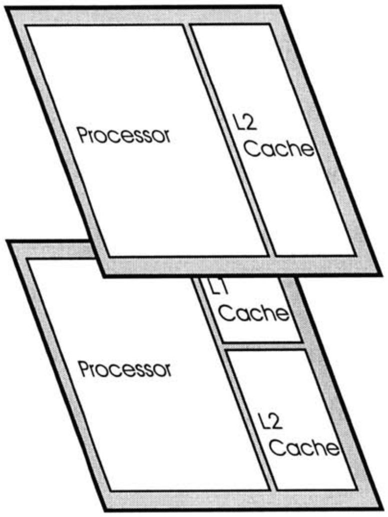


Figure 4-3: A 3-D implementation of microprocessor.

4.3 Chip Area and Interconnect Parameter Estimation of Microprocessors Implemented in Scaled Technologies

In this section, chip area and interconnect parameters of microprocessors implemented in 2-D and 3-D scaled technologies will be estimated based on system-level modeling and simulations. We consider a 3-D implementation with two strata by distributing logic gates and cache memory transistors on both strata as shown in Fig. 4-3. If desired, the main memory can be integrated on additional strata (i.e. third and fourth strata), resulting in faster main memory access time and higher bandwidth compared to conventional off-chip implementation. An alternative approach by placing logic gates and cache memory transistors on separate strata can also be considered. However, we find that distributing logic gates in every stratum would be more beneficial for increasing the device density compared to alternative approaches.

To estimate interconnect delay, we use the values of resistivity, dielectric constant and interconnect parameters listed in the SIA roadmap [7], and also presented in Table 4.4.

Technology Node (nm)	250	180	150	130	100	70	50
Transistor Count (million)	11	21	40	76	200	520	1400
Metal Effective Resistivity ($\mu\Omega - cm$)	2.7	2.2	2.2	2.2	2.2	1.8	1.8
IDL Dielectric Constant (κ)	3.9	3.5	3.5	3.5	2.2	1.5	1.5
Number of Metal Levels	6	6	7	7	8	9	9

Table 4.4: Metal resistivity, dielectric constant, and interconnect parameters in scaled technology nodes for microprocessors [7].

In our critical path model only the canonical logic gate delay (i.e. delay of f_{ld} stages of 3-input NAND gate with f.o. = 3 and connected by average length wires) is used to estimate the cycle time. It is assumed that by full-custom design, long wire-delay can be avoided in the critical path. For system-level evaluation of 3-D ICs, we also impose the constant cost function constraint as described in Chapter 3.

Technology Node (nm)	250	180	150	130	100	70	50
Number of Logic Gates (million)	1.3	2.44	3.5	7.34	22.7	67.5	161
L1 D-Cache/ L1 I-Cache (KB)	32/32	64/64	64/64	64/64	128/128	128/128	256/256
L2 Cache (KB)	256	256	256	512	1024	2048	8192
L1 Cache Access Time (FO4 Delay)	16.5	18.5	18.4	23	26	23	27
L2 Cache Access Time (FO4 Delay)	25	26.2	26.5	40	44	86	161

Table 4.5: The number of logic gates in the processor core and cache memory size for optimum system performance. In 250 nm and 180 nm technology nodes, L2 cache is implemented off-chip and beyond 180 nm technology node, both L1 and L2 cache are implemented on-chip. FO4 delay stands for fan-out of 4 inverter delay, and for first order calculation, 16FO4 inverter delay can be used as a measure of cycle time [18].

We begin by estimating the number of logic gates and cache memory size that would minimize CPI as represented by Eq. 4.2. Based on a recent study of conventional microprocessor design from Intel family, it has been found that $E_e = -.7725$, $E_c = 51829$, $\alpha_m = 14.5$, $\beta_m = -.69$, $m_{penalty} = 17$ and $m_{ref} = 1$ [50]. Using these values in Eq. 4.2, an optimum logic-memory partition can be found. However, this methodology suggests a relatively large on-chip L1 memory size beyond 150 nm technology node which may not be a realistic design choice. The cache memory access time increases significantly as a function of its size, specially for large memory, as shown in Fig. 4-4. Often L1 cache memory is in the critical path, and it is desirable to set its size to minimize the access time [98]. In our analysis, once the optimum logic-memory partition is known, we divide the cache memory into on-chip L1 and L2 cache such that L1 cache access time is within 1-2 clock cycles. We use 16 stages of FO4 inverter delay as an approximate value of the cycle time, and the SRAM cache access time is determined by extending Wada's cache access time models to scaled technology nodes [96]. The cache access time is estimated taking into account decoding delay, word line delay, bit line and sense amplifier delay, and data bus/output delay. The projected number of logic gates and cache sizes for x86 microprocessors in scaled technology nodes are shown in Table 4.5.

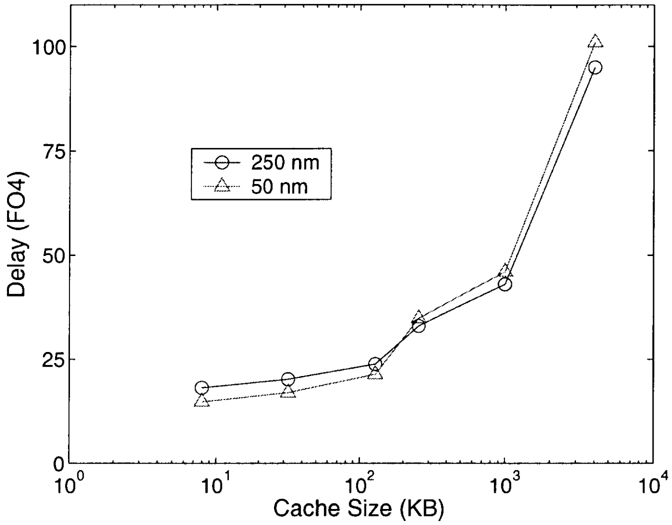


Figure 4-4: Cache access time, in terms of FO4 inverted delay, as a function of cache size in two scaled technology nodes. For small cache size, access time increases as the *log* of cache size [96]. For larger cache size, access time scales as the cache size due to the dominance of interconnect delay [97]. It is assumed that global wiring pitch is 8F, where F is the minimum feature size.

Once the total number of logic gates and memory size are known, system-level modeling and analysis is conducted to estimate chip area, wiring pitch, and clock frequency. In Table 4.6, estimated values of wiring pitch and number of interconnect levels for optimum clock frequency are presented. It is assumed that logic depth is 14, average fan-out is 3, and Rent's parameters are k = 4 and p = .6.

Technology Node (nm)	250	180	150	130	100	70	50
2-D Integration							
Number of Local Wiring Levels	2	2	2	2	3	3	3
Pitch (μm)	.5	.18	.3	.26	.2	.14	.1
Number of Semi-Global Wiring Levels	2	2	2	2	2	2	2
Pitch (μm)	2.1	1.4	1.15	.9	.6	.4	.25
Number of Global Wiring Levels	2	2	3	3	3	4	4
Pitch (μm)	4.2	3.4	2.7	1.75	.7	.47	.3
3-D Integration (2 strata)							
Number of Local Wiring Levels	2	2	2	2	3	3	3
Pitch (μm)	.5	.18	.3	.26	.2	.14	.1
Number of Semi-Global Wiring Levels	1	1	2	2	2	2	2
Pitch (μm)	1.8	1.2	1.3	1.1	.7	.43	.3
Number of Global Wiring Levels	2	2	2	2	2	3	3
Pitch (μm)	6	3.5	3.3	2.15	1.35	.65	.37

Table 4.6: Number of interconnect levels and wiring pitch for 2-D and 3-D implementation of microprocessors.

In Fig. 4-5, estimated values of clock frequency for 2-D and 3-D implementation with 2 strata are shown. The intrinsic value of clock frequency, as shown in Fig. 4-5, does not include the contribution of interconnect delay and the other two curves include interconnect delay. We find that increasing the W/L ratio of transistors in the critical path initially reduces overall delay (due to smaller value of $R_g C_{int}$, where R_g is gate output resistance and C_{int} is interconnect's capacitance). However, one quickly reaches a point of diminishing return, and increasing W/L ratio beyond $15 - 20$ does not seem to reduce overall delay significantly. In a critical path that includes capacitive loading due to average length wires, with $W/L = 30$, interconnect delay accounts for roughly $30\% - 35\%$ of total delay. In other words, even if interconnect delay could be eliminated, the improvement in clock frequency would be 50%. However, if the the critical path delay is limited by interconnect delay,

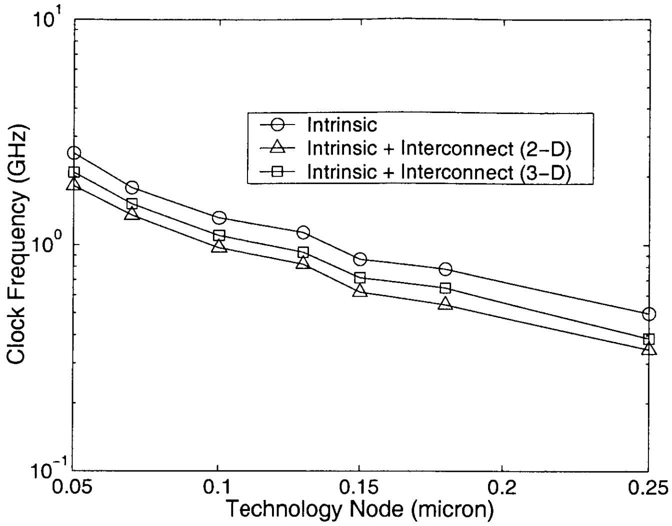


Figure 4-5: The clock frequency in 2-D and 3-D implementation of microprocessors in scaled technology nodes. The critical path consists of 14 stages of 3-input NAND gates with fan-out of 3, and they are connected by average length wires.

a much higher improvement in system performance can be achieved by minimizing the contribution of interconnect delay. We also observe that our estimation of clock frequency is smaller than the projection in SIA road map [7] which suggests that a more aggressive architecture with smaller logic depth would be required in future technologies to meet the projected system performance.

The estimated values of chip-edge length global wire delay are shown in Fig. 4-6 for

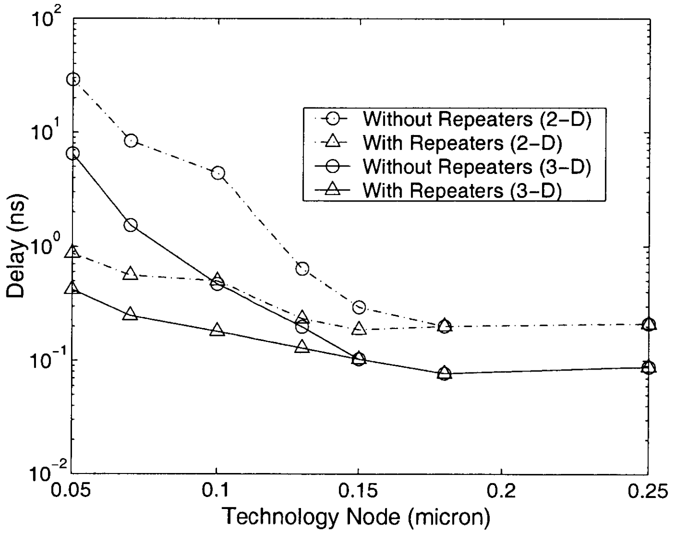


Figure 4-6: Interconnect delay of chip-edge length global wires with and without repeaters in 2-D and 3-D implementation with 2 strata.

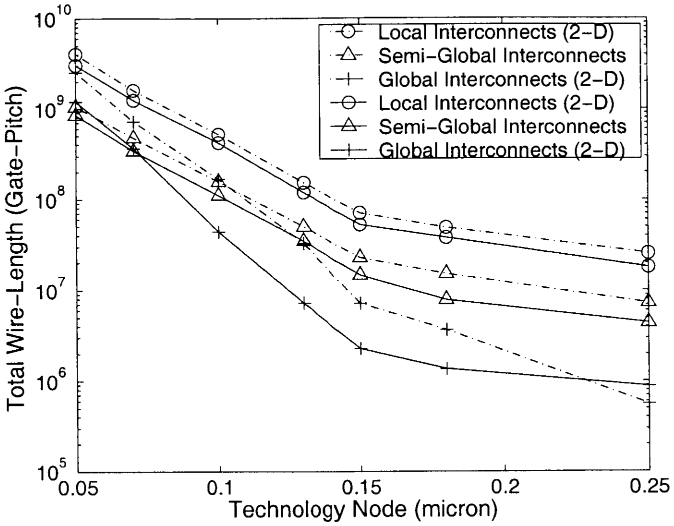


Figure 4-7: Total wire-length of local, semi-global, and global interconnects in 2-D and 3-D implementation.

both without repeaters and with optimum number and size of repeaters. We find that 2.5× to 5× reduction in long wire delay can be achieved by 3-D integration with two strata. The reduction in long wire delay in 3-D ICs is due to both shorter wire-length and wider interconnect pitch (see Table 4.6).

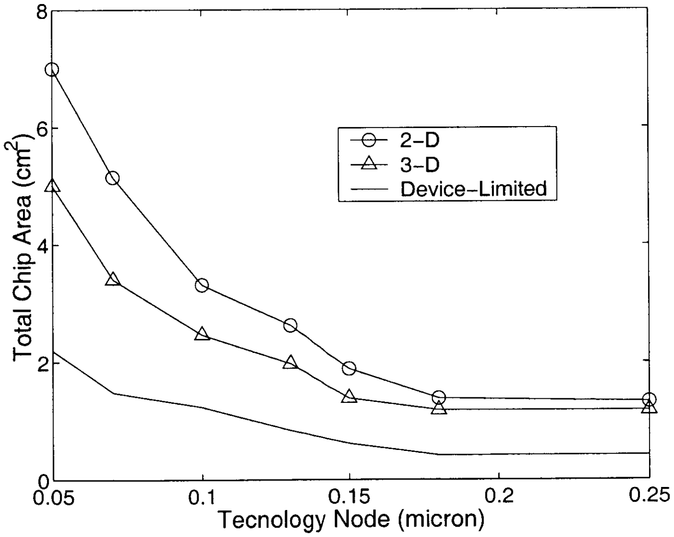


Figure 4-8: Total chip area in 2-D and 3-D implementation. The device-limited chip area is estimated assuming minimum size transistors.

In Fig. 4-7, simulation results of total wire-length of local, semi-global, and global interconnects for the random logic section are shown. The reduction in wire-length is due to the narrower wire-length distribution in 3-D integration. As a result of shorter total wire-length, the wiring-limited chip area in 3-D implementation is smaller, as shown in Fig. 4-8. As a frame of reference, device-limited chip area, assuming minimum size transistors, is also shown in the same figure. As much as 30% reduction in total chip area can be achieved by 3-D implementation with just 2 strata. If the number of strata is increased, at some point the chip area will be comparable to device-limited chip area and further reduction in total chip area will not be feasible.

4.4 CPI and Throughput Estimation

In this section, we examine the contribution of logic and memory elements on CPI and throughput. It has been suggested by many authors that in future technology nodes, the impact of a cache miss would be more severe as it would require higher number of clock cycles to access requested data/instruction from L2/L3 cache or main memory [84, 85]. It is speculated that L1 cache size is not going to increase significantly because a large L1 cache cannot be accessed at the fast processor cycle time [98]. As more transistors are available per chip in future technologies, on-chip implementation of L2 cache would be feasible. L2 cache size can be rather large to minimize the delay penalty associated with a L2 cache miss. Based on our system-level analysis, we find that it would be feasible to integrate 8MB on-chip L2 cache in 50 nm technology node. However, the access time of such a large cache can be as high as 10 clock cycles (see Table 4.5). A significant fraction of the access time in a large cache memory is due to the RC delay of long (global) word line and data output bus. By distributing L2 cache memory in multiple strata, it would be feasible to reduce the RC delay associated with long interconnects. Based on our cache access time models, we find that access time of a 8MB cache (in 50 nm node) can be reduced by 2-3 clock cycles in 3-D implementation with two strata.

To make analytical projections of CPI, first, the cache miss rates, L2 cache access time, and main memory access time are estimated. The cache miss rates can be found in numerous published tables of cache miss rates as a function of cache size and organization [61, 99]. As

Technology Node (nm)	250	180	150	130	100	70	50
Main Memory Speed, t_{mm} (ns)	50	43	37	32	26	21	17
L2 Access time in 2-D Number of Clock Cycles	4	4	4	4	5	6	10
L2 Access time in 3-D Number of Clock Cycles	3	3	3	3	4	4	7
Block Size (Bytes)	32	32	32	32	64	64	64
Bus Width, B_{bus}	128	128	128	256	256	512	512

Table 4.7: Assumptions on memory access time, bus frequency, etc. based on empirical observation and projections from the SIA roadmap [7].

mentioned earlier, we have extended Wada's model to scaled technology nodes to estimate SRAM cache access time [96]. The access time of main memory is estimated assuming DRAM access time decreases at an annual rate of 7% [88]. Some of the assumptions for analytical projection of CPI are listed in Table 4.7

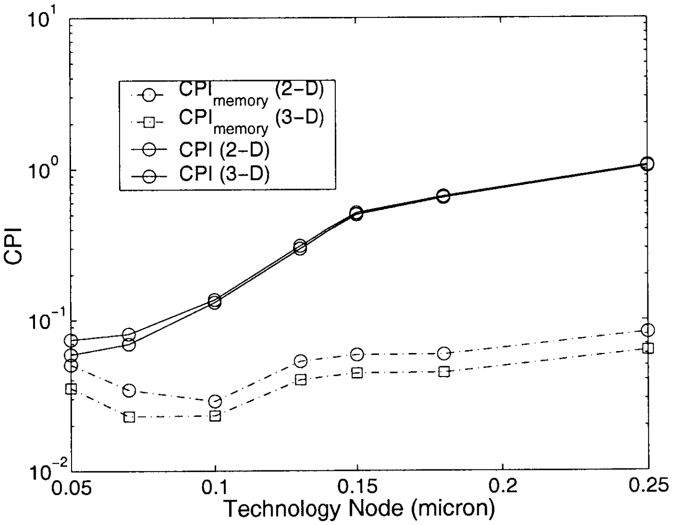


Figure 4-9: CPI and CPI_{memory} in scaled technology nodes for 2-D and 3-D implementation of microprocessors.

In Fig. 4-9 and Fig. 4-10, CPI and CPI_{memory}/CPI are presented for x86-like architecture implemented in 2-D and 3-D scaled technology nodes. With higher number of logic gates and more computation power available per chip in future technology nodes, it is expected that CPI would be smaller. This is observed clearly in Fig. 4-9. Beyond 100 nm technology generation, CPI_{memory} begins to have a significant impact on overall CPI as

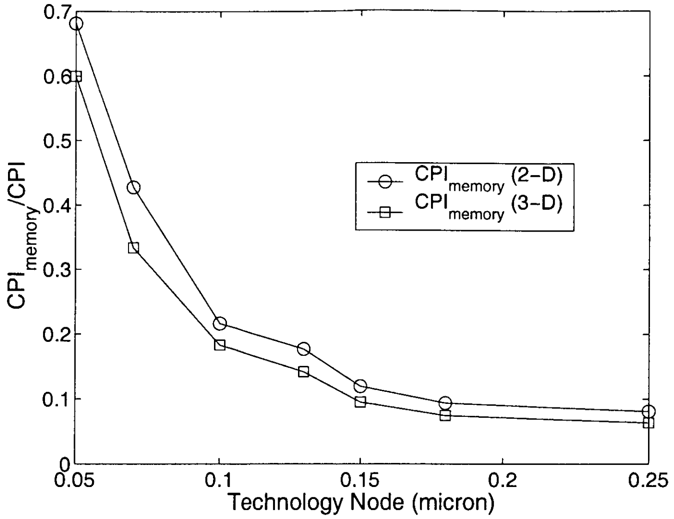


Figure 4-10: The ratio of CPI_{memory} and CPI in scaled technology nodes.

shown in Fig. 4-10. If the L2 cache memory access time could be reduced by 3-D integration, 20% – 30% reduction in CPI can be achieved. For completeness, throughput for 2-D and 3-D implementation of microprocessor has been estimated using Eq. 4.3, and it is shown in Fig. 4-11.

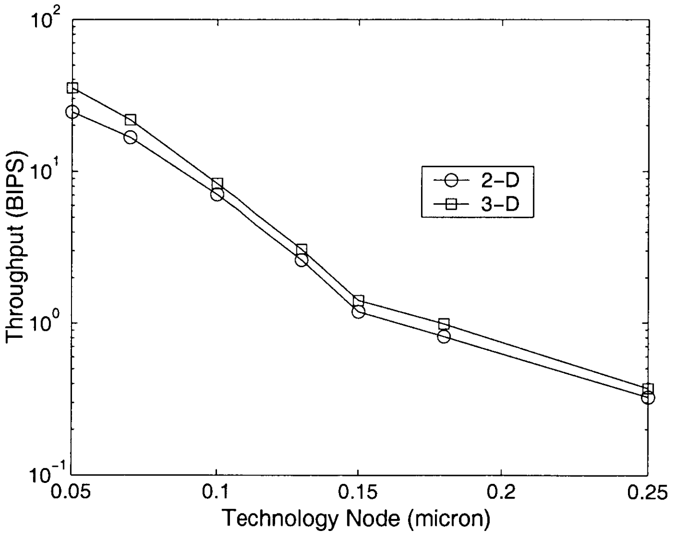


Figure 4-11: Throughput (BIPS) in 2-D and 3-D implementation of microprocessor.

4.5 Summary

In this chapter, opportunities for 3-D implementation of microprocessors have been examined. Based on system-level modeling and analysis, we find that by 3-D integration with two strata, 20% – 30% improvement in system performance, 2.5× to 5× reduction in long wire delay, and 30% reduction in total chip area are feasible. We have also conducted simulation results with higher number of strata. It appears that beyond three strata, chip area becomes comparable to device-limited area, and further reduction in total chip area would not be feasible. In a recent study we have also found that in future technology nodes, the number of repeaters in microprocessors can be reduced by more than an order of magnitude by 3-D integration [21]. This could easily translate into additional reduction in total chip area and power dissipation for 3-D ICs. Beyond 100 nm technology nodes, a significant fraction of the overall system performance, measured by CPI, would be determined by the performance of the memory system. By 3-D integration, it would be feasible to reduce memory access time, increase memory bandwidth, and improve system performance.

Chapter 5

Stochastic Routability Prediction and Three-Dimensional Integration of Field-Programmable Gate Arrays (FPGAs)

5.1 Introduction

There are several options to a system designer for implementing a digital integrated circuit. In one end of the spectrum, one can use full-custom design that requires time-intensive design, verification, and optimization to achieve maximal performance. On the other end of the spectrum, field-programmable gate array (FPGA) based design can be used. In FPGA based implementation, a design is mapped onto an array of programmable logic blocks that are connected by programmable interconnections [100, 101]. A field-programmable gate array based implementation requires fewer design iterations and has the advantage of shorter time-to-market. However, the system performance and density in FPGA based implementation is not as high as full-custom designs.

The fine-grain architecture in FPGA is suitable for bit- and byte-level computation [100]. They can also be used as flexible logic resources for encryption, error corrections, address generations, etc. Since the programming of FPGAs is determined by a configuration memory, the design can be changed quickly and easily if needed without any extra cost. Due to the flexibility in reconfiguring the hardware, FPGAs are becoming very popular for implementing networking and communication ICs where protocols and standards are often revised and updated [102, 103, 104]. As FPGA vendors accelerate their use of advanced deep submicron technologies, higher logic density and better performance will be achievable, and FPGA based designs will be more competitive with traditional ASICs for implementing low-cost general purpose ICs [102, 103, 104].

The implementation of a design using FPGA consists of several steps. First, a high-level description of a circuit/system is converted to a set of boolean equations. These equations are optimized to minimize the number of logic gates and then mapped to a programmable logic array architecture. After the logic mapping, placement and routing are conducted to determine the value of configuration memory bits for connection switches. Depending on the programming technology, interconnections are programmed permanently at the end of the design cycle (in anti-fuse based technology) or configured when the power is turned on (in SRAM based FPGAs) [100].

One of the key design elements in FPGA is the design of interconnect architecture [100, 101, 105]. System performance, chip area requirements, and power dissipation in FPGAs are generally determined by the programmable wiring need [100, 105, 106]. In a programmable gate array design, interconnects account for $40\% - 80\%$ of the overall design delay [100, 105], and interconnects and programmable configuration memory account for more than 90% of the chip area [105]. Recently, it was also found that more than 80% of total power in SRAM-based FPGAs could dissipate in interconnects and clock wiring networks [106]. Since 3-D integration eliminates long wiring need and reduces the total wire-length compared to 2-D integration, 3-D implementation of FPGAs offers significant improvements in system performance, density, and power dissipation.

The wiring architecture (number of wiring tracks, switching configuration, etc.) in FPGAs is determined by placement and routing of a set of benchmark circuits that represent typical wiring need [107]. Analytical models to predict the wiring architecture in FPGAs have also been proposed [64, 108, 109]. In this chapter, existing analytical models for predicting wiring architecture in FPGAs will be examined, and a refined model based on stochastic wire-length distribution will be proposed to estimate the wiring need in 2-D and 3-D FPGAs. System-level modeling and simulation work will also be conducted to assess the impact of 3-D integration on system performance and power dissipation of FPGAs.

5.1.1 Approaches to FPGA Design

The implementation of FPGAs in silicon falls into three groups: SRAM-programmed, antifuse-programmed, and EPROM-programmed. The configurable logic blocks in different implementations are very similar. The primary difference in these implementations is in the programmable routing architecture and the way it is configured.

In SRAM-based FPGAs, pass transistors or tri-state buffers are used as programmable switches to establish interconnections between configurable logic blocks (CLBs) or look up tables (LUTs). The on- or off-state of the pass transistor is determined by the memory configuration of a static memory (SRAM) cell which is connected to the gate of the pass transistor [100, 101], as shown in Fig. 5-1(a). In antifuse-based FPGAs, as shown in Fig. 5-1(b), antifuse device irreversibly changes from a high to a low resistance when a programming voltage is applied across its terminals. The programming pulse melts the dielectric, creating a conductive link of polycrystalline silicon between the electrodes. A programmable connection based on antifuse has low resistance of 100 to 600 ohms, and it also requires much smaller area compared to a static memory based switch. The drawback in antifuse-based FPGA is that it cannot be reprogrammed [100, 101]. In erasable-programmable read-only memory (EPROM)-based FPGA, as shown in Fig. 5-1(c), links between interconnect segments are established by programmable EPROM devices. EPROM cells operate via charge injection in floating gates. When appropriate drain and gate bias voltages are applied, charges are injected in the floating gate. Due to the injected charge in floating gate, there is a shift in threshold voltage. In the unprogrammed state, the EPROM-based switch opens and closes like a transistor, and in the programmed state, the switch is always open, due to a shift in threshold voltage, regardless of whether 0 or V_{DD} volts is is applied to the control gate [100].

In this chapter, routability and opportunities for 3-D integration of SRAM based FPGA will be examined. Static RAM based FPGA is probably the most popular and widely used approach to FPGA design, being manufactured by Xilinx, Altera, Lucent, Atmel and many other companies [102, 103, 104].

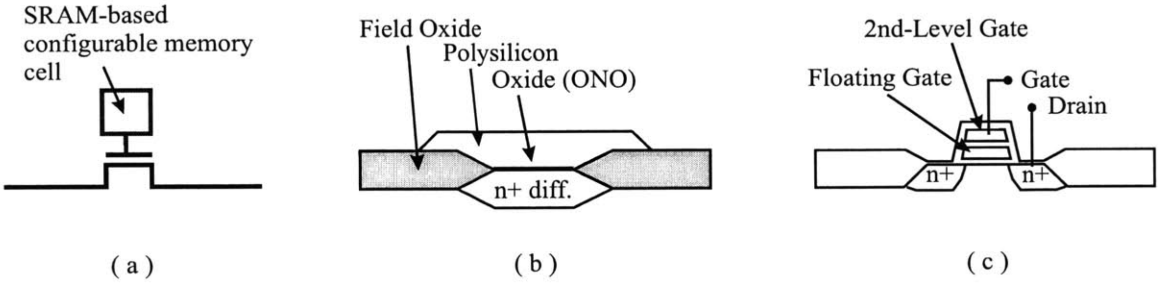


Figure 5-1: Various approaches for implementing programmable interconnections in (a) SRAM-based (b) antifuse-based and (c) EPROM based FPGAs.

5.2 SRAM Based FPGA

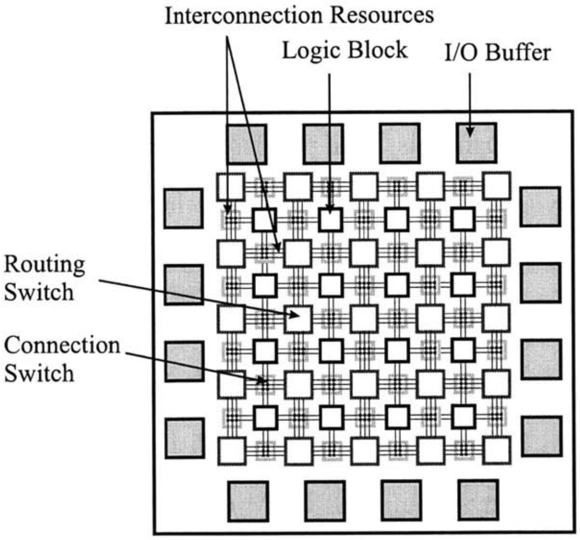


Figure 5-2: A static memory-based FPGA.

A generic SRAM-based FPGA architecture is depicted in Fig. 5-2. It consists of a two-dimensional array of logic blocks and horizontal and vertical routing channels. This type of architecture is also known as island based design and resembles closely with FPGA designs by Xilinx [102], Lucent Technologies [103], and Altera [104]. The configurable logic block (CLB) can be programmed to perform a variety of functions for a set of input variables. It is implemented using NAND, NOR, and XOR gates and multiplexers, or it can be implemented using look up tables (LUTs). The number of unique inputs, K, to a CLB or LUT can range anywhere from 2 to as high as 16. It has been found that for the most area-efficient design, optimum value of K is approximately $3 - 4$ [101]. Approaches for implementing configurable logic blocks and look up tables are shown in Fig. 5-3.

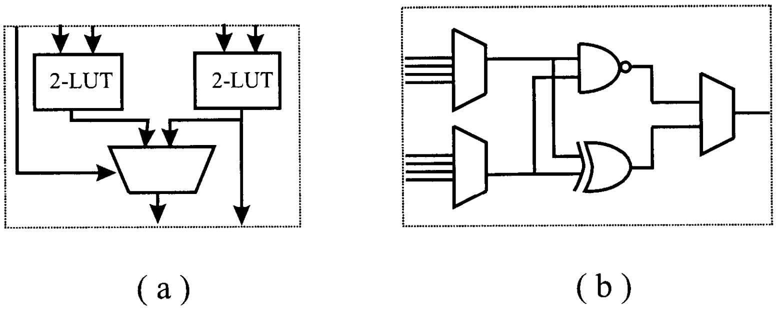


Figure 5-3: Programmable logic cells based on (a) look up tables and (b) configurable logic gates in SRAM-based FPGA.

The programmable interconnections in SRAM-based FPGA consist of routing switches, connection switches, and interconnect segments. In Fig. 5-4, schematics of conventional routing and connection switches are shown. The routing switch is generally implemented by pass transistors or tri-state buffers. The flexibility of a routing switch is determined by the maximum allowable fan-out, F_s, provided to an incoming wiring segment by the routing switch. A popular 2-D routing switch allows each incoming wiring segment to connect to wiring segments on three other sides of the routing switch box, resulting in $F_s = 3$ [102]. An extension of this routing switch topology to 3-D will result in $F_s = 5$. The connection switches are used to establish input or output connections between programmable logic blocks and wire segments. The flexibility of the connection switch is determined by the number of tracks, F_c, each logic block pin can connect to. By routing a set of benchmark circuits, it has been found that for complete routability it is sufficient to have $0.7W \leq F_c \leq W$, where channel density, W, is the maximum number of tracks per CLB or LUT in horizontal or vertical direction [101].

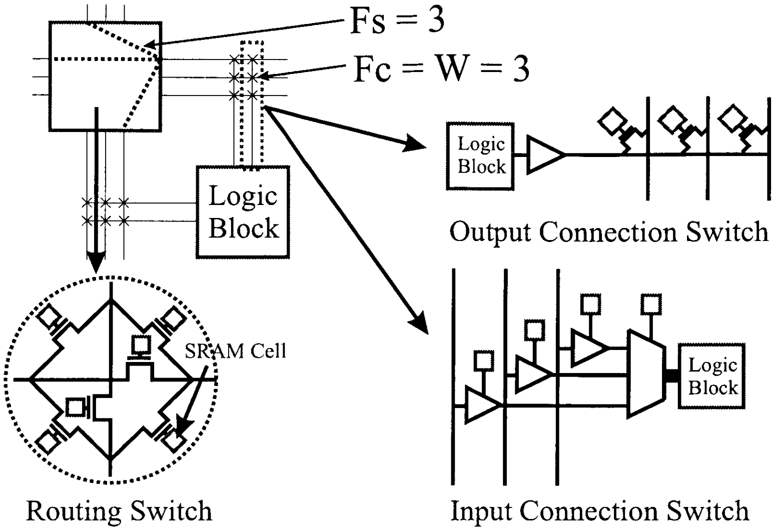


Figure 5-4: Programmable routing and connection switches in SRAM-based FPGAs.

In early FPGAs, wiring tracks consisted mostly of short wire segments that spanned one LUT or one unit. Longer wires could be formed by connecting short wire-segments using pass transistor routing switches. Though one unit long wire segments provide good wiring utilization, they degrade the performance of long interconnections. This is because long interconnections are formed by connecting many pass transistors which add significant series resistance and capacitance. To reduce the number of pass transistors in long interconnections, wire segments of various lengths are used in high-performance FPGAs. For example, Xilinx XC4000X series FPGA contains 25% length 1 wires, 12.5% length 2 wires, 37% length 4 wires, and 25% one-quarter of chip edge long wires [102, 110]. Though an assortment of wire segments reduces signal delay in long interconnections, it can result in under-utilization of many long wire segments and an increase in channel density and chip area [110].

Some of the key performance metrics for SRAM-based FPGAs are:

- Equivalent Gates or Density
- Logic Utilization
- Routability
- Speed

Equivalent Gates, density or device capacity in FPGA-based design is generally expressed in terms of gate counts per chip. It is the equivalent number of 2-input NAND gates that would be required to implement the same functionality. However, FPGAs don't consist of 2-input NAND gates. They have logic components such as LUTs, multiplexers, flip flops, etc. A more accurate methodology for measuring logic density is based on the concept of logic cells [102]. A logic cell can be defined as the combination of a 4-input LUT and dedicated registers. For example, Xilinx's XC3000 FPGAs have 1.5 logic cells/CLB; the XC4000, 2.375 logic cells per CLB. *Logic utilization* in FGPA-based design is generally $10\% - 50\%$ lower than ASIC based designs. The inefficiency in logic utilization arises from several factors. Routing congestion may result in LUTs or CLBs that are not accessible. Sometimes, complex logic cells are used to implement simple functions that in a custom design would require only a few gates.

Routability describes the effectiveness in utilizing the programmable routing resource. Routability of a design in FPGAs depends strongly on the configuration of wire segments as well as on the values of W, F_s, and F_c. These parameters are determined heuristically by the wiring need of representative benchmark designs. The *speed or performance* in FPGAs is limited by interconnect delay, and it can account for $40\% - 80\%$ of overall design delay [105]. The wiring nets in FPGA are more resistive and capacitive compared to the wiring nets with similar length in custom design. This is due to the high resistance and capacitance of pass transistors in the programmable interconnects as well as the large wiring resistance and capacitance.

Since a large fraction of the chip area in FPGAs is dedicated to programmable interconnects, it is not surprising that most of the power in FPGAs is dissipated in reconfigurable interconnects. Recently, a detailed analysis of power consumption in Xilinx XC4003A was conducted, and it was found that 80% of total power dissipation was due to driving interconnect and clock wiring capacitance [106].

In the following sections, a methodology based on system-level modeling will be presented to estimate some of the performance metrics for SRAM-based 2-D FPGAs. By extending the models for 2-D FPGAs to three-dimension, key advantages for 3-D implementation of FPGAs will also be examined.

5.3 Stochastic Models for Routability Prediction in FPGAs

The routability of a design in FPGA-based implementation depends on the configuration of the LUTs and routing resource. Typically, optimum configurations of LUTs and routing resource are determined by placement and routing experiments with benchmark circuits. In this section, we address the routability of a design in FPGAs, before place-and-route, based on analytical models. These models are useful for providing an early feedback for various design trade-offs without having to go though many iterations of time consuming and laborious placement and routing process.

A popular analytical model for predicting routability in FPGAs is based on a two-dimensional stochastic model for channel density by Gamal [64]. His analysis suggests that the channel density (i.e. number of wiring tracks per channel), W, in array based FPGAs follows Poisson distribution, and the average channel density, $\overline{W}$, is given by

$$\overline{W} = \frac{\gamma \overline{L}}{2}, \tag{5.1}$$

where $\overline{L}$ is the average wire-length and γ is the average number of wires emanating from each logic block [64]. An enhancement of Gamal's model has also been proposed that takes into account multi-terminal nets in predicting channel density [68]. Recently, it has been also found that routability is best predicted by estimating the total wire-length in a circuit, not by the mean wire-length times pins per cell as described in Gamal's model [108]. We also find this to be consistent with our routing and placement experiments with benchmark circuits. In [108] random net lists were generated based on a set of input parameters such as pins per LUT, Rent's parameters, etc. In deriving our model for predicting channel density, we also follow a similar methodology; however, the net lists and total wire-length are estimated based on the stochastic wire-length distribution [17].

5.3.1 Calibration and Validation

We use Rent's rule for estimating the number of interconnects and wire-length distribution in FPGAs [17]. Knowing Rent's parameters (k and p) and the number of LUTs needed to implement a design, wiring complexity of the design can be estimated. In a K-input LUT table based FPGA, Rent's parameter, $k \leq K+1$, and a typical value of p is in the range of 0.7-0.8 [68]. We estimate the channel density in FPGAs that consist of one unit long wire segments by

$$W = \frac{\sum_{l=1}^{2\sqrt{N}-2} l f(l) \chi_{fpga} + l_r}{2Ne_t}, \tag{5.2}$$

where l is the wire-length, $f(l)$ is the wire-length distribution, χ_{fpga} is a point-to-point to net-length conversion factor, l_r is the additional length of occupied wiring tracks due to non-ideal locations of input/output terminals in the LUT (see Fig. 5-5), and e_t is the utilization of wiring tracks. The derivation of Eq. 5.2 is based on the assumption that for a design, the available length of wiring tracks, $W2Ne_t$, is equal to the required total wire-length, $\sum_{l=1}^{2\sqrt{N}-2} l f(l) \chi_{fpga} + l_r$. The wire-length distribution, $f(l)$, can be found using the methodology presented in Chapter 2. The values of χ_{fpga} and e_t can be estimated and calibrated by placement and routing of a set of benchmark circuits.

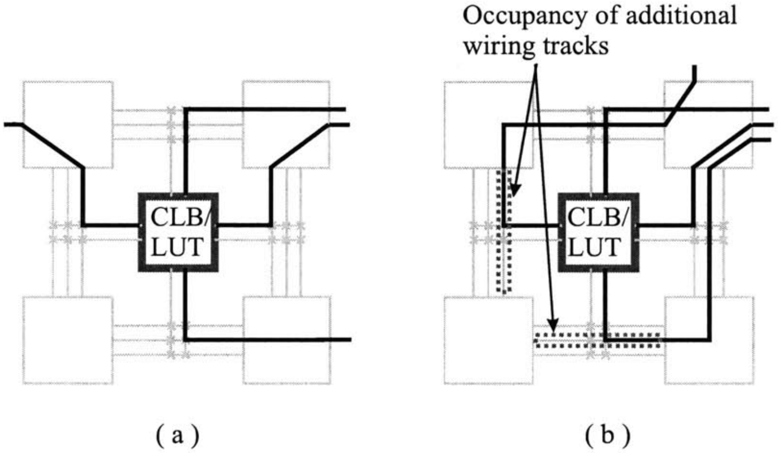


Figure 5-5: (a) An ideal scenario where connections to input/output terminals of a LUT do not require additional wire-segments. (b) A non-ideal scenario that requires utilization of 2 additional wire segments for making input/output connections.

We begin by examining the wiring complexity of a set of benchmark circuits using a place-and-route tool, SEGment Allocator (SEGA), developed at University of Toronto [111]. SEGA performs routing and placement to optimize for speed-performance. It is assumed that all routing channels have equal number of tracks, W. For simplicity, we also assume all wire segments are one unit long, $F_c = W$, and $F_s = 3$. Some of the properties of

Circuit Name	Number of Nets	Number of Graphs	Estimated Average Fan-Out
alu2	153	511	3.3
alu4	256	851	3.3
9symml	79	259	3.27
c499	145	360	2.48
c880	174	427	2.45
k2	404	1256	3.1
z034	608	2135	3.51

Table 5.1: Benchmark circuits for validating stochastic routing models.

the benchmark circuits [111] for validation and calibration of our models are shown in Table 5.1. The logic functions for these circuits are mapped to 4-input LUT based FPGAs. Using SEGA, we estimate the minimum value of W that will result in 100% routability for the benchmark circuits. One of the outputs of the routing tool is the number of shared wire-segments used by nets with fan-out more than one. We find that only 5% – 10% of wire-segments are shared, resulting in $\chi_{fpga} = 90\% - 95\%$. The routing tool also outputs the number of occupied wire-segments which can be used to estimate utilization factor of wiring tracks, e_t. Based on the placement and routing of benchmark circuits we find that $e_t = 40\% - 50\%$.

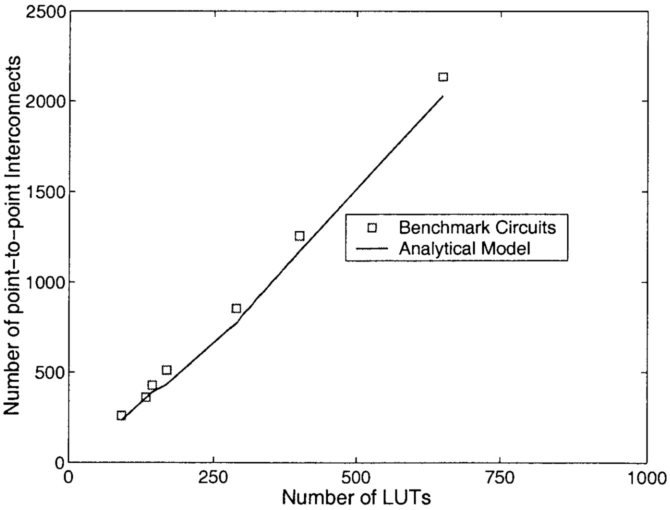


Figure 5-6: The number of point-to-point interconnects in benchmark circuits in 4-input LUT based FPGAs.

In Fig. 5-6, the number of graphs or point-to-point interconnects for the benchmark circuits and the projection based on Rent's rule are shown. By hierarchically partitioning a logic graph, it can be shown that the total number of point-to-point interconnects in an IC is given by [62]

$$I_{total} = \frac{fo}{fo+1} kN(1 - N^{(p-1)}). \tag{5.3}$$

We find that the analytical model, based on Donath's methodology, for estimating total number of interconnects agrees very well with the routing and placement results from benchmark circuits. For this analysis, it has been assumed that k = 5 and p = .75. Though individual circuits may have slightly different Rent's parameters, the interconnection complexity as a function of number of LUTs is quite accurately modeled by $k = K + 1$ and $p = .75$.

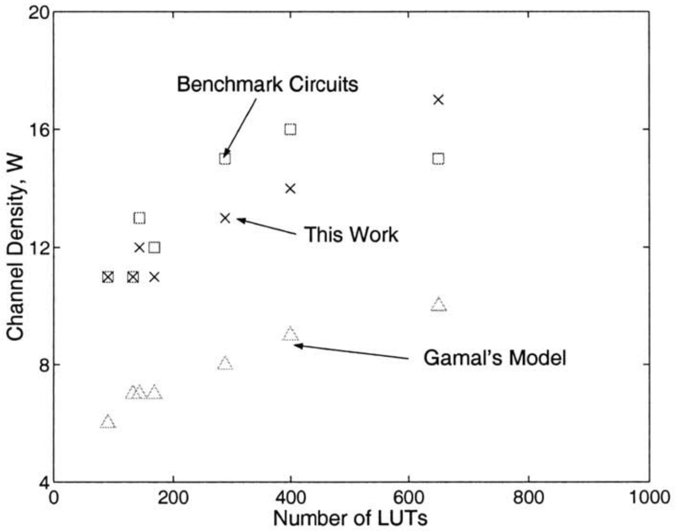


Figure 5-7: Channel density for the benchmark circuits implemented in 4-input LUT based FPGAs.

Next, we estimate the channel density, W, using Eq. 5.2. We assume 50% input/output terminals in LUTs could be located on the undesirable sides, and they contribute to additional wire-length of $l_r = 50\% \times I_{total} \times one\ unit\ wire-length$. In Fig. 5-7, estimated values of channel density for the benchmark circuits are compared with the results from SEGA and Gamal's analytical model. Our analytical model predicts the channel density within ±15% error. We also find that estimated average channel density based on Gamal's model is much smaller compared to the results from SEGA. We believe, Gamal's model can be improved, if a parameter reflecting the effective utilization of wiring tracks is incorporated in his model.

5.4 Opportunities for 3-D Implementation of FPGAs

In the earlier section, it has been shown that channel density in array-based FPGAs is proportional to the total wire-length. Three-dimensional integration can result in a significant reduction in total-wire length. As a result, the channel density in FPGA can be reduced by 3-D integration which may lead to reduction in chip area and improvements in system performance. In the past, various approaches for implementing 3-D FPGAs have been considered [112, 113, 114, 115, 116]. They include FPGAs based on 3-D routing switches with electrical or optical inter-stratum interconnections [112, 113, 116] or partitioning of memory and routing functions in different strata [115]. These earlier works were focused on either routing architectures or FPGA technologies. In this section, both routability and technology related issues in 3-D FPGAs will be examined.

5.4.1 Channel Density in 3-D FPGAs

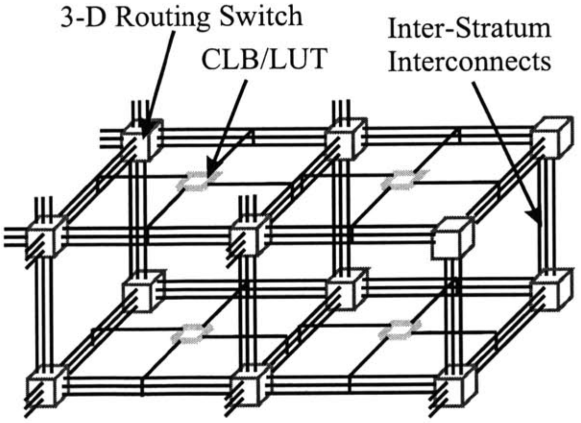


Figure 5-8: A 3-D implementation of FPGAs with 3-D routing switches.

We consider a 3-D implementation of FPGA architecture with 3-D routing switches, as shown in Fig. 5-8 and also discussed in [112, 113]. In [112, 113], MCM-based packaging approaches to 3-D are considered which may not be suitable for 3-D FPGAs with high logic density. We assume $F_c = W$ and $F_s = 5$, where each incoming wire-segment can connect to other wire-segments on five sides of a cubic switch box. We also assume the wiring track utilization in 3-D FPGA is comparable to that of 2-D implementation. One of the drawbacks in implementing 3-D routing switch is that it would require more pass transistors and SRAM cells per routing switch box per channel. However, if channel density can be reduced by 3-D integration, it will be possible to reduce the total number of routing switches and configurable memory bits. In Fig. 5-9, estimated values of channel density are shown for 2-D and 3-D implementation of FPGAs as a function of number of LUTs. Channel density in 3-D FPGAs is estimated by Eq. 5.4

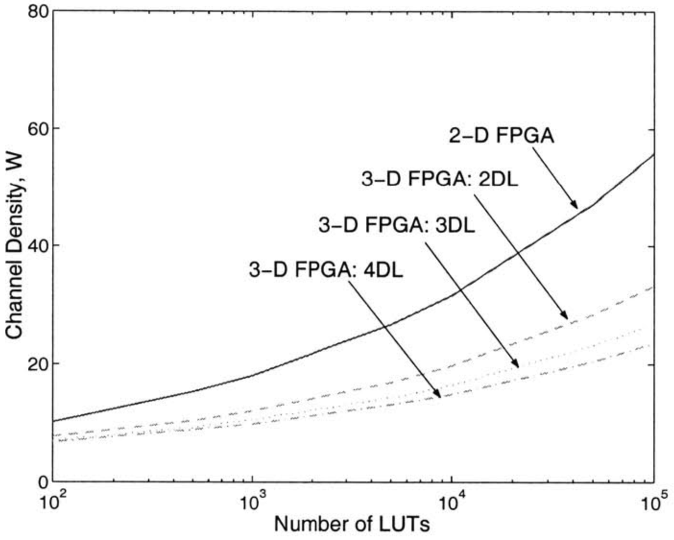


Figure 5-9: Channel density in 2-D and 3-D 4-input LUT based FPGAs as a function of number of LUTs. The routing resource consists of one segment long wires. It has been assumed that Rent's parameters $k = 5$ and $p = 0.75$, average fan-out $= 3.5$, $\chi_{fpga} = 0.95$, and $e_t = 0.4$. In the figure, DL stands for number of device layers or strata.

$$W = \frac{\sum_{l=1}^{2\sqrt{N/N_z}-2+(N_z-1)t_z} l f_{3D}(l)\chi_{fpga} + l_r}{(2N + \frac{(N_z-1)N}{N_z})e_t}, \tag{5.4}$$

where $f_{3D}(l)$ is the 3-D wire-length distribution, N_z is the number of strata, and all other parameters are defined the same way as in Eq. 5.2. The higher value of the denominator reflects the availability of more wiring tracks per LUT in 3-D configuration. The 3-D wire-length distribution is derived using the methodology described in Chapter 3. In Fig. 5-9, as more strata are integrated, the average and total wire-length become shorter, resulting in a significant reduction in channel density.

5.4.2 Logic Density

In FPGAs, the number of usable logic gates is much smaller compared to that of an ASIC-based design with the same die size. It is desirable to increase the logic density by scaling or innovative system architecture so that more functionality can be integrated within a field-programmable gate array chip. In SRAM-based FPGAs, chip area is primarily limited by the area dedicated to programmable interconnects and the configurable memory bits. In [105], by empirical observation, it has been found that 80% – 90% of the area in FPGA is dedicated to switches and wires making up the reconfigurable interconnect. The CLBs/LUTs account for only a few percent of the total area. Due to the programmable interconnect overhead, there is roughly a $20\times - 50\times$ density disadvantage in FPGAs compared to a full-custom design [105].

The number of switches and programmable memory bits in FPGAs is roughly proportional to the channel density. In Section 5.4.1, it has been shown that 3-D integration results in a significant reduction in channel density which could lead to smaller chip area and higher logic density. In this section, logic density in FPGAs will be estimated by modeling the area dedicated to LUTs, connection switches, routing switches, multiplexers, SRAMs, and buffers. The area model is based on counting the number of minimum-width transistor areas needed to implement FPGAs. We follow the methodology presented in [110] to estimate the chip area. The definition of a minimum width transistor area and a method to increase the drive strength of a transistor are shown in Fig. 5-10. In some recent works, dependencies of chip area and performance on transistor sizing have been investigated [110, 117]. Based on these studies, as well as our own HSPICE-based analysis, we find that to minimize delay or power-delay product, the optimum value of pass transistor size in switches and buffers is roughly $10\times - 15\times$ the minimum-width transistor's size.

To assess improvements in logic density, measured by the number of LUTs per unit area per stratum, we consider FPGAs implemented with 4-input LUTs. The LUT consists of a

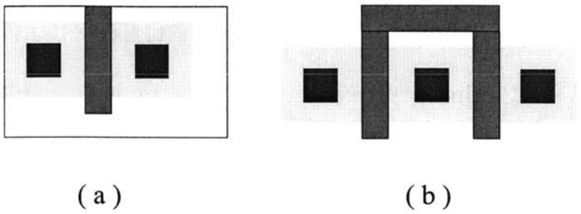


Figure 5-10: (a) A minimum-width transistor including the area needed to satisfy design rules. (b) A method to increase the drive strength of a transistor by folding the polysilicon gate.

Components	Area
4-input LUT	235
An Input Connection	$(6log_2W + 2W) + 33.5W$
An Output Connection	$20 + 13.5W$
A Routing Switch Box	$13.5WF_s(F_s + 1)/2$

Table 5.2: Models for chip area dedicated to various components that make up SRAM-based FPGAs.

pass transistor multiplexer, a register, and a set/reset logic block [110]. We estimate the area of input and output connection switches and routing switches using the switch configuration shown in Fig. 5-4. We assume the buffer and pass transistor have 10× minimum drive strength. The 3-D routing switch box requires $F_s(F_s+1)/2 = 15$ pass transistors per channel compared to 6 transistors per channel in 2-D implementation. As a result, for small number of LUTs, there is a higher area-overhead due to routing switches in 3-D implementation of FPGAs. The area models of various components in SRAM-based FPGAs, measured in units of minimum width NMOS area, are listed in Table 5.2. In addition to the models listed in Table 5.2, the 3-D routing switch boxes on top and bottom most strata require fewer pass transistors, and the corresponding saving in (pass and SRAM transistor) area is $\sim 2 \times (N/N_z) \times 4 \times 13.5W$. We also assume Si efficiency is 60% [110].

In Fig. 5-11, estimated values of improvements in logic density (number of LUTs per unit area per stratum) are presented as a function of number of LUTs which have been calculated using the area model presented in Table 5.2. In 3-D FPGAs, as the number of LUTs is increased, significant reduction in channel density can be achieved. As a result, the area dedicated to connection and routing switches will be smaller. Based on our analysis, for example, in Xilinx's high-density Virtex family FPGAs, XCV3200E, that is projected to have 73K logic cells[1] [102], the improvement in LUT density could be 25% – 60% in 3-D implementation with 2 – 4 device layers.

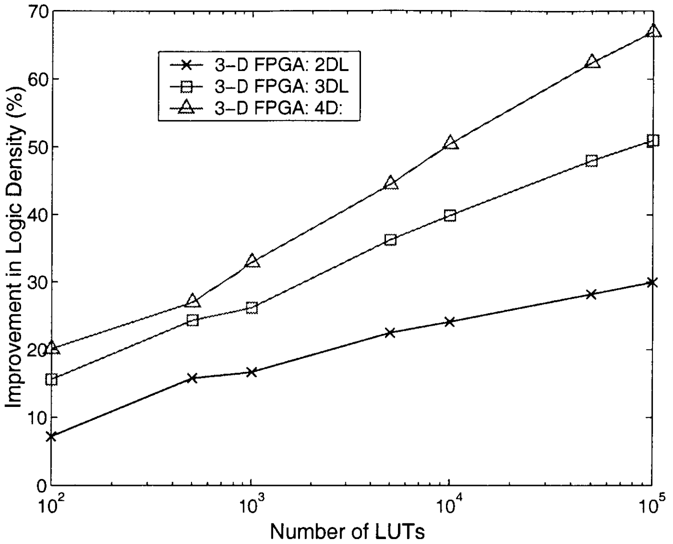


Figure 5-11: Improvements in logic density as function of number of LUTs in 3-D FPGAs.

5.4.3 Interconnect Delay

In this section, interconnect delay of average length and chip-edge length interconnects in 2-D and 3-D FPGAs will be examined. Using the models presented in Table 5.2, we estimate the LUT area and wire-length assuming all wire segments are one unit long. Wire segments of various lengths are often used in FPGAs to reduce delay for long interconnections. When long wire segments area used, as a result of under utilization of many wiring tracks, the chip area may increase [118]. In our analysis for estimating interconnect delay in long wire segments, for simplicity and illustration purposes, the increase in chip area due to the under-utilization of long wire segments is not taken into account.

We consider a 4-input LUT based FPGAs with 20K LUTs and implemented in .25 μm technology. The estimated area per stratum for 2-D implementation and 3-D implementation with 2, 3, and 4 strata are 7.84 cm^2, 3.1 cm^2, 1.77 cm^2, and 1.21 cm^2, respectively.

[1] The concept of logic cell is used to measure the logic density in FPGAs. A logic cell is defined as the combination of a 4-input LUT and a dedicated register. For example, Xilinx's XC3000 FPGAs have 1.5 logic cells/CLB; the XC4000, 2.375 logic cells per CLB.

The corresponding values of channel density are 41, 24, 20, and 18, and the average wire-lengths are 8.3, 6.22, 5.4, and 4.9. It is assumed that M3 and M4 interconnect levels are used for routing programmable interconnects; the wiring pitch is 8λ, where λ is half of the the minimum feature size. The estimated wiring capacitance and resistance are 2.8 pF/cm and 540 Ω/cm, respectively.

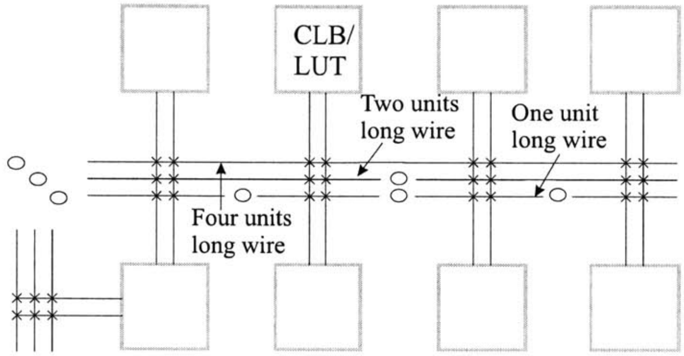


Figure 5-12: Alternative routing choices using various length wire segments in SRAM-based FPGAs. Wire-length is measured in units of LUT pitch, the average separation between neighboring LUTs.

We estimate interconnect delay from an output terminal of a LUT to an input terminal of another LUT, using HSPICE, as a function of interconnection length. We assume the interconnections are implemented with various length wire segments, as shown in Fig. 5-12. Configurations of input and output terminals are the same as shown Fig. 5-4. The buffers and pass transistors have 10× minimum drive strength, and a gate voltage of $V_{dd} + V_t$ is applied in pass transistors to eliminate a V_t drop across drain-to-source. The simulation results are shown in Fig. 5-13. Estimated interconnect delay is roughly proportional to n_{tr}^2, where n_{tr} is the number of pass transistors in the interconnection path. As a result, for the same wire-length, interconnections implemented with longer wire segments, which require fewer pass transistors in series, have better performance. By examining Fig. 5-13, one can easily appreciate the advantage of reducing the wire-length and the number of pass transistors in an interconnection.

To compare interconnect delay between 2-D and 3-D implementations of FPGAs, HSPICE simulations have been performed to estimate interconnect delay of average-length and chip-edge length wires. We assume the average length connections are implemented using one

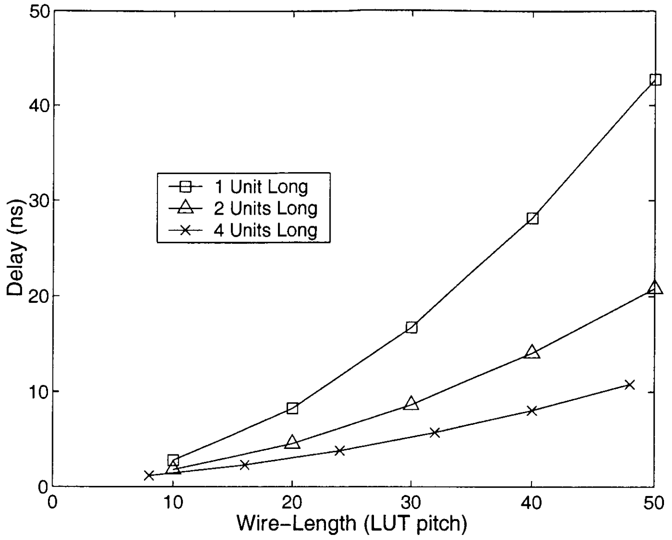


Figure 5-13: Interconnect delay as a function of length for various wiring architecture implemented using 1, 2, and 4 unit long wire segments. The delay is roughly proportional to n_{tr}^2, where n_{tr} is the number of pass transistors in the interconnection path. LUT pitch is the separation between adjacent LUTs on the same row or column.

unit long wire segments. In 3-D implementation, 2-D routing switch boxes are replaced by 3-D routing switches. Simulation results of interconnect delay of average-length wire in 2-D and 3-D FPGAs are shown in Fig. 5-14. We find that the drain junction capacitance (at the output terminals of a LUT and in routing switches) and interconnect capacitance are comparable for short interconnections. Both shorter wire length and reduction in channel density, W, result in lower capacitance for short interconnections in 3-D FPGAs, and subsequently, the reduction in interconnect delay.

Similarly, we have also examined interconnect delay of chip-edge length wires. We assume these wires are routed in 1/4 chip-edge length wire segments. Two cases are considered: in the first case there is no buffer driving the routing switches and in the second case 10× buffers are inserted to drive 2-D and 3-D routing switches. We find that delay in chip-edge length connection is limited by interconnect's RC delay. The significant reduction in chip-edge length interconnection delay in 3-D FPGAs is primarily due to the lower wiring capacitance and resistance. Though the capacitive loading due to 3-D routing switches is higher compared to that of 2-D routing switches, they (3-D routing switches) do not seem to have a significant impact on overall delay.

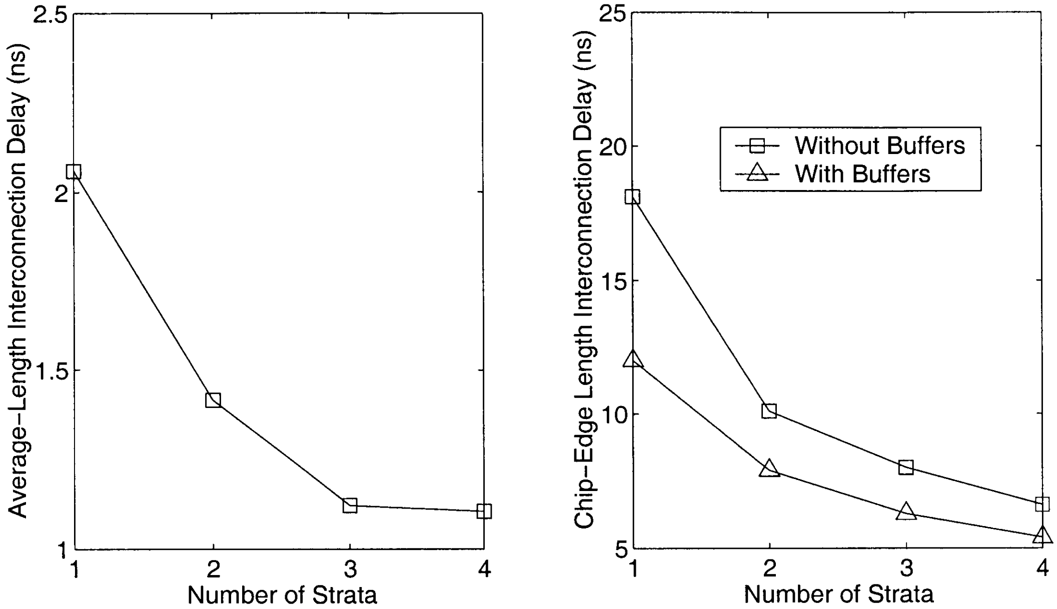


Figure 5-14: Interconnect delay as a function of number of strata in 2-D and 3-D FPGAs. The average-length wires are implemented by connecting one unit long wire segments, and the chip-edge length wires are implemented by connecting four 1/4 of chip-edge length wire segments.

5.4.4 Power Dissipation

In field-programmable gate arrays, a significant fraction of total power is dissipated in the programmable interconnects and clock network. By 3-D integration, both wire-length and channel density can be reduced, resulting in tremendous savings in power dissipation in interconnects and clock network. In this section, power dissipation in 2-D and 3-D FPGAs will be estimated based on analytical models. It will be assumed that the primary contribution to total power dissipation is due to dynamic power dissipation in the programmable interconnects, clock network, and LUTs. Power dissipation in I/O pads is generally less than 10% of total power dissipation, and for simiplicity, it will not be included.

Assumptions

To estimate the total power dissipation, we consider a 4-input LUT and SRAM-based FPGAs. The configurations of input and output connections and routing switches are the same as shown in Fig. 5-4. We assume the buffer associated with an input connection can be shared between adjacent LUTs to reduce power dissipation and chip area [110]. The power dissipation in programmable interconnects is estimated by taking into account power dissipation associated with one unit long signal wires, routing switches, connection switches, and buffers for input connections. Based on power-delay analysis, W/L = 10 seems to be the optimum size for buffers and pass transistors. We assume only a fraction of the programmable interconnect resource, determined by the utilization factor of wiring tracks, e_t, is utilized and contributes to total dynamic power dissipation. The total capacitance associated with a LUT includes the capacitance of a 4-input pass transistor based multiplexer, latch, and set-reset logic, and output buffers [110].

Some of the popular methods for global clock distribution in FPGAs are based on variations H-spine clock network, being used in Xilinx's Spartan-II and Lucent's ORCA4 architecture [102, 103]. This type of clock distribution network does not result in zero skew. However, by inserting of buffers, skew can be kept at a tolerable level. To estimate the capacitance associated with clock distribution network, we assume the topology shown in Fig. 5-15 which is used in Xilinx's Spartan-II FPGAs. The global clock signal is routed to the center of the chip, and it drives a horizontal backbone. From the horizontal backbone, the clock signal is distributed to global clock columns, and from global clock columns it is distributed to LUTs. The size of the buffers is generally chosen such that their input capacitance is comparable to the load capacitance they have to drive [119]. Total power dissipation in clock distribution network is estimated by taking into account power dissipation in the horizontal backbone, global clock columns, and latches within the LUTs. It is assumed that the wiring pitches of horizontal backbone and global clock columns are $32F$ and $16F$, respectively, where F is the minimum feature size.

In 3-D integration, we assume the clock distribution network is located on the bottom most stratum, and inter-stratum interconnects originating from global clock columns will be used to distribute clock signals to top strata. Based on these assumptions, in FPGAs with N LUTs and N_z strata, the number of global clock columns is $2\sqrt{}(N/N_z)$, and their length is proportional to $\sqrt{}(N/N_z)$.

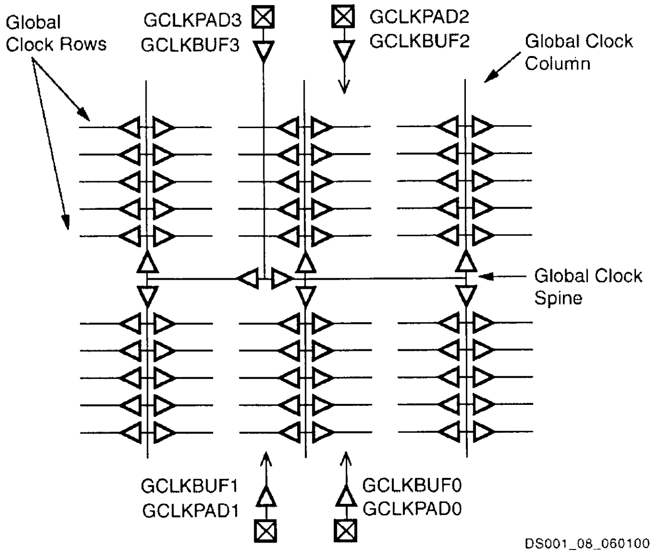


Figure 5-15: A typical global clock distribution network in FPGAs [102].

Simulation Results

Based on the models for LUTs, programmable interconnects, and clock network, we have estimated the capacitance associated with them and their total power dissipation. In a typical 2-D FPGA implemented in .25 μm technology node, we find that power dissipation in programmable interconnects is 50% – 60% and in clock network 37% – 45% of total power dissipation. The rest of the power dissipation is in LUTs. Within the programmable interconnection, the total power dissipation in connection switches, routing switches, and buffers is comparable to that of signal interconnects.

In Fig. 5-16, simulation results of total capacitance associated with programmable interconnects, clock distribution network, and LUTs are shown for 2-D and 3-D implementation of FPGAs. The number of strata in 3-D FPGAs is two. Further reduction in interconnect and clock network capacitance can be achieved by integrating more strata. In FPGAs with 20K 4-input LUTs and 2-4 strata, reductions in programmable interconnect capacitance and clock network capacitance are 24% – 43% and 50% – 75%, respectively. We find that clock network capacitance is dominated by the capacitance of interconnects and buffers. The significant reduction in clock network capacitance is due to the lower capacitance associated with global clock columns in 3-D implementation.

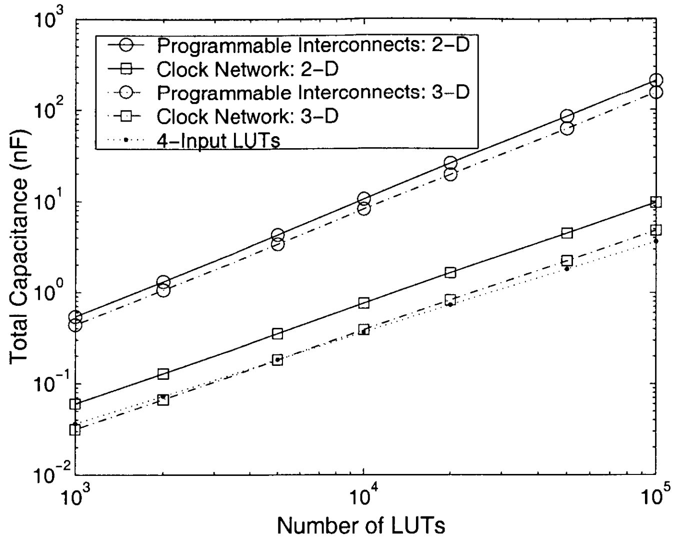


Figure 5-16: Comparison of total capacitance associated with programmable interconnects, clock distribution network, and LUTs in 2-D and 3-D FPGAs with 2 strata and implemented in .25 μm technology generation.

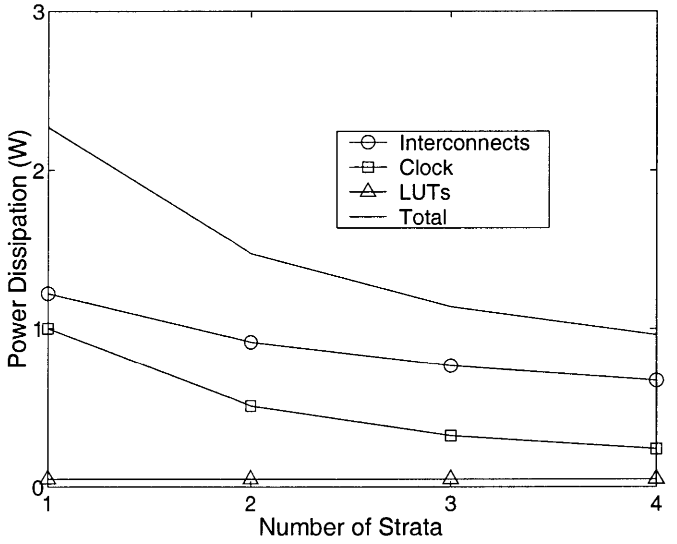


Figure 5-17: Power dissipation in 2-D and 3-D FPGAs with 20K logic cells and implemented in .25 μm technology node with 2.5 V supply voltage and 100 MHz clock frequency.

In Fig. 5-17, estimated values of total power dissipation are presented for 2-D and 3-D FPGAs as a function of number of strata. We assume the number of 4-input LUTs (or logic cells) is 20K, and the FPGA is implemented in .25 μm technology with 2.5 V supply voltage. We assume the clock frequency is 100 MHz which is typical for FPGAs such as XC4000XV with comparable number of equivalent logic cells [102]. By 3-D integration with

2-4 strata and the same clock frequency as 2-D FPGAs, the reduction in power dissipation is 35% – 55%.

5.5 Summary

In this chapter, applications of stochastic wire-length distribution model to estimate channel density and chip area have been presented. By extending the models for 2-D integration to 3-D, key advantages for implementing 3-D FPGAs have been discussed. Based on stochastic routability and HSPICE modeling, we find that 3-D implementation of FPGAs results in a significant improvement in LUT density and interconnect delay. With 2 – 4 strata, 20% – 40% improvement in LUT density is achievable in 3-D FPGAs with 20K LUTs. Improvements in interconnect delay can be as much as 45% for short interconnects and 60% for long interconnects. We also find that due to the lower capacitance in signal wires, clock distribution network, and smaller channel density, power dissipation in 3-D FPGAs with 2-4 strata is 35% – 55% smaller than that of 2-D FPGAs.

Chapter 6

Thermal Analysis of Three-Dimensional Integrated Circuits

6.1 Introduction

In the previous chapters, some of the key advantages of 3-D ICs have been presented for various applications. By 3-D integration, significant reduction in interconnect delay and chip area can be achieved. As more strata are integrated, there is a point of diminishing return beyond which improvements in system performance begin to slow down. Based on our models, this is attributed to the cost or complexity for manufacturing 3-D ICs and the reduction in wiring efficiency due to via blockage. Moreover, compliance with thermal design guidelines could also limit the number of strata in 3-D ICs. The chip or junction temperature in integrated circuits depends on the total power dissipation and the heat removal capabilities of ICs and packaging technologies. Integration of multiple strata results in higher thermal resistance from junction[1] to ambient, leading to higher chip or junction temperature. For reliable operation of devices and interconnects, the chip temperature is

[1] In the context of thermal analysis, "junction" is used commonly to describe the channel region of a device.

typically required to be less than 100^0C [7]. As today's high-performance systems such as microprocessors already operate at a temperature very close to 100 ^0C, 3-D implementation of such systems, complying with thermal design guidelines, would be quite challenging.

To assess thermal issues in 2-D and 3-D ICs, it is essential to examine both power dissipation and heat removal. Generally, the dominant contribution of total power dissipation comes from dynamic power dissipation which is proportional to $C_L V_{DD}^2 f_c$, where C_L is the load capacitance, V_{DD} is the supply voltage, and f_c is the clock frequency. It is expected that in future technology generations, as much as 60% of load capacitance (C_L) would be due to interconnects [120]. By 3-D integration, it will be feasible to reduce interconnect capacitance significantly which may lead lower power dissipation (for system performance comparable to that of 2-D ICs). Also, the size of the transistors is often determined by the value of capacitive loading they have to drive. If the capacitive loading due to interconnects could be reduced by 3-D integration (due to shorter interconnect length), transistor sizing could optimized to reduce the power dissipation as well. To make accurate predictions of chip temperature in 2-D and 3-D ICs, various components of thermal resistance should be modeled accurately. Due to the increasing levels of power dissipation in future ICs, innovative cooling techniques such as thermal vias, two-phase cooling, etc. may also have to be considered [121].

From thermal design standpoint, heat removal issues in microprocessors are the most challenging compared to those of other applications such as ASIC, DSP, or FPGA. In some of these (ASIC, DSP, FPGA, etc.) applications there is still room for reducing the overall thermal resistance using conventional cooling techniques (such as heat sinks). However, that is not the case with microprocessors. The thermal resistance of heat sinks for microprocessors, in production today, is in the range of $1.0^0C/W - 1.5^0C/W$. Using the value of projected power dissipation in 100 nm scaled technology (as shown in Fig. 6-1), the corresponding junction temperature is 175 ^0C – 250 ^0C. This is much higher than the desirable junction temperature of 100 ^0C which illustrates the importance of thermal issues in high-performance ICs.

In the past, thermal analyses were conducted mainly for packaging approaches to 3-D

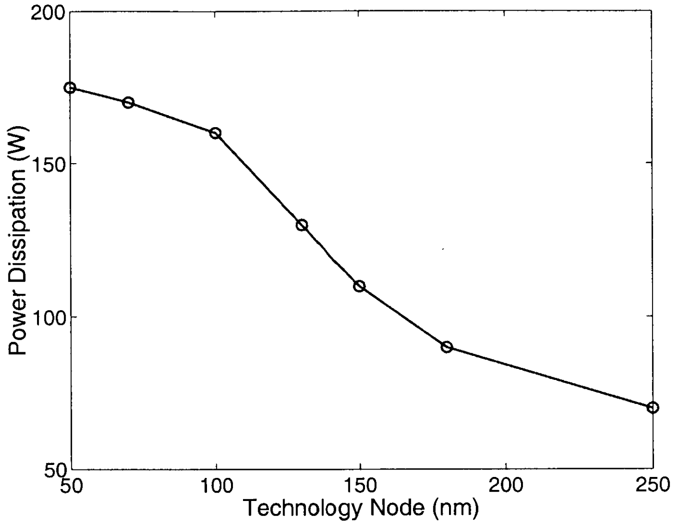


Figure 6-1: Projected total power dissipation in microprocessors [7].

integration [122, 123]. Though in [124], thermal analysis was conducted for monolithic implementation of 3-D ICs, only device-level thermal analysis was conducted by modeling the temperature increase with respect to the bottom of the wafer for uniform and point heat sources. In this chapter, a more detailed thermal analysis will be conducted for 3-D implementation of microprocessors. The methodology presented here can be extended easily to other applications as well. In our study, power dissipation in 2-D and 3-D ICs will be estimated by modeling different components of total power dissipation such as static and dynamic power associated with logic gates, and dynamic power due to driving signal interconnects and clock networks, etc. Detailed studies will be carried out to compare these components of total power dissipation in 2-D and 3-D implementation. Heat removal issues will also be examined by thermal analysis at device- and package-levels. For thermal analysis, analytical and finite element methods (FEM) will be used to study thermal design trade-offs between various 3-D integration schemes.

6.2 Estimation of Power Consumption in Microprocessors

There are many commercially available simulators and analytical modeling frameworks for estimating power consumption in VLSI chips [50, 125, 119]. Though the commercially available simulators are more accurate for predicting power dissipation, they require descriptions of complete net lists. Since we are interested in a priori estimation of power consumption, analytical models are more suitable for our studies. With proper calibration, analytical models can be used to estimate power consumption in ICs within 85% – 90% accuracy [50].

In some recent works, analytical models have been used for estimating power dissipation in microprocessors and VLSI circuits [50, 119]. The total power dissipation was estimated by combining power dissipation associated with logic gates, interconnects, memory transistors, and clock networks. Typically, power dissipation in clock networks in 2-D implementation of microprocessors is a major contributor to total power dissipation [126, 127, 128, 119], and it can be as high as 40% of total power dissipation [126, 127, 128]. Power dissipation in datapaths (execution units, instruction issue units, etc.), which include the power dissipation in logic gates and signal interconnects, is around 30% – 35%. Rest of the contribution to total power dissipation comes from power dissipation in memory, I/O buffers, control logic, etc. In Fig. 6-2, breakdown of total power dissipation is illustrated.

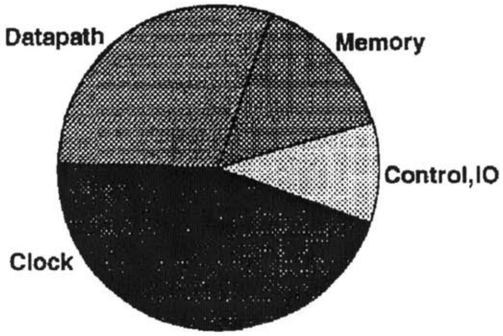


Figure 6-2: Breakdown of power in a high-performance microprocessor [126].

In 3-D implementation of microprocessors, total capacitance associated with both signal and clock wiring can be reduced significantly, resulting in lower total power dissipation. Based on published results and also from our simulations, it is found that power consumption in on-chip memory is generally less than 10%–15% of total power dissipation. For simplicity, only the power dissipation in the logic section will be derived in our analysis, and it will be assumed that power dissipation in on-chip memory is 15% of total power dissipation.

6.2.1 Power Dissipation in Logic Gates

The sources of power dissipation in a logic gate can be summarized by the following equation [129, 130]:

$$\begin{aligned} P_{avg} &= P_{dynamic} + P_{short-circuit} + P_{leakage} \\ &= \frac{1}{2}\alpha C_L V_{DD}^2 f_c + I_{sc} V_{DD} + I_{leakage} V_{DD} \end{aligned} \qquad (6.1)$$

The first term is the switching component of power, where α is the activity factor, C_L is the load capacitance, V_{DD} is the supply voltage, and f_c is the clock frequency. The second term represents the short-circuit power, where I_{sc} is the short-circuit current. The third term corresponds to leakage power, where $I_{leakage}$ is the leakage current.

The total power dissipation in logic gates of a microprocessor is given by

$$P_{logic} = N_g[\gamma_{cp} P_{avg\_cp} + (1 - \gamma_{cp}) P_{avg\_ncp}], \qquad (6.2)$$

where N_g is the total number of logic gates, γ_{cp} is the fraction of logic gates in critical paths, P_{avg_cp} is the average power dissipation per logic gate in critical paths, and P_{avg_ncp} is the average power dissipation per logic gate in non-critical paths. To enhance system performance, W/L ratio of logic gates in critical paths is increased compared to non-critical path logic gates [50, 15]. Based on interconnect delay analysis, we find that increasing the W/L ratio of logic gates in the critical path initially reduces overall delay. However, as the W/L ratio is increased beyond 15-20, there is only a minor reduction in critical path delay and a significant increase in power dissipation. Considering a conservative design guideline, in our analysis $W/L = 30$ has been assumed for logic gates in the critical path. We also assume that roughly 25% logic gates are in the critical path. [2]

The load capacitance and activity factors can vary depending on the logics styles [56, 130]. For simplicity, we assume the microprocessor under consideration is synthesized with CMOS NAND gates, and the average fan-in and fan-out is 3 [50, 15]. We also assume the

[2] This assumption is consistent with a similar assumption used in GENESYS [50] and also leads to comparable total capacitance for logic gates and interconnects in ICs as postulated in [120].

minimum gate width of an nMOS transistor is 1.5F, where F is the minimum feature size, and the gate width of a pMOS transistor is twice the gate width of an nMOS transistor. In our analysis total capacitance of an arbitrary size transistor is represented in units of C_{tr}, where C_{tr} is a minimum size nMOS transistor's gate capacitance; we assume, the contribution of drain diffusion capacitance for a unit size transistor is $.5C_{tr}$ [18]. By examining typical standard cell NAND gate layout, total load capacitance of a minimum size gate is $C_L = 3C_{tr}f.o. + 2.5C_{tr}$, where the first term is due to gate oxide and overlap capacitance and the second term is due to drain diffusion capacitance. The total interconnect capacitance is estimated separately using the wire-length distribution function.

The short circuit current arises when there is a direct current path from V_{DD} to GND (ground) due to finite rise and fall times of the input wavefront. Typically, short circuit power is less than 10% of the total dynamic power [130].

The leakage current arises from the reverse biased drain-substrate region and the subthreshold leakage through the channel of an off transistor. Subthreshold leakage is generally the dominant source of leakage current. The drain current in the subthreshold region is given by

$$I_{DS} = \kappa e^{(V_{GS}-V_t)q/nkT}(1 - e^{V_{DS}q/kT}), \tag{6.3}$$

where κ and n are technology dependent parameters. With smaller threshold voltage, V_t, in scaled technologies and high chip temperature, subthreshold leakage power can be quite significant. Using dual-V_t CMOS technology, dynamic threshold voltage control, etc., subthreshold leakage current can be reduced significantly [131].

6.2.2 Power Dissipation in Clock Network

In today's high-performance systems, power dissipation in clock network is by far is the most dominant contributor to total power dissipation. The clock network is designed in a way that clocking signals can be distributed to various points within the chip with "ideally" zero skew. H-tree, grid, and length-matched serpentine clock networks are some of the commonly used topologies for clock-signal distribution [132]. Depending on the design styles, the total load capacitance of a clock network can vary from several hundred pF to few nF [132].

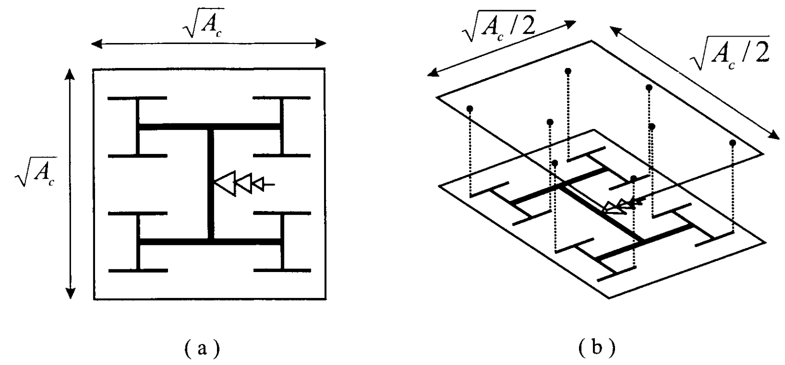


Figure 6-3: H-tree clock network in (a) 2-D and (b) 3-D implementation of microprocessors. The 3-D implementation consists of two strata and the clock network is located on the bottom stratum. Short inter-stratum vias are used to distribute clock signals to the top stratum.

In our analysis for estimating total power dissipation, H-tree clock network has been assumed. H-tree clock network, as shown in Fig. 6-3, consists of several hierarchical levels of progressively smaller H-shaped wiring levels. The terminal nodes of an H-tree level expand into additional H-tree levels until there are enough terminal nodes for all latches or regions of interest. In H-tree network with n_{level}, the number of terminals is $4^{n_{level}}$. If each terminal is connected to a latch, the number of H-tree levels in a clock network is given by

$$n_{level} = log_4 N_{latch}, \tag{6.4}$$

where N_{latch} is the number of latches. We assume in 3-D ICs, the clock network is located on the bottom most stratum and inter-stratum vias, as shown in Fig. 6-3(b), connect the output terminals of the clock network to latches on top stratum. In a flip-chip based packaging technology with heat sinks, the bottom most stratum would have the least thermal resistance from junction to ambient. Placing the clock wiring network and clock drivers on the bottom most stratum would results in a better thermal design compared to distributing clock wiring and clock drivers to all stratum.

The total capacitance of the clock network, C_{clock}, consists of three components:

$$\begin{aligned} C_{clock} &= C_{clkwire} + C_{latch} + C_{driver} \\ &= (1 + k_{driver})(C_{clkwire} + C_{latch}), \end{aligned} \tag{6.5}$$

where $C_{clkwire}$ is the clock wiring capacitance, C_{latch} is the load capacitance at output terminals of clock network due to latches, C_{driver} is clock driver's capacitance, and k_{driver} is clock driver ratio. We assume, clock wires at the lowest level of H-tree network is pitch matched to global wiring pitch. In each hierarchical level above the lowest level, wiring pitch is increase by 2× for impedance matching [49]. The total capacitance of the clock network is given by

$$C_{clkwire} = 1.5\sqrt{A_c}[\sum_{i=1}^{n_{level}} 2^{i-1}c_i], \tag{6.6}$$

where A_c is the chip area per stratum and c_i is wiring capacitance per unit length at ith hierarchical level of H-tree network. To estimate clock driver's capacitance, we assume it consists of a chain of progressively larger inverters, and the final inverter that drives the clock network has gate capacitance comparable to $C_{clkwire}+C_{latch}$. For example, if the final driver's input capacitance is half of $C_{clkwire} + C_{latch}$, and the size of inverters is increased by 2×, k_{driver} in Eq. 6.5 is equal to $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + ... = 1$. The load capacitance due to latches, $C_{latch} = N_{latch}k_{latch}C_{latch}$, where N_{latch} is the total number of latches, k_{latch} is the size factor, and C_{latch} is the latch capacitance. The power dissipation in the clock network can now be written as

$$P_{clock} = f_c C_{clock} V_{dd}^2. \tag{6.7}$$

Using the methodology presented here, estimated values of clock network capacitance in scaled technology nodes are shown in Fig. 6-4 for both 2-D and 3-D implementation. We assume k_{driver} for the clock network is 1.5 and dynamic simple latches [119], with W/L ratio of 10, are used. The values of chip area and wiring pitches are taken from the scaling study presented in Chapter 4. As shown in Fig. 6-4, roughly 20% reduction in clock network capacitance can be achieved by 3-D integration. If total clock network capacitance is dominated by wiring capacitance, the reduction in clock network capacitance by 3-D integration could be even higher.

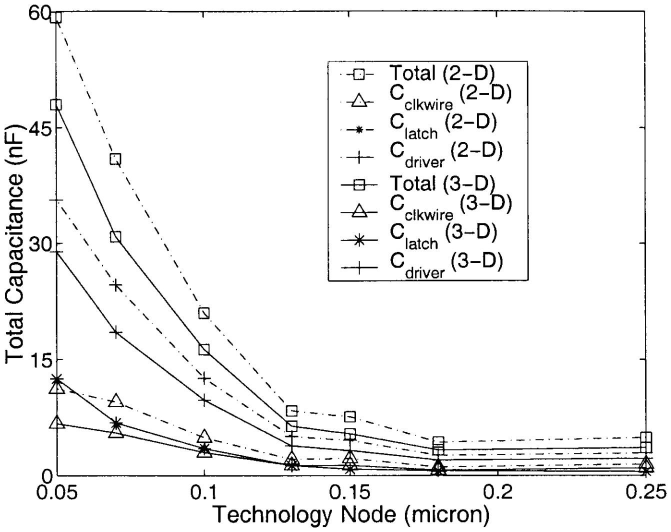


Figure 6-4: The clock network capacitance in 2-D and 3-D implementation (with 2 strata) of microprocessors in scaled technology nodes.

6.2.3 Power Dissipation in Interconnects

The stochastic wire-length distribution becomes quite useful in estimating the total interconnect capacitance. By partitioning the wire-length distribution into local, semi-global, and global regions using interconnect delay criteria, total wire-length of local, semi-global, and global interconnects can be found. The power dissipation due to driving interconnect capacitance is given by

$$P_{int} = \frac{1}{2}\alpha[\chi(L_{tl}c_l + L_{tsg}c_{sg} + L_{tg}c_g)]V_{DD}^2 f_c, \tag{6.8}$$

where, χ is a point-to-point to net-length conversion factor; L_{tl}, L_{tsg}, and L_{tg} are total point-to-point wire-length of local, semi-global, and global interconnects, respectively, and c_l, c_{sg}, and c_g are their respective wiring capacitance per unit length.

6.2.4 Power Dissipation in I/O Pads

The pad capacitance, C_{pad}, is determined from the pad pitch, p_{pad}, using an empirical observation by Bakoglu [49], who reports that typical pad capacitance is around 10 pF for a 1000 μm pad pitch in printed wiring board type package. Assuming the pad edge is half the pad pitch and C_{pad} scales with pad area, an expression for pad capacitance can be derived in terms of pad pitch [50]:

$$C_{pad}\ (pF) = 4 \times 10^{-5} (\frac{p_{pad}}{2})^2, \tag{6.9}$$

where the units of the constant is pF per square micron. The power dissipation in I/O pads is given by

$$P_{pad} = \frac{1}{2}\alpha(1 + k_{pad})C_{pad}V_{DD}^2 f_c, \tag{6.10}$$

where k_{pad} is pad driver ratio which is typically around $\frac{1}{2} - 1$.

6.2.5 Simulation Results: Estimation of Total Power Dissipation

The total power dissipation in 2-D and 3-D implementation of microprocessors can be estimated by combining the power dissipation in logic gates, clock network, signal interconnects, and I/O pads. Though only microprocessors are considered in this study, similar methodology can be applied to other types of ICs. We have validated our methodology by comparing published results of total power dissipation with our estimated values. To find the number wiring levels in H-tree clock network, n_{level}, we assume number of latches, N_{latch}, connected to the output terminals of a clock network, is 2% of the total number of logic gates [50]. An activity factor $\alpha = 15\%$ has also been assumed.

To validate our methodology, we have compared our simulation results of total power dissipation with the published results for Alpha 21164 and 21264 processors [93]. The number of transistors in Alpha 21164 microprocessor is 9.3 million, and the clock frequency is 300 MHz. It is fabricated in .5 μm process technology and uses 3.3 V supply voltage, and the total power dissipation 50 W. Using system-level modeling and analysis, the estimated clock frequency is 345 MHz and the total power dissipation is 47 W; the power dissipation in logic gates, clock network, signal interconnects, I/O pads, and on chip memory are 16 W, 12 W, 10 W, 2 W, and 7 W, respectively. Similar analysis was also conducted for Alpha 21264 microprocessor. The Alpha 21264 processor operates at 600 MHz with 2−2.5 V supply voltage, and the power dissipation is 72 W. Using system-level analysis, the estimated clock frequency is 515 MHz and the total power dissipation is 82 W. The power dissipation in logic gates, clock network, signal interconnects, I/O pads, and on chip memory are 34 W, 22 W, 12 W, 2 W, and 12 W, respectively. In general, our estimated values of total power dissipation are comparable to published results and reasonably accurate for making first order calculations. Some of the factors that may lead to discrepancies between published and estimated values of total power dissipation are: **i**) generalization of logic gates in microprocessors by a collection of static CMOS gates rather than a mixture of static and dynamic logic gates, **ii**) assumption of the same activity factor in all logic gates and interconnects, and **iii**) inaccuracy in projecting the number of transistors in the critical and non-critical paths.

After the rough validation, total power dissipation in 2-D and 3-D implementation of microprocessors, in scaled technology nodes, is estimated. Simulation results of total power dissipation are shown in Fig. 6-5. The values of chip area, interconnect capacitance, total wire-length, etc. are taken from the scaling study presented in Chapter 4. Based on our scaling study, chip area does not increase significantly between .25 μm to .15 μm technology nodes, and the increase in wiring capacitance and clock frequency is offset by the reduction in power supply voltage. As a result, there is a small drop in total power dissipation around .15 μm technology node. However, beyond that point, there is a significant rise in total power dissipation for both 2-D and 3-D implementation. Based on the scaling study presented in Chapter 4, in 3-D implementation (with two strata) the increase in clock frequency is roughly 10% – 20%. However, there is also a significant reduction in clock network and signal wiring capacitance in 3-D implementation. As a result, we find that total power dissipation in 3-D ICs, for comparable system performance, is smaller than that of 2-D ICs (see Fig. 6-5).

We have also examined the impact of SOI technology on total power dissipation, by assuming there is negligible drain junction capacitance in logic gates. The simulation results of total power dissipation in 2-D bulk technology and 3-D technology implemented with SOI wafers are compared in Fig. 6-6. Using SOI technology, reduction in both gate and interconnect capacitance are feasible. As a result, even with 10% – 20% improvement in clock frequency, $\sim$ 30% reduction in total power dissipation can be achieved by 3-D

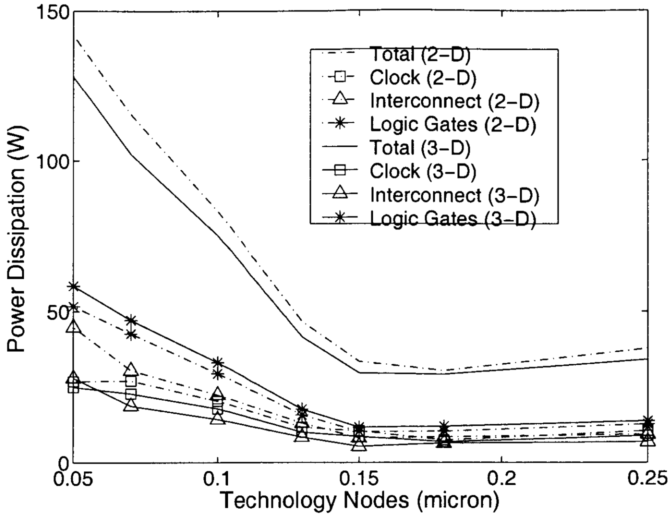


Figure 6-5: The total power dissipation and various components of total power dissipation in 2-D and 3-D implementation of microprocessors in scaled technology nodes.

integration (with SOI technology). For simplicity and illustration purposes, the impact of smaller parasitic capacitance in SOI technology has not been included in estimating clock frequency.

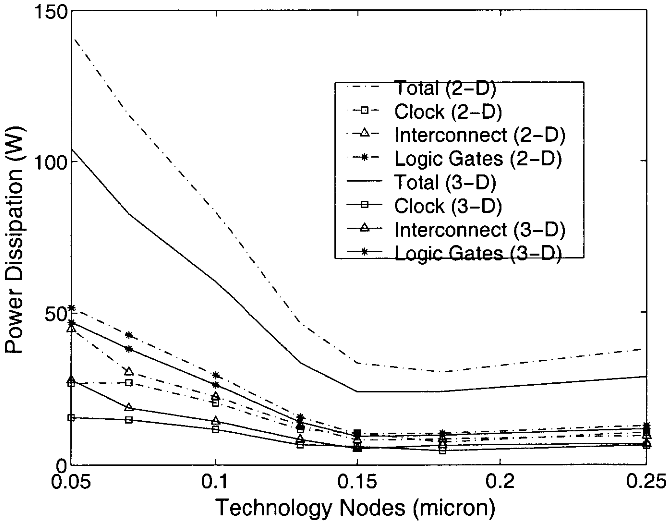


Figure 6-6: The total power dissipation and various components of total power dissipation in 2-D and 3-D implementation of microprocessors in scaled technology nodes. The 3-D integrated circuit is implemented using SOI technology.

6.3 Device-Level Thermal Analysis

In the previous section, a system-level approach has been applied to estimate total power dissipation in ICs which is essential to project the average chip temperature. However, due to non-uniform heat generation across the chip, there could be localized hot spots that would have much higher junction temperature than the average chip temperature. In this section, thermal analysis is conducted to examine the impact technology dependent parameters on localized heat dissipation. Localized heat generation is modeled by a point heat source and the dependencies of junction temperature on bonding technology, Si layer's thickness, etc. are examined using a Finite Element Method (FEM) based thermal simulator, ANSYS [133]. The junction temperature in ANSYS is calculated by solving the heat equation

$$\frac{d^2T}{dx^2} + \frac{d^2T}{dy^2} + \frac{d^2T}{dz^2} + \frac{\dot{q}}{k} = 0 \tag{6.11}$$

in a layered medium, where T is the temperature, $\dot{q}$ is the heat generation per unit volume, and k is the thermal conductivity of the material [134].

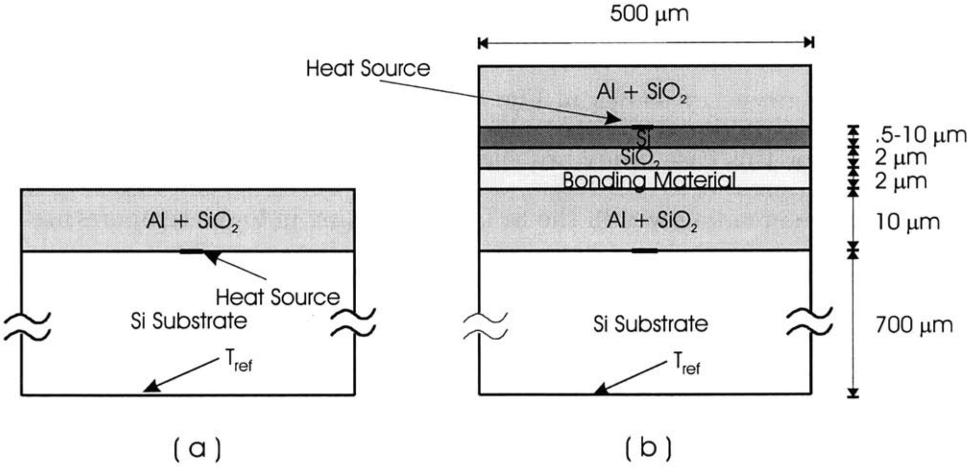


Figure 6-7: Simplified cross section of (a) 2-D and (b) 3-D ICs with two strata for FEM based thermal analysis.

We use the geometries shown in Fig 6-7, where point heat sources are placed at the $\mathrm{Si} - \mathrm{SiO}_2$ interface. Assuming heat removal is through the back side of the wafer, adiabatic boundary conditions are applied on the side walls and the top surface, and a constant reference temperature, T_{ref}, is applied on the bottom surface. Then the temperature at the

Materials	Thermal Conductivity (W/cmK)
Aluminum	2.37
Copper	4.01
SiO_2	0.014
Polyemide	0.004
HSQ (low-k dielectric)	0.006
Silicon	1.41

Table 6.1: The thermal conductivity of commonly used materials in ICs.

$\mathrm{Si} - \mathrm{SiO_2}$ interface is calculated using ANSYS. In our analysis, interconnect metal and ILD layers are represented by a 10 μm thick $\mathrm{SiO_2}$ layer. The thermal conductivities of commonly used materials in front-end and back-end of the line processes are shown in Table 6.1.

We begin by conducting thermal analysis for a point heat source in 2-D ICs as shown in Fig. 6-7(a). Then point heat sources with the same heat flux are placed on both strata in 3-D ICs, as shown in Fig. 6-7(b). By examining the isothermal contours and maximum temperature as a function of various parameters, we find that heat spreading on top stratum depends strongly on top stratum's Si thickness and the thermal conductivity of the underlying of bonding material. To examine this dependency, temperature rise on top stratum is estimated for the geometry shown in Fig. 6-7(b) by varying top stratum's Si thickness. Simulation results for this case study are shown in Fig. 6-8. Thicker Si layer allows more efficient lateral heat spreading within the Si layer, resulting in lower temperature rise compared to thinner Si layer. Also, the Cu bonding layer assists in heat spreading. As a result, the temperature rise in Cu-based bonding is smaller compared to that of polyemide-based bonding. The rate of temperature increase as a function of Si layer's thickness is also smaller for Cu-Cu bonding technology. This simple case study shows that Cu bonding layer can be useful as heat spreader/thermal conduit in 3-D ICs for localized heating. Based on this study, we also find that 1-D analytical models don't capture the heat spreading effect as a function Si layer's thickness, and one must use caution in using 1-D models.

Simulations are also conducted to asses the role of interconnects, vias, etc. on heat removal. Integration of thermal vias in addition to electrical vias can reduce the effective thermal conductivity of the ILD layer. To examine the role of vias on heat removal, we

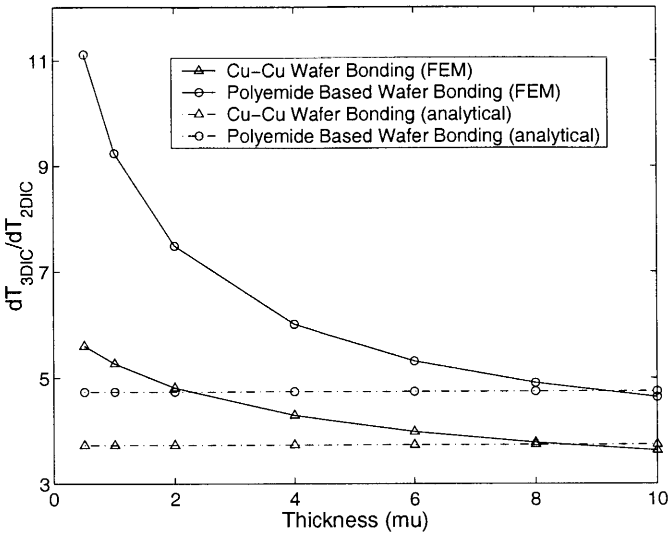


Figure 6-8: Normalized temperature rise in top stratum for the geometry show in Fig. 6-7 as a function of top stratum's Si thickness. dT_{3DIC} and dT_{2DIC} are the maximum temperature rise in 2-D and 3-D ICs with respect to the reference temperature on the bottom most surface, T_{ref}. The temperature rise based on analytical models is calculated by solving 1-D heat equation.

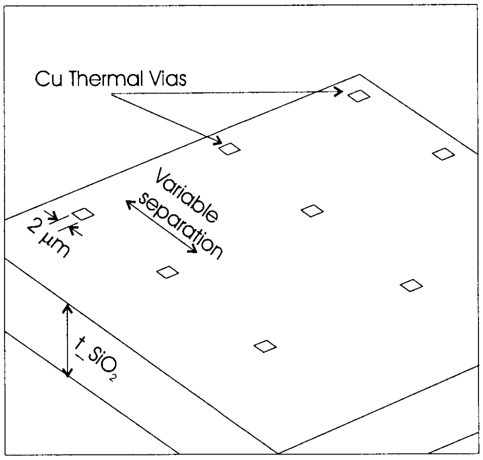


Figure 6-9: A simplified geometry to assess the role of vias on heat removal.

perform thermal analysis using a simplified structure as shown in Fig. 6-9. A uniform heat source is applied to the top surface, and the bottom surface is kept at a reference temperature. In this analysis, multi-level interconnects and ILD layers are represented by a 10 μm thick SiO_2 layer. Then the temperature rise on top surface with respect to the bottom surface is estimated as a function of via density, assuming uniform via distribution, and the reduction in effective thermal resistance of the SiO_2 layer is calculated. Simulation

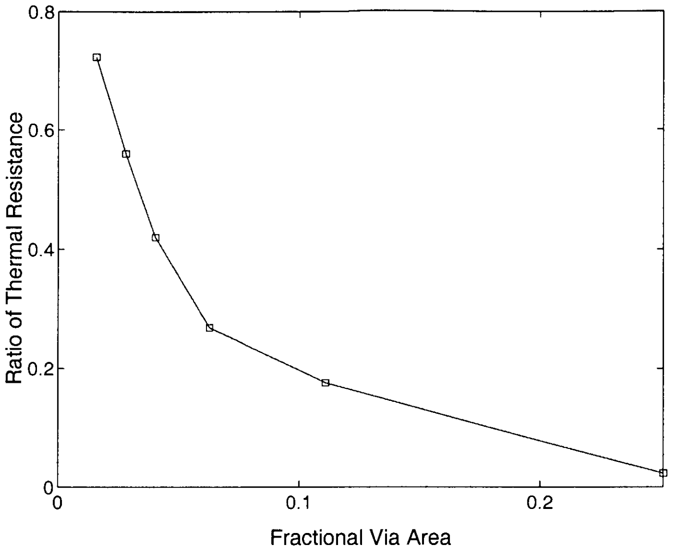


Figure 6-10: The ratio of thermal resistance of a 10 μm thick SiO_2 layer with and without thermal vias as a function of fractional area occupied by thermal vias.

results for this case study is shown in Fig. 6-10. From thermal design perspective a very high via density is desirable, but high via density also introduces routing congestion and reduces the wiring efficiency in interconnect levels. In a multi-level interconnect structure if the via blockage is 15%, considering design rules, the fractional area occupied by vias is $\sim$ 4%. If these electrical vias also serve as thermal vias, based on Fig. 6-10, the corresponding reduction in thermal resistance is $\sim$ 2.5×. One could also consider introducing thermal vias through the bottom most bulk Si layer. Unless the thermal conductivity of the via material is much higher than that of Si, this approach may not be very useful because Si itself is a very good thermal conductor. The planar metal interconnects can be useful for heat removal from localized areas of heat generation because heat is removed both laterally and vertically. However, for uniform heat generation, the primary heat removal path is along the vertical direction through the backside of the bottom most Si wafer, and the planar metal interconnects have negligible thermal resistance (or thermally short) along the vertical direction.

6.4 Package-Level Thermal Issues in 3-D IC

In the previous sections, power dissipation and device-level thermal issues have been examined. To have a complete understanding of heat removal capabilities in 2-D and 3-D ICs, it is necessary to examine package-level thermal issues because the heat removal capability in ICs is primarily determined by the packaging technology and the cooling methods associated with it. From device-level thermal analysis, we find that the temperature difference from the back of the wafer to junction in 3-D ICs is $(4\times$ to $5\times)$ or higher than that of 2-D ICs. However, in 2-D ICs this temperature difference corresponds to only $5\% - 10\%$ of the temperature difference from junction to ambient. In other words, $90\%-95\%$ of temperature difference from junction to ambient is due to the package itself.

The dominant heat removal mechanism from package to ambient is by convection. Depending on the packaging technology, heat transfer due to heat conduction from chip to next level of interconnect substrate may also be substantial compared to the convective heat transfer. Other factors that can also effect the heat removal capability include: the amount of power dissipation in neighboring chips, lateral and vertical spacing between boards that are used by other packages, conductivity of the printed wiring boards (PWB), etc. Neglecting the secondary heat removal mechanisms, a simplified heat transfer model of a packaged chip is illustrated in Fig. 6-11, where T_j, T_c, and T_a are junction, chip, and ambient temperature, respectively. Using the definitions in Fig. 6-11, thermal resistance from junction to ambient, R_{ja} $(^0C/W)$, is given by

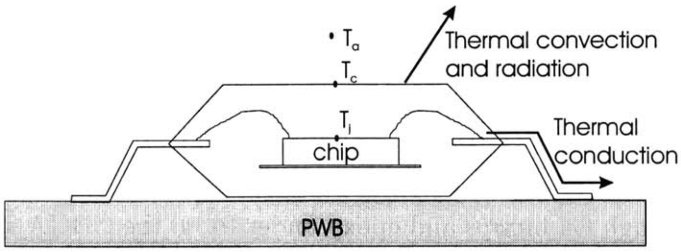


Figure 6-11: A simplified heat transfer model of a packaged chip.

$$R_{ja} = (T_j - T_a)/P, \tag{6.12}$$

where P is the average power dissipation and $T_j > T_a$. R_{ja} can be written as $R_{ja} = R_{jc}+R_{ca}$, where R_{jc} is the thermal resistance from junction to case and R_{ca} is the thermal resistance from case to ambient. As discussed in earlier sections, heat removal issues in microprocessors is more challenging compared to that of DSP chips, ASICs, etc. Besides, there is a significant room for reducing the packaging thermal resistance in these (DSP, ASICs, etc.) applications. So, in the following sections package-level thermal issues will be examined for 2-D and 3-D implementation of microprocessors.

6.4.1 C4 Packaging Technology

In future high-performance ICs such as microprocessors, to meet the I/O count requirements, area-based I/O configuration will be necessary [7]. The controlled-collapse-chip-connection (C4), also known as flip-chip packaging technology, offers significant reduction in signal delay and high I/O count. The C4-based ceramic ball grid array (CBGA) package is ideal for meeting future pin count requirements in high-performance circuits such as microprocessors. The cross section of a typical C4-CBGA package along with a heat sink attachment and the equivalent thermal model of the package are shown in Fig. 6-12. In a C4 package, the flipped die is connected to a ball grid array (BGA) substrate by C4 bumps, and the BGA substrate is connected to the next interconnect level, printed wiring board (PWB), by solder balls. The back side of the chip is attached to an aluminum lid by thermal paste. Depending on the thermal design requirements, heat sinks can be attached to increase heat removal capabilities. As discussed in [135, 136], the primary heat transfer path from the chip to ambient is through the back side of the wafer. Heat is then convectively removed from the heat sink by natural or forced convection. A secondary heat transfer path exists in parallel through C4 bumps and epoxy under-fill to the CBGA substrate. From the CBGA substrate, heat conduction to the next-level substrate, PWB, is via pins or solder balls; then the primary mode of heat transfer is by convection or radiation. In thermal analysis of C4 packages, the error in neglecting heat transfer through the frontside of the die (C4 bumps, CBGA substrate, etc.) is quite negligible [135, 136]. At the back side of the chip, when heat sinks and fans are used, both coefficient of heat convection and the effective area (total area of heat sink fins) through which convection occurs are much higher than those at the front side of the chip. As a result, consideration of only the heat transfer path through the back side of is sufficient for thermal analysis.

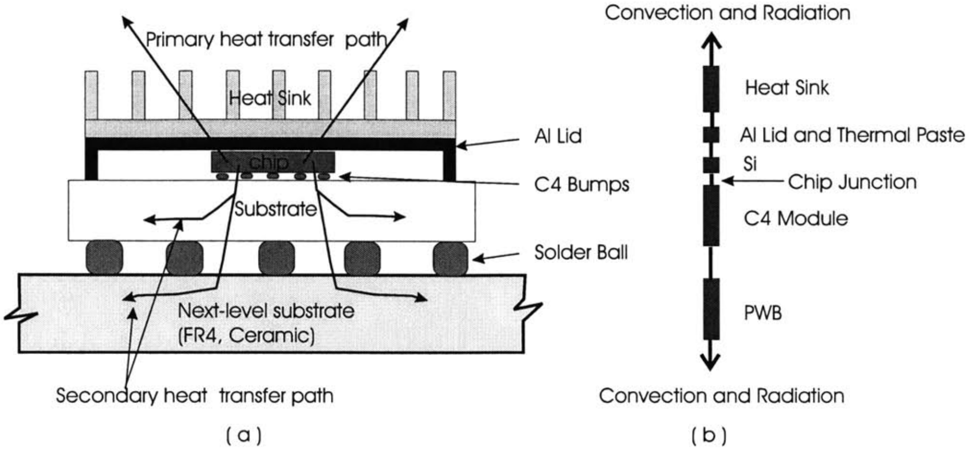


Figure 6-12: (a) The cross section of a C4 package. (b) The thermal model of a C4 package.

If only the primary heat transfer path through the back side of the die is considered, R_{ja} for C4 package can be written as $R_{ja} = R_{jc} + R_{ca}$, where R_{jc} is the thermal resistance from junction to Al lid and R_{ca} is the thermal resistance from Al lid to ambient. R_{jc} consists of the thermal resistance of Si and ILD layers and thermal paste. R_{ca} consists of the thermal resistance of the heat sink. Thermal resistance of the heat sink depends on heat sink's design parameters (i.e. number of fins, their area and separation, etc.) and the air velocity within the heat sink [137]. Commercially available heat sinks for Pentium or Celeron processors have thermal resistance in the range of $.75^0C/W$ to $1.5^0C/W$ for air velocity of 2 m/s. Heat sinks with thermal resistance as low as $.45^0C/W$ are also available.

6.4.2 Estimation of Packaging Thermal Resistance

The thermal conductivity of some of the commonly used materials in microelectronic packaging is shown in Table 6.2. For first-order analysis, thermal resistance from junction to Al lid, R_{jc}, can be estimated using 1-D analytical models[3]. We use the geometry shown in Fig. 6-13 for estimating R_{jc}. In Figure 6-14, R_{jc} is plotted as a function of chip size

[3] Using 1-D models, thermal resistance of a geometry with thickness t, cross sectional area A_c, and thermal conductivity k is given by $t/(kA_c)$. The direction of heat flow is normal to the cross sectional area.

Materials	Thermal Conductivity (W/cmK)
Aluminum	2.37
Copper	4.01
Solder Ball	0.35
Alumina Ceramic	.18 to .2
AlN Ceramic	.7 to 1.2
Glass Ceramic	0.05
EPX Underfill	0.185e-2
FR-4, Printed Wiring Board	0.2e-2

Table 6.2: Thermal conductivity of commonly used materials in micro electronic packaging.

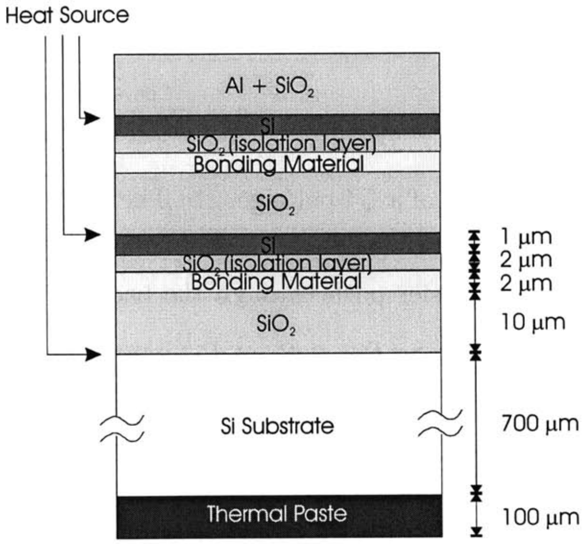


Figure 6-13: Cross section of the geometry used for estimating R_{jc} (not drawn to scale).

and number of strata. It is assumed that total power dissipation and total chip area (i.e. the area per stratum × number of strata) is constant in 2-D and 3-D implementation, and in 3-D IC the total power is equally distributed in all strata. We also assume a thermal conductivity of .038 W/cmK for the thermal paste, and Cu-based bonding technology is used. One of the major advantages of 3-D integration is that by integrating multiple strata, footprint (area/stratum) size can be reduced. However, it leads to higher values of R_{jc} as shown in Fig. 6-14. In particular, thermal resistance of the thermal paste increases quite significantly as multiple strata are integrated, and it can even be the limiting factor in R_{jc}.

The thermal resistance from Al lid to ambient, R_{ca}, can be estimated knowing the ther-

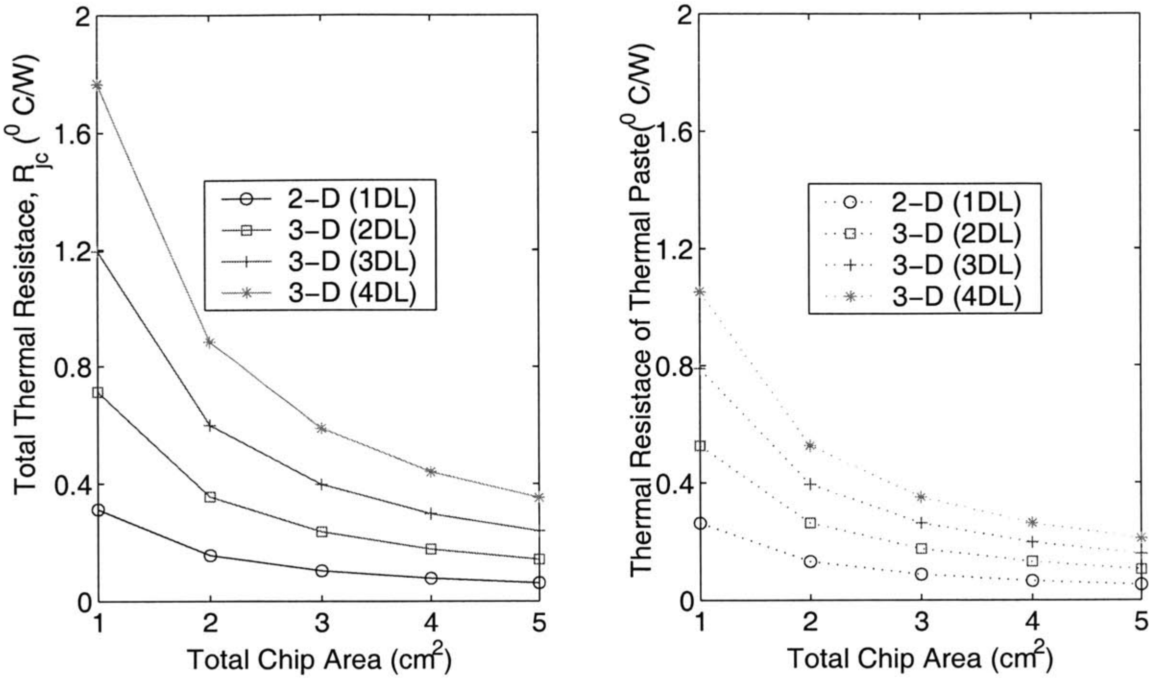


Figure 6-14: Thermal resistance from junction to Al lid, R_{jc}, and thermal resistance of the thermal paste as a function of total chip area and number of strata (device layers). DL stand for device layers.

mal resistance of the heat sink, $R_{heat\ sink}$. It ($R_{heat\ sink}$) depends on the design parameters such as number of fins, their separation and area, air flow rate inside the fins, etc. For a set of design parameters and heat flow rate, thermal resistance of the heat sink can be found from manufacturer's data sheet. A simplified representation and equivalent thermal model for estimating R_{ca} are shown in Fig. 6-15. There are two components of R_{ca}.

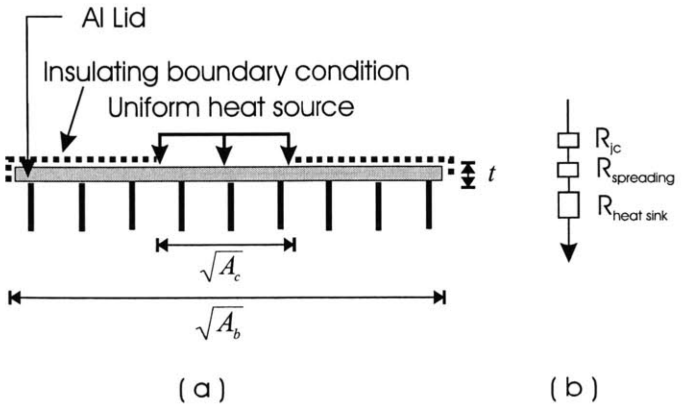


Figure 6-15: (a) A simplified representation of Al lid and heat sink and (b) their thermal models for estimating R_{ca}.

One of them, $R_{heat\ sink}$, is heat sink's intrinsic thermal resistance. The other component, $R_{spreading}$, arises due to spreading resistance. Whenever heat flows through one or more solids involving a change in cross-sectional area, there is an additional component of thermal resistance associated with it. It is commonly known as spreading or constriction resistance. The change in cross-sectional area from chip (of area A_c) to Al lid (of area A_b) gives rise to spreading resistance. Consideration of spreading resistance in 3-D ICs is very crucial. If the footprint size is reduced by 3-D integration and heat sink's size remains unchanged, spreading resistance will increase, resulting in higher packaging resistance for 3-D ICs.

Accurate models for spreading resistance can be found by solving differential equations governing heat flow [138, 139]. For the simplified geometry shown in Fig. 6-15, $R_{spreading}$ is given by [138, 139]

$$R_{spreading} \simeq \frac{\psi}{\sqrt{\pi} k_{Al} a}, \tag{6.13}$$

where ψ is a dimensionless parameter and k_{Al} is the thermal conductivity of the Al lid. Average and maximum values of ψ and rest of the variables are given by following equations:

$$\begin{aligned} \psi_{average} &= \frac{\epsilon\tau}{\sqrt{\pi}} + \frac{1}{2}(1-\epsilon)^{\frac{3}{2}}\phi_e \\ \psi_{maximum} &= \frac{\epsilon\tau}{\sqrt{\pi}} + \frac{1}{\sqrt{\pi}}(1-\epsilon)\phi_e \\ \phi_e &= \frac{tanh(\lambda_c\tau) + \lambda_c/B_i}{1 + \frac{\lambda_c}{B_i} tanh(\lambda_c\tau)} \\ \lambda_c &= \pi + \frac{1}{\sqrt{\pi}\epsilon} \\ \epsilon &= \frac{a}{b} \\ \tau &= \frac{t}{b} \\ B_i &= \frac{1}{\pi k b R_{heatsink}} \\ a = \sqrt{\frac{A_c}{\pi}} \quad &, \quad b = \sqrt{\frac{A_b}{\pi}} \end{aligned} \tag{6.14}$$

The value of spreading resistance depends strongly on $\sqrt{\frac{A_c}{A_b}}$, and the thickness of the Al lid, t. In Fig. 6-16, $R_{spreading}$ is plotted as a function of $\sqrt{\frac{A_c}{A_b}}$ and the thickness of the Al lid, t. As $\sqrt{\frac{A_c}{A_b}}$ approaches 1, $R_{spreading}$ approaches to 0. Also, for thinner Al lid there is less heat spreading along the lateral direction of the lid which results in higher spreading resistance. As discussed earlier, one of the advantages of 3-D integration is the reduction in footprint size. However, it leads to smaller values of $\sqrt{\frac{A_c}{A_b}}$ and higher spreading resistance. If the intrinsic thermal resistance of a heat sink is very small ($< .5\ ^0C/W$), the spreading resistance could even limit the thermal resistance of a heat sink in 3-D ICs. Placement of a heat spread with high thermal conductivity, such as diamond substrate, in between the bulk Si layer and the heat sink may be helpful for reducing the spreading resistance.

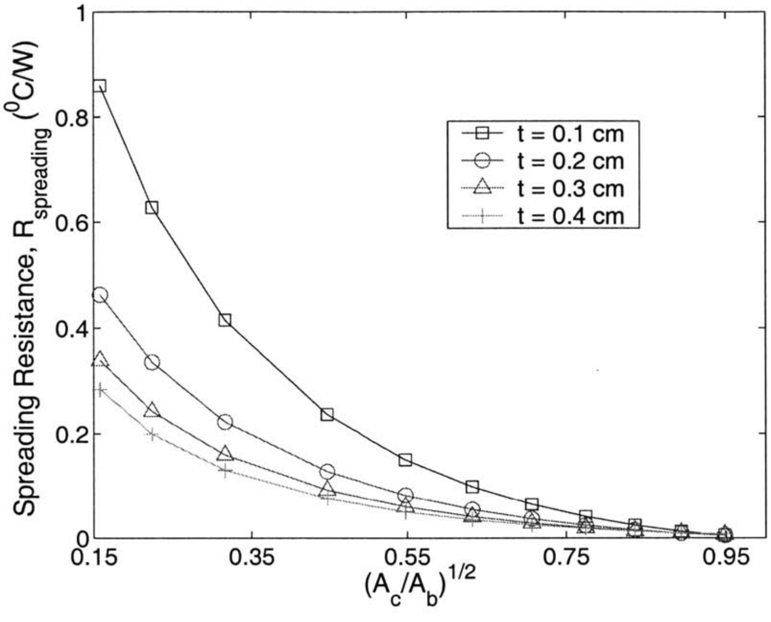


Figure 6-16: Spreading resistance, $R_{spreading}$, of a typical heat sink for high-performance systems such as microprocessors as a function of $\sqrt{\frac{A_c}{A_b}}$ and the thickness of the Al lid. The base area of the heat sink $A_b = 5 \times 5\ cm^2$, and $R_{heat\ sink} = 1\ ^0C/W$.

6.5 Estimation of Die Temperature

In this section, using the methodology presented in earlier, the chip temperature of 2-D and 3-D ICs will be estimated. Though our stochastic system-level modeling and analysis suggest that only 10% – 20% improvement in clock frequency can be achieved by 3-D integration with two strata, we believe in a real application, overall system performance could be much higher. By taking advantage of the third dimension, logic gates in the critical path can be rearranged in multiple strata to minimize the capacitive loading due to interconnects. As system-level modeling captures an average scenario, the best possible arrangement of logic gates in multiple strata cannot be modeled accurately by stochastic modeling and analysis. In this Section, to estimate the chip temperature and power dissipation, it is assumed that in a real application the improvement in clock frequency by 3-D integration can be much higher than the prediction based on system-level modeling. So the clock frequency is varied to reflect possible design enhancements by 3-D integration.

6.5.1 Assumptions and Methodology

We use the projected values of total power dissipation, presented in the SIA Roadmap [7], as our starting point for estimating the die temperature. Based on the observation from published results and our own system-level simulation, we make the following assumptions:

- Total power dissipation is roughly proportional to clock frequency.

- The power dissipation in on-chip memory is 10% of total power dissipation. The rest of total power dissipation is equally associated with logic gates, signal interconnects, and clock networks.

- If the total chip area is conserved, signal interconnect capacitance can be reduced by 30%, 45%, and 52% by 3-D integration with 2, 3, and 4 strata, respectively. This is attributed to shorter wire-length in 3-D integration.

- In clock network, total capacitance due to clock wiring and latches are comparable. Similarly, by 3-D integration with 2, 3, and 4 strata, the reduction in clock network capacitance is roughly 15%, 19%, and 25%, respectively.

- Power dissipation in all strata due to logic gates and interconnects is comparable. However, the bottom most stratum also includes additional component of power dissipation due to the clock network.

6.5.2 Simulation Results

We use a geometry similar to the one shown in Fig. 6-13 to estimate R_{jc}. In scaled technology nodes, the total thickness of ILD layers (without the metal interconnects) can be found knowing the number of interconnect levels, wiring pitch, and aspect ratio. For uniform heat generation, heat flow is along the vertical direction and the metal interconnects have negligible thermal resistance. In Table 6.3, input parameters for estimating average chip temperature are presented. We also assume the base area of the heat sink is twice the chip area of 2-D ICs. The thicknesses of SiO_2 isolation layer separating adjacent strata, bulk Si wafer, and thermal paste are 2 μm, 700 μm, and 100 μm, respectively. The chip temperature is estimated for 2-D and 3-D implementation of ICs by keeping the total chip area, $A_c N_z$, constant. Simulation results of temperature drop, $T_j - T_a$, within the ILD layers and thermal paste, and the package are shown in Fig. 6-17. The total temperature drop can be found by adding up these components. In 3-D ICs, the temperature rise on the topmost stratum is considered.

Technology Node (nm)	250	180	150	130	100	70	50
Power Dissipation (2-D IC)	70	80	90	110	160	170	175
ILD Thermal Conductivity (W/cmK)	.014	.012	.0054	.0054	.0019	.0012	.0007
Total Thickness of ILD layers (μm)	10	9.3	11	8	4.7	4.2	3
Total Chip Area, $N_z A_c$ (cm^2)	3	3.4	3.85	4.3	5.2	6.2	7.5

Table 6.3: Input parameters for estimating chip temperature [7, 140].

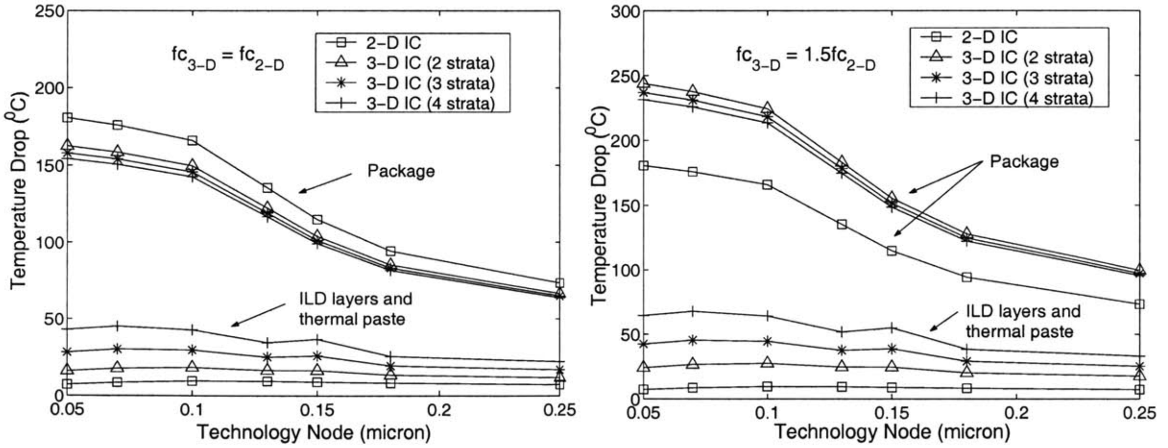


Figure 6-17: The temperature drop with the ILD layers and thermal paste and the package in 2-D and 3-D implementation of ICs. fc_{2-D} and fc_{3-D} are the clock frequency of 2-D and 3-D ICs. It has been assumed that the thermal resistance of the heat sink is 1 $^0C/W$, and Cu-based bonding technology has been used.

By examining Fig. 6-17, we find that for comparable system performance, ($fc_{3-D} = fc_{2D}$), temperature drop within the package is much higher (4×) than that of ILD layers and thermal paste. When the number of strata is increased, total power dissipation can be reduced due to lower power dissipation associated with interconnects and clock network. By examining the set of curves for 2-D and 3-D ICs, we find that the impact on temperature drop within the the package due to the increase in spreading resistance is offset by the reduction in total power dissipation. As a result, when $(fc_{3-D} = fc_{2D})$, the temperature drop within the package is reduced when multiple strata are integrated. However, the impact on temperature drop within the ILD and bonding layers and thermal paste due to the reduction in power dissipation is offset by the increase in thermal resistance of ILD layers and thermal paste. As a result, the temperature drop within the ILD layers and thermal paste deteriorates when multiple strata are integrated. For higher system performance in 3-D ICs, $(fc_{3-D} = 1.5fc_{2D})$, temperature drop within the ILD layers, thermal paste, heat sink in 3-D ICs increases quite significantly.

Though the suggested chip temperature for reliable operation of devices and interconnects in the SIA and ITRS roadmaps is around 100^0C [2, 7], we find if the packaging solutions available today are used (for example, heat sinks with thermal resistance 1 $^0C/W$), chip temperature will reach an unacceptable level in future technology generations. However, with thermal resistance of heat sinks in the range of $.3 - .5$ $^0C/W$, reasonable chip temperature can be obtained. Also, if the thermal resistance of heat sinks is reduced to $.3 - .5$ $^0C/W$, the total temperature drop in the ILD layers and thermal paste may become comparable to the temperature drop in the package in 3-D ICs, and integration of thermal vias and the use of higher thermal conductivity ILDs could be beneficial to reduce the chip temperature.

6.6 Summary

In this section, simulation results of power dissipation and thermal analysis have been presented for 2-D and 3-D implementation of microprocessors. We find that for similar system performance, the amount of power dissipation in 3-D ICs is comparable to that of 2-D ICs. It may even be feasible to reduce total power dissipation in 3-D ICs due to the reduction in signal interconnect and clock network's capacitance. However, if 3-D ICs operate at higher clock frequency compared to that of 2-D ICs, power dissipation could reach an unacceptable level. For reliable operations of devices and interconnects at a reasonable chip temperature, low-power circuit techniques will be integral to system design. At the same time, heat sinks with better cooling technology such as two-phase cooling may be needed [121].

Chapter 7

Conclusion

7.1 Summary

In this thesis work, based on system-level modeling and analysis, opportunities for 3-D implementation of ICs have been presented. A stochastic wire-length distribution for 3-D ICs has been derived which shows that significant reduction in wire-length and wiring-limited chip area can be achieved by 3-D integration. Using the stochastic wire-length distribution model, interconnect delay criteria, and simple models representing system architecture, key performance metrics and technology requirements for 2-D and 3-D ICs have been estimated. We have found that the wiring-limited chip area in ASIC-based designs, represented by random logic networks, and FPGAs can be reduced by $30\% - 60\%$ by 3-D integration with 2-4 strata. Similar improvements in clock frequency are also feasible. We have also found that for 3-D implementation of microprocessors in scaled technology generations, $15\% - 45\%$ improvement in throughput, $2.5\times - 5\times$ reduction in long wire delay, and 30% reduction in total chip area are feasible by 3-D integration with two strata. We speculate that due to the complexity for integrating multiple strata, via blockage, heat removal issues, etc. it may not be suitable to integrate more than 3-4 strata for monolithic integration.

We have also conducted thermal analysis by modeling power dissipation and device- and package-level heat removal issues. We find that for comparable system performance, power dissipation in 3-D ICs can be reduced due to lower signal interconnect and clock network's capacitance. However, for higher system performance in 3-D ICs, power dissipation could reach an unacceptable level. Similar trends are also observed for 2-D ICs in scaled technology nodes. To reduce the chip temperature to an acceptable level, it will be necessary to reduce the power dissipation by integrating innovative low-power circuit and system design techniques. Considering the projections for total power dissipation in high performance systems [7], it will be also necessary to reduce the packaging thermal resistance by $2\times - 4\times$ beyond 100 nm technology node compared to the thermal resistance of packages available today.

Based on examining various opportunities and challenges for 3-D integration, the following conclusions can be made about 3-D integrated circuits:

- In general, any IC where the chip area and system performance are interconnect-limited could benefit from 3-D integration. It appears that the potential for improving system performance by 3-D integration is higher in systems that require less custom designs such as ASICs and FPGAs.

- Considering realistic values of inter-stratum via density in 3-D integration and the reduction in wire-length, there seems to an optimum partitioning of a random logic network in 3-D architecture. It is likely that logic networks have to be partitioned at logic block levels, where a logic block consists of several logic gates, and distributed in multiple strata to reduce the number of inter-stratum interconnects and their via density.

- The success of 3-D technology depends not only on the performance gain but also on the opportunities for innovative applications by 3-D integration. One such area that deserves special attention is 3-D systems-on-a-chip. Using 3-D technology based on wafer bonding, functional blocks fabricated with various technologies such as bulk CMOS, BiCMOS, MEMS, optical, etc. can be combined to form compact 3-D systems-on-a-chip for sensing or communication applications. There is simply on other alternative to fabricate such SOC monolithically using 2-D CMOS-compatible technologies.

- Though there are trade-offs in bonding wafers or known good dies for fabricating 3-D ICs, die-to-die bonding does not seem to be a realistic choice for large volume manufacturing. To improve overall yield in 3-D ICs, uniformity of Cu-Cu bonding across the wafer would be crucial.

- The challenges involved in heat removal in 3-D ICs will not be very different from that of 2-D ICs. Since the heat removal capability in 2-D and 3-D ICs is generally limited by the packaging technology, thermal resistance of the package has to be reduced significantly using innovative cooling techniques for reliable operation of devices and interconnects.

7.2 Future Work

The topic "3-D ICs" is so broad that one can easily feel directionless along the line of research work. In order to maximize the payoff from future research work, research efforts have to be focused on topics that are very fundamental to the development of 3-D ICs.

In the technology development side, future direction of research is probably very well defined. Once the demonstration of working 3-D devices or logic gates is complete, research efforts have to be focused on yield mapping and yield enhancement of the bonding interface. Other topics of interest are: characterization of inter-stratum interconnects, experimental investigation of thermal issues in 3-D ICs, heat removal techniques using thermal vias, etc.

In the modeling side, it would be beneficial to associate future work in system architecture, tool development, or design methodology with specific applications in mind. Though it is very difficult to speculate what would be the "killer" application in 3-D integration, the trade-offs between 2-D and 3-D implementation of various applications can be compared to select a few applications. It is expected that future SOC would contain both digital and analog functionalities. For mixed signal designs, digital and analog circuits can be fabricated on separate wafers, without having to make sacrifices in individual process steps, and vertically stacked to form 3-D mixed-signal ICs. The Cu bonding layer can be used as a ground plane to provide signal isolation between digital and analog components, and the substrate-coupled noise between them, as found in conventional 2-D implementation, can be eliminated. In order to explore opportunities for 3-D mixed-signal chip, design methodologies have to be developed for partitioning and placement of circuits in multiple strata. The signal integrity issues should also be examined based on numerical or analytical modeling and experimentally validated.

Besides SOC, 3-D ASIC- and FPGA-based applications also have tremendous prospects for improving system performance and gate density. Partitioning and placement methodologies for SOC-based applications can be extended easily to ASIC- and FPGA-based designs. Another application that may benefit from 3-D integration is monolithic CMOS imagers. By 3-D integration, functionally partitioned components in an imager can be fabricated on different strata to form 3-D imaging system with $\sim$ 100% fill factor. The short and high-bandwidth inter-stratum interconnects in 3-D integration could replace the long interconnects between the detectors and read out circuits in a conventional implementation of an imager, resulting in higher system performance and lower power dissipation.

Appendix A

Fabrication Process Flow of 3-D ICs Based on Wafer Bonding

In this section, the process flow for fabricating monolithic 3-D ICs based on wafer bonding [47] will be presented. The necessary fabrication steps are illustrated in Fig. A-1 and Fig. A-2 and they include:

1: Fabrication of bulk Si wafers which will be used as the bottom most stratum and pattering of Cu bumps and dummy fills for wafer bonding. The thickness of the Cu layer is 400 nm, and it is deposited by e-beam evaporation.

2: Fabrication of SOI wafers which will be used as the 2^{nd} and/or 3^{rd} stratum. Deposition of the Cu layer, pattering of Cu bumps and dummy fills.

3: Attachment of the SOI wafer with a handle wafer using a sacrificial layer, a sandwich of Ti/Cu/Ti. Cu is used to bond the handle wafer with SOI wafer, and Ti is used as a sacrificial layer which can be selectively etched/dissolved by HF.

4: Mechanical grind back and chemical etching (using KOH) of the back side of SOI wafers

5: Via opening through the thin Si layers

6: Via filling

7: Deposition of Cu film and pattering of Cu bumps for wafer bonding

8: IR alignment of bulk Si and SOI wafer and low-temperature wafer bonding. The bonding process is based on thermo-compression at $400^0C - 450^0C$ for 30 minutes, followed by a 30-minute anneal at $400^0C - 450^0C$. The IR alignment tolerance is roughly ±3 μm.

9: Handle wafer release by dissolving the sacrificial layer. After releasing the handle wafer, another SOI wafer can be integrated by following Steps 1-9. Unlike other approaches to 3-D integration such as front-to-front bonding, our approach can be repeated to integrate

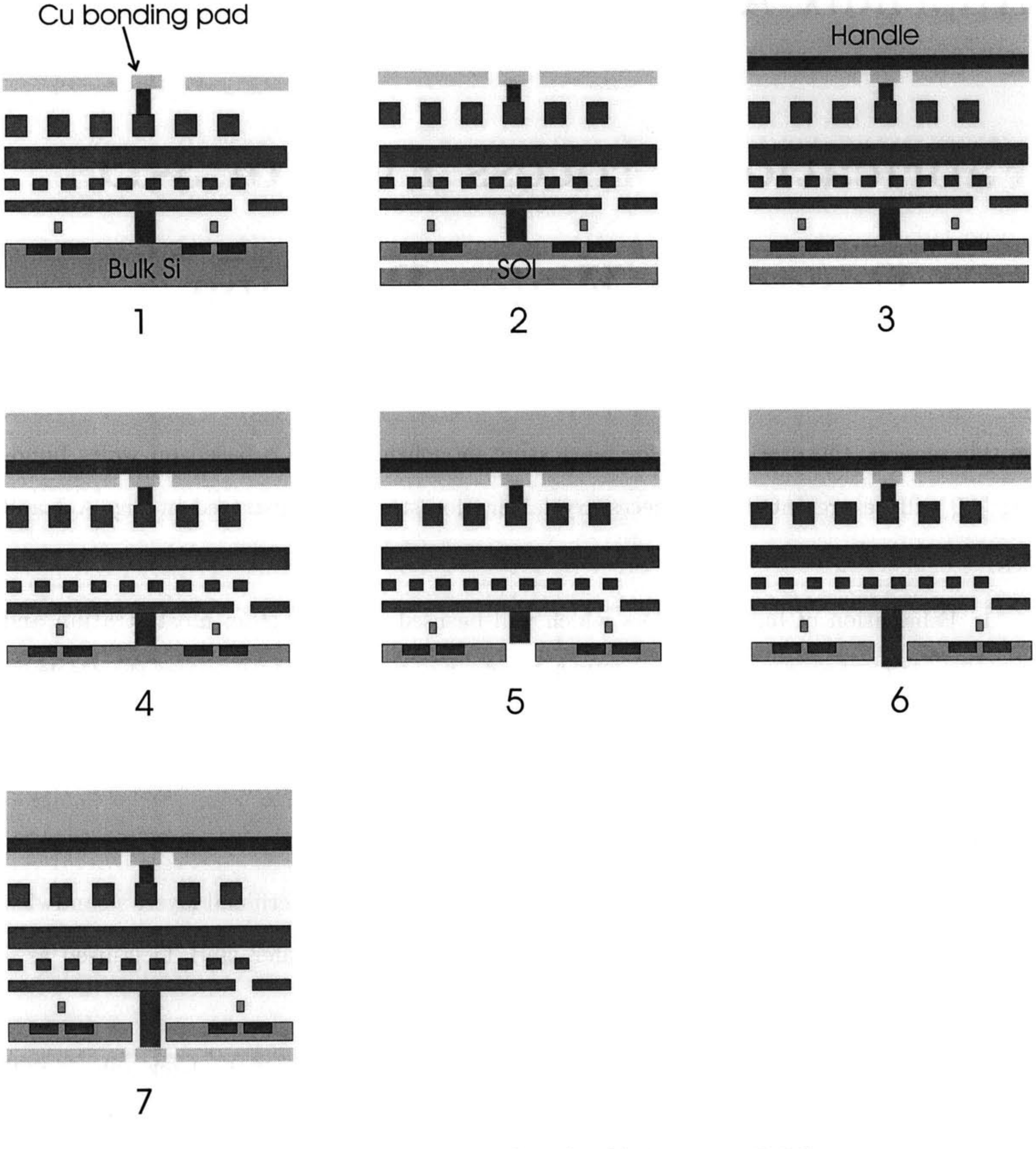


Figure A-1: The process flow for fabricating 3-D ICs

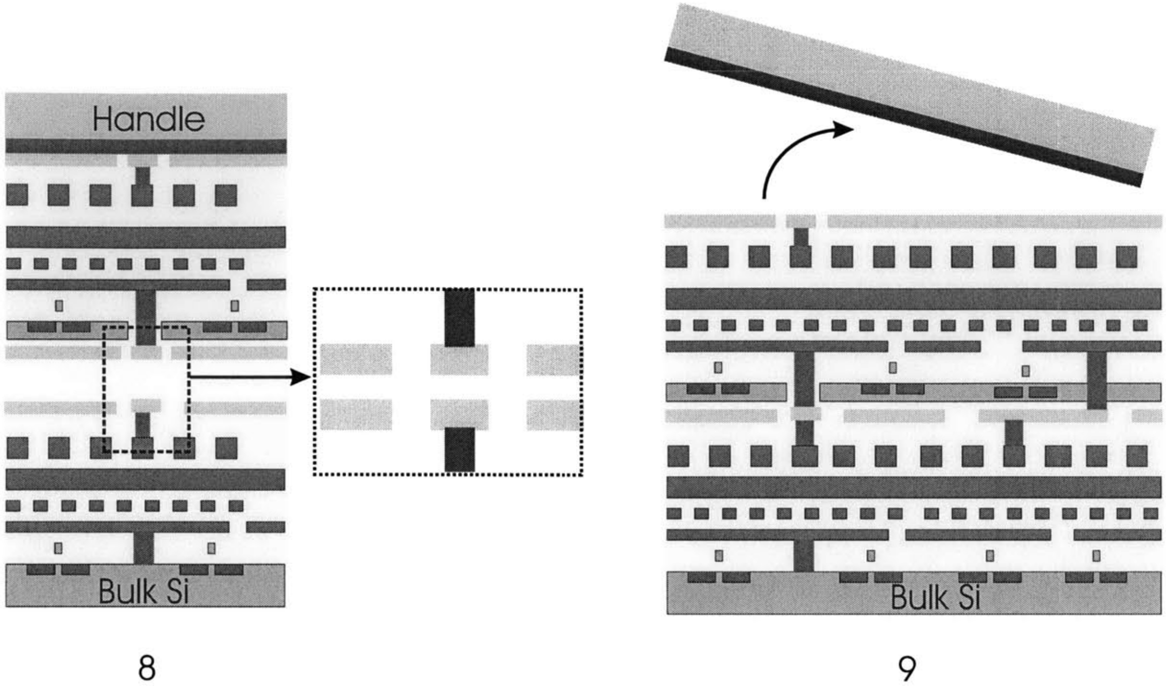


Figure A-2: The process flow for fabricating 3-D ICs

as many stratum as desired. However, due to the cost/complexity of the bonding process, thermal issues, yield, etc. there could be a limit on the number of wafers that can be integrated profitably.

Appendix B

Stochastic Net-Length Distribution of Global Interconnects in 3-D Heterogeneous System-on-a-Chip

A system-on-a-chip (SOC) consists of heterogeneous megacells or functional blocks that have been designed, routed, and verified for optimal performance. The Rent's parameters for heterogeneous megacells in a SOC could be same or different, and the length of global wires between megacells may vary from a megacell's semi-perimeter to chip semi-perimeter. The communication between megacells and system performance in a SOC is often limited by the interconnect delay of global wiring network. It is desirable to reduce the global wire-length between megacells to improve overall system performance. Recently, a stochastic net-length distribution model was proposed for 2-D heterogeneous system-on-a-chip (SOC) [82]. In stochastic wire-length prediction for SOC, the number of multi-terminal nets is estimated using Rent's Rule. Then placement and routing models are used to estimate global net-length distribution. In this section an extension of the 2-D net-length distribution model to three dimension for SOC applications will be presented.

B.1 Distribution of Multi-Terminal Nets

B.1.1 Heterogeneous Rent's Rule

Recently, a heterogeneous Rent's rule has been proposed for a system that consists of logic blocks or modules with various Rent's parameters. By empirical observation and curve fitting, it has been found that if a system consists of megacells with N_1, N_2, $\cdots$, N_m logic gates that have Rent's parameters (k_1,p_1), (k_2,p_2), $\cdots$, (k_m,p_m), respectively, equivalent Rent's parameters can be estimated to describe the interconnection complexity of the entire system. The equivalent Rent's rule for a heterogeneous system is given by [72]

$$T_{eq} = k_{eq} N^{p_{eq}}, \tag{B.1}$$

where T_{eq}, k_{eq}, and p_{eq} are the equivalent number of I/O terminals, Rent's constant, and Rent's exponent, respectively; N is total number of logic gates. The equivalent Rent's parameters are given by

$$\begin{aligned} k_{eq} &= (\prod_{j=1}^{m} k_j^{N_j})^{\frac{1}{N}} \\ p_{eq} &= \frac{\sum_{j=1}^{m} p_j N_j}{N} \\ N &= \sum_{j=1}^{m} N_j \end{aligned} \tag{B.2}$$

B.1.2 Multi-Terminal Nets

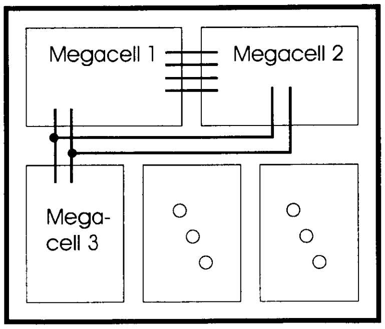


Figure B-1: Multi-terminal nets in a SOC.

The heterogeneous Rent's rule can be used to estimate the distribution of multi-terminal nets. For example, consider the two megacells in a SOC as shown in Fig. B-1. The number of two terminal nets, N_{12}, between megacells 1 and 2 is given by [72]

$$N_{12} = (k_1 N_1^{p_1} + k_2 N_2^{p_2} - k_{1,2} N_{1,2}^{p_{1,2}})/2, \tag{B.3}$$

where the first two terms on the right hand side of Eq. B.3 are the number of I/O terminals of megacells 1 and 2, respectively. The third term in Eq. B.3 is the equivalent number of I/O terminals of a system combined with megacells 1 and 2 which can be found using Eq. B.1. Similarly, the number of three-terminal nets between megacells 1, 2, and 3, N_{123}, can be found, and it is given by

$$N_{123} = (k_1 N_1^{p_1} + k_2 N_2^{p_2} + k_3 N_3^{p_3} - 2N_{12} - 2N_{23} - 2N_{31} + k_{1,2,3} N_{1,2,3}^{p_{1,2,3}})/3, \tag{B.4}$$

This process can be repeated by conserving the number of I/O terminals to estimate the number of nets with t terminals in a SOC with m megacell, and it is given by

$$\begin{aligned} N_{123\cdots t} &= [\sum_{i}^{m} k_i N_i^{p^i} - \sum_{i=1,\cdots,m;j=1,\cdots,m;i\neq j}^{m} 2N_{ij} \\ &+ \sum_{i=1,\cdots,m;j=1,\cdots,m;k=1,\cdots,m;i\neq j\neq k}^{m} 3N_{ijk} \\ &\cdots + (-1)^{t+1} k_{1,2,3,\cdots,t} N_{1,2,3,\cdots,t}^{p_{1,2,3,\cdots,t}}]/t \end{aligned} \tag{B.5}$$

B.1.3 Stochastic Placement and Routing Information

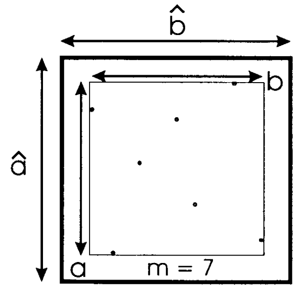


Figure B-2: The bounding are of terminals in a multi-terminal net.

Megacell Number	N	k	p	Area (mm^2)
1, 2	300K, 300K	4, 4	0.55, 0.55	11.2
3, 4	200K, 200K	4, 4	0.6, 0.6	7.6
5, 6	300K, 300K	3, 3	0.65, 0.65	11.8

Table B.1: The assumed Rent's parameters and the estimated chip area of the megacells under consideration.

Once the number of terminals involved in multi-terminal nets is known, assuming random placement of terminals, the average bounding area and wire-length for multi-terminal nets can be found [82]. The average bounding area dimensions of m-terminal net (as shown in Fig. B-2) is given by

$$a = \frac{m-1}{m+1} \cdot \hat{a} \qquad b = \frac{m-1}{m+1} \cdot \hat{b}, \tag{B.6}$$

where $\hat{a}$ and $\hat{b}$ are the dimensions of active area of the megacells, and they are given by [82]

$$\hat{a} = \hat{b} = \sqrt{\bar{A}_{meg}/N_z[m\eta_p + N_m(1-\eta_p)]}. \tag{B.7}$$

In Eq. B.7, $\bar{A}_{meg}$ is the average area of a megacell, N_z is the number of strata, η_p is the placement efficiency, and N_m is the total number of megacells. Assuming the routing of multi-terminal nets is based on Minimal Rectilinear Steiner Trees (MRST), the lower bound L_l and upper bound L_u of net-length for m-terminal nets are given by [141, 142]

$$L_l = (a+b) \qquad L_u = (a+b) \cdot \frac{\sqrt{m}+1}{2} \tag{B.8}$$

B.2 Three-Dimensional Placement and Routing

To examine the advantages of 3-D placement and routing, we consider a 3-D IC with two strata. We assume through wafer vias are used for inter-mega cell connections. The hypothetical SOC is composed of six megacells and their connectivity information is provided in Table B.1. It is implemented in 0.18 μm technology node. For interconnections within the megacell, two local and two semi-global interconnect levels are used. Their wiring pitches are $2F$ and $4F$, where F is the minimum feature size. The area of the mega-cells is estimated using the methodology presented in Chapter 2. To minimize 3-D SOC's area, we consider the scenario where megacells 1, 3, and 5 are placed on the lower stratum and megacells 2, 4, and 6 are placed on the upper stratum. Based on the methodology described earlier, and using the upper bound of MRTS for wire-length models, the global wire-length distribution in a SOC is estimated. The simulation results of wire-length distribution of multi-terminal nets and total wire-length distribution are shown in Fig. B-3 and Fig. B-4. We find that

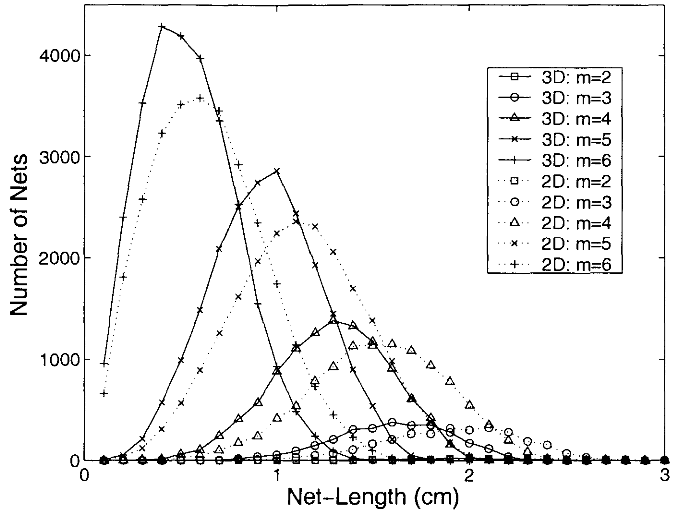


Figure B-3: The stochastic wire-length distribution of m-terminal nets in 2-D and 3-D implementation of SOC.

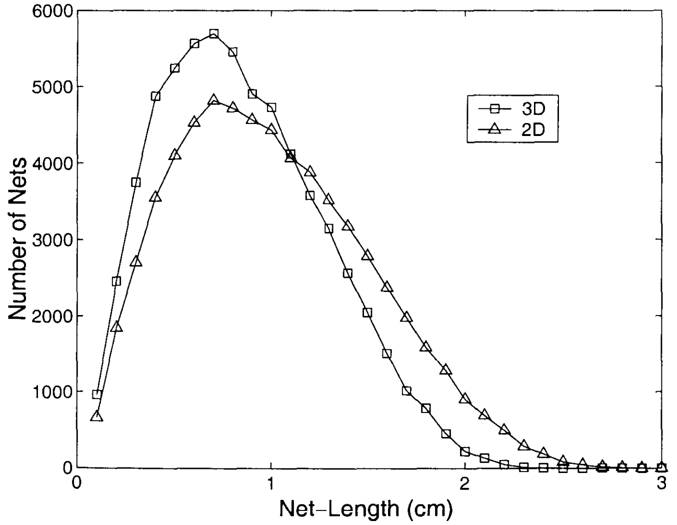


Figure B-4: The stochastic global (inter-megacell) wire-length distribution in 2-D and 3-D implementation of SOC.

3-D placement and routing result in a narrower global wire-length distribution compared to 2-D implementation for SOC applications. By 3-D integration, the length of high fan-out nets can be reduced significantly. The reduction in average global wire-length with two strata is roughly 18%. We find that the overall reduction in wire-length based on megacell-level partitioning is not as high as gate-level partitioning. To achieve further reduction in wire-length in a SOC, each megacell should be partioned and placed in multiple strata and inter-stratum interconnections will be needed for both within and between megacells.

Bibliography

[1] Mark T. Bohr and Youssef A. El-Mansy. Technology for Advanced High-Performance Microprocessors. *IEEE Transactions on Electron Devices*, 45(3):620–625, 1998.

[2] 1999 International Roadmap for Semiconductors.

[3] H. Kurino, T. Matsumoto, K. H. Yu, N. Miyakawa, H. Itani, H. Tsukamoto, and M. Koyanagi. Three-Dimensional Integration Technology for Real Time Micro-Vision System. In *Proceedings of the Innovative Systems in Silicon Conference*, pages 203–212, 1997.

[4] Chenming Hu. Future CMOS-Scaling and Reliability. In *Proceedings of the IEEE*, volume 81, pages 682–689, 1993.

[5] Bijan Davari, Robert H. Dennard, and Ghavam G. Shahidi. CMOS-Scaling for High Performance and Low Power - The Next Ten Years. In *Proceedings of the IEEE*, volume 83, pages 595–606, 1995.

[6] T. Sakurai. Closed-Form Expressions for Interconnect Delay, Coupling, and Crosstalk in VLSI. *IEEE Transactions on Electron Devices*, 40(1):118–124, 1993.

[7] 1997 SIA Roadmap.

[8] J. A. Davis and R. Venkatesan and A. Kaloyeros and M. Bylansky and S. J. Souri and K. Banerjee and K. Saraswat and A. Rahman and R. Reif and J. D. Meindl, Gigascle Integration (GSI) Interconnect Limits in the 21st Century, submitted to the Proceedings of the IEEE.

[9] David A. B. Miller. Rationale and Challenges for Optical Interconnects to Electronic Chips. In *Proceedings of the IEEE*, volume 88, pages 728–749, 2000.

[10] Yoichi Akasaka. Three-Dimensional IC Trends. In *Proceedings of the IEEE*, volume 74, pages 1703–1714, 1986.

[11] Arifur Rahman and Rafael Reif. System-level Performance Evaluation of Three-Dimensional Integrated Circuits. *IEEE Transactions on VLSI Systems*, 8(6):xx, 2000.

[12] RF/Wireless Interconnect, M. Frank Chang, Interconnect Focus Center Program Review, Georgia Institute of Technology, August 15, 1999.

[13] William J. Dally. Interconnect-Limited VLSI Architecture. In *Proceedings of IITC*, pages 15–17, 1999.

[14] Anant Agarwal, Raw Computation, Scientific American, 1999, www.sciam.com/1999/0899issue/0899agarwal.html.

[15] George A. Sai-Halasz. Performance Trends in High-End Processors. In *Proceedings of The IEEE*, volume 83, pages 20–36, 1995.

[16] D. Sylvester and K. Keutzer. A Gobal Wire Paradigm for Deep Submicron Design. *IEEE Tran. on Computer Aided Design of Integrated Circuits and Systems*, 19(2):242–252, 2000.

[17] Jeffery A. Davis, Vivek K. De, and James D. Meindl. A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) - Part I: Derivation and Validation. *IEEE Transaction on Electron Devices*, 45(3):580–589, 1998.

[18] M. Horowitz, R. Ho, and K. Mai, "The Future of Wires", SRC/SEMATECH/MARCO Workshop on Interconnects for Systems on a Chip, May 22, 1999, Standford University.

[19] S. Takahashi, M. Edahiro, and Y. hayashi. Interconnect Design Strategy: Structures, Repeaters and Materials toward 0.1 μ m ULSI with a Giga-hertz Clock Operation. In *Proceedings of the IEDM*, pages 833–836, 1998.

[20] Jeffery A. Davis and James D. Meindl. Length Scaling and Material Dependence of Crosstalk between Distributed RC Interconnects. In *Proceedings of IITC*, pages 227–229, 1999.

[21] Arifur Rahman, Andy Fan, and Rafael Reif. Comparison of Key Performance Metrics in Two- and Three- Dimensional Integrated Circuits. In *Proceedings of IITC*, pages 18–21, 2000.

[22] The Collaborative Node, Anantha Chandrakasan, Interconnect Focus Center Program Review, Georgia Institute of Technology, August 15, 1999.

[23] K. F. Lee, J. F. Gibbons, K. C. Saraswat, and T. I. Kamins. Thin Film MOSFET Fabricated in Laser-Annealed Polycrystalline Silicon. *Journal of Applied Physics Letters*, 35:173–175, 1979.

[24] K. Oyama, T. Kunio, Y. Hayashi, K. Kajiyana, and K. Tsunenari. High Density Dual-Active-Layer (DUAL) CMOS Structure with Vertical Tungsten Plug-In Wirings. In *Proceedings of The IEDM*, pages 59–62, 1990.

[25] Y. Hayashi, K. Oyama, S. Takahashi, S. Wada, K. Kajiyana, R. Koh, and T. Kunio. A New Three Dimensional IC Fabrication Technology, Stacking Thin Film DUAL-CMOS Layers. In *Proceedings of The IEDM*, pages 657–660, 1991.

[26] T. Kunio, K. Oyama, Y. Hayashi, and M. Morimoto. Three Dimensional ICs, Having Four Stacked Active Device Layers. In *Proceedings of The IEDM*, pages 837–840, 1989.

[27] T. Nishimura, Y. Inoue, K. Sugahara, S. Kusonoki, T. Kumamoto, S. Nakagawa, Y. Horiba, and Y. Akasaka. Three Dimensional IC, for High Performance Image Signal Processor. In *Proceedings of The IEDM*, page 111, 1987.

[28] Said F. Al-sarawi, Derek Abbott, and Paul D. Franzon. A Review of 3-D Packaging Technology. *IEEE Transactions on Components, Packaging, and Manufacturing Technolgy- Part B*, 21(1):2–14, 1998.

[29] Christan VAL. The Future of 3D Packaging. In *Proceedings of IEMT/IMC*, volume 1, pages 261–271, 1998.

[30] Nobauki Takahashi, Naoji Senba, Yozo Shimada, Ikushi Morisaki, and Kenichi Tokuno. Three-Dimensional Memory Module. *IEEE Trans. on Components, Packaging, and Manufacturing Technoloy-Part B*, 21(1):15–19, 1998.

[31] Christan VAL and T. Lemoine. 3D Interconnection for Ultra-Dense Multichip Modules. In *Proceedings of 40th Electronic Components and Technology Conference*, volume 1, pages 540–557, 1990.

[32] Robert Bruns, Warren Chase, and Dean Frew. Utilizing Three-Dimensional Memory Packaging and Silicon-on-Silicon Technology for Next Generation Recording Devices. In *ICMCM*, pages 327–333, 1992.

[33] Claude L. Bertin, David J. Perlman, and Stuart N. Shanken. Evaluation of a Three-Dimensional Memory Cube System. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 16:1006–1011, 1993.

[34] Yee L. Low and Kevin J. O'Connor. Electrical Performance of Chip-on-Chip Modules. *IEEE Transactions on Advanced Packaging*, 22:321–325, 1999.

[35] Michael L. Campbell and Scott T. Toborg. 3-D Wafer Scale Architectures for Neural Network Computing. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 16(7):646–655, 1993.

[36] J. F. McDonald. An Overview and Analysis of 3D WSI. In *International Conference on Wafer Scale Integration*, pages 223–234, 1991.

[37] W. J. Howell, D. W. Brouillette, J. W. Konejwa, S. J. Sprogis, E. J. Yankee, and J. M. Wursthorn. Area Array Solder Interconnection Technology for The Dimensional Silicon Cube. In *Proceedings of 45th Electron. Comp. Technol. Conf.*, 1995. Las Vegas, NV.

[38] K. Hatada, H. Fujimoto, T. Kawakita, and T. Ochi. A New LSI Bonding Technology: "Micron Bump Technology". In *Proceedings of the 5th IEEE/CHMT Int. Electron. Manufact. Technol. Symp.*, 1995. Lake Buena Vista, FL.

[39] J. M. Segelken, L. J. Wu, M. Y. Lau, K. L. Tai, R. R. Shively, and T. G. Grau. Ultra-Dense: An MCM-Based 3-D Digital Signal Processor. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 15:438–443, 1992.

[40] Gerold W. Neudeck. Three-Dimensional CMOS Integration. *IEEE Circuits and Devices Magazine*, 6:32–38, 1990.

[41] J. P. Denton and G. W. Neudeck. Fully Depleted Dual Gate Thin-Film SOI P-MOSFET's Fabricated in SOI Islands with an Isolated Buried Polysicon Backgate. *IEEE Electron Device Lett.*, 17:509–510, 1996.

[42] J. O. Borland. Novel Device Stuctures by Selective Epitaxial Growth (SEG). In *IEDM Technical Digest*, pages 12–15, 1987.

[43] V. Subramanian, M. Toita, N. R. Ibrahim, S. J. Souri, and K. C. Saraswat. Low-Leakage Germanium-Seeded Laterally-Crystallized Single-Grain 100-nm TFTs for Vertical Integration Applications. *IEEE Electron Device Letters*, 20:341–343, 1999.

[44] Sangwoo Pae, Taichi Su, John P. Denton, and Gerold W. Neudeck. Multiple Layers of Silicon-on-Insulator Island Fabrication by Selective Epitaxial Growth. *IEEE Electron Device Lett.*, 20(5):194–196, 1999.

[45] Philip M. Sailer, Piyush Singhal, Jeffrey Hopwood, David R. Kaeli, Paul M. Zavracky, Keith Warner, and D. P. Vu. Creating 3D Circuits Using Transfeerred Films. *Circuit and Devices*, pages 27–30, 1997.

[46] P. Ramm, D. Bollman, R. Braum, R. Buchner, U. Cao-Minh, M. ngelhardt, G. Ermann, T. Grabl, K. Hieber, H. Hubner, G. Kawala, M. Kleiner, A. Klumpp, S. Kuhn, C. Landesberger, H. Lezec, W. Muth, W. Pamler, R. Popp, E. Renner, G. Ruhl,

A. Sanger, U. Scheler, A. Schertel, C. Schimdt, S. Schwarzl, J. Webber, and W. Webber. There Dimensional Metallization for Vertically Integrated Circuits. *Microelectric Engineering*, 37(38):39–47, 1997.

[47] Andy Fan, Arifur Rahman, and Rafael Reif. Copper Wafer Bonding. *Electrochemical and Solid State Letters*, 2:534–536, 1999.

[48] Mark R. Pinto. Atoms To Applets: Building Systems ICs in the 21st Century. In *Proceedings of ISSCC*, pages 26–30, 2000.

[49] H. B. Bakoglu. *Circuits, Interconnections and Packaging for VLSI.* Addison-Wesley, 1990.

[50] John Eble, A Generic system simulator with novel on-chip cache and throughput models for gigascale integration, PhD Thesis, Georgia Institute of Technology, 1998.

[51] D. Sylvester and K. Keutzer, BACPAC- Berkeley Avanced Chip Performance Calculator, www-device.eecs.berkeley.edu/ dennis/bacpac/models/.

[52] Bibiche Geuskens and Kenneth Rose. Modeling Limits of Multilevel Interconnect Tecnology. *SPIE*, 2636:317–325, 1995.

[53] Jeffery A. Davis, Vivek K. De, and James D. Meindl. A Stochastic Wire-Length Distribution for Gigascale Integration (GSI) - Part II: Application to Clock Frequency, Power Dissipation, and Chip Size Estimation. *IEEE Transaction on Electron Devices*, 45(3):590–597, 1998.

[54] H. B. Bakoglu and J. D. Meindl. A System Level Circuit Model for Multi- and Single-Chip CPUs. In *ISSCC Digest of Technical Papers*, pages 223–234, 1991.

[55] T. Sakurai and A. R. Newton. Alpha-Power Law MOSFET Model and Its Application to CMOS Inverter Delay and Other Formulas. *IEEE Journal of Solid State Circuits*, 25(2):584–594, 1990.

[56] Jan M. Rabaey. *Digital Integrated Circuits.* Prentice Hall, NJ, 1996.

[57] T. Sakurai. Approximation of Wiring Delay in MOSFET LSI. *IEEE Journal of Solid-State Circuits*, 18:418–426, 1983.

[58] A. Deutsch, G. V. Kopcsay, P. J. Restle, H. H. Smith, G. Katopis, W. D. Becker, P. W. Coteus, C. W. Sorovic, B. J. Rubin, R. P. Dunne, T. Gallo, K. A. Jenkins, L. M. Terman, R. H. Dennard, G. A. Sai-Halasz, B. L. Krauter, and D. R. Knebel. When are Transmission-Line Effect Important for On-Chip Interconnects. *IEEE Tran. on MTT*, 45(10):1836–1846, 1997.

[59] Y. I. Ismail and E. G. Friedman. Effects of Inductance on the Propagation Delay and Repeater Insertion in VLSI Circuits. *IEEE Tran. on VLSI Systmes*, 8(2):195–206, 2000.

[60] J. H. Chern, J. Huang, L. Arledge, P. C. Li, and P. Yang. Multilevel Metal Capacitance Models for CAD Design Synthesis Systmes. *IEEE Electron Device Lett.*, 13:32–33, 1992.

[61] Michael J. Flynn. *Computer Architecture.* Jones and Barlett Publishers, 1995.

[62] Wilm E. Donath. Placements and Average Interconnection Lengths of Computer Logic. *IEEE Transactions on Circuits and Systems*, 26(4):272–277, 1979.

[63] Akira Masaki and Minoru Yamada. Equations for Estimating Wire Length in Various Types of 2-D and 3-D System Packaging Structures. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 10:190–198, 1987.

[64] A. Gamal. Two Dimensional Model for Interconnections in Master Slice Integrated Circuits. *IEEE Trnas. Circuits Syst.*, 28:127–138, 1981.

[65] W. Heller, W. Mikhail, and W. Donath. Prediction of Wiring Space Requirements for LSI. *Design Automation and Fault Tolerant Computing*, pages 117–144, 1978.

[66] P. Christie. A Fractal Analysis of Interconnect Complexity. *Proceedings of the IEEE*, 81:1492–1499, 1993.

[67] Bernard S. Landmand and Roy L. Russo. On a Pin Versus Block Relationship For Partitions of Logic Graphs. *IEEE Transactions on Computers*, 20(12):1469–1479, 1971.

[68] P. K. Chan, M. D. F. Schlag, and Jason Y. Zien. On Routability Prediction for Field-Programmable Gate Arrays. In *Proceedings for DAC*, pages 326–330, 1993.

[69] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, 1982.

[70] D. Stroobandt and V. Campenhout. Estimating Interconnection Length in Three Dimensional Computer Systems. *IEICE Trans. on Information and Systems*, 80(10):1024–1031, 1997.

[71] H. M. Ozaktas and J. W. Goodman. The Limitations of Interconnections in Providing Communication Between an Array of Points. *Frontiers of Computing Systems Research, Edited by S. K. Tewksbury, Plenum Press, NY*, 2:61–130, 1991.

[72] Payman Zarkesh-Ha, Jeffery A. Davis, William Loh, and James D. Meindl. On a Pin Versus Gate Relationship for Heterogeneous Systems: Heterogeneous Rent's Rule. In *Proceedings of Custom Integrated Circuit Conference*, pages 93–96, 1998.

[73] Tom Knight, personal communications.

[74] A. L. Rosenberg. Three-Dimensional VLSI: A Case Study. *Journal of the Association of Computer Machinery*, 30(3):397–416, 1983.

[75] Chao Chi Tong and Chuan lin Wu. Routing In a Three-Dimensional Chip. *IEEE Trans. on Computers*, 44(1):106–117, 1995.

[76] V. K. Jain and S. Horiguchi. A Gobal Wire Paradigm for Deep Submicron Design. *IEEE Transaction on VLSI Systems*, 6(3):346–353, 1998.

[77] Michael Feuer. Connectivity of Random Logic. *IEEE Trans. on Computers*, 31(1):29–33, 1982.

[78] Arifur Rahman, Andy Fan, James Chung, and Rafael Reif. Wire-Length Distribution of Three-Dimensional Integrated Circuits. In *Proceedings of IITC*, pages 233–235, 1999.

[79] J. Joyner, P. Zarkesh-Ha, J. Davis, and J. Meindl. A Three-Dimensional Stochastic Wire-Length Prediction for Variable Separation of Strata. In *Proceedings of IITC*, pages 123–125, 2000.

[80] Jeffery A. Davis, Vevek K. De, and James D. Meindl. A Priori Wiring Estimation and Optimal Multilevel Wiring Networks for Portable ULSI Systems. In *Proceedings of Electronic Components and Technology Conference*, pages 1002–1008, 1996.

[81] D. A. Antoniadis, A. Wei, and A. Lochtefeld. SOI Devices and Technologies. In *Porc. of ESSDERC*, pages 81–87, 1999.

[82] Payman Zarkesh-Ha and James D. Meindl. Stochastic Net Length Distribution for Global Interconnects in a Heterogeneous System-on-a-Chip. In *IEEE Symposium on VLSI Technology*, pages 44–45, 1998.

[83] M. B. Kleiner, S. A. Kuhn, P. Ramm, and W. Weber. Performance Improvement of the Memory Hierarchy of RISC- Systems by Application of 3-D Technology. In *IEEE Trans. on CPMT-B*, volume 19, pages 709–718, 1996.

[84] Doug Burger, James R. Goodman, and Alain Kagi. Limited Bandwidth to Affect Processor Design. *IEEE Micro*, pages 55–62, 1997.

[85] David Patterson, Thomas Anderson, Neal Cardwell, and Richard Fromm. A Case for Intelligent RAM:IRAM. *IEEE Micro*, pages xx–xx, 1997.

[86] Tadaaki Yamauchi, Lance Hammond, and Kunle Olukotun. A Hierarchical Multi-Bank DRAM: A High-Performance Architecture for Memory Integrated with Processors. *Proceedings of Advanced VLSI Research*, pages xx–xx, 1997.

[87] Lee Higbie. Quick and Easy Cache Performance Analysis. *Computer Architecture News*, 18:33–44, 1990.

[88] John L. Hennessy David A. Patterson. *Computer Architecture: A Quantitative Approach.* Morgan Kaufmann Publishers, 1996.

[89] H. Nambu, K. Kanetani, K. Yamasaki, K. Higeta, M. Usami, T. Kusunoki, K. Yamaguchi, and N. Homma. A 1.8 ns Access, 550 MHz 4.5 Mb CMOS SRAM. In *Proceedings of ISSCC*, pages 360–361, 1998.

[90] B. Bateman, C. Freeman, J. Halbert, K. Hose, G. Petrie, and E. Reese. A 450 MHz 512kB Second-Level Cache with a 3.6GB/s Data Bandwidth. In *Proceedings of ISSCC*, pages 358–359, 1998.

[91] R. Stephany, K. Anne, J. Bell, G. Cheney, J. Eno, G. Hoeppner, G. Joe, R. Kaye, J. Lear, T. Litch, J. Meyer, J. Montanaro, K. Patton, T. Pham, R. Reis, M. Sillia, J. Slaton, K. Snyder, and R. Witek. A 200 MHz 32b 0.5W CMOS RISC Microprocessor. In *Proceedings of ISSCC*, pages 238–239, 1998.

[92] C. Lage, J. D. Hayden, and C. Subramanian. Advanced SRAM Technology - The Race Between 4T and 6T Cells. In *Proceedings of IEDM*, pages 271–274, 1996.

[93] Paul Gronowski. Designing High Performance Microprocessors. In *Proceedings of Symposium of VLSI Circuits*, pages 51–54, 1995.

[94] L. Gwennap. Microprocessors Lead the Way to .35 microns. *Microprocessor Report*, pages 1–5, 1995.

[95] http://bwrc.eecs.berkeley.edu/People/Grad_students/burd/cic.

[96] T. Wada, S. Rajan, and S. A. Przybylski. An Analytical Access Time Model for On-Chip Cache Memory. *IEEE Journal of Solid State Circuits*, 27(8):1147–1156, 1992.

[97] B. S. Amrutur and M. A. Horowitz. Speed and Power Scaling of SRAM's. *IEEE Journal of Solid State Circuits*, 35(2):175–185, 2000.

[98] Michael J. Flynn. Basic issues in microprocessor architecture. *Journal of System Architecture*, 45:939–948, 1999.

[99] J. D. Gee, M. D. Hill, D. N. Pnevmatikatos, and A. J. Smith. Cache Performance of the SPEC92 Benchmark Suite. *IEEE Micro*, 13(4):17–26, 1993.

[100] Edited by Stephen M. Trimberger. *Field-Programmable Gate Array Technology.* Kluwer Academic Publishers, 1994.

[101] Stephen D. Brown, Robert J. Francis, Jonathan Rose, and Zvonko G. Vranesic. *Field-Programmable Gate Arrays.* Kluwer Academic Publishers, 1992.

[102] WWW.Xilinx.com.

[103] WWW.Lucent.com/micro/netcom/.

[104] WWW.Altera.com.

[105] Andre DeHon. *Reconfigurable Architectures for General-Purpose Computing.* PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1996.

[106] Eric Kusse. Ananlysis and Circuit Design for Low Power Programmable Logic Module, M. S. Thesis, Department of Electrical Engineering and Computer Science, University of California at Berkeley, 1997.

[107] Stephen D. Brown. An Overview of Technology, Architecture and CAD Tools for Programmable Logic Devices. In *Porceedings of Custom Integrated Circuits Conference*, pages 69–76, 1994.

[108] Joel Darnauer and Wayne Wei ming Dai. A Method for Generating Random Circuits and Its Application to Routability Measurements. In *Proceedings of International Sysposium on Field Programmable Gate Arrays*, pages 66–72, 1996.

[109] Stephen D. Brown, Jonathan Rose, and Zvonko G. Vranesic. A Stochastic Model to Predict the Routability of Field-Programmable Gate Arrays. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 12(12):1827–1838, 1993.

[110] Vaughn Betz, Jonathan Rose, and Alexander Marquardt. *Architecture and CAD for Deep-Submicron FPGAs.* Kluwer Academic Publishers, 1999.

[111] WWW.eecg.toronto.edu/EECG/RESEARCH/FPGA.html.

[112] Michael J. Alexander, James P. Cohoon, Jared L. Colflesh, John Karro, Edward L. Peters, and Gabriel Robins. Placement and Routing for Three-Dimensional FPGAs. In *Porceedings of Canadian Workshop on Field-Programmable Devices*, pages 11–18, 1996.

[113] Michael J. Alexander, James P. Cohoon, Jared L. Colflesh, John Karro, , and Gabriel Robins. Three-Dimensional Field-Programmable Gate Arrays. In *Proceedings of Eight Annual IEEE International ASIC Conference and Exhibit*, pages 253–256, 1995.

[114] Miriam Leeser, Waleed M. Meleis, Mankuan M. Vai, Silviu C hiricescu, Weidong Xu, and Paul M. Zavracky. Rothko: A Three-Dimensional FPGA. *IEEE Design and Test of Computers*, 15(1):16–23, 1998.

[115] Silviu M. S. A., Chiricescu, and Michael Vai. A Three-Dimensional FPGA with an Integrated Memory for in-Application Reconfigurable Data. *Procedings of IEEE International Symposium on Circuits and Systmes*, 2:232–235, 1998.

[116] Jan Van Campenhout, Herwig Van Marck, Jo Depreitere, and Joni Dambre. Optoelectronic FPGA's. *IEEE Journal of Selected Topics in Quantum Electronics*, 5(2):306–315, 1999.

[117] M. Khellah, S. Brown, and Z. Vranesic. Modeling Routing Delays in SRAM-Based FPGAs. *Proc. of CCVLSI, Banff, Canada*, pages 13–18, 1993.

[118] Muhammad Khellah, Stephen D. Brown, and Zvonko Vranesic. Minimizing Interconnection Delays in Array-Based FPGAs. In *Porceedings of Custom Integrated Circuits Conference*, pages 181–184, 1994.

[119] Dake Liu and Christer Svensson. Power Consumption Estimation in CMOS VLSI Chips. *IEEE Journal of Solid-State Circuits*, 29:663–670, 1994.

[120] T. Sakurai. Design Challenges for 0.1 μm and Beyond. In *Proceedings of ASP-DAC*, pages 553–558, 2000.

[121] C. Gillot, L. Meysenc, C. Schaeffer, and A. Bricard. Integrated Single and Two-Phase Micro Heat Sinks Under IGBT Chips. *IEEE Transaction on CPMT-A*, 22(3):384–389, 1999.

[122] C. Cahill, T. Compagno, J. O. Donavan, O. Slattery, J. Barrett S. Mathuna, I. Serthelon, C. Val, J. P. Tigneres, P. Ivey J. Stern, M. Masgrangeas, and A. C. Vera. Thermal Characterization of Vertical Multichip Modules (MCM-V). In *IEEE Trans. on CPMT-A*, volume 4, pages 765–772, 1995.

[123] J. Barrett, C. Cahill, T. Compagno, M. O. Flaherty, W. Lawton T. Hayes, J. O. Donavan, C. O. Mathuna, G. McCarthy, O. Slattery, and F. Waldron. Performanc and Reliablility of a Three-Dimensional Plastic Moulded Vertical Multichip Module (MCM-V). In *Proceedings of the Electronic Components and Technology Conference*, pages 656–663, 1995.

[124] Michael B. Kleiner, Stefan A. Kuhn, Peter Ramm, and Werner Weber. Thermal Analysis of Vertically Integrated Circuits. In *Proceedings of The IEDM*, pages 487–490, 1995.

[125] www.synopsys.com/products/etg/etg.html.

[126] V. Tiwari, D. Singh, S Rajgopal, G. Mehta, R. Patel, and F Baez. Reducing Power in High-Performance Microprocessors. In *Proceedings of the Design Automation Conference*, pages 732–737, 1995.

[127] M. K. Gowan, L. L. Brio, and D. B. Jackson. Power Consideration in the Design of the Alpah 21264 Microprocessors. In *Proceedings of the Design Automation Conference*, pages 726–731, 1995.

[128] MTL VLSI Seminar, Feb. 15, 2000.

[129] N. Weste and K. Eshragian. *Principles of CMOS VLSI Design: A Systems Perspective*, chapter 4. Addison-Wesley, Reading, MA, 1992.

[130] A. P. Chandrakasan and R. W. Brodersen. Minimizing Power Consumption in Digital CMOS Circuits. *Proceedings of the IEEE*, 83:498–523, 1995.

[131] I. Y. Yang, C. Vieri, A. Chandrakasan, and D. A. Antoniadis. Back-Gated CMOS on SOIAS For Dynamic Threshold Voltage Control. *IEEE Trans. on Electron Devices*, 44:822–831, 1997.

[132] P. J. Restle and A. Deutsch. Designing the Best Clock Distribution Network. *Digest of Technical Papers for Symposium of VLSI Circuits*, pages 2–5, 1998.

[133] www.ansys.com.

[134] B. Gebhart. *Heat Transfer.* McGraw-Hill Book Company, 1971.

[135] Gary B. Kromann, R. David Gerke, and Wayne Wei-Xu Huang. A Hi-Density C4/CBGA Interconnect Technology for a CMOS Microprocessor. *IEEE Trans. on CPMT, Part B*, pages 166–173, 1996.

[136] Gary B. Kromann. Thermal Modeling and Experimental Characterization of the C4/Subrface-Mount-Array Interconnect Technologies. *IEEE Trans. on CPMT, Part A*, pages 87–93, 1995.

[137] Seri Lee. Optimum Design and Selection of Heat Sinks. *IEEE Trans. on CPMT, Part A*, pages 812–817, 1995.

[138] Seri Lee, Seaho Song, Seri Lee, Van Au, and Kevin P. Moran. Constriction/Spreading Resistances Model for Electronics Packaging. In *Proceedings of The ASME/JSME Thermal Engineering Conference*, volume 4, pages 199–206, 1995.

[139] S. Song, S. Lee, and Van Au. Closed-Form Equation for Thermal Constriction Spreading Resistance with Variable Resistance Boundary Condition. In *Proceedings of IEPS Conference*, pages 111–121, 1994.

[140] S. Im and K. Banerjee. Full Chip Thermal Analysis of Planar (2-D) and Vertically Integrated (3-D) High-Performance ICs. In *Proceedings of IEDM*, page xx, 2000.

[141] F. R. K. Chung and F. K. Huang. The Largest Minimal Rectilinear Steiner Trees for a Set of n Points Enclosed in a Rectangle with Given Parameter. *Networks*, 9:19–36, 1979.

[142] F. K. Hwang and D. S. Richards. *The Steiner Tree Problem*, chapter 4. North Holland, NY, 1992.